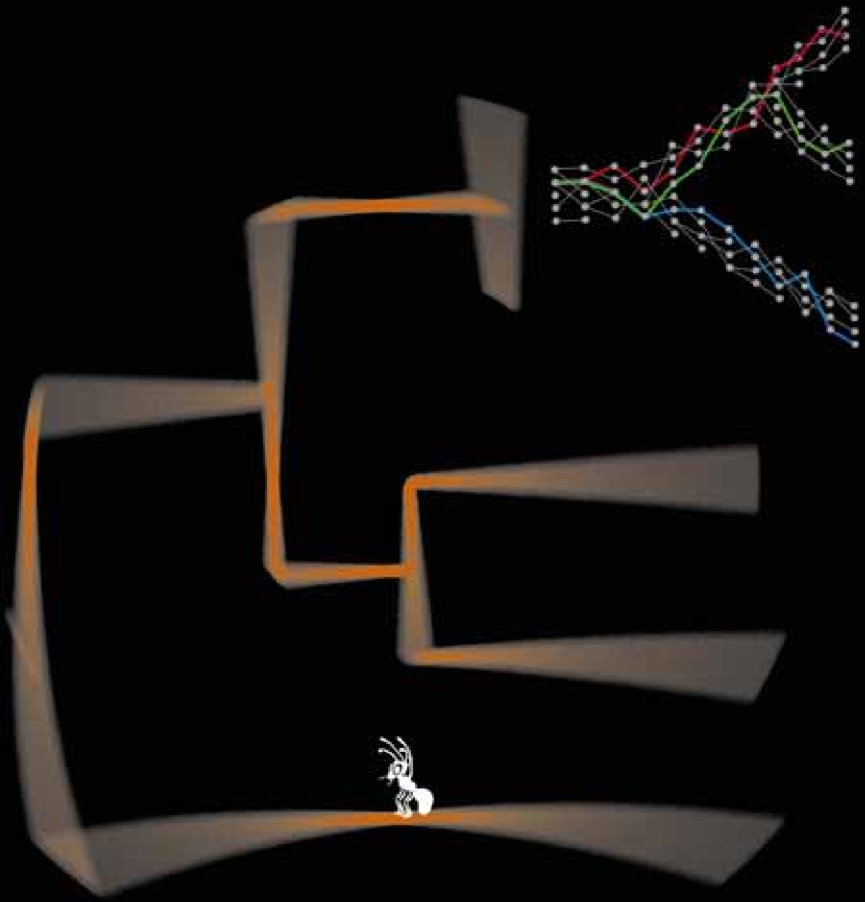
Phylogenetic inference at different insect taxonomic levels



Gerard Talavera Mor

· Doctoral Thesis ·

Phylogenetic inference at different insect taxonomic levels

Inferència filogenètica a diferents nivells taxonòmics d'insectes

- TESI DOCTORAL -

Gerard Talavera Mor

Bellaterra, Octubre del 2012

Facultat de Biociències – Departament de Genètica i Microbiologia

Memòria presentada per optar al grau de doctor per la Universitat Autònoma de Barcelona – Programa de Doctorat en Genètica

Vist-i-plau del director de la tesi	Vist-i-plau del tutor de la tesi
Dr. Roger Vila Ujaldón	**Dr. Antonio Barbadilla Prados**
Investigador Científic	*Professor Titular*
Institut de Biologia Evolutiva (CSIC-UPF)	Universitat Autònoma de Barcelona

PHYLOGENETIC INFERENCE AT DIFFERENT INSECT TAXONOMIC LEVELS

Gerard Talavera Mor

Agraïments

No ens enganyem, escriure unes línies que desprenguin certa emoció després d'elaborar una tesi d'estructura científica, no és gens fàcil. Ja havia sentit a dir que ciència i prosa no eren compatibles. És una qüestió de registres, tons i ús de les paraules. Poc o molt, perds la capacitat d'escriure línies sense realment dir gaire res. Sobre el que ha suposat tot aquest temps, resulta complexe discriminar entre el que és causa i conseqüència, o el que són mètodes i el que són resultats, com si el "com" no fos tant o més important que el "perquè". En realitat no hi ha massa a introduïr, més que les circumstàncies d'un moment sumades a un xic de curiositat em va dur a trucar a una porta.

Clar i directe, vull agraïr a en ROGER VILA per haver aconseguit convertir aquest temps de tesi en uns anys estimulants. Tinc dades: 1426 mails rebuts en 4 anys, amb una mitjana de gairebé un al dia. Gràcies Roger per fer-ho tot tant fàcil, per haver confiat en mi i per haver-me ensenyat tantes i tantes coses. Quatre anys ho canvien tot, què més puc dir...

I want to thank NAOMI PIERCE, who adopted this shy guy is her work group for such a long time. Thanks Naomi for all your support, for taking care of me while being far from home, for your mentorship and always catching enthusiasm, and for inspiring me new perceptions on biological understanding.

My special gratitude to VLADIMIR LUKHTANOV, a fantastic mentor and friend, thanks for showing to this phylogeneticist the particular vision of a taxonomist. It has been a real pleasure learning from you.

Per a tantes d'altres persones que m'han ensenyat coses. A en José Castresana, que em va ensenyar l'art de la filogènia i em va introduïr en el món de la recerca. A en Hafid Laayouni, que ja fa força anys em va deixar remenar un laboratori per primer cop. A en Xavier Espadaler, per aquesta sempre predisposició entusiasta i per haver-me pacientment passejat pel món de la mirmecologia. A l'Alfredo Ruiz, l'Antonio Barbadilla i tots els membres del grup de Genòmica, Bioinformàtica i Evolució, per ser tant acollidors.

I és que primer a la UAB i després a l'IBE, el que sempre va importar van ser els companys del dia a dia, malgrat hàbilment ens veiessim un cop per setmana. Gràcies als membres del Butterfly Diversity and Evolution Lab, perque sempre he tingut la fantàstica sensació de fer treball en equip. Vlad, per tots els riures i projectes plegats. Marga, què faríem sense tu. Claudia, perque sent tant diferents ens portem tant bé. Raluca, Martha, per no haver aborrit l'únic noi del grup.

To everyone in the MCZ, for all the great personal and scientific moments together. Many thanks Rod, Lukas, Chris, Gabe, Daniel, Sarah, Jignasha, Jon, Ben, Line, Ian, Ada, Milan, Jack, Marianne, James, Wenfei, Chris, Leonora, Petra, Mark, Tiago, and a long list of sensational people I met at Harvard. To my other friends in Boston, specially Tina, Natassia and Bel, e para o resto dos meus amigos brasileiros, muito obrigado. Para Germán y su família, por acogerme siempre como a uno más.

Als companys dels laboratoris veïns de l'IBE, per fer-ho tot més fàcil i divertit, i per estar sempre a punt per si calia un cop de mà. Gràcies Víctor pels anys d'incalculable suport, tècnic i moral. Javi, veterà company de batalles, Ana R, Joan, Txus i Anna P, per fer de l'Institut un lloc més agradable i humà. Al personal d'administració de l'IBE, per ser els funcionaris més eficients i entregats que conec, gràcies Rita, Emiliano, Blanca i Anna, per ajudar-me amb tots els maldecaps burocràtics.

I res no és sinó hi ha recolzament de base, que sinó rutllés, segurament ens quedaríem tancats a casa. Som el que ens envolta:

Als meus PARES, que van donar-me el valor de la curiositat. Perque em van fer pujar una muntanya, entrar a una cova i em van comprar un caçapapallones. Al meu germà, que va entendre que conegués el meu nebot a distància, i a tota la família, que sovint no esperant gaire resposta, insistien en preguntar com anava tot.

A tu, LAURA, que sempre vas tenir aquest paràgraf reservat. Perque vas ser la millor amiga, perque em vas inspirar, perque et trobo a faltar. Sempre junts.

El meu record també avui per algú que va tenir algo a veure amb que veiés apassionant això de la biologia. Gràcies JORDI, perquè em vas donar una visió diferent de tot plegat. No t'oblido.

A tots els AMICS, per sort una llarga llista de qualitat d'una i altra procedència, amb els qui el temps simplement no passa, sinó que és. Muntanyencs, biòlegs, companys de vida, de pis, de clubs, de corda, de festes, de viatges... gràcies Guillem, Marsi, Dani, Vicenç, Ferran, Aritz, Núria T, Jordi S, Jovita, Cristina, Martí, Mònica, Roger, Pere, Dídac, Sandra, Isis, Eli, Anna, Joan, Jep, Montse, Iona, Marcos, Iuli, Xus, Núria M, David, Albert, Laura R, Àlex, Mar, Carol, Jordi M..., perque hem anat creixent junts.

Finalment, a vosaltres, muntanyes, per ser tant blanques.

Contents

Introduction

1. Tree thinking and its importance in evolutionary theory

Living organisms share **homologous characters** that are inherited from **common ancestors**, which are the entities allowing to link organisms according to their evolutionary relationships along a temporal scale. Both homology and common ancestry are essential concepts for understanding evolution and both are the basic units represented in a **phylogenetic tree.** A phylogeny, or evolutionary tree, represents the evolutionary relationships among groups of organisms (i.e. taxa), where the tips represent descendent taxa and the nodes the common ancestors of those descendants. Their relative branching order is known as **topology** and **branch lengths** can represent amount of character change or evolutionary time.

Such a *tree-thinking* approach has been the basis for the modern understanding of biological diversity, which was laid out by Darwin (1859) with his concept of *descent with modification*. In 1837, shortly after his voyage on the Beagle, Darwin sketched a tree diagram in a notebook representing his emerging idea (Figure 1a). Few years later he published *On the Origin of the Species* (1859), which included a single illustration: a phylogenetic tree (Figure 1b):

"...the great Tree of Life, which fills with its dead and broken branches the crust of the earth, and covers the earth with ever-branching and beautiful ramifications" (Darwin 1859, p.159)

The emergence of Darwin's ideas was certainly influenced by other naturalists of his time, such as Jean Baptiste de Lamarck (1744-1829), who is actually credited with the first, but incorrect, scientific theory of evolution. He used a tree-like form to explain his thoughts, although he assumed spontaneous generation and posterior inheritance of changes and not common ancestry. Erasmus Darwin (1731-1802), Charles Darwin's grandfather, put forward some first hints on the idea of common ancestry:

"*would it too bold to imagine, that all warm-blooded animals have arisen from one living filament..."* (Darwin 1794, Sect IV.8)"

And the geologist Charles Lyell (1797-1985):
"*We know that individuals which are mere varieties of the same species, would, if their pedigree could be traced back far enough, terminate in a single stock; so according to the train of reasoning before described, the species of a genus, and even the genera of a great family, must have had a common point of departure. What then was the single stem from which so many varieties have ramified?* (Lyell 1832, p. 10)".

Tree-thinking also states that the natural system of classification of the organisms might be strictly genealogical, that is, argumentations might be based on vertical but not horizontal comparisons of living organisms (Baum *et al*, 2005; Baum and Smith, 2012). Linnaeus' *Systema Naturae* (1735), which forms the baseline of the current nomenclature system, reflected similarity of organisms in natural groups based on horizontal thinking (i.e. without a causal explanation). The first classification of living beings based on Darwin's evolutionary thought was presented by Haeckel (1866) (Figure 1c). He notably introduced important concepts such as **monophyly**, although he denoted to have an implicit *ladder thinking*, where living species represented a continuum of forms of different degrees of advancement or complexity (Baum and Smith, 2012; Omland *et al*, 2008). Other forms of ladder thinking attempted for example to organize living organisms into "phyletic series" (Bessey, 1915), where some extant groups were identified as ancestors of others. These examples illustrate only a few of several misinterpretations of Darwin's ideas during the past two centuries. During the second half of the 20th century a fervent debate took place regarding the appropriate use of phylogenetic trees in taxonomy. On one side, the classic evolutionary systematics school (or Dawinian systematics) defended that biological classifications might reflect both history and evolutionary processes, therefore a combination of inferred relationships and degree of similarity should be considered. On the other side, phenetics (or numerical taxonomy) argued that classification should only consider phenotypic similarity (Sokal and Sneath, 1963; Sneath and Sokal, 1973; Hull, 1998). From the 1930s to the 1960s, Willi Hennig (1913-1976) and Walter Zimmerman (1892-1980) developed the field of **phylogenetic systematics** (Hennig, 1966). This was founded on the premise that biological classifications might be based merely on common ancestry and not on similarity, as these do not necessarily co-occur. Thus, classification of organisms moved slowly from using similarity, as a subjective but straightforward criterion towards adopting phylogenetic kinship, as an objective but more complex criterion to be measured. With the great statistical and methodological improvements in the 1980s and 1990s, phylogenetic systematics expanded and finally prevailed (Baum and Smith, 2012). Since then, the field of phylogenetics has experienced an exponential growth, thanks to the great advances in molecular evolution, the consequent use of molecular-based phylogenies, the computer boom and the specific methodologies and software developed. Nowadays, phylogenetics has become an essential tool in many biological fields, in addition to systematics.

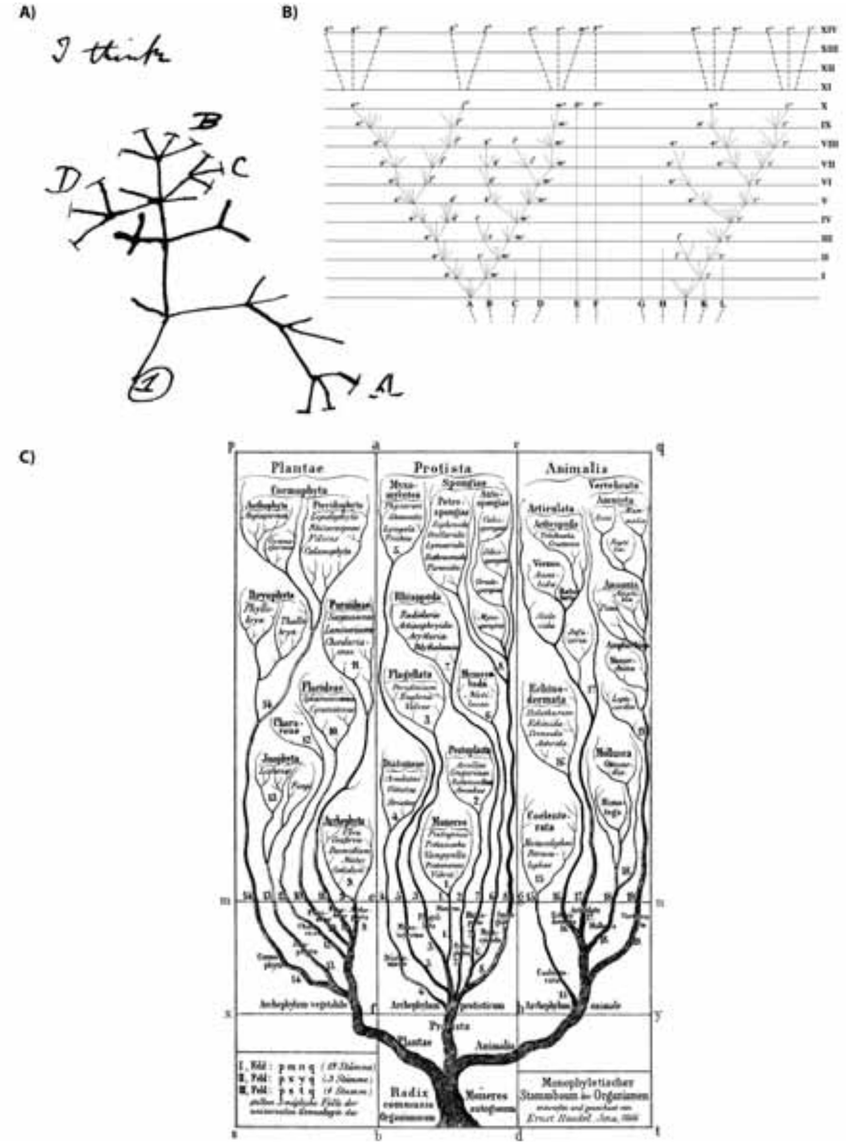


Figure 1. A) A first tree sketch from Charles Darwin in 1837, B) The only illustration in On the Origin of Species (1859) and C) Haeckel´s classification of living organisms (1866).

2. The phylogenetic process

There is a long way to go before obtaining a phylogenetic tree, and it cannot be accomplished in a single step (Figure 2). Firstly, it is required to identify homologous characters that can be comparable among specimens and, if the characters are based on molecular sequences, homologous positions (from homologous sequences) need to be aligned. Once a data matrix or molecular alignment is available, an appropriate model of character evolution needs to be selected by statistical methods. Subsequently, phylogenetic inference can be performed under a variety of algorithms, and its reliability might be evaluated by assessing phylogenetic confidence. Once the tree is available, it can be used as a framework for testing evolutionary hypotheses using phylogeny-based methodologies.

Admittedly, a phylogenetic tree is an oversimplification that does not entirely represent reality. Consequently, phylogenetic inference has been developed from a statistical point of view, by incorporating methods for testing phylogenetic confidence. However, these methods assume that all previous steps (sampling, homology assignments, etc.) conform to reality. If this is not fulfilled, strongly supported erroneous trees can be obtained (without being detected as incorrect). Eventually, resulting trees can be misleading due to accumulative sources of error in the entire phylogenetic process. First, biological sampling can be somehow biased, due to surveying only a limited amount of variability that does not truly represent a population or species, to collecting rare specimens or hybrids, or to ignoring intermediate lineages. The assessment of homologous characters can be a source of problems as well. For molecular data, only orthologous genes should be used to build phylogenies, and the discrimination of true orthologs (homologous genes separated by speciation) from paralogs (genes related by duplication events) has resulted in serious challenges especially for the most ancient phylogenies potentially accumulating large numbers of gene duplications and gene losses (Fitch, 1970; Remm *et al*, 2001; Hulsen *et al*, 2006; Chen *et al*, 2007; Hubbard *et al*, 2007; Creevey *et al*, 2012). Next, the alignment process may involve important sources of error too that will increase progressively according to the relative divergences and total length of gene sequences. In fact, aligning molecular sequences, either DNA or protein, has arguably a strong impact on the phylogenetic reconstruction process, albeit it is the piece of the puzzle on which less attention is usually paid (Castresana, 2000; Ogden and Rosenberg, 2006; Talavera and Castresana, 2007; Penn *et al*, 2010a,b; Wu *et al*, 2012). Before running the phylogenetic algorithm, a proper selection of the best-fitting evolutionary model needs to be performed (Bos and Posada, 2005) and, once more, the choice of model of evolution may substantially change the results of a phylogenetic analysis (Sullivan and Joyce, 2005). Finally there are also several artifacts associated with the phylogenetic reconstruction itself, mainly due to the degree of sequence divergence (i.e. when sequences are either too conserved and contain very few substitutions, or too diverged and experience substantial substitution saturation (Rodriguez-Ezpeleta *et al*, 2007)), due to base-compositional and heterotachy biases (Hassanin *et al*, 2005; Philippe *et al*, 2005a; Pagel and Meade, 2008), or due to the artificial grouping of the most rapidly evolving lineages (long-branch attraction) (Felsenstein, 1978; Bergsten, 2005; Brinkmann *et al*, 2005). By all means, the phylogenetic process as a whole represents a serious challenge and its accuracy highly depends on the correct resolution of the different steps, which have and additive influence on the final result (Figure 2).

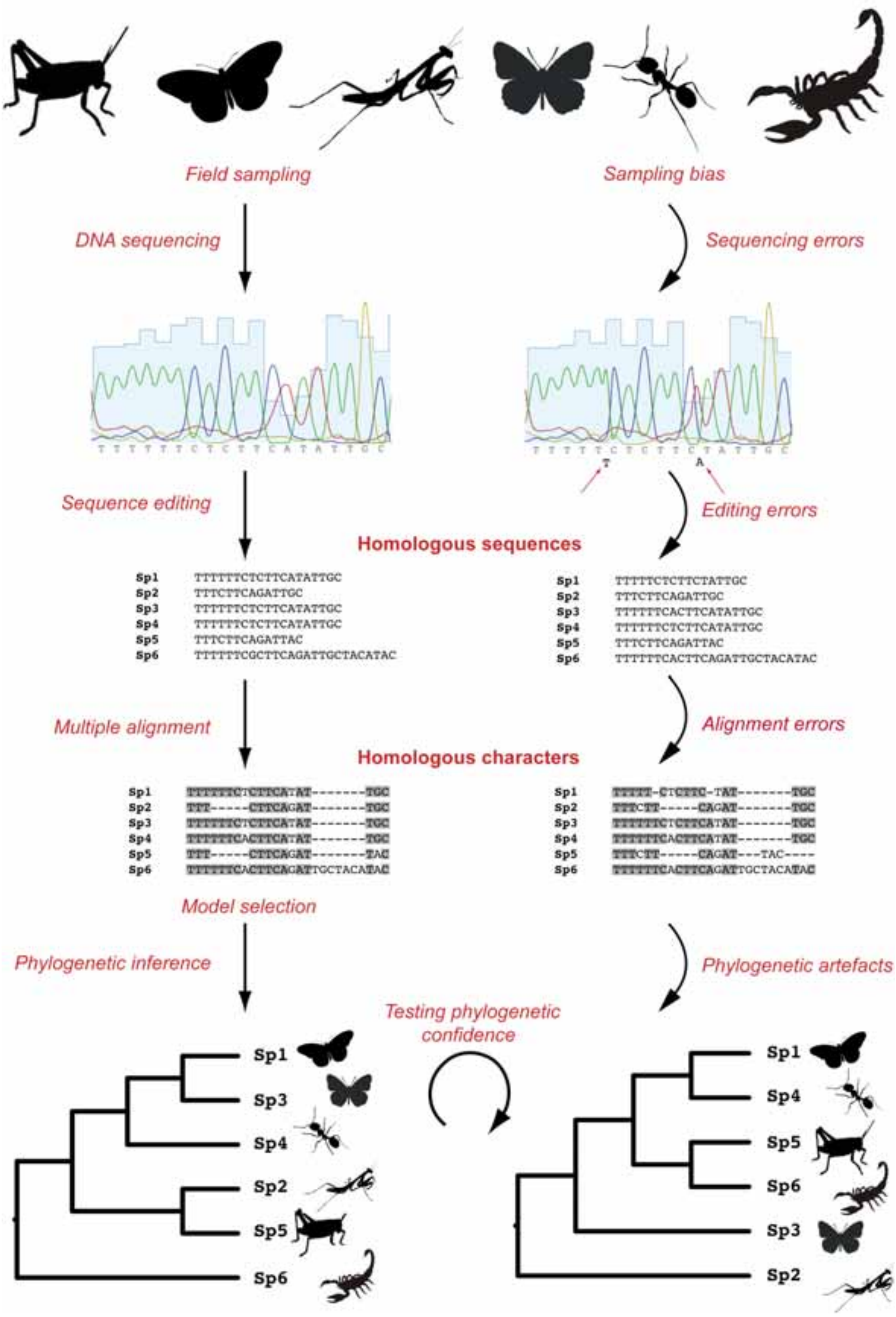

Field sampling
Sampling bias
DNA sequencing
Sequencing errors
T T T T T T C T C T T C A T A T T G C
T T T T T C T C T T C T A T T G C
T
A
Sequence editing
Editing errors
Homologous sequences
Sp1 TTTTTTCTCTTCATATTGC
Sp2 TTTCTTCAGATTGC
Sp3 TTTTTTCTCTTCATATTGC
Sp4 TTTTTTCTCTTCATATTGC
Sp5 TTTCTTCAGATTAC
Sp6 TTTTTTCGCTTCAGATTGCTACATAC
Sp1 TTTTTCTCTTCTATTGC
Sp2 TTTCTTCAGATTGC
Sp3 TTTTTTCACTTCATATTGC
Sp4 TTTTTTCTCTTCATATTGC
Sp5 TTTCTTCAGATTAC
Sp6 TTTTTTCACTTCAGATTGCTACATAC
Multiple alignment
Alignment errors
Homologous characters
Sp1 TTTTTTCTCTTCATAT-------TGC
Sp2 TTT-----CTTCAGAT-------TGC
Sp3 TTTTTTCTCTTCATAT-------TGC
Sp4 TTTTTTCACTTCATAT-------TGC
Sp5 TTT-----CTTCAGAT-------TAC
Sp6 TTTTTTCACTTCAGATTGCTACATAC
Sp1 TTTTT-CTCTTC-TAT-------TGC
Sp2 TTTCTT-----CAGAT-------TGC
Sp3 TTTTTTCTCTTCATAT-------TGC
Sp4 TTTTTTCACTTCATAT-------TGC
Sp5 TTTCTT-----CAGAT---TAC----
Sp6 TTTTTTCACTTCAGATTGCTACATAC
Model selection
Phylogenetic inference
Phylogenetic artefacts
Testing phylogenetic confidence
Sp1
Sp3
Sp4
Sp2
Sp5
Sp6
Sp1
Sp4
Sp5
Sp6
Sp3
Sp2

Figure 2. Summary of the different steps that needs to be accomplished to obtain a phylogeny. Potential sources of errors and how they can derive in erroneous phylogenies are shown in the right column.

2.1. Methodological challenges in phylogenetics: from saturation to reticulation

Phylogenetics is a broad field that can eventually face very different levels of evolutionary relationships. Thus, depending on the taxon-level analyzed, the phylogenetic process previously described needs to be accordingly adapted. The main goal of any phylogenetic study is to skillfully make use of the available tools to extract the maximum **phylogenetic signal** from the data. Phylogenetic signal can be defined as the amount of phylogenetically significant information (i.e. synapomorphies or shared traits among taxa and their most recent common ancestor) that is not blurred by phylogenetic noise (i.e. homoplasies or common characters among unrelated taxa), which does not represent real evolutionary relationships. Generally, phylogenetic noise increases at the two extremes of the divergence spectrum. On one side, datasets including too divergent organisms may display recurrent character changes or molecular substitutions that will finally become random, and will not allow inferring the total number and the entire succession of changes. This phenomenon is known as **substitution saturation**, and it can dramatically decrease the phylogenetic signal contained in molecular sequences. The ultimate reason is that as homoplasies increase, the substitution model underestimates evolutionary rates, and in the extreme case when sequences experience full substitution saturation, sequence similarities will entirely depend on the similarity in nucleotide frequencies (Steel *et al*, 1993; Xia, 2001; Xia *et al*, 2003). The divergences at which saturation arise may vary depending on the organisms, their inherent evolutionary history, and the genetic markers used. Additional troubles arise when speciation events are closely spaced in time, in the form of very short internal tree branches, where it is rather difficult to detect phylogenetic signal (Philippe *et al*, 2011; Philippe *et al*, 1994; Saitou and Nei, 1986). In general terms we can relate substitution saturation to basal (or ancient or deep-level) phylogenies and, to highlight some examples, we can certainly mention the conflicting phylogenetic reconstruction of the major arthropod groups (Regier *et al*, 2008; Meusemann *et al*, 2010; Rota-Stabelli *et al*, 2011; von Reumont *et al*, 2012) or, even deeper, that of the early metazoan origins (Medina *et al*, 2001; Philippe *et al*, 2005b; Baurain *et al*, 2007; Dunn *et al*, 2008; Philippe *et al*, 2009; Schierwater *et al*, 2009; Philippe *et al*, 2011).

On the other hand, datasets including small amounts of divergence or closely related taxa will push the phylogenetic inference onto completely different obstacles. In an extreme case, the lack of phylogenetic signal in the genetic markers (or other characters), i.e. the absence of parsimony-informative changes, will result in polytomic or unresolved phylogenies. Moreover, in this type of datasets, it is common to find conflicting signals in the form of **genealogical discordance** among markers, **reticulation** between lineages or both and, as a consequence, there might be multiple equally correct trees relating any given set of tips. This phenomenon needs to be explained under the coalescent theory framework, where different genes have rather different genealogical histories because they are not yet fixed across recently diverged clades. The fact of sampling genes before they coalesce is known as **incomplete lineage sorting** (ILS) or retention of ancestral polymorphisms. In the absence of ILS the process is strictly treelike, with population lineages that split but never reconnect. However, population lineages sometimes adopt a netlike or reticulate pattern that track genealogies differing from the history of the population, wherein different genes sampled from the same set of tips exhibit different trees. Apart from ILS, other phenomena can cause population reticulation as lateral gene transfer, genetic introgression or lineage fusion (Figure 3). *Lateral gene transfer* (LGT) occurs when a small piece of the genome is transferred between organisms by a process other than sexual reproduction. While it is commonly cited as a regular process in prokaryotes (but see Soria-Carrasco and Castresana, 2008), it seems to be very rare in animals and plants. *Introgression* occurs when organisms from distinct, already coalesced, population lineages come into contact and sexually reproduce, producing hybrid offspring that then breed with members of one or the other parental populations. *Lineage fusion* could be considered as an extreme case of introgression, and it is often called hybrid speciation. In this case two formerly distinct populations are merged into a single descendant lineage.

Trying to deal with all these challenges, phylogenetic methodologies are being continuously updated. The sophistication of each step of the process aims to avoid biases and to recover phylogenetic signal. The adopted strategies need to adjust to the specific datasets and hypotheses, as the use of more realistic evolutionary models, the filtering and selection of appropriate phylogenetically informative markers or the implementation of the population genetic models through the coalescent theory. Examples within the entire spectrum, including both extremes, are analyzed in this thesis.

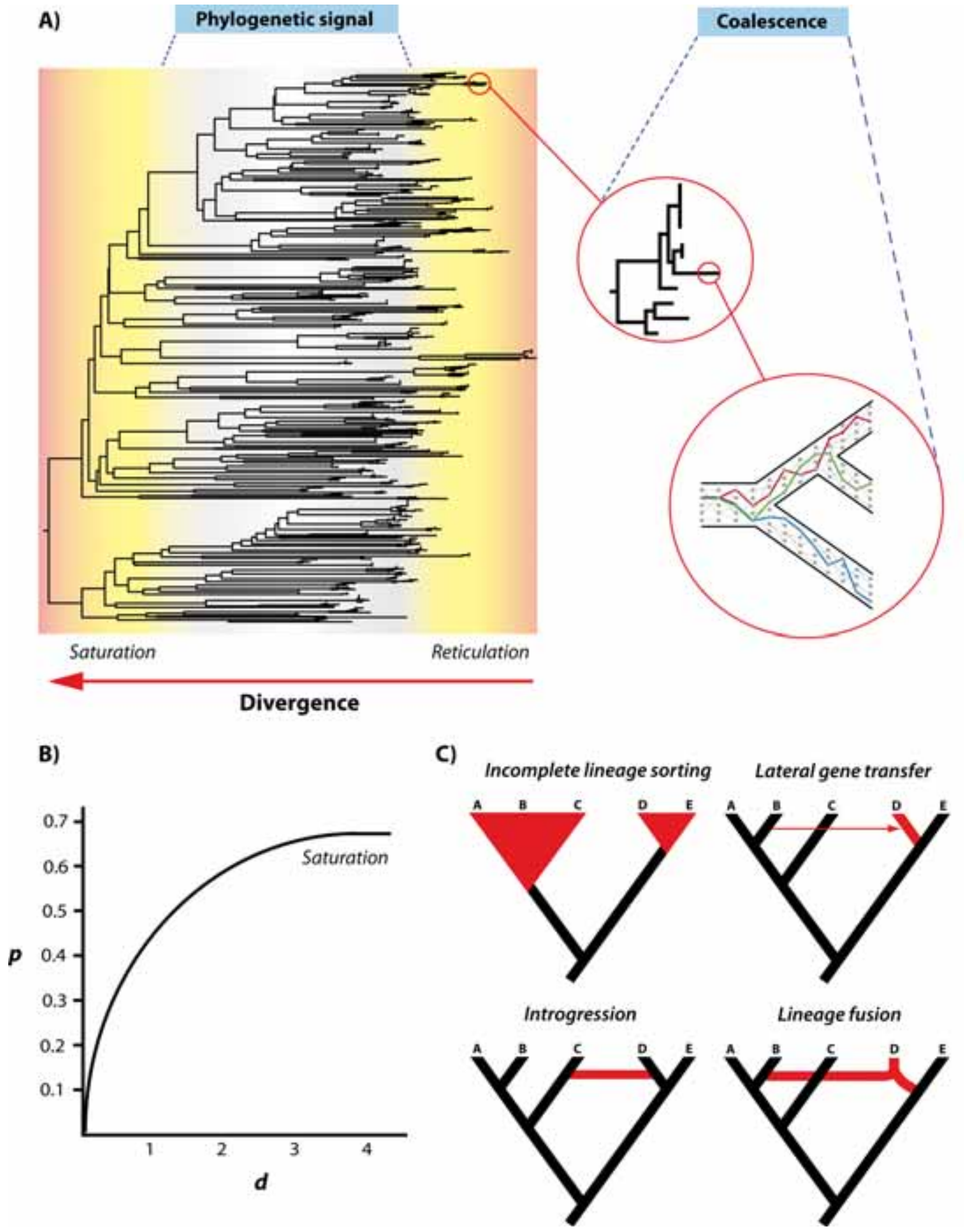


Figure 3. The phylogenetic spectrum: A) The range of optimal phylogenetic signal along a tree depends on the divergence level that it will be blurred either by reticulation at low divergences or by saturation at deep divergences. B) Substitution saturation arises when the observed *p*-distance underestimates the true genetic distance *d*. C) Different origins of reticulation at low genetic divergences.

2.2. Data sources for phylogenetic inference

Character selection for phylogenetic analysis can be a complex issue. It is fundamental that any set of characters accomplishes the principle of homology (i.e. they are inherited from a common ancestor). Different kinds of data have become useful for phylogeneticists. The classical way of estimating species relationships is to compare their **morphological characters** and, indeed, the vast majority of the current taxonomic classification system is entirely based on morphology. Nowadays, though, molecular data is the most regularly used source for inferring phylogenies, even if morphology remains highly relevant for specific evolutionary questions, and indeed both sources combined have been sometimes proved to increase phylogenetic signal (Giribet *et al*, 2002; Wiens, 2004; Nylander *et al*, 2004; Wahlberg *et al*, 2005; Wortley and Scotland, 2006). Also, other types of information have been occasionally used in phylogenetics, such as ecological or behavioral traits (e.g. host plants, lifestyle condition, habitat preferences, animal song, etc.), which in certain cases have been described to accurately reflect evolutionary history (Atz, 1970; Wenzel, 1992; Miller and Wenzel, 1995; Grandcolas *et al*, 2001; Cap *et al*, 2008; Ord and Martins, 2010).

Whether the morphological or molecular approach is preferable for any particular evolutionary question was greatly debated during the 90s (e.g. Patterson *et al*, 1993), when the generation of molecular data grew exponentially. It is now evident that molecular data offer several advantages over morphological data. First, molecular data provide, a priori, a greater confidence of homology, since changes occur at the most basic level of organization. In contrast, morphological changes are undeniably a product of countless changes at the lower levels of organization that are difficult to measure. Second, gathering molecular data has become an extremely fast and accessible way of generating information, something that has encouraged researchers from a wide variety of fields to make use of the growing phylogenetic techniques, including biochemistry, ecology, molecular evolution or virus evolution, and as a consequence the use of morphological characters has been practically restricted to the field for which phylogenetics was initially build, the reconstruction of the tree of life. Morphological characters play a fundamental (or almost exclusive) role, in phylogenetic studies of extinct lineages from fossils or museum specimens. Even in this field, although only very recently, morphology is shadowed by modern laboratory techniques capable of recovering **ancient DNA** (aDNA) (Shapiro and Hofreiter, 2012; Huynen *et al*, 2012). Indeed, the entire genome information for legendary extinct species like the mammoth (Miller *et al*, 2008) or the neanderthal (Green *et al*, 2010), or mitochondrial genomes of recently extinct taxa (Cooper *et al*, 2001; Willerslev *et al*, 2009; Krause *et al*, 2010) has been obtained, as well as much more ancient molecules corresponding to undetermined insect DNA contained in the sediments from arctic permafrost (Thomsen *et al*, 2008). Target types of molecular data for phylogenetics are pretty variable, either in the source or in the method of obtention. We can cite: complete or partial codifying sections of the genome (or their translated protein sequences), non-coding DNA markers, microsatellites, AFLPs, randomly genomic ultrasequenced SNP variation, entire mitogenomes, and others. In addition, different mechanisms of acquiring molecular variation have led to the actual biodiversity, and although nucleotide single nucleotide mutations are the most commonly used, other sources as gene duplications, genomic reorganizations or genetic exchanges as recombination, introgression, reassortment or horizontal gene transfer are also susceptible to be used as phylogenetic criteria.

2.3 Phylogenetic algorithms and tree robustness

After more than three decades of methodological improvements, several tools are nowadays available to infer phylogenetic trees. Conventionally it is required to implement three consecutive steps: 1) constructing a data matrix, 2) identifying the trees mostly compatible with the data and 3) evaluating the confidence of the resulting phylogeny by statistical assessment.

Phylogenetic methods can be classified into three main categories: distance-based, parsimony, and probabilistic methods, which include Maximum Likelihood and Bayesian approaches.

Distance methods first convert the character matrix into a distance matrix that represents the evolutionary distances between all pairs of species. Next, a tree is calculated, generally by applying clustering methods as Neighbor Joining (NJ) (Saitou and Nei, 1987) or UPGMA (Sokal and Michener, 1958), minimum evolution (ME) (Rzhetsky and Nei, 1992) or least square (Fitch and Margoliash, 1967; Cavalli-Sforza and Edwards, 1967). Pairwise distances (p-distances) can be converted to genetic or evolutionary distances by employing an evolutionary model that corrects for multiple hits. The big advantage of distance methods is that they are computationally inexpensive, but they perform poorly at high levels of genetic diversity (Felsenstein, 2004; Yang, 2006), so their use is mostly restricted to datasets including large amounts of taxa.

Maximum parsimony (MP) is a discrete-character-based method that selects the tree that requires the minimum number of character changes to explain the observed data (Kluge and Farris, 1969; Farris, 1970; Fitch, 1971). In other words, parsimony methods operate by selecting trees that minimize the total tree length: the number of evolutionary steps required to explain a given set of data. To infer maximum parsimony trees, the minimum number of substitutions at each informative site is calculated for each possible tree, to finally sum the number of changes over all the informative sites for each tree and choose the tree associated with the smallest number of substitutions. Parsimony analysis was the predominant approach used to construct phylogenetic trees from the early 1970s until relatively recently, to be progressively replaced by probabilistic model-based methods. Since the parsimony principle in biology assumes minimum evolutionary change, it has been often justified mostly on philosophical rather than statistical grounds over model-based methods, which introduce mathematical models of character evolution. Also, parsimony presents some methodological limitations (e. g. misleading results if homoplasy is common, underestimation of branch lengths), however it is a fast method that has been demonstrated to be quite effective in specific situations (e.g. Hillis, 1996), very versatile to perform analyses combining diverse character types, and still offering an alternative when model-based methods cannot be used due to computational limitations.

Because of their statistical power, probabilistic methods like **Maximum Likelihood** (ML) and Bayesian inference (BI) are the major methods currently used for molecular phylogenetics. The likelihood of a tree is the probability of observing the data (the sequences matrix) given the hypothesis (the tree and the model of DNA evolution). The maximum likelihood criterion involves searching for the tree and the model parameters that maximise the likelihood of obtaining the observed data (Lewis, 1998). The likelihood of a tree is then defined as:

$$L(T) = \Pr(D \mid T, Q)$$

Where the probability of the data (D) can be efficiently calculated given a phylogenetic tree (T) and a probabilistic model of molecular evolution (Q). Thus, the model parameters, topology and branch lengths that maximize the likelihood of the data are searched.

In a different way, **Bayesian inference** (Rannala and Yang, 1996; Huelsenbeck *et al*, 2001) follows Bayes' theorem to calculate the **posterior probability** that a hypothesis (tree topology and evolutionary model) is explained by the data (D), which will depend on the likelihood of the tree and its prior probability. The prior probability of a particular tree is the probability (before looking at the data) that among all the possible trees this would be the true tree, and the denominator (prior probability of the data) is the probability of the data averaged over all possible parameter values weighted by their prior distribution.

$$\overset{\text{Posterior probability}}{\Pr(T \mid D)} = \frac{\overset{\text{Likelihood}}{\Pr(D \mid T, Q)}\,\overset{\text{Prior probability of the hypothesis}}{\Pr(T, Q)}}{\underset{\text{Prior probability of the data}}{\Pr(D)}}$$

Since Bayesian inference implies complex parameter-rich estimations and consider entire distributions of parameters values, the methods rely on the Markov chain Monte Carlo (MCMC) algorithm by subsampling the posterior distribution landscape. The main advantage of Bayesian inference over ML is that it allows accommodating phylogenetic uncertainty by providing entire posterior distributions rather than a single phylogeny.

Thus, posterior probabilities in BI provide a straightforward way to evaluate phylogenetic robustness. Another commonly employed procedure for testing phylogenetic confidence is the use of bootstrapping, which can be applied on any method of phylogenetic inference and data source. The **non-parametric bootstrap** is a technique based on sampling characters with replacement (Felsenstein 1985a; Efron *et al* 1996). The method involves initially sampling with replacement of *n* sites among *n* sites, and then tree reconstruction based on the resampled data matrices. This procedure is repeated a large number of times and the results are summarized using a consensus tree with a percentage indicating the proportion of the sampled trees in which each clade is present. This percentage is the bootstrap value and it provides a measure of statistical support for each branch of the tree, with values greater than 75% typically considered as satisfactory. Whether or not there is some conflict among characters, bootstrapping reflects the characteristics of the phylogenetic signal in a given data set. Following a different procedure, support values can be calculated from a large number of replicates obtained by using Monte Carlo simulation (Goldman, 1993; Huelsenbeck and Crandall, 1997), a measure known as **parametric bootstrap**. As an alternative to bootstrapping and Bayesian posterior probabilities for branch support, a young group of parametric methods for ML based on **approximate likelihood-ratio tests** (aLRT) have been developed (Anisimova and Gascuel, 2006). These methods basically perform comparisons between the likelihoods of two given hypotheses from the distributions of generated trees, and their main advantage is a much faster speed, which makes them especially useful for large datasets. Finally, among the morphologist community, mostly integrated by paleontologists, a popular parsimony-based method is to calculate the **Bremer index** (Goodman *et al*, 1982; Bremer, 1998), which estimates the number of steps that need to be added to the most parsimonious tree in order to find a contradicting clade. Thus, the greater the number of steps, the stronger the support for a clade.

It is important to note that none of the methods described is exempt from controversy. In general, statistical indices only assess sampling effects, and give an indication of tree reliability that is conditional on the data and the model assumptions (Huelsenbeck and Rannala, 2004; Simmons *et al*, 2004; Yang and Rannala, 2005; Cummings *et al*, 2003; Buckley and Cunningham, 2002; Lewis *et al*, 2005). To add some uncertainty to the phylogenetic process, thus, the possibility that incorrect trees receive strong statistical supports needs to be considered, even if those cases are extremely rare.

2.4 Models of evolution

Any model in science represents an abstraction of complex natural processes in order to make them mathematically tractable and hence useful to make reasonable predictions (extrapolations) about the outcome of the studied process or system under different scenarios. In the context of molecular phylogenetics, models are used to make predictions about the substitution process in molecular sequences along the branches of a tree. In phylogenetics, the use of correct sequence substitution models in every case is crucial (Bos and Posada, 2005; Posada, 2009). Sequence substitution models describe in probabilistic terms the process by which a sequence of characters, either nucleotides or aminoacids, changes into another set of homologous character states over time. Current available models have been built 1) empirically, where properties are inferred from comparisons of a large data amount of data and parameters are fixed and applicable to all the analyses, or 2) parametrically, by using chemical or biological properties of the molecules that allow to derive the parameters from our data (Lio and Goldman, 1998). Empirical models have been developed mainly for aminoacids by analyzing the frequencies of observed substitutions in sets of alignments of conserved protein domains with varying degrees of evolutionary divergence. These have resulted in general substitution matrices like Dayhoff (Dayhoff *et al*, 1978), BLOSUM (Henikoff and Henikoff, 1992), JTT (Jones *et al*, 1992; Gonnett *et al*, 1992), WAG (Whelan and Goldman, 2001), LG (Le and Gascuel, 2008), MtRev for mitochondrial evolution or group-specific models like MtMam or MtArt. More recently empirical profile mixture models (Quang *et al*, 2008; Le *et al*, 2008; Lartillot and Philippe, 2004) have been developed, and have been shown to outperform standard replacement matrices (Le *et al*, 2008). Parametric substitution models have been developed more intensively for nucleotide sequences. Such models consider the nucleotide frequencies and the instantaneous rates of change among them. There is a large list of possible substitution models depending on the degree of parameterization. The JC69, K80 or K2P, HKY85, TN93 and the General Time-Reversible (GTR) are among the most widely used, with GTR being the most parameter-rich model.

The sophistication of the models can be increased relaxing some of the implicit assumptions. For example, the inference of the proportion of invariable sites (p-inv, +I) can reflect the fact that some characters do not change over time. Also, the rate of nucleotide substitution is assumed to be the same for all positions in the models considered, however this rarely holds, and rates varies from site to site. Thus a rate heterogeneity correction, where the variance of the substitution rate among sites becomes more flexible, can be done if using a gamma distribution (+G).

The procedure for selecting the best fitting model for particular datasets requires statistical evaluation between two alternative models, which can be done using techniques like hierarchical Likelihood Ratio Test (hLRT), information criteria (as Akaike Information Criterion - AIC) or bayesian approaches (as Bayes Factors, posterior probabilities or the Bayesian Information Criterion –BIC).

2.5. Phylogenomics and species trees: making sense of multiple loci

Seminal phylogenetics was conceived to deal with a limited amount of characters, and phylogenetic reconstruction was originally often based on single molecular markers. This rapidly proved to introduce new limitations and the use of multiple markers was progressively incorporated as a solution, to ultimately, try to accommodate the full genome sequencing techniques, which are nowadays nearly prevailing. In this context of multigene studies emerged the term phylogenomics and the multi-locus coalescent-based methods, also known as species trees inference.

Phylogenomics can be defined as an extension of phylogenetics that uses its main principles in order to integrate and make sense of genomic data (Eisen and Fraser, 2003; Delsuc *et al*, 2005). The term ‘*genomic data*’ in this context is rather broad. Some authors refer to phylogenomics when using any subjectively-defined large number of genetic markers obtained by classic Sanger sequencing. Others tend to make use of it exclusively when dealing with data generated from modern next-generation sequencing techniques. That commonly includes complete genomes (Henz *et al*, 2005; Gorecki and Tiuryn, 2007; Rannala and Yang, 2008; Sims *et al*, 2009). To date, 8373 complete genomes have been sequenced or projected (2229 eukaryotes (213 insects), 14259 prokaryotes and 3219 viruses), and the list grows rapidly. But there are also other approaches more phylogeny-directed that allow the retrieval of large amount of optimal informative markers from non-model organisms, such as RADtag sequencing (Baird *et al*, 2008; Peterson *et al*, 2012; Rubin *et al*, 2012), or sequence capture (Lemmon *et al*, 2012; McCormack *et al*, 2012; Faircloth *et al*, 2012). Because it uses many characters, the phylogenomics approach drastically reduces stochastic or sampling errors associated with a finite length of single genes in traditional phylogenetic analyses (Rokas *et al*, 2003; Delsuc *et al*, 2005; Rokas and Carroll, 2005; Ciccarelli *et al*, 2006). What had to be a straightforward improvement, though, opened the door to new discussions and methodological challenges. An earlier debate highlighted that differences in individual gene rates of evolution or particular gene histories cast doubt on the convenience of concatenating genes in a **supermatrix** (de Queiroz and Gatesy, 2007), under the principle of total evidence (Kluge, 1998; Barrett *et al*, 1991; Rieppel, 2005). Consequently, the necessity of using partitioned datasets with independent evolutionary models fitting each locus has become the regular procedure. Even so, phylogenomics are not immune to systematic errors, which highly depend on data quality and the inference methods employed (Jeffroy *et al*, 2006; Rokas *et al*, 2006; Rodriguez-Ezpeleta *et al*, 2007; Leigh *et al*, 2011). Efforts to improve data quality of large phylogenomic datasets emphasize on filtering methods that allow the selection of genes containing the minimum amount of non-phylogenetic signal, including the proper detection of ortholog genes or the performance of saturation tests (Steel *et al*, 1993, 1995; Xia and Xie, 2001; Xia *et al*, 2003). Additionally, other approaches follow the so-called *gap-method*, based on removing specific fast-evolving sites or problematic taxa from supermatrices and filling the holes with gaps or missing data. Another well-known, but controversial (Gatesy *et al*, 2002; Bininda-Emonds *et al*, 2003; Bininda-Emonds, 2004; Gatesy *et al*, 2004) methodology in this context is the construction of **supertrees**, where a single phylogenetic tree results from the combination of several smaller, and independently inferred trees, from different datasets (either different taxa and/or markers) (Bininda-Emonds *et al*, 2002; Bininda-Emonds, 2004). Last but not least, other recent improvement on evolutionary models are resulting very efficient for large phylogenomic datasets (see chapter 2.4). It is worth to mention that, although sequence-based methods are the most commonly employed, **comparative genomics** approaches can also be of phylogenetic utility. Methods comparing gene content (specific genes found in a genome) or gene order (Wolf *et al*, 2002), as well as methods focusing on rare genomic changes (RGCs) like insertions and deletions (indels) or retroposon (SINE and LINE) integrations and gene fusion and fission events (Rokas and Holland, 2000; Gribaldo and Philippe, 2002) are some of the possibilities.

While phylogenomics mostly aim to resolve relatively deep phylogenetic relationships, at high taxonomic levels, **species trees** methodologies deal with inferring relationships among closely related species or populations. It was expected that the lack of resolution due to gene tree conflicts would be reduced or totally removed by massive concatenation of independent loci (Chen and Li, 2001; Rokas *et al*, 2003; Gadagkar *et al*, 2005; Rokas and Carroll, 2005; Wu and Eisen, 2008), but it still remained unsound in some cases (Carstens and Knowles, 2007; Kolaczkowski and Thornton, 2004; Kubatko and Degnan, 2007; Mossel and Vigoda, 2005). Certainly, the greatest contribution arises when fully integrating the coalescent theory to coestimate multiple gene trees embedded in a shared species tree using the so-called **multispecies coalescent** model (Rannala and Yang, 2003; Maddison and Knowles, 2006; Carstens and Knowles, 2007; Liu and Perl, 2007; Ané *et al*, 2007; Kubatko *et al*, 2009; Heled and Drummond, 2010; Larget *et al*, 2010), which extends the basic coalescent model by incorporating populations of multiple species connected by an evolutionary tree..

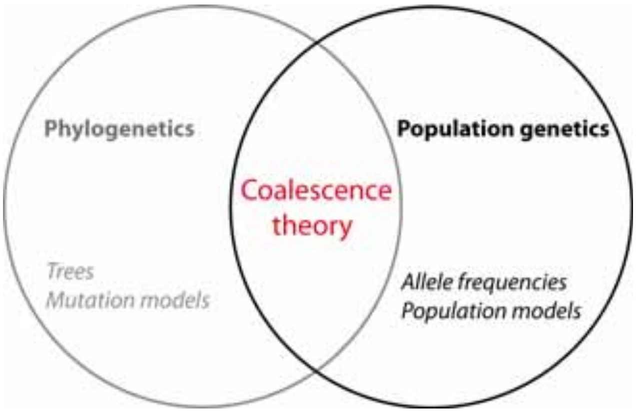


Figure 4. The coalescent theory integrates elements from phylogenetics and population genetics.

The *n-coalescent* is based on a mathematical theory that was conceived by Kingman (1982) to describe the genealogy of genes by looking backwards in time. He showed the behavior of *n* gene copies, when two randomly chosen copies coalesce to their most recent common ancestor at each successive preceding generation.

$$\textit{coalescence rate} = \binom{n}{2} = \frac{n(n-1)}{2}$$

The theory provides the equations that describe the distribution of the coalescent times (coexistent times for a particular set of lineages) in a genealogy. In a simplistic situation of a haploid population assuming no selection, no recombination and a constant population size, the average time (t, in number of generations) required so that *n* lineages to coalesce into *n*-1 for a total population size N (in number of genes) is:

$$E(t_i) = \frac{2N}{[n(n-1)]}$$

And the average time (T) for *n* genes to be coalesced into their most common recent ancestor (MRCA) is:

$$E(T) = 2N\left[1 - \left(\frac{1}{n}\right)\right]$$

Since the time to the most common recent ancestor (T) is directly proportional to N (total number of genes), the coalescent theory allow to estimate population size parameters and divergence times in its simplest version, while recent extensions of the basic coalescent improve the sophistication of the equations and the software accommodating them, and allow to account for other crucial factors as migration rates, fluctuating population sizes (expansions and bottlenecks), selection, recombination, or mutation rates (Hey and Nielsen, 2004, 2007; Nielsen and Wakeley, 2001; Buckley *et al*, 2006; Joly *et al*, 2009, Meng and Kubatko, 2009; Kuhner, 2006; Beerli, 2004, Heled and Drummond, 2008).

To better illustrate the coalescent, some properties are exemplified in Figure 5. First, populational sampling effects can result in different topologies and TMRCA estimations for a gene tree, while different gene trees may provide conflicting topologies and divergence times. So, a coalescent-based species tree reconstruction will integrate genealogies from different gene trees into one species tree, providing accurate estimates of TMRCA and population sizes. It is important to note that in small-sized populations genes will coalesce earlier in time due to stochastic effects.

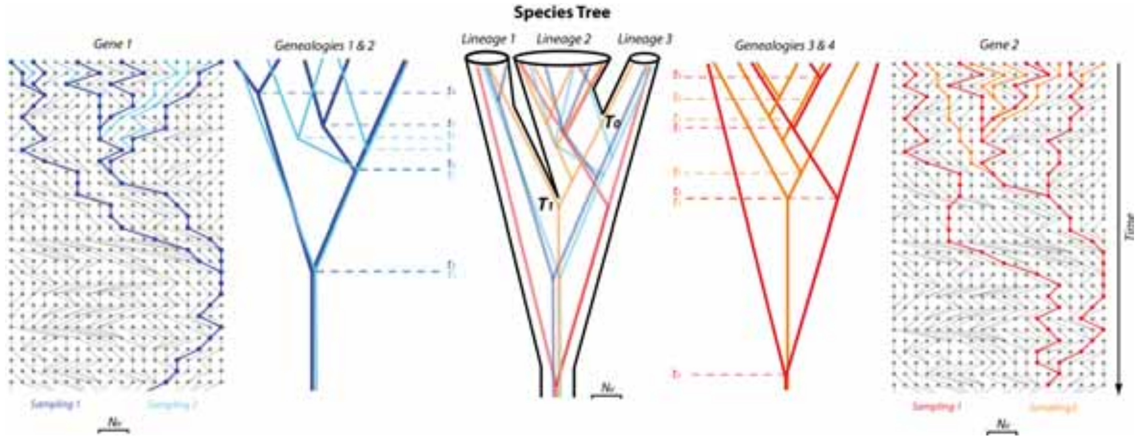


Figure 5. The multi-locus/multi-individual methods accommodate different genealogies from gene trees into a species tree.

2.6. The use of phylogenetic trees for natural history studies

Phylogenetic trees represent a unique evolutionary framework to test multiple biological hypotheses. Therefore, any phylogenetic scenario opens the door to multiple applications beyond systematic classifications, like phylogeographic studies, reconstruction of character evolution, comparative methods between evolutionary traits, inference of divergence times or studies of epidemiological dispersion, among others.

2.6.1 Inferring divergence times

Similarity among sequences provides, intuitively, some kind of information about their relatives divergence times, however, it would be practicable to obtain absolute divergence time values by applying what is known as a molecular clock (Hillis *et al*, 1996; Li, 1997; Page and Holmes, 1998). This concept was initially introduced in the 1960s (Zuckerkandl and Pauling, 1962), and was based on the idea that molecular changes occur at a constant rate. If true, that would allow to date the unknown divergence times among species just by comparing their DNA or protein sequences. However, early criticisms against the application of a **strict molecular clock** in phylogenies appeared (Simpson, 1964), and finally this idea was totally rejected (Kumar, 2005) in favor of relaxed models, where each branch displays independent evolutionary rates. Indeed, evolutionary rates can vary between different lineages (Bousquet *et al*, 1992; Pereira and Baker, 2006; Thomas *et al*, 2006; Nabholz *et al*, 2008, 2009; Smith and Donoghue, 2008), among different genomic regions or even the within the same gene (Rausher *et al*, 1999; Kumar and Subramanian, 2002; Yang and Nielsen, 1998; Aguileta *et al* 2006; Bininda-Emonds, 2007; Elango *et al*, 2009; Pons *et al*, 2010). Thus, the development of relaxed molecular clock models that allow variation of evolutionary rates through time, and their bayesian implementation, became an essential tool and set the basis for modern phylogenetics (Figure 6). As a consequence, new models accommodating for rate variation have been developed like the **local molecular clock** (Yoder and Yang, 2000), allowing constant rate within a clade but assuming different rates for different clades, the **autocorrelated relaxed clock** (Thorne *et al*, 1998; Kishino *et al*, 2001; Rannala and Yang, 2007) which permits the rate of evolution to vary but it assumes a tendency for branches to share a similar rate of evolution with their immediate ancestral and immediate descendant branches, and the **uncorrelated relaxed clock** (Drummond *et al*, 2006) in which no correlation between adjacent branches is assumed.

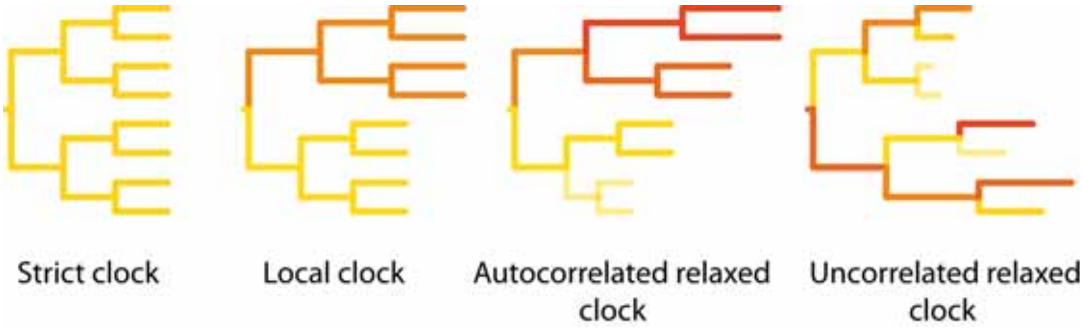


Figure 6. Different models of molecular clock applied to phylogenies [Adapted from Posada, 2012]

A chronogram inferred by applying a molecular clock model will have branch lengths in relative units others than time. To convert them to absolute times it is necessary to calibrate the tree. A direct method involves assigning ages to specific nodes of the tree. These dates usually arise from fossils of known age, although other sources like palaeobiogeographic events can be used. Also, if the information about the divergence time between two species is known, then the rate of molecular evolution of specific genes can be inferred and be directly applied to the molecular clock models to get divergence times in absolute units.

Nevertheless, calibrating trees is a challenging task since **calibration points** cannot be known with confidence and there can be considerable uncertainty as to where to attach the point on the tree. For example, fossils generally only serve to indicate minimum ages, since they represent lineages that diverged from one of the tree branches some time before the date of fossilization (Benton and Donoghue, 2007). On the other hand, palaeobiogeographic events as island emergence, mountain range formation or continent fragmentation provide evidence about maximum ages, because speciation events could only have occurred after the population splits (Weir and Schluter, 2008). To reduce error rates, it is this important to include multiple calibration points and prior distributions of possible ages in a bayesian framework, which incorporates uncertainty from a more realistic point of view and provides certain flexibility that might help to increase the confidence of the dating inferences.

2.6.2 Character evolution, Diversification and the Comparative Method

Evaluating evolutionary hypotheses concerning adaptation, behavioral ecology functional morphology or others can be done in a phylogenetic framework. It can be of interest to reconstruct how a specific trait (morphological, ecological, geographic, molecular...) that we observe in the present has evolved along the phylogenetic history, and therefore understand better the natural history of the group of study. Such an exercise is commonly known as **ancestral states reconstruction** and can be performed using principles as the previously explained maximum parsimony or the probabilistic methods ML and BI. MP minimizes the total number of changes for the character and makes the simplistic assumption that only one change can occur in every branch, or in other words, branch lengths are ignored. Searching for more realistic approaches, the use of model-based methods take into account the branch lengths, as well as a rate of character evolution, to finally output probabilities as a measure of confidence. Thus, the analysis aims to find the rate of character evolution that maximizes the likelihood (the probability of the observed trait data). An advanced probabilistic tool called *stochastic mapping* (Huelsenbeck *et al*, 2003), that uses bayesian inference, even incorporates phylogenetic uncertainty and allows to estimate, by the performance of simulations, specific branch portions where the changes most likely occurred. The reconstruction of ancestral states can be inferred either for discrete binary characters or for continuous characters, and, in the models employed, it is important to be aware about directionality of character changes (*trends*), since evolution can be reversible. In addition, likelihood methods permit to consider asymmetric models in which the rate of gains and losses are allowed to be different. To study correlation among continuous traits is often more complex, and are usually modeled using Brownian motion or the Ornstein-Uhlenbeck (OU) model. Reconstruction of ancestral biogeographic history can be performed too, since geographical location can be treated as a heritable trait, although this is a challenging implementation due to the fact that spatial distributions are constantly changing. This procedure is generally called **ancestral range reconstruction** and specific methods incorporate the evaluation of different models of vicariance and/or dispersal (Bremer, 1992, 1995; Hausdorf, 1998; Ronquist, 1997; Ree *et al*, 2005; Ree and Smith, 2008; Clark *et al*, 2008; Lamm and Redelings, 2009).

The relative shape of a phylogenetic tree provides essential tools for exploring the clade-to-clade variation in lineage **diversification rate**. The mode of lineage diversification through time (or diversification patterns) can be studied since it is possible to assess distances among cladogenetic events using specific models (Hey, 1992; Nee *et al*, 1992, 1994a,b; Harvey *et al*, 1994; Kubo and Iwasa, 1995; Paradis, 1997). The simplest model is the pure birth or Yule model (Yule, 1925), where lineages arise at a constant rate and extinctions are not considered. Next, a birth-death model accounts for rates of both speciation (λ) and extinction (μ), which can be constant or relaxed, thus allowing for changes along the tree. Depending on the distribution of the inferred cladogenetic events, phylogenetic trees can be symmetric or asymmetric (i.e. having a balanced or unbalanced distribution of tips between sister clades) or they can present long or short internal or terminal branches. When having closely spaced cladogenetic events in form of successive very short branches we usually refer to the occurrence of a rapid **evolutionary radiation.**

By reconstructing a phylogenetic scenario, it is also possible to look for significant correlations between the possession of a certain trait and the occupation of a particular selective regime (or a second trait that may be indicative of a selective regime) in order to assess whether a trait variant is favored by selection. This is called the **comparative method**, and has become a powerful tool in the study of adaptation (Brooks and McLennan, 1991; Harvey and Pagel, 1991; Maddison and Maddison, 1992). Tests like the *concentrated changes test* (Maddison, 1990), based on parsimony, the model-based *Pagel test* (Pagel, 1994) for discrete characters, the *phylogenetic independent contrasts* (Felsenstein, 1985b), that assume Brownian motion, and the generalized least squares (PGLS) (Butler and King, 2004), that can implement the OU model for continuous characters, or the estimation of scaling parameters that allow tests of the tempo (δ), mode (κ) and phylogenetic association of trait evolution (λ) (Pagel, 1999a), are among the most popular tools to test for correlated evolution.

The study of groups that have concordant phylogenies can be of interest in order to test hypotheses of **coevolution** (Huelsenbeck and Rannala, 1997; Huelsenbeck *et al*, 1997; Charleston 2002). Species certainly do not evolve disconnected from other species, communities and ecosystems, but plenty of interactions occur, in the form of ecological relationships such as trophic (Farrell and Mitter, 1990; Percy *et al*, 2004), mutualistic or symbiotic (Chenuil and McKey, 1996; Pierce *et al*, 2002; Kawakita *et al*, 2004), or parasitic associations (Weiblen and Bush, 2002; Jackson, 2004). Some of these relationships may be long-term interactions than leave some common traces in the phylogenetic trees and even in some cases they appear to reciprocally accelerate speciation rates, promoting thus cospeciation (Moreau *et al*, 2006). Also, species inhabiting geographical areas with a particular history can also coevolve.

2.6.3. Phylogeography

While biogeography is the study of the distribution of organisms and ecosystems across the geographic space and through geological time, phylogeography is the study of the genetic distribution of populations and species. The term phylogeography was first introduced by Avise *et al* (1987), as a "field of study concerned with the principles and processes governing the geographic distributions of genealogical lineages, specially those within and among closely related species" (Avise, 2000). Thus, the aim is to redraw the factors that have most influenced the distributions of current genetic variation by comparing the evolutionary relationships of genetic lineages to their geographical locations. Phylogeography, thus, integrates divergences times (evolutionary relationships) in a spatial context (geographic distributions). It also integrates classic population genetic concepts with the coalescent theory, which represents a powerful way to estimate phylogeographic parameters as historical population sizes, divergence times and migration rates (Wakeley, 2008) from gene genealogies, likewise allowing to test hypotheses regarding the causal relationship among geographic phenomena, species distributions, and the mechanisms driving speciation.

Classic genetic markers used in phylogeography for animals mainly involves mtDNA, due to its lack of recombination, putative neutrality and smaller effective population size, and as a consequence shorter times to get reciprocal monophyly between geographic regions. Classic analytical procedures involve the reconstruction of phylograms or haplotype networks, which represent genealogical relationships and the number of mutational steps among haplotypes (Templeton, 1992; Posada and Crandall, 2001), or the Nested Clade Phylogeographic analysis (NCPA; Templeton *et al*, 1995; Templeton, 1998), which provides a statistical framework for using gene genealogies to infer demographic history. Although the reliability of this last method has been intensely debated (Knowles and Maddison, 2002; Panchal and Beaumont, 2007; Petit and Grivet, 2002; Beaumont and Panchal, 2008; Petit, 2008; Garrick *et al*, 2008; Templeton, 2009a,b), it set the basis for the development of *statistical phylogeography* (Kuhner *et al*, 1998; Pritchard *et al*, 1999; Beerli and Felsenstein, 2001; Rosenberg and Nodborg, 2002; Hey and Nielsen, 2004, 2007; Knowles, 2004; Kuhner, 2006; Knowles, 2009; Hickerson *et al*, 2010; Nielsen and Beaumont 2009; Beaumont *et al*, 2002) which is more related to coalescent model-based methodologies (Beaumont *et al*, 2010). In addition, next-generation sequencing is revolutionizing the potential of the field, as it can provide thousands of unlinked markers for several populations (Gompert *et al*, 2010a,b; Nosil *et al*, 2012), most of them from non-model organisms. Statistical phylogeography permits to identify specific historical events such as founder effects, long-term barriers to dispersal and to distinguish the relative roles of selection and genetic drift.

A next step of complexity in phylogeography refers to the combination of genetic data from multiple codistributed taxa to go deep into the geographic, geological and ecological circumstances that have generated the observed distribution of biodiversity, something known as *comparative phylogeography*. By analyzing multiple species distributed in the same geographic areas, more light may be shed on the influence that historical events exert on the evolution of populations and species, sometimes in the form of regional concordance (e.g. by vicariant events) (Avise, 1992; Joseph *et al,* 1995; Schenider *et al,* 1998; McCulloch *et al,* 2010) or continental concordance (e.g. Pleistocene glaciations and postglacial recolonization routes) (Hewitt, 2004; Willis *et al*, 2000; Deffontaine *et al*, 2005). Also, identifying common patterns similarly influencing multiple taxa can be relevant for conservation biology. Indeed, the classic operational criteria used for delimiting conservation areas have been richness, endemism and phylogenetic diversity (Brooks *et al*, 2006; Lamoreux *et al*, 2006; Orme *et al*, 2005). However, such observations are not made under a historical perspective that define the reasons that promoted the observed biodiversity, something that potentially reports new criteria to be evaluated in conservation management, because it allows to conserve not only extant biodiversity but the processes that generate this diversity. In addition to conservation planning, comparative phylogeography also informs about evolutionary changes that have been affected by past climate changes and, as a consequence, can be also used as a possibility to predict how climate change will genetically, demographically and spatially influence regional biodiversity in the future (Ferrier and Guisan, 2006; Taberlet and Cheddadi, 2002; Williams *et al*, 2007).

Overall, phylogeographic approaches provide new sources of evidence for species delimitation (Leavitt *et al*, 2007; Sites and Marhall, 2004; Hudson and Coyne, 2002; Knowles and Carstens, 2007; Pons *et al*, 2006), contribute with powerful tools to identify historical hybridization events, hybrid zones or the occurrence of introgression (Gonzalez-Rodriguez *et al*, 2004; Hewitt, 2001; Swenson and Howard, 2005), to recognize geographic origins of isolation, to track global movements of pests or to detect invasive species. Phylogeography has a great potential to integrate with other fields, like Ecological Niche Modeling (Peterson *et al*, 2002; Waltari *et al*, 2007), spatial analysis of genomic signatures in natural selection (Joost *et al*, 2007, 2008), spatial analysis of morphological and functional trait evolution, studies of ecological speciation (Shcluter, 2009) or studies of community assembly (Jabot and Chave, 2009). As a summary, it is probably not an exaggeration to say that phylogeography constitutes the culmination and the integration of a large array of advances and methodologies involving phylogenetics and population genetics, and that it paves the way for testing complex evolutionary hypotheses previously unthinkable.

3. Biodiversity and classification: The Phylogenetic Systematics

Taxonomy is the science in charge of the discovery, identification, description and classification of organisms (Mayr and Ashlock, 1991), while phylogenetics study the evolutionary relationships among them. That is, taxonomy proposes hypotheses of kinship through classification, and phylogenetics test them by applying the principle of monophyly. Both disciplines together conform the body of systematics, or phylogenetic systematics (Hennig, 1966).

Between 1.5 and 1.9 millions of species on Earth have been described so far, although broad estimations of global biodiversity range between 3 and 15 millions, or even up to 100 millions (Hawksworth and Kalin-Arroyo, 1995; May, 2000; Chapman, 2009; IUCN, 2011). Therefore, a vast mass of biological diversity is waiting to be discovered. Largely, then, the principal challenge remains on finding mechanisms to explore and taxonomically assess either poorly known regions and habitats or non-popular groups of organisms. Although there should not be need to justify the importance of discovering world's remaining biodiversity, to laid out correct taxonomic hypotheses are a prerequisite for any posterior credible research. The targets of several other research fields could be seriously threatened without taxonomy and phylogenetics, and we have already highlighted the variety of fields (most of crucial relevance for society) that are based on taxonomy: agriculture, conservation, ecology, fisheries, forestry, environmental science, among others (Kim and Byrne, 2006). For incomprehensible reasons, though, taxonomy not only is unpopular in academia, but it is declining in terms of specialists, training and financial support (Mallet and Willmott, 2003; Wheeler and Valdecasas, 2007; Drew, 2011). In detriment of classical taxonomy, which focused on descriptions of diagnosable morphological characters, DNA now plays a fundamental role in taxonomy and promotes new procedures and criteria based on molecular diagnosability. DNA-taxonomy (Tautz *et al*, 2003) takes advantage of the rapid accumulation of molecular data that is greatly exceeding the speed of morphological descriptions performed by the limited number of specialized taxonomists and, as a consequence, a new period for biodiversity assessment seems to be starting. Systematic research has, so far, arguably yielded a reasonable idea of how the main branches of the Tree of Life are arranged, at least for eukaryotic organisms (Bininda-Emonds, 2011). Nevertheless, numerous problematic branches still remain unresolved and decoding the most ancient relationships is specially challenging, while taxonomists also have to deal with another big responsibility: the arduous task to decide what is a species and what is not.

3.1. Species concepts and phylogenetics

Species are the fundamental unit of biological organization (Mayr, 1982), and biodiversity is essentially measured in terms of species richness (Wilson, 1992). A large number of species definitions have been proposed with the objective to find criteria satisfying a universal species concept (Mayden, 1997; Mallet, 2001; de Queiroz, 1998, 2005, 2007; Hey, 2006a), falling on what seems to be closer to a dialectic than to a conceptual problem. Biologists do not generally disagree on concepts, they disagree on the properties explaining their universality. To emphasize some of these properties, the biological species concept (Wright, 1940; Mayr, 1942; Dobzhansky, 1950) is based on reproductive isolation, the ecological species concept (Van Valen, 1976; Andersson, 1990) on the occupation of distinct niches, the evolutionary species concept (Simpson, 1951; Wiley, 1978; Mayden, 1997) on the occurrence of unique historical evolutionary roles, the genotypic cluster species concept (Mallet, 1995) on the absence of genetic intermediates and the phylogenetic species concept (Hennig, 1966; Rosen, 1979; Cracraft, 1983; Donoghue, 1985; Baum and Shaw, 1995) focuses on reciprocal monophyly and/or diagnosability of fixed differences as fundamental properties.

To simplify, de Queiroz (1998, 2005, 2007) identified a common element and propose the **unified species concept**, for which the only necessary property of a species is to be a separately evolving metapopulation lineage (i.e. an inclusive population made up of connected subpopulations extended through time). Indeed, a vast majority of evolutionary biologists accepts that species are lineages. This proposal attempts to separate the concept (separately evolving metapopulation lineages) from the properties or criteria (lines of evidence) that are used to delimit the boundaries of species (Sites and Marshall, 2003, 2004; Hey, 2006b; Wiens, 2007; Braby *et al*, 2012).

Figure 7 shows the timeframe in which new lineages acquire different properties, which serve as operational criteria for recognizing species boundaries under the different concepts. It is possible to deduce that some of the properties (or species concepts) are more restrictive than others. For example, the phylogenetic species concept criterion (reciprocal monophyly) would result in much higher estimates of species diversity (taxonomic inflation) in comparison to the biological species concept, but it would still be conservative than adopting only a few properties from the early phase of lineage divergence, as the observance of phenotypic differentiations (sometimes applied to define categories as subspecies, varieties, races or forms) (Braby *et al*, 2012). The unified species concept, then, does not really supply any consensus definition with practical and general criteria allowing to take decisions, but it clarifies the issue of species delimitation by separating the conceptual problem from the methodological problem of inferring boundaries and numbers of species (de Queiroz, 2007). Admittedly, then, we must be aware of the uncertainty of the species rank, since it can have different meanings in different taxonomic groups (Hey, 2001).

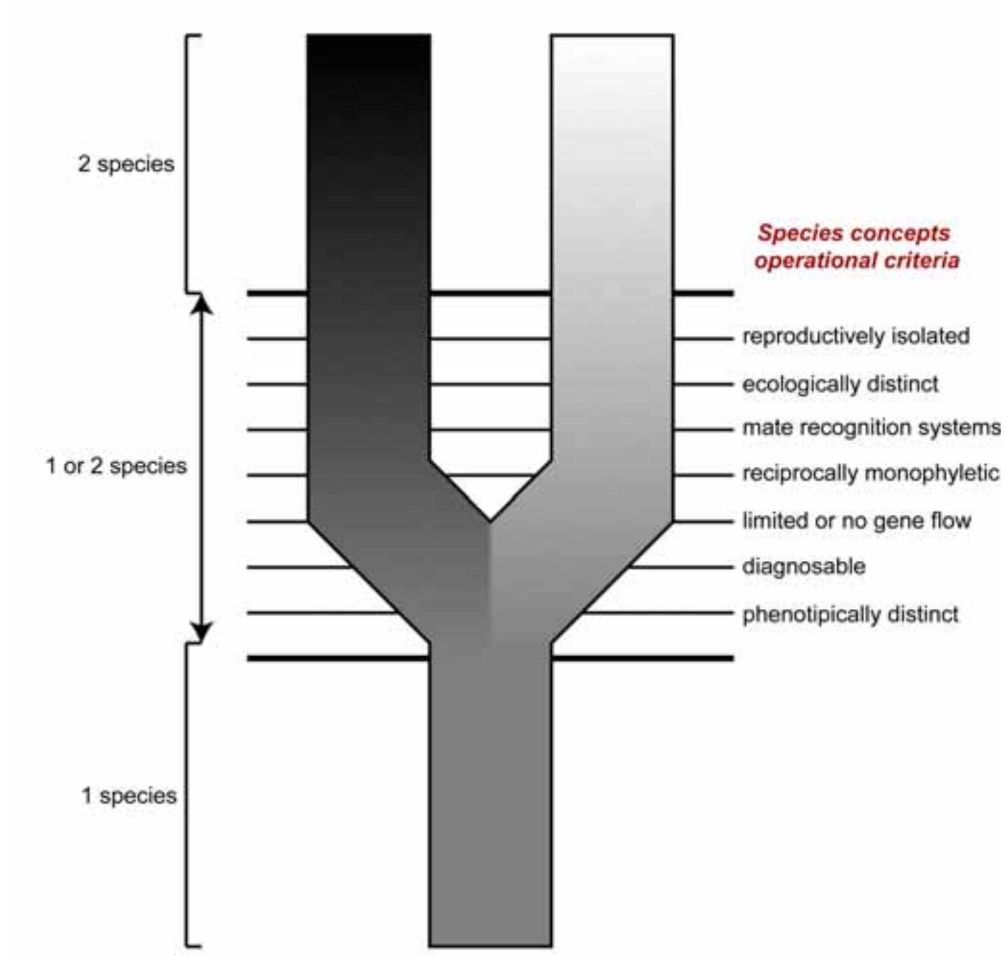


Figure 7. Speciation and species concepts. One lineage is split in two through speciation and both progressively acquire properties that are the foundations for different species concepts. Disagreements among taxonomists are frequently given in this area. [Adapted from de Queiroz (2007) and Braby *et al* (2012)]

3.2 Species delimitation

Two steps are required to delimit species: first, detecting separately evolving lineages and, second, delimiting boundaries between incipient species and genetic complexes of populations. Lineages can be detected by phylogenetic inference, while establishing boundaries among species requires examining the criteria adopted under different species concepts.

The majority of the described species have been delimited based on patterns of morphological variation, and species identification is traditionally based on the use of morphological keys (Wiens, 2004; Ebach and Holdredge, 2005). Typically, non-overlapping patterns of morphological variation (or morphological gaps) have been used as a criterion for describing and delimiting species. It can be argued that evolutionary forces may prevent two distinct lineages, separated by a morphological gap, from homogenizing (Davis and Nixon, 1992; Wiens and Severdio, 2000; Sites and Marshall, 2004). Therefore, the morphological gap has been shown to be useful as a criterion, although not perfect. For example, it can often fail in detecting cryptic species or result in uncertainty when dealing with discontinuities among allopatric populations. Another classic operational criterion to delimit species consists in testing the existence of gene flow, looking for evidence for the biological species concept. By performing mating experiments it is possible to evaluate the level of pre- and postzygotic reproductive isolation between lines (e.g., Ackermann *et al*, 2008; Lukthanov *et al*, 2005). This approach, although powerful, requires a hard experimental process that not always results operative or even feasible to reproduce in natural systems.

DNA-based tools for species delimitation, in parallel, give response to the massive amount of molecular data that is currently being gathered. The potential and utility of DNA sequences for species delimitation is undeniable. Current proposals range from clustering methods applying genetic distance thresholds (Hebert *et al*, 2003a; Puillandre *et al*, 2011), statistical approaches for delimitation of phylogenetic species like population aggregation analysis (PAA) (Davis and Nixon, 1992) or cladistic haplotype analysis (CHA) (Brower, 1999), to the use of models for detection of branching-rate shifts on time-calibrated trees (GMYC) (Pons *et al*, 2006), the inference of the *gsi* statistic testing the differentiation of species and populations (Cummings *et al*, 2008), or to methods assessing the probability of monophyly in a coalescent process (Masters *et al*, 2011). More recently, coalescence methods for species delimitation based on multilocus data are being developed. Despite the fact that for recently diverged species incomplete lineage sorting can be widespread throughout the genome, coalescent theory can model the relationship between gene trees and species history even before lineages have become reciprocally monophyletic (Knowles and Carstens, 2007; Carstens and Knowles, 2007; Yang and Rannala, 2010) and thus succeed in providing evidence for species delimitation. The overcome of these methodologies over other DNA-based tools is widely accepted although it probably represents a very resource-consuming methodology for large-scale biodiversity assessments.

It is now is usually agreed is that a highly corroborated hypothesis of lineage separation (i.e. existence of two species) should be based on multiple lines of evidence (de Queiroz, 2007; Will and Rubinoff, 2004; Rubinoff and Holland, 2005; Will *et al*, 2005; Carstens and Knowles, 2007; Knowles and Carstens, 2007). That is, the more properties (evidence) to distinguish between two lineages, the more confidence one has in rejecting the null hypothesis of a single species and accepting the alternative of two or more species (Mallet, 1995). In that sense, biologists often combine strategies to test their hypotheses. To list some examples, several studies have combined ecological niche models with other sources of evidence (e.g., Raxworthy *et al*, 2007; Wiens and Graham, 2005; Stockman and Bond, 2007; Bond and Stockman, 2008; Leache *et al*, 2009), morphometrics with geographic data (Zapata and Jimenez, 2012), phylogeography with morphometrics and numerical association tests (Puorto, 2001), molecular with morphological variation (Wiens and Penkrot, 2002; Dincă *et al*, 2011a), coalescent model-based approaches with simulations (Knowles and Carstens, 2007; Carstens and Dewey, 2009), molecular data with morphometrics and karyological data (Dincă *et al*, 2011b), behaviour (song variation) with other sources (Cadena and Cuervo, 2010), etc.

3.3 Biodiversity surveys: the DNA barcoding approach

As briefly mentioned in the beginning of this chapter, a major percentage of biodiversity on Earth most probably remains to be discovered. The DNA recourse has provided hope to accelerate exploration and global biodiversity assessments by massively sequencing living organisms. Two alternative strategies are generally followed: first, massive single-locus sequencing projects allow global scale discoveries of unknown diversity and second, comprehensive genomic approaches enable a finer way to study already known diversity. There is a compromise between the two strategies, regarding quantity and quality. It is generally agreed that multiple genomic markers allowing coalescence-based studies are optimal for studying species boundaries. However, such a fine knowledge not always responds to real needs. One approach for quantifying global biodiversity is the **environmental sequencing**, that focuses on sequencing single markers for all living taxa of a specific area (Venter *et al*, 2004; Acinas *et al*, 2004). It produces interesting ecological outputs like the possibility to calculate biodiversity indexes for specific regions. However, it does not relate sequences to any specific specimen, which makes it useless for DNA-taxonomy. Probably the most successful tool for exploring biodiversity in the last years has been the use of **DNA barcoding** (Hebert *et al*, 2003a,b). This is based on the use of a short, standarized and highly variable genomic segment as a marker for species identification. For animals, this barcode corresponds to a mtDNA fragment of ca. 650 bp belonging to the *cytochrome c oxidase subunit 1* (COI) and, at present, more than two million sequences (corresponding to 162194 species) are deposited in the Barcode of Life Data System (BOLD; Ratnasingham and Hebert, 2007), which is organized into large projects for specific living groups and managed by taxonomic specialists. New query sequences are contrasted against the databases and if they yield a good match (usually no more than X differences, but depending on the group) they are identified to already known species or, otherwise, a new specific assignment needs to be done based on extra evidence provided by expert taxonomists. Therefore a catalogue of life based on an objective, universal and a quickly accessible criterion is increasing in size. This approach is particularly useful for studying levels of diversity among challenging or poorly known taxa. Barcodes have also enabled useful applications as linking the different animal life stages, discerning males and females of sexually dimorphic species (Janzen *et al*, 2005), discriminating cryptic biodiversity (Hebert *et al*, 2004; Brower, 2006; Nielsen and Matz, 2006), identifying degraded samples or studying diets from semi-digested food fragments from stomach contents (Jurado-Rivera *et al*, 2009).

The implications of DNA barcoding initially generated a fervent debate focusing on to the oversimplification of considering only a single-gene sequence to discriminate species reliably (Mallet and Willmott, 2003; Janzen, 2004; Moritz and Cicero, 2004; Will *et al*, 2005; Hebert and Gregory, 2005; Smith, 2005; Rubinoff, 2006; DeSalle, 2006; Hajibabaei *et al*, 2007). The real controversy, though, arose due to a mix-up of two parallel nascent proposals: 1) the DNA-taxonomy (Tautz *et al*, 2003), which pretends to construct a DNA-based taxonomic system and 2) the DNA barcoding (Hebert *et al*, 2003a), which focuses on species identification. As a consequence, species identification and species delimitation were occasionally misinterpreted (Hebert *et al*, 2004; Brower, 2006), but nowadays DNA barcoding has been broadly rejected as a tool for delimiting species without being accompanied by independent evidence. On one side, it has been pointed out that barcodes cannot discriminate recently diverged taxa (Mallet and Willmott, 2003; Meyer and Paulay, 2005; Rubinoff *et al*, 2006; Hickerson *et al*, 2006; Meier *et al*, 2006; Whitworth *et al*, 2007), while establishing any divergence threshold as a species limit would be arbitrary and without any biological foundation. Moreover, mitochondrial (or organellar) sequences are more likely to be affected by reticulation events, a fact that may artifactually identify sequences from one individual as belonging to some other species.

From a phylogenetic point of view, as it has been previously explained, individual gene trees may not necessarily represent the species trees, which makes the use of DNA barcodes inappropriate for phylogenetic inference. However, phylogenies are commonly inferred from barcodes with different objectives other than reconstructing evolutionary relationships, as for example, placing unknown specimens in a broader taxonomic framework, or detecting independent lineages within a known taxon category that deserve further study.

To sum up, DNA barcoding comes out as an excellent tool for conducting species inventories and biodiversity assessments of specific areas. In parallel, it also allows to assess putative unknown diversity by detecting separately evolving lineages (but not relating them phylogenetically), which will need to be further explored by deeply studying species delimitations. Overall, it arises as a useful, fast, cheap and operative tool for conducting large-scale studies of poorly known areas and/or groups of taxa.

4. Insects as a model system

4.1 Insect biodiversity and phylogeny

Insects constitute the most diverse group of animals on the planet, with more than 1 million described species (1.004.898) (Foottit and Adler, 2009) in a total biota of 1.4 to 1.9 million (Stork 1988, 1993; May, 1990; Hammond, 1992). They represent, then, more than 50% of known biodiversity. Furthermore, it is believed that millions of species are still waiting to be discovered (Grimaldi and Engel, 2005). Estimates of total insect diversity range from 2 to 50 million species (Hodkinson and Casson, 1991; Gaston, 1991; Hammond, 1992; Stork, 1993; Erwin, 1993; Grimaldi and Engel, 2005), although 5 million is considered a more reasonable digit. If so, an **80% of insect species would remain to be described**. From a temporal perspective, insects have become the dominant component of the known diversity on earth. Since their origin ca 400 Mya at around the Silurian-Ordovician boundary (Gaunt and Milles, 2002), close to 100 million species may have arisen (Grimaldi and Engel, 2005). In terms of biomass, insects are probably the most important group of terrestrial animals too, with gross approximations from 1 to 10 quintillions (10^{18}-10^{19}) of living insects at any specific moment (Johnson, 2003).

The class Insecta (=Ectognatha) is systematically divided into 28 or 30 orders, depending on the authors (Cranston and Gullan, 2010; Traitwein *et al*, 2012) (Figure 8). Two orders conform the basal groups: Archaeognatha and Zygentoma. Although both are apterygotes (i.e. primitively wingless) and had been traditionally clustered as Thysanura, they almost certainly are not sister groups. The oldest lineage of the extant Insecta is the Archaeognatha, which displays only one condyle (articulation point) in the mandibles instead of two, a character that differentiates them from the rest of the insects (Dicondylia). The Pterygota are the winged or secondarily apterous insects and comprise four main groups: the Palaeoptera, Polyneoptera, Paraneoptera and Holometabola (=Endopterygota). The Palaeoptera cannot fold their wings against the body although the rest of Pterygota (Neoptera) can. They include Odonata and Ephemeroptera, but the monophyly of this group is controversial (Traitwain *et al*, 2012). Polyneoptera are the most basal of the extant members of Neoptera, and their internal phylogenetic relationships are highly uncertain, as well as their monophyly (Gullan and Cranston, 2010). Paraneoptera include insects that have primarily sucking mouthparts and some of their ordinal relationships are also controversial (Grimaldi and Engel, 2005; Johnson *et al*, 2004), although their monophyly is not debated. Holometabola comprise the insects that undergo complete metamorphosis, which actually constitute the most diverse lineage of life on Earth including four of the five megadiverse orders (Hymenoptera, Coleoptera, Diptera and Lepidoptera). Some of their internal relationships are debated too, such as the relative position of Hymenoptera or Strepsiptera (Grimaldi and Engel, 2005).

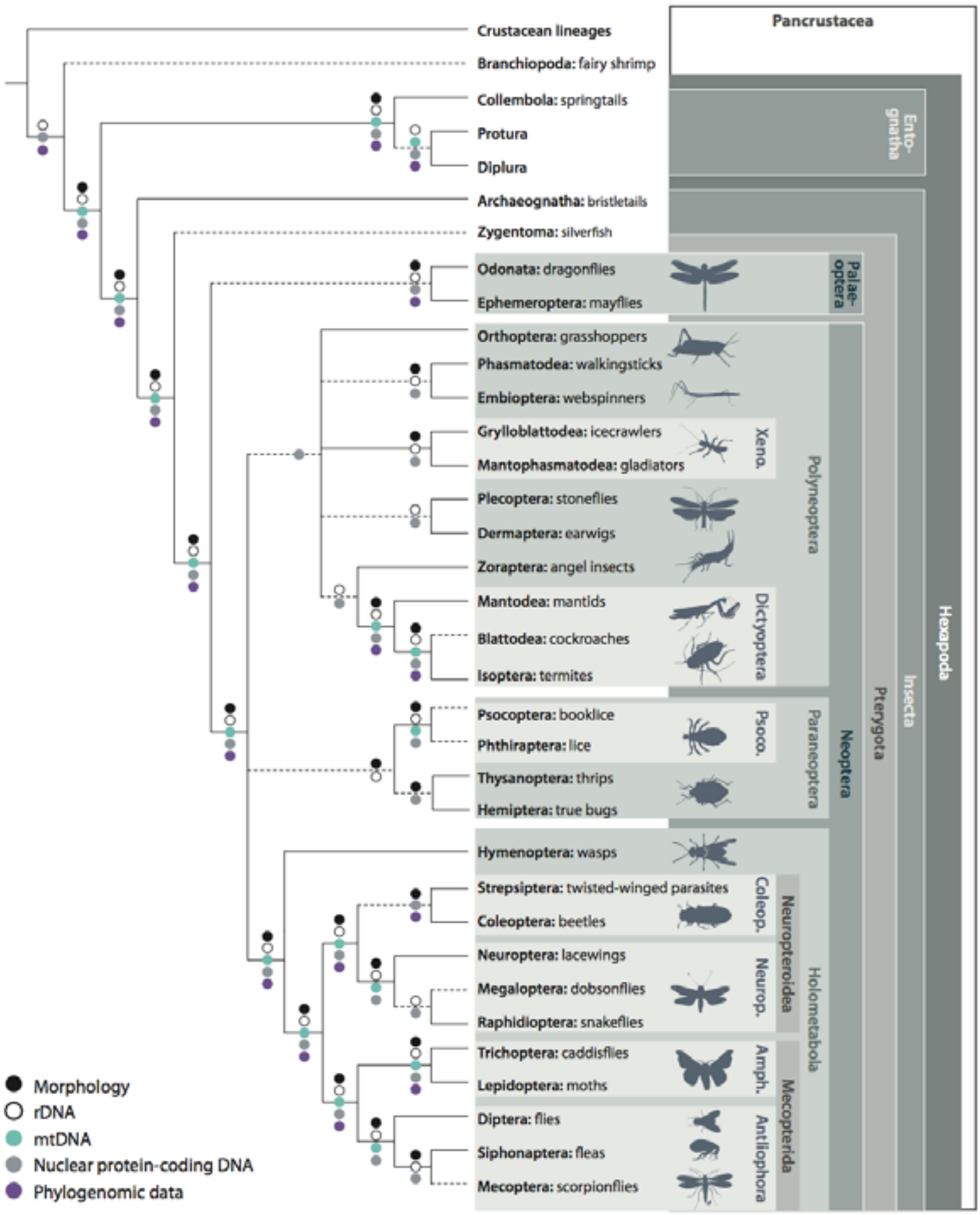


Figure 8. Current knowledge on higher-level phylogenetic relationships of insects based on a review of recent literature. Dashed lines indicate tenuously supported relationships or possible non-monophyly (in the case of terminal branches). The types of data supporting each node are displayed if a node was recovered by a particular line of evidence alone or in a combined analysis. Phylogenomic data refer to a molecular data set of at least 20 kb, to data collected through EST harvests, or to large-scale genome

comparison. Abbreviations: EST, expressed sequence tag; mtDNA, mitochondrial DNA; rDNA, ribosomal DNA; Amph., Amphiesmenoptera; Coleop., Coleopterida; Neurop., Neuropterida; Psoco., Psocodea; Xeno., Xenonomia. [Extracted from Trautwein *et al*, 2012]

Coleoptera are by far the most species rich group of insects, representing approximately a 40% of their biodiversity, followed by Lepidoptera, Diptera, Hymenoptera and Hemiptera (Grimaldi and Engel, 2005; Beutel and Pohl, 2006; Foottit and Adler, 2009) (Figure 9). Insects have invaded every niche in the planet, apart from the oceanic benthic zone (Grimaldi and Engel, 2005) and the greatest concentration of insect species lies in tropical areas (May, 1998; Erwin, 2004). They can be conspicuous, mimics or concealed, and have diurnal, crepuscular or nocturnal behavior. Their lifestyles can be solitary, gregarious, subsocial or highly social. They can live in water, on land or in soil during part or all of their lives and they can survive under a wide range of conditions, such as extremes of heat and cold, wet and dry, and unpredictable climates. Insects are, undeniably, the most evolutionaryly and ecologically successful lineage across the tree of life. Evolutionaryly, if measured in terms of species diversity, geological duration and/or geographic spread, and ecologically, if measured in terms of impact upon the ecosystems (Bradley *et al*, 2009).

Several synergistic factors may have triggered the enormous diversification and the wide range of adaptations of the insects, starting by a major adaptive feature during the origin of hexapods: terrestriality. Traditionally, certain key innovations have been associated with their evolutionary success, as for example the development of wings during the Carboniferous (ca 370 Mya). The flight ability provided important advantages to improve their dispersal capacity and defense against predators. In fact, there is a great asymmetry in diversity when flightedness is contrasted to aptery, since the pterygota are vastly more speciose than the apterous groups. Additionally, the arthropod exoskeleton provided protection and strength to life on land, reduced desiccation and allowed to produce defense exotoxines. Also, metamerism and repetitive pairs of appendages led to divergent specialized adaptations, such as raptorial, fossorial or mouthparts modifications (Grimaldi and Engel, 2005). Importantly, the appearance of metamorphosis provided further advantages: it allowed the adult and larval stages to differ or overlap in phenology depending upon timing of suitable conditions (Gullan and Cranston, 2010), permitted to adopt strategies like diapause during the early stages of development in holometabolans or to have more permissive diet ranges between adult and larval stages. The evolution to eusociality, during the Late Jurassic or Early Cretaceous (150-140 Mya), might also have been of great importance as it allowed insects to have a huge collective impact on terrestrial environments (Bradley *et al*, 2009).

Being the insects among the earliest terrestrial animals, they were already in position to exploit new resources when some of the most important new terrestrial radiations appeared. The evolution of seed plants and, subsequently, the angiosperm radiation provided a great variety of new resources and ecological niches for insects, promoting increased diversification rates when shifting from aphytophagy to phytophagy (Farrell, 1998; Barraclough *et al*, 1998). Today, it is estimated that half of the insect species are phytophagous, but only 9 of the 28-30 extant insect orders are primarily phytophagous. This imbalance suggests that when a barrier to phytophagy (e.g. plant defenses) is breached, an asymmetry in species number occurs, with the phytophagous lineage being much more speciose. To cite some examples, the phytophagous Lepidoptera have undergone a tremendous diversification in comparison to their relatively species-poor sister group, the non-phytophagous Trichoptera. Likewise, the major herbivorous beetle superfamilies Chrysomeloidea and Curculionoidea are overwhelmingly more diverse than their mostly fungivorous sister groups, while within each of these superfamilies the angiosperm-feeding clades are far more speciose than the gymnosperm-feeding ones (Hunt *et al*, 2007; McKenna *et al*, 2009). Likewise, the radiations of birds and mammals could have greatly favored insects too by providing new niche opportunities, in particular to those groups adopting parasitic lifestyles, although this hypothesis has been debated (Wiegmann *et al*, 1993; Tilmon, 2008). As a consequence of such a cascade of innovations and ecological opportunities, insects exhibit either high speciation or low extinction rates (or both) in comparison to other terrestrial animals, which has given rise to a unique case of extreme lineage accumulation that started more than 400 millions ago (Grimaldi and Engel, 2005).

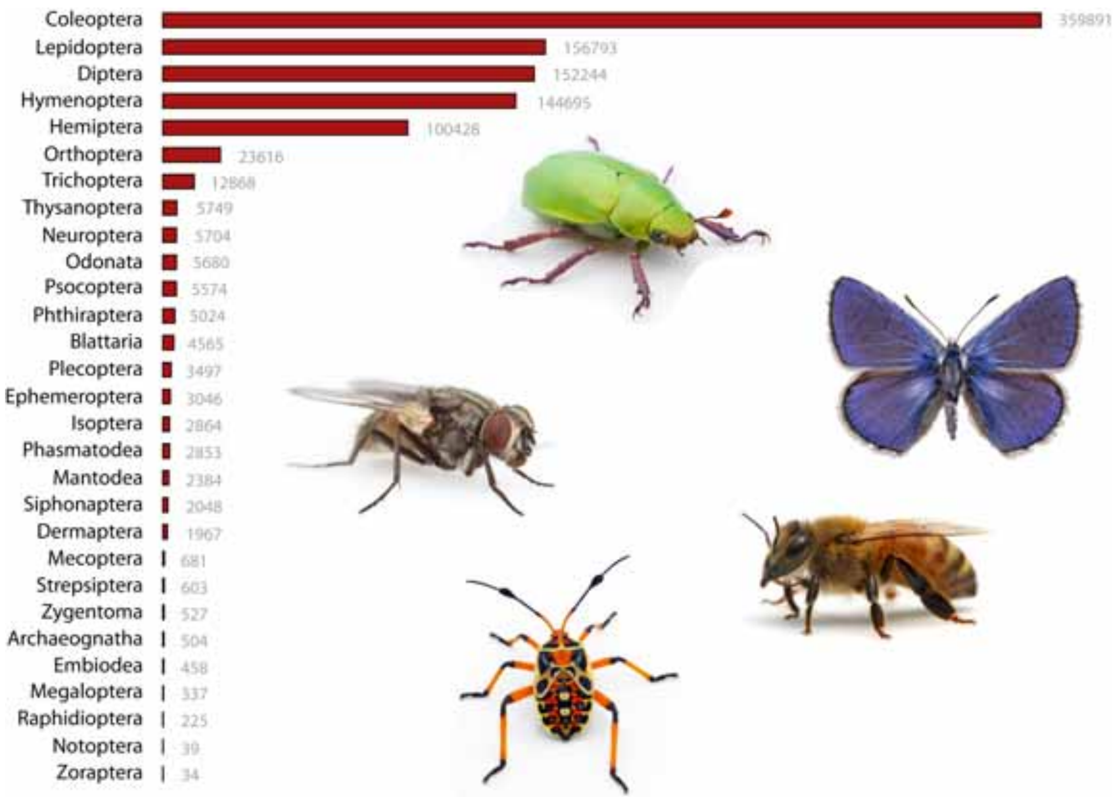


Figure 9. Described number of species for each order of insects. Pictures representing the five megadiverse orders are shown. [Photos: Alex Wild]

4.2 Why are insects important?

The very truth of the exceptional insect diversity and the important gap in biodiversity knowledge should serve as sufficient reasons to invest seriously in entolomological research. But, in addition to their interest for natural history, insects have huge implications for human society that prioritize their study.

Insects constitute a major biological component of all terrestrial ecosystems and may dominate food chains and food webs in both volume and numbers. They have become the most important multicellular heterotrophs thanks to several key physiological traits and their developmental and evolutionary plasticity. They participate in cycling nutrients via leaf-litter and wood degradation, dispersal of fungi, decomposing plants or animals, disposal of carrion and dung, and soil turnover. Insects are essential for plant propagation, including pollination and seed dispersal, which make possible the production of many agricultural crops.

They perform the maintenance of plant community composition and structure, via phytophagy and/or including seed feeding. They provide a major food source for other taxa, like insectivorous vertebrates such as many birds, mammals, reptiles and fish. They also perform maintenance of animal community structure, through transmission of diseases of large animals, and predation and parasitism of smaller ones. Certain insects can also provide human society with food (e.g. honey) or commercial chemical compounds (e.g. chitin, silk, shellac). Some can also be detrimental, either damaging human health by acting as disease vectors or adversely affecting agriculture in form of pests. Last but not least, they have become key in scientific research, thanks to their short generation time, high fecundity and ease of laboratory manipulation. Thus, insects have been used in landmark studies in medical research, biomechanics, climate change, developmental biology, ecology, evolution, genetics, paleolimnology and physiology. By all means, insects are so important that if all were to disappear, humanity probably could not last more than a few months (Wilson, 1992).

4.3 Challenges in insect phylogenetics: six study cases.

Mapping the evolutionary relationships of insects, with their stunning diversity, remains a challenge, even in light of new theory and technology (Whitfield and Kjer, 2008; Ronquist, 2010; Trautwein *et al*, 2012). As previously described, the insects have had a complex and long evolutionary history, replete with novelties, interactions, adaptations, radiations and other events that may left strong but sometimes confusing signatures either on their phenotypes or genotypes.

Such a chain of facts may complicate phylogenetic reconstruction, due to the difficulties of accommodating evolutionary models or interpreting data. One of the most severe problems for insect phylogenetic analyses has been to deal with rapid radiations (Rokas and Carroll, 2006; Whitfield and Kjer, 2008). It is common to find lineages of insects that have diversified very rapidly within a relatively short time span, generating patterns of molecular and morphological change that are difficult to discern phylogenetically. While many insect datasets are plagued by short, ancient internodes, it is also very frequent to come up against lineage-specific substitution rate biases (substitutions accumulate at different rates among lineages), or lineage-specific base compositional biases (nucleotides at different proportions among lineages and/or among regions of genes) (Whitfield and Kjer, 2008). Although ancient radiations have described as especially problematic for tree inference, these biases may apply to phylogenies among closely related taxa too, in addition to certain particularities involved in insect speciation (i.e. large population sizes, short generation times, reticulation, high dispersion rates, co-interactions, chromosome evolution), which may challenge discerning species trees inference from gene tree topologies. Moreover, most of these problems are frequently combined with the difficulty of selecting data that are relevant for the questions to be addressed.

Since the implementation of molecular data into insect phylogenetics, there have been great improvements in understanding insect evolutionary history, such as a new statement on the origin of insects within Arthropoda (Meusemann *et al*, 2010; Regier *et al*, 2008; Rota-Stabelli *et al*, 2011), new evidence for the phylogenetic placement of insect orders (Savard *et al*, 2006; Wiegmann *et al*, 2009; Ishiwata *et al*, 2010), and hundreds of works unraveling internal relationships of lower level taxa, all greatly contributing to a better knowledge of the insect tree of life. In this thesis I present six study cases, which address phylogenetic complexity in insects at different taxonomic levels and aim to resolve several questions related to systematics, evolution and species limits:

4.3.1. Insect phylogenomics and the limits of mitochondrial genomes

If inferring deep level phylogenies are by definition problematic, those are even harder when using mitochondrial sequences. In insects, while some inter-order relationships are fully accepted, others have become classic examples of phylogenetic complexity. Indeed, most of the debate arises from results obtained using mitochondrial data. Mitochondrial sequences are widely used in insect systematics, but even if they of great utility at low taxon levels, they can be misleading at deep level relationships because they reach saturation levels much before than nuclear sequences (Burger *et al*, 2003; Lin and Danforth, 2004). This is specially true for groups that have had accelerated evolutionary rates likely related to big phenotypic changes or to parasitic lifestyle (Castro *et al*, 2002), most notably

Hymenoptera, Strepsiptera, Phthiraptera, Hemiptera and Thysanoptera. In this chapter I test the utility by using a phylogenomic approach with entire mitochondrial genomes for resolving deep insects inter-ordinal relationships and evaluate the potential phylogenetic artifacts involved.

4.3.2. Polyommatina butterflies: complex taxonomy

Butterflies are a diverse and charismatic group of insects that have received significant taxonomic and systematic attention. The *Polyommatus* blues belongs to the family Lycaenidae and its taxonomy has been highly controversial at the genus level. It includes about 460 species in as many as 81 genera described, but their morphological delineations are generally unclear and a wide array of taxonomic combinations are currently in use. This situation highlights a need to propose taxonomy-friendly systems that accommodate existing classifications in a phylogenetic context to maximize the preservation of taxonomic stability. In this chapter I propose to organize *Polyommatus* genera classifications on a phylogenetic framework relying on monophyly and an age interval as a criteria for genera delineation.

4.3.3. Large biodiversity surveys in Rophalocera

Several methods have been proposed in order to find objective and biologically meaningful systems for species delimitation. The General Mixed Yule Coalescent (GMYC) (Pons *et al*, 2006) model uses molecular phylogenies, usually inferred from a single marker, to detect significant switches from branching diversification patterns reproducing a Yule or a coalescent model. It is a fast and practical method for assessing levels of biodiversity from large surveys in specific areas or poorly known local faunas/floras. Even if the approach is widely used, the real performance of the method has never been assessed with credible empirical data. A better understanding of the performance of the model by contrastable results may be important to understand the real rate of success and the set of factors mostly affecting the results. Using a large dataset including all butterfly species for a entire country, in this chapter I evaluate a variety of phylogenetic methodologies and sampling effects on the GMYC behavior and propose a list of best-practices when dealing with similar situations.

4.3.4. Agrodiaetus *butterflies: an extraordinary recent insect radiation*

It is undeniably that butterflies of the large Palaeartic subgenus *Agrodiaetus* are one of the most impressive fast animal radiations (Coyne and Orr, 2004). Close to 130 species have been recognized for a group starting to diversify at 2.51-3.85 Mya (Kandul *et al*, 2004). Such a radiation most probably occurred due to a chain of newly rearranged chromosome fixations establishing barriers and population isolation (Kandul *et al*, 2007). Indeed, only within *Agrodiaetus*, a range of haploid chromosome numbers from n=10 to n=134 are found. Chromosome number, thus, have been used as a diagnostic criterion for describing and delimiting new taxa that highly complement molecular phylogenies within such a complex group that, in fact, is extremely uniform and exhibit few distinguishable morphological characters. One of the main taxonomic problems of this system relies on the credibility of the specific status for small isolated populations. These dot-liked distributions may represent either relicts of species that had broader distributions in the past or young species recently originated. During the last 40 years, approximately 50 species of *Agrodiaetus* have been described. In western Europe, 11 species present extremely dot-liked distributions limited to particular mountains or valleys. The fact of having an important taxonomical oversplitting in *Agrodiaetus* in Europe, then, is meritorious to be tested, due to the relevant implications in conservation.

4.3.5. Lasius *ants: the link with conservation*

Lasius (Hymenoptera: Formicidae: Lasiini) is one of the most abundant ant genera in the Holarctics, comprising around 86 extant species (Wilson, 1955; Janda *et al*, 2004). In this chapter I front the task to assess and describe what potentially represents a new species of ant from the genus *Lasius.* This taxon present some particularities that makes it potentially vulnerable in terms of conservation, since it is restricted to the mountain summits of the island of Mallorca, representing an ecological island within an island. By the application of phylogenetic inference it may be possible to assess the levels of genetic distinctness of this taxon as well as placing it into a phylogenetic context within *Lasius*. In addition, the comparative study of its morphology and its ecological preferences might provide the evidence to finally describe a new species as well as to assess the levels of conservation priorities that it would need.

4.3.6. Lysandra *butterflies: real-time speciation*

Similarly to the *Agrodiaetus*, the also Polyommatina butterfly species of the genus *Lysandra* display great differences among their chromosome numbers. However, while *Agrodiaetus* presents karyotype homogeneity within populations, in the case of *Lysandra* is not always true. For example, the species *Lysandra coridon* exhibit a variability of n=87 to n=93. On the other side, most of the *Lysandra* taxa have controversial specific status and internal phylogenetic relationships have never been studied using comprehensive molecular data. By all means, this represents an ideal system to study the limits and interactions among coexisting species. I tackle the complex *Lysandra* evolutionary history under the strategy of the multispecies coalescent model, to finally study the implications of karyotype instability into the speciation processes within *Lysandra*.

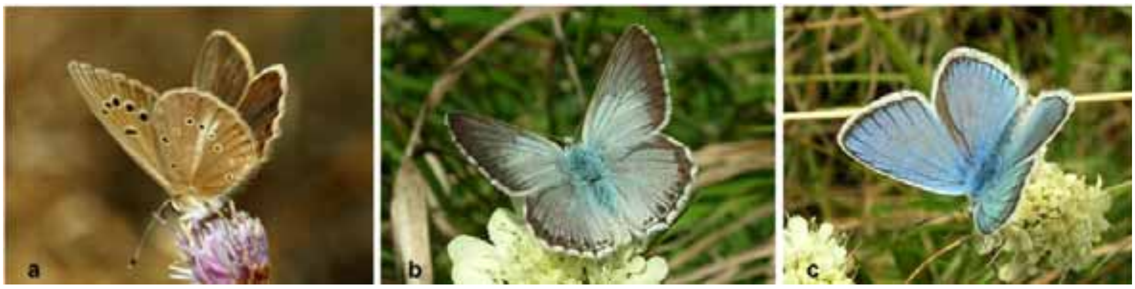


Figure 10. Butterflies of the genera/subgenera *Agrodiaetus*, *Lysandra* and *Plebicula* exhibit remarkable variability in chromosome numbers. (**a**) *Polyommatus* (*Agrodiaetus*) *violetae subbaeticus* (n=90); (**b**) *Lysandra coridon* (n=87-93); (**c**) *Polyommatus* (*Plebicula*) *dorylas* (n=146-151). [Photos: V. Dincă]

Objectives

OBJECTIVES

The relevance and evolutionary interest of the insects, as well as the gap of knowledge that they represent have been discussed in the introduction. This thesis has been structured to face a variety of challenging questions on systematics and evolution of the insects, methodologically conducted by phylogenetic inference. Given the enormous range of divergences at which tree reconstruction can be done, I propose to address six study cases at different taxonomic levels, covering a wide spectrum of phylogenetic divergence. The main objectives of this thesis are divided in three subcategories:

At deep taxon levels:

1) To assess the feasibility of using complete mitochondrial genomes as phylogenetic markers to resolve inter-ordinal relationships of insects. An exhaustive approach accounting for possible biases on the data and resulting phylogenetic artifacts, most notably saturation and long branch attraction, will allow exploring the limits of the phylogenetic signal of mitogenomes. Therefore, contrasting inferred mitogenomic phylogenies with current phylogenetic knowledge of insects will provide new evidence about controversial placements as the relative position of the Strepsiptera, the Hymenoptera or the Phthiraptera.

At intermediate divergences:

2) To perform a taxonomic revision of the butterflies from the *Polyommatus* section *sensu* Eliot 1973 based on constructing a reliable phylogenetic framework. Establishing a flexible temporal threshold to accommodate phylogenetic uncertainty and current taxonomic delineations, and defining a set of practical criteria to consolidate the systematics of such a complex group of butterflies.

At recent divergences among closely-related taxa:

3) To explore the factors affecting the Generalized Mixed Yule Coalescent (GMYC) model for species delimitation when dealing with phylogenetic trees from large-scale biodiversity surveys. A test with an adequate control dataset including all butterfly species for an entire country with a solid taxonomic framework will allow to analyze the general performance and the relevance of several factors to finally develop a list of best-practices for users.

4) To study the specific status of a potential new species of ant of the genus *Lasius* with interesting morphological and ecological traits. An exhaustive sampling and habitat inspection will allow studying morphological and molecular variation between populations and among putatively closely related taxa, documenting its placement in a phylogenetic context within the genus, and the assessment of its niche properties by species distribution modeling.

5) To evaluate the specific status of eight taxa in the butterfly subgenus *Agrodiaetus* with dot-like distributions and test for a potential taxonomic oversplitting of butterflies in western Europe. A joint analysis combining molecular phylogenetic singularity with a potential isolation by chromosomal rearrangements will permit revising the taxonomy and the conservation categories assigned.

6) To improve the knowledge on the phylogenetic history of the recent species radiation of the *Lysandra* butterflies in the context of chromosomal changes. By using the multispecies coalescent model, it will be possible to assess the limits among species, the internal relationships drawing the current geographic distributions and the role of chromosomal rearrangements on the evolution of the group.

Results and Discussion

Chapter I

Talavera, G., Vila, R. **2011**. What is the phylogenetic signal limit from mitogenomes? The reconciliation between mitochondrial and nuclear data in the Insecta class phylogeny. *BMC Evolutionary Biology* **11**:315

BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

What is the phylogenetic signal limit from mitogenomes? The reconciliation between mitochondrial and nuclear data in the Insecta class phylogeny

Gerard Talavera[1,2*] and Roger Vila[1*]

Abstract

Background: Efforts to solve higher-level evolutionary relationships within the class Insecta by using mitochondrial genomic data are hindered due to fast sequence evolution of several groups, most notably Hymenoptera, Strepsiptera, Phthiraptera, Hemiptera and Thysanoptera. Accelerated rates of substitution on their sequences have been shown to have negative consequences in phylogenetic inference. In this study, we tested several methodological approaches to recover phylogenetic signal from whole mitochondrial genomes. As a model, we used two classical problems in insect phylogenetics: The relationships within Paraneoptera and within Holometabola. Moreover, we assessed the mitochondrial phylogenetic signal limits in the deeper Eumetabola dataset, and we studied the contribution of individual genes.

Results: Long-branch attraction (LBA) artefacts were detected in all the datasets. Methods using Bayesian inference outperformed maximum likelihood approaches, and LBA was avoided in Paraneoptera and Holometabola when using protein sequences and the site-heterogeneous mixture model CAT. The better performance of this method was evidenced by resulting topologies matching generally accepted hypotheses based on nuclear and/or morphological data, and was confirmed by cross-validation and simulation analyses. Using the CAT model, the order Strepsiptera was recovered as sister to Coleoptera for the first time using mitochondrial sequences, in agreement with recent results based on large nuclear and morphological datasets. Also the Hymenoptera-Mecopterida association was obtained, leaving Coleoptera and Strepsiptera as the basal groups of the holometabolan insects, which coincides with one of the two main competing hypotheses. For the Paraneroptera, the currently accepted non-monophyly of Homoptera was documented as a phylogenetic novelty for mitochondrial data. However, results were not satisfactory when exploring the entire Eumetabola, revealing the limits of the phylogenetic signal that can be extracted from Insecta mitogenomes. Based on the combined use of the five best topology-performing genes we obtained comparable results to whole mitogenomes, highlighting the important role of data quality.

Conclusion: We show for the first time that mitogenomic data agrees with nuclear and morphological data for several of the most controversial insect evolutionary relationships, adding a new independent source of evidence to study relationships among insect orders. We propose that deeper divergences cannot be inferred with the current available methods due to sequence saturation and compositional bias inconsistencies. Our exploratory analysis indicates that the CAT model is the best dealing with LBA and it could be useful for other groups and datasets with similar phylogenetic difficulties.

* Correspondence: gerard.talavera@uab.cat; roger.vila@ibe.upf-csic.es
[1] Institut de Biologia Evolutiva (CSIC-UPF), Pg. Marítim de la Barceloneta 37, 08003 Barcelona, Spain
Full list of author information is available at the end of the article

Background

From the seminal comprehensive study of Hennig [1], to the impressive descriptive work of Kristensen [2,3], to the increasingly common molecular approaches [4-14], Insecta class systematics has been a challenging field of study. Molecular phylogenies have become a powerful tool that shed light on many parts of the Tree of Life. At the same time, due to the increasing number of sequences and genomes published, methodological questions are broadly explored by researchers in order to correctly and fully infer evolutionary relationships and patterns. In fact, it is widely accepted that many factors can influence final tree topologies, not to mention supports. Among these factors, we can cite 1) the quality of the sequences and the alignment; 2) the amount of phylogenetic information present in the sequences; 3) the presence of evolutionary biases that are not taken into account by most used evolutionary models (compositional heterogeneity, heterotachy...); 4) the use of markers whose evolution does not reflect the species evolutionary history (paralogs, xenologs); 5) the accuracy of the evolutionary model and the efficiency of the tree search algorithm used for the study [15-19]. Thus, different strategies in the analyses can often lead us to arrive at mutually contradictory conclusions starting from the same dataset. This seems to be particularly true when comparing the studies of relationships among the main taxonomic groups of Arthropoda [20-26]. Intra- and inter-ordinal insect relationships are not an exception and represent a ceaseless source of debate. They have been commonly explored using different types of molecular data: rDNA 18S and 28S, mitochondrial genes, complete mitochondrial genomes, nuclear protein coding genes, the presence of shared intron positions [12] or mitochondrial gene rearrangements [27]. Among the most controversial insect groups with regard to systematic position we can mention the Strepsiptera, an order of obligate endoparasitic and morphologically derived insects. The most basal relationships within the holometabolous and the paraneopteran insects are another example of long-debated relationships.

Mitochondrial genomes have been successful in recovering intra-ordinal phylogenetic relationships concordant with other sources of data, with convincing levels of support, such as in Diptera [28], Hymenoptera [29], Orthoptera [30] and Nepomorpha (Heteroptera) [31]. Nevertheless, mitogenomes proved so far to be generally inadequate to study inter-ordinal relationships of insects and deeper levels of Arthropoda, frequently resulting in strong incongruence with morphological and nuclear data, poor statistical supports, and high levels of inconsistency among different methods [16,24-26,32]. Indeed, comparative studies that contrast nuclear and mitochondrial datasets suggest that nuclear markers are better suited to deal with deep arthropod relationships, as the mitochondrial genome is on average more saturated, biased, and generally evolves at a much faster rate than the nuclear genome [33,34]. Thus, knowing the specific limits for each set of mitogenomes analyzed, i.e. when substitution rates result in saturation that distorts the phylogenetic signal at deeper nodes, is crucial to assess their usefulness in phylogenetics [35].

It is well known that arthropod mitochondrial genomes present some anomalous characteristics, like very high percentage of AT content, frequent gene rearrangements [36] or accelerated evolutionary rates likely related to phenotypic changes in body size or to parasitic lifestyle [37], all of which can limit their applicability in phylogenetic reconstruction. These biases in the data can result in systematic errors when the evolutionary model used for phylogenetic inference does not take them into account. Thus, homogeneous models of substitution or replacement where all sites evolve under the same substitution process [38] and constantly through time [19,39] are not adequate for Arthropoda. One of the most usual artefacts, especially in deep relationships where mutational saturation exists [40], is the long-branch attraction (LBA), a systematic error where two or more branches tend to cluster together producing false relationships [41]. Also, models not accounting for heterogeneity in nucleotide composition among taxa [16] can lead to artefactually group unrelated taxa with similar base composition [42-45].

For all these reasons, artropods in general and insects in particular, constitute an excellent model to tackle challenging questions of phylogenetic methodological interest. Several strategies have been designed to minimize potential biases: 1) Increasing the taxon sampling as far as possible, although generally counteracted by the removal of taxa with an evidently incorrect placement disturbing the reconstruction. 2) Filtering genes in large phylogenomic analyses to avoid paralogy problems and unexpected effects of missing data [45-47]. 3) The use of more specific substitution/replacement models. For example, matrices of amino acid replacement have been designed for Arthropoda (MtArt) [48,49] and Pancrustacea (MtPan) [25]. 4) Removing fast-evolving sites according to discrete gamma category [40,50-52]. 5) Removing third codon position or recoding them as purines and pyrimidines (RY-coding) in DNA alignments [23,53] to reduce the effects of saturation. 6) Using a site-heterogeneous mixture model (CAT) to allow flexible probabilities of the aminoacid replacement equilibrium frequencies, in order to minimise LBA effects [38,54,55].

In this work, we test the performance of different phylogenetic methodological strategies, using mitochondrial genomes of the Class Insecta as a model and including long-branched problematic taxa within Hymenoptera, Strepsiptera, Thysanoptera and Phthiraptera orders that have been usually excluded from mitochondrial datasets.

We address controversial taxonomical questions at three different levels of divergence, for which solid hypotheses based on nuclear phylogenies and morphological data exist. Our results show strong differences among the methods tested in their power to resolve inter-ordinal relationships. Using both real and simulated data (see Additional file 1), we confirm the capacity of the site-heterogeneous mixture model (CAT) under a Bayesian framework, currently implemented in software PhyloBayes [38,56], to substantially avoid the LBA artefacts. We show for the first time that the reconciliation between mitochondrial and previous nuclear and morphological knowledge is possible in the cases studied.

Results and Discussion

About the exploratory phylogenetic framework

After applying a variety strategies for phylogenetic inference, we compared the trees obtained to the most widely accepted nuclear DNA and morphology-based hypotheses for Holometabola and Paraneoptera systematics (Figure 1). These hypotheses were carefully selected from bibliography based on a great variety of data sources. As a result, we grouped the proposed relationships within Holometabola in two main hypotheses mainly disagreeing in the position of the order Hymenoptera, and a single general hypothesis for Paraneoptera, although this group has been much less intensely studied.

Methodologies that produced topologies identical or very similar to these hypotheses were considered better than those that resulted in very different trees. We noticed a strong susceptibility of our data to the type of analyses performed, confirming once more the instability of phylogenies based on insect mitochondrial genomes. Indeed, almost each approach resulted in a different topology and only the Bayesian inference using the CAT model with amino acid sequences (BI-AA-CAT) was able to obtain trees fitting potentially correct hypotheses. A better performance of the CAT model was confirmed using simulations (Additional file 1; Figure S1). No differences were observed between the MtRev and MtArt models in ML trees for any dataset. Cross-Validation statistics were performed to test the fit of the replacement models for proteins MtRev and CAT to the data. A better fit of the CAT model for Paraneoptera (mean score = 9.216 ± 12.038) and Eumetabola (mean score = 5.75, SD = ± 32.3539), and a similar fit for Holometabola (mean score = -0.055, SD = ± 32.3539) were detected.

The site-heterogeneous mixture model CAT [38] assumes the existence of distinct substitution processes, which usually results in a better fit to the data than site-homogeneous models based on empirical frequencies of amino acid or nucleotide substitutions, like MtRev or GTR [57-60]. In fact it has already been shown in other taxonomical groups that the CAT model is very powerful to overcome LBA artefacts [45,54,61-63]. Thus, the use of models accounting for compositional heterogeneity in the replacement process seems to be more effective than strategies focused on the removal of saturated positions in the case of Insect mitogenomes. Combining the CAT model with the use of amino acid sequences instead of DNA, which should reduce saturation biases, under a Bayesian framework produced the most satisfactory results. The topologies resulting from the analyses with different methods are discussed further on.

Holometabola phylogeny and the Strepsiptera problem

In our dataset for the Holometabola we included one strepsipteran and taxa of the Hymenoptera usually removed from mitochondrial analyses because of their long branches. We observed strong discrepancies among the methodologies used, ML-AA (Figure 2A), BI-DNA (Figure 2B) and BI-AA-CAT (Figure 2C) (see methods for details), confirming the difficulties introduced by such groups. The Strepsiptera species *Xenos vesparum* [64] appeared within Hymenoptera in the ML-AA tree, being completely trapped by the longest branches of the hymenopterans. The same happened with BI-DNA, although in that case, *Xenos* appeared in a more basal position within the hymenopterans, apparently slightly reducing the LBA effect. Finally when applying the BI-AA with the CAT model the LBA was suppressed, revealing completely different positions for the very long branches of Hymenoptera clade and the Strepsiptera (Figure 2C). The topology obtained in this case indicated a sister group relationship between Strepsiptera and Coleoptera (the composite clade being known as Coleopterida), and supported Diptera + Lepidoptera (Mecopterida), which represents the first evidence that mitochondrial data supports these groups.

Since their discovery, Strepsiptera has been associated with Diptera, Siphonaptera, Odonata, Ephemerida, Hymenoptera and Lepidoptera [65]. More recently, four different placements have been suggested as possibilities: membership in the Coleoptera [66], sister group to the Coleoptera [67,68], outside the Holometabola [2] and sister to Diptera [4,69]. Based on molecular studies of 18S rDNA, Whiting et al [4] proposed the grouping of Diptera + Strepsiptera under the name Halteria. Chalwatzis et al [70,71] reached similar conclusions about the relationships between Diptera and Strepsiptera using a larger dataset of 18S rDNA sequences. Later, other authors [72-74] attributed that grouping to an artefact due to LBA, becoming one of the best examples of LBA ever, called "the Strepsiptera problem".

Further phylogenetic evidence using other nuclear data contradicted the Halteria hypothesis, and supported associations between Strepsiptera and Coleoptera. Rokas et al (1999) [75] pointed to an intron insertion in *en* class homeoboxes of Diptera and Lepidoptera but not in

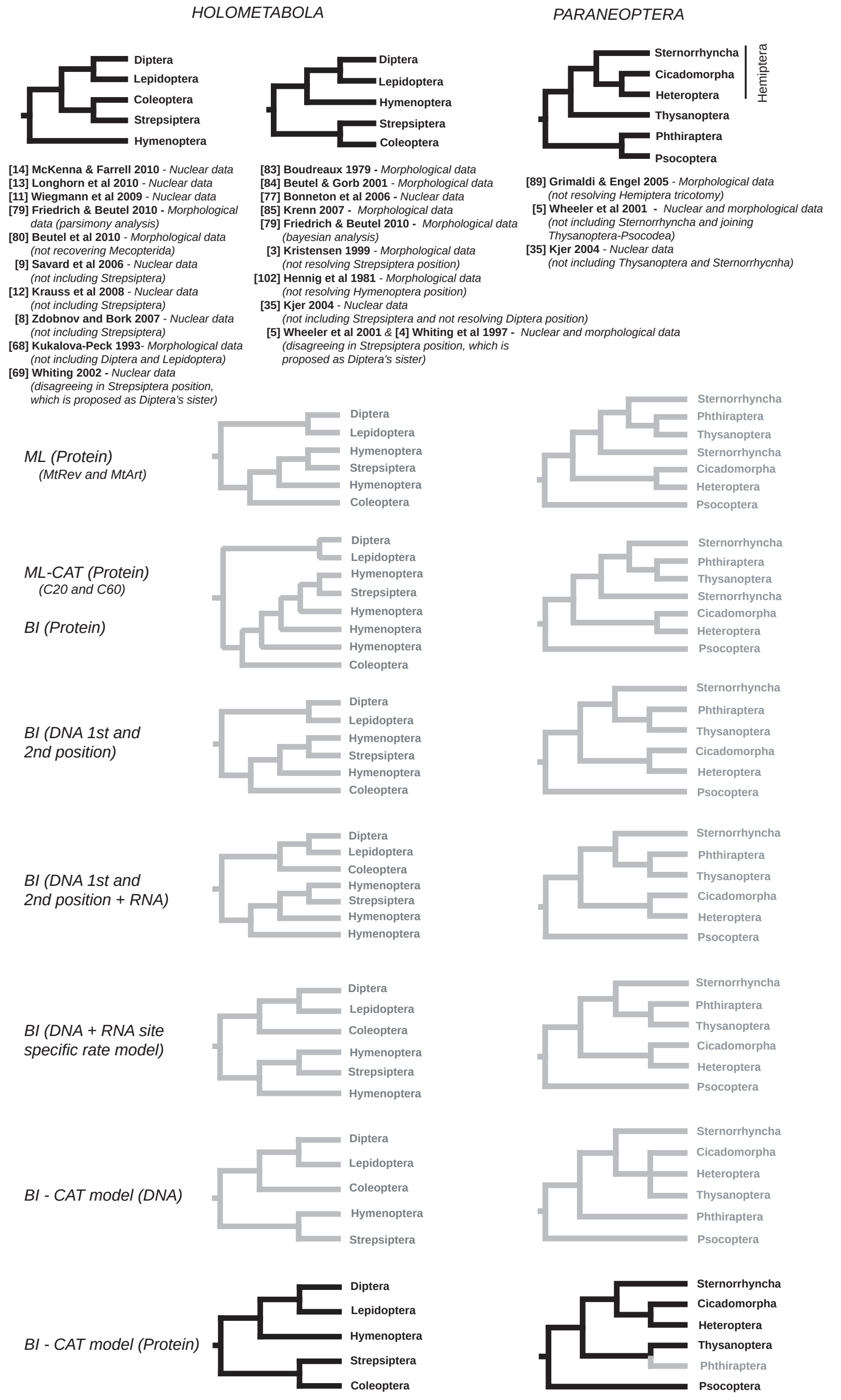

HOLOMETABOLA
PARANEOPTERA
Diptera
Lepidoptera
Coleoptera
Strepsiptera
Hymenoptera
[14] McKenna & Farrell 2010 - Nuclear data
[13] Longhorn et al 2010 - Nuclear data
[11] Wiegmann et al 2009 - Nuclear data
[79] Friedrich & Beutel 2010 - Morphological data (parsimony analysis)
[80] Beutel et al 2010 - Morphological data (not recovering Mecopterida)
[9] Savard et al 2006 - Nuclear data (not including Strepsiptera)
[12] Krauss et al 2008 - Nuclear data (not including Strepsiptera)
[8] Zdobnov and Bork 2007 - Nuclear data (not including Strepsiptera)
[68] Kukalova-Peck 1993- Morphological data (not including Diptera and Lepidoptera)
[69] Whiting 2002 - Nuclear data (disagreeing in Strepsiptera position, which is proposed as Diptera's sister)
Diptera
Lepidoptera
Hymenoptera
Strepsiptera
Coleoptera
[83] Boudreaux 1979 - Morphological data
[84] Beutel & Gorb 2001 - Morphological data
[77] Bonneton et al 2006 - Nuclear data
[85] Krenn 2007 - Morphological data
[79] Friedrich & Beutel 2010 - Morphological data (bayesian analysis)
[3] Kristensen 1999 - Morphological data (not resolving Strepsiptera position)
[102] Hennig et al 1981 - Morphological data (not resolving Hymenoptera position)
[35] Kjer 2004 - Nuclear data (not including Strepsiptera and not resolving Diptera position)
[5] Wheeler et al 2001 & [4] Whiting et al 1997 - Nuclear and morphological data (disagreeing in Strepsiptera position, which is proposed as Diptera's sister)
Sternorrhyncha
Cicadomorpha
Heteroptera
Hemiptera
Thysanoptera
Phthiraptera
Psocoptera
[89] Grimaldi & Engel 2005 - Morphological data (not resolving Hemiptera tricotomy)
[5] Wheeler et al 2001 - Nuclear and morphological data (not including Sternorrhyncha and joining Thysanoptera-Psocodea)
[35] Kjer 2004 - Nuclear data (not including Thysanoptera and Sternorrhycnha)
ML (Protein)
(MtRev and MtArt)
Diptera
Lepidoptera
Hymenoptera
Strepsiptera
Hymenoptera
Coleoptera
Sternorrhyncha
Phthiraptera
Thysanoptera
Sternorrhyncha
Cicadomorpha
Heteroptera
Psocoptera
ML-CAT (Protein)
(C20 and C60)
BI (Protein)
Diptera
Lepidoptera
Hymenoptera
Strepsiptera
Hymenoptera
Hymenoptera
Hymenoptera
Coleoptera
Sternorrhyncha
Phthiraptera
Thysanoptera
Sternorrhyncha
Cicadomorpha
Heteroptera
Psocoptera
BI (DNA 1st and 2nd position)
Diptera
Lepidoptera
Hymenoptera
Strepsiptera
Hymenoptera
Coleoptera
Sternorrhyncha
Phthiraptera
Thysanoptera
Cicadomorpha
Heteroptera
Psocoptera
BI (DNA 1st and 2nd position + RNA)
Diptera
Lepidoptera
Coleoptera
Hymenoptera
Strepsiptera
Hymenoptera
Hymenoptera
Sternorrhyncha
Phthiraptera
Thysanoptera
Cicadomorpha
Heteroptera
Psocoptera
BI (DNA + RNA site specific rate model)
Diptera
Lepidoptera
Coleoptera
Hymenoptera
Strepsiptera
Hymenoptera
Sternorrhyncha
Phthiraptera
Thysanoptera
Cicadomorpha
Heteroptera
Psocoptera
BI - CAT model (DNA)
Diptera
Lepidoptera
Coleoptera
Hymenoptera
Strepsiptera
Sternorrhyncha
Cicadomorpha
Heteroptera
Thysanoptera
Phthiraptera
Psocoptera
BI - CAT model (Protein)
Diptera
Lepidoptera
Hymenoptera
Strepsiptera
Coleoptera
Sternorrhyncha
Cicadomorpha
Heteroptera
Thysanoptera
Phthiraptera
Psocoptera

Figure 1 Phylogenetic strategies tested. Current knowledge on main holometabolan (two competing hypotheses) and paraneopteran relationships, based on nuclear and morphological data. Topologies obtained with the phylogenetic strategies tested are represented below (relationships matching the currently accepted hypotheses are highlighted in black)

Strepsiptera, Coleoptera or Hymenoptera, arguing that an intron loss is an improbable event. Based on a different approach, Hayward et al (2005) [76] used the structure of the USP/RXR hormone receptors, which showed a strong acceleration of evolutionary rate in Diptera and Lepidoptera, to reject the Halteria clade and to provide strong

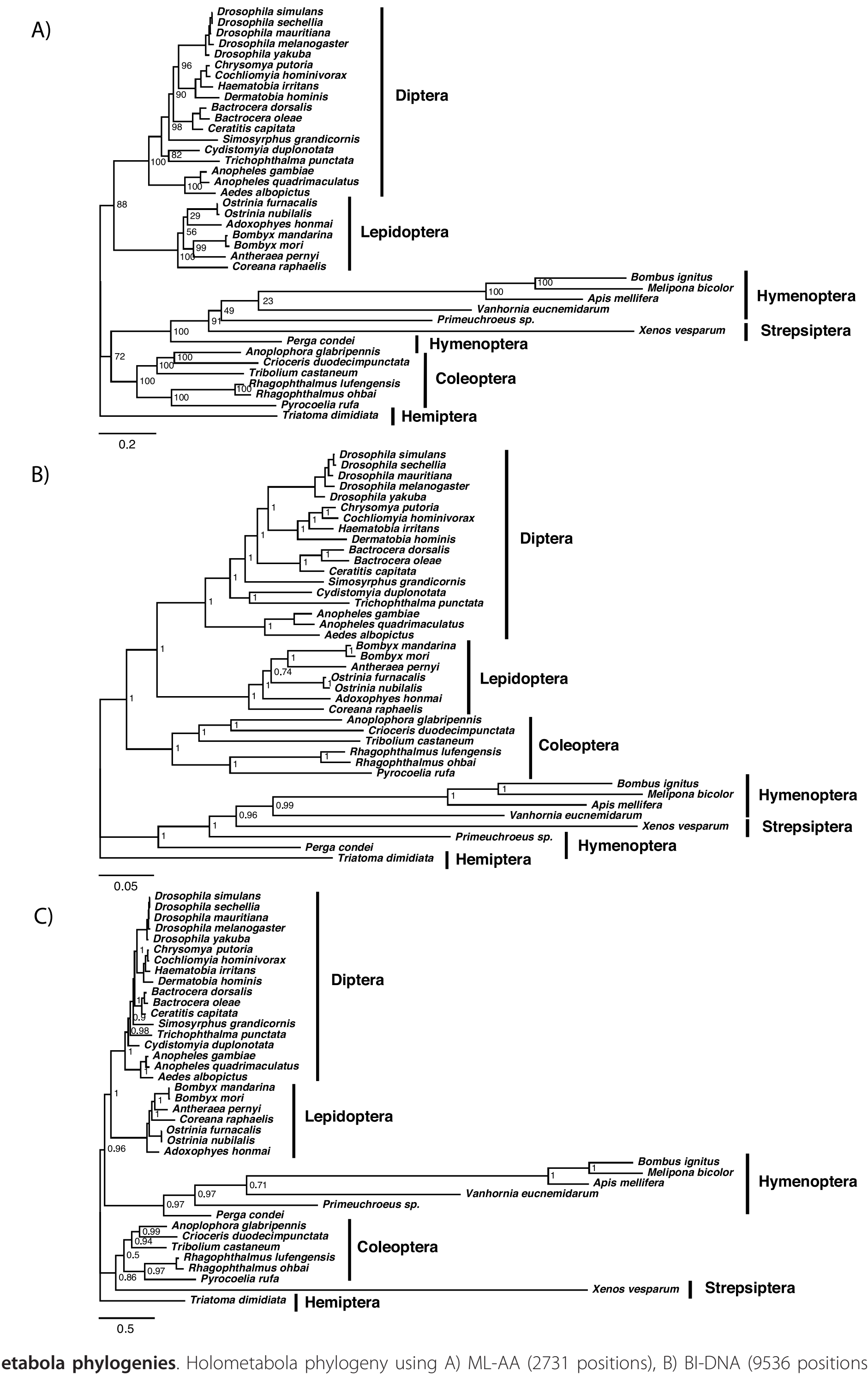


Figure 2 Holometabola phylogenies. Holometabola phylogeny using A) ML-AA (2731 positions), B) BI-DNA (9536 positions) and C) BI-AA-CAT (2731 positions). Values at nodes show bootstrap or posterior probabilities, and scale bar represents substitutions/site.

evidence for Mecopterida. Bonneton et al (2003, 2006) [77,78], confirmed the USP/RXR approach of Hayward et al and added the ecdysone receptor (ECR; NR1H1) to the analysis, confirming a Mecopterida monophyletic group. However, these two studies were not able to define clear associations for Strepsiptera. Misof et al (2007) [10] published a large phylogeny of Hexapoda using 18S rDNA and applying mixed DNA/RNA substitution models.

Although they recovered well-supported hexapod basal relationships, they obtained very low resolution and unclear relationships within Holometabola.

Recent molecular studies using extensive nuclear data seemed to contradict the Halteria hypothesis again, recovering a close relationship between Coleoptera and Strepsiptera [11,13,14]. First, Wiegmann et al (2009) [11] used a complete dataset of six nuclear protein coding genes including all holometabolan orders. They recovered Coleopterida and provided statistical evidence discarding LBA effects. They found some conflicting signal using individual genes like *cad*, which recovered Halteria, a result that was attributed to LBA because this is a rapidly evolving locus. Longhorn et al (2010) [13] used a total of 27 ribosomal proteins and tested several nucleotide-coding schemes for 22 holometabolan taxa, including two strepsipteran species, where a majority of the schemes tested recovered Coleopterida. McKenna and Farrell (2010) [14] raised identical conclusions using a total of 9 nuclear genes for 34 holometabolan taxa. Also, the Coleopterida have been recently recovered when using large morphological datasets [79,80]. Thus, evidence supporting Coleopterida has grown in recent years, suggesting that the phylogenetic placement of Strepsiptera has been definitely identified.

In summary, classical and most recent morphological and molecular studies based on nuclear data support the Mecopterida and Coleopterida hypotheses. Until now no mitochondrial evidence backed these hypotheses and our results are the first to fully agree with the most generally accepted point of view.

The Hymenoptera position and the basal splitting events of Holometabola

Depending on algorithm conditions, we observed inconsistencies among analyses in the Hymenoptera position (Figure 2). For example, the fact of using six gamma rate categories instead of four in ML-AA, or simply performing 5000000 instead of 1000000 runs (each with chain stability checked with Tracer) for BI-DNA, or assigning different partitions for DNA, tRNA and rRNA produced alternative results, either ((Diptera + Lepidoptera) Hymenoptera) Coleoptera) or (Diptera + Lepidoptera) + (Coleoptera + Hymenoptera) (not shown). Similar problems when using mitochondrial data have been previously described by Castro and Dowton (2005, 2007) [81,82] regarding this question, namely inconsistencies depending on the ingroup and outgroup selection and the analytical model. Overall, they described a tendency in their analyses to group Hymenoptera as sister taxa to Mecopterida, but they also found Hymenoptera or Hymenoptera + Coleoptera as the most basal lineages in some of their trees.

When using the BI-AA-CAT method our mitochondrial overview suggests a sister relationship of Hymenoptera with Mecopterida, placing Coleopterida outside a clade comprising the other examined holometabolan insects. This result coincides with one of the classical morphological points of view [3,83,84], some nuclear evidence [77], and with morphological and nuclear combined analyses [4,5] that recovered Coleoptera at the base of Holometabola (but not Strepsiptera). A phylogeny inferred from 356 anatomical characters by Beutel et al (2010) [80] placed Hymenoptera as the basal holometabolous insects and recovered a paraphyletic Mecopterida, although these groups were not strongly resolved. A morphological study based on characters of the thorax contributed by Friedrich and Beutel (2010) [79] offered two scenarios depending on the phylogenetic algorithm used: Coleopterida as the most basal group in the Bayesian analysis, but Hymeoptera as the most basal when using parsimony. Several hypotheses based on morphology situate hymenopterans as sister to Mecopterida [85,86], grouping coleopterans with the basal Endopterygota [see references in [87]], or with Neuroptera (not present in our dataset) [2,3,5,84-90]. Also, based on the analysis of wing characters, Kukalova-Peck & Lawrence (1993) [68] proposed an alternative phylogenetic hypothesis consisting in a most basal position for the Hymenoptera. Such discrepancies enhance the view that morphological characters are rather useless in order to determine the phylogenetic position of Hymenoptera within the Holometabola [69].

Our results do not support the most recent molecular studies based on nuclear data, all of them reporting Hymenoptera as the most basal holometabolan insects, for example, the phylogenomic results contributed by Savard et al (2006) [9] using a total of 185 nuclear genes. Since these authors were using emerging genome projects to assemble and analyze all the genes, they only were able to use 8 taxa with 4 orders of holometabolan insects represented (Diptera, Lepidoptera, Coleoptera and Hymenoptera). Their phylogeny resulted in a supported Coleoptera sister to Mecopterida clade, leaving Hymenoptera at the base. Zdobnov and Boork (2007) [8] obtained the same conclusions in another phylogenomic approach, using 2302 single copy orthologous genes for 12 genomes representing the same 4 holometabolous insect groups. Based on a dataset with similarly limited taxon sampling, and using the gain of introns close to older pre-existing ones as phylogenetic markers, Krauss et al (2008) [12] arrived to the same conclusion identifying 22 shared derived intron positions of Coleoptera with Mecopterida, in contrast to none of Hymenoptera with Mecopterida. Additionally, phylogenies with a large number of markers and a complete taxon sampling also gave rise to the same conclusions [11,13,14]. Therefore, mitochondrial data under the CAT model avoids obviously wrong relationships caused by LBA and recovers one of the two main current hypotheses. This hypothesis has been proposed mostly based on morphological evidence and differs from most recent nuclear and genomic results. This issue remains thus an open question deserving deeper study.

Paraneoptera phylogeny and the position of Phthiraptera

For Paraneoptera we observed once more an array of topological changes depending on the method used. In ML-AA (Figure 3A), Sternorrhyncha was recovered as paraphyletic with respect to Phthiraptera and Thysanoptera, which evidences the tendency of the method to join lineages according to relative branch length. Indeed, the white flies clade (Sternorrhyncha: Aleyrodoidea) displays a faster substitution rate than their relatives *Daktulosphaira vitifoliae*, *Schizapis graminum* and *Pachypsylla venusta*, and it seems to attract other long-branched clades: Phthiraptera and Thysanoptera. Using BI-DNA (Figure 3B), the topology improved and grouped all Sternorrhyncha representatives, although a paraphyletic Hemiptera remained. Only when using BI-AA-CAT (Figure 3C) a topology with most long-branched taxa not

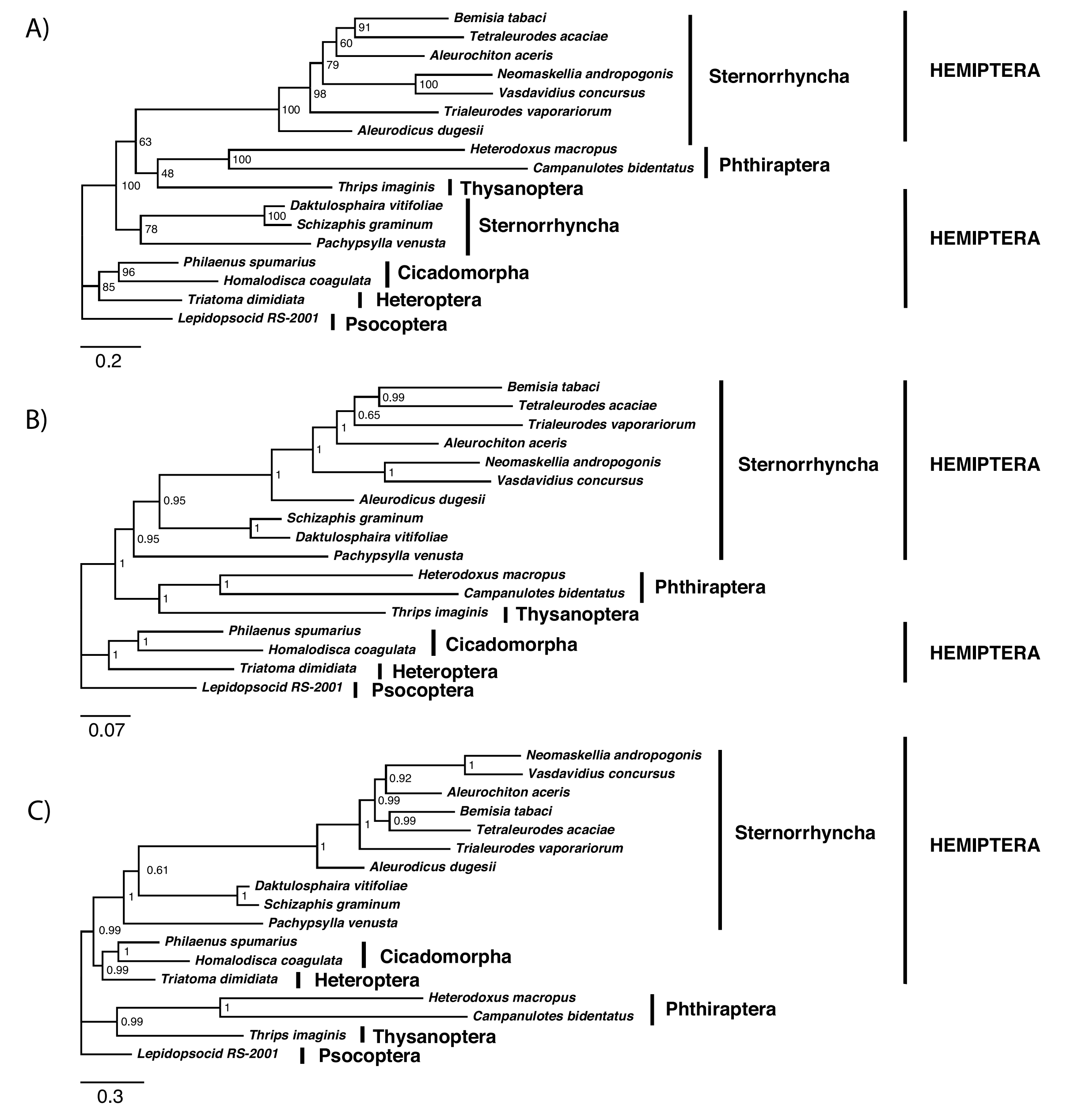


Figure 3 Paraneoptera phylogenies. Paraneoptera phylogeny using A) ML-AA (3501 positions), B) BI-DNA (10202 positions) and C) BI-AA-CAT (3501 positions). Values at nodes show bootstrap or posterior probabilities, and scale bar represents substitutions/site.

clustered and with a monophyletic Hemiptera was recovered.

Classically, Hemiptera is divided in two suborders: Homoptera and Heteroptera. Homoptera includes Sternorrhyncha and Auchenorrhyncha (Cicadomorpha + Fulgoromorpha). However, according to inferred phylogenies from 18S rDNA, Euhemiptera (Heteropterodea (Heteroptera + Coleorrhyncha) + Auchenorrhyncha) were proposed as sister group of Sternorrhyncha, leaving Homoptera as paraphyletic [91-93], which is currently the most accepted hypothesis. Our mitochondrial analyses with BI-AA-CAT produced the same conclusions as 18S rDNA datasets. Thus, Euhemiptera was recovered as a robust clade formed by Cicadomorpha plus Fulgoromorpha (a group known as Auchenorrhyncha) plus Heteropterodea, while Homoptera (Sternorrhyncha + Cicadomorpha + Coleorrhyncha) was paraphyletic with respect to the heteropteran *Triatoma dimidiata*, which appeared in all the tested methods as sister to Cicadomorpha with high support. In this study we were not able to test the Auchenorrhyncha paraphyly due to the lack of a Fulgoromorpha genome when the analyses were performed.

A sistergroup relationship between the Hemiptera and Thysanoptera, jointly known as Condylognatha [94,95], has been proposed based on morphological characters and supported by 18S rDNA data [96]. Moreover, the closest relatives of this group seem to be the Psocodea (= 'Psocoptera' + Phthiraptera). Although Homoptera paraphyly is fully accepted, at molecular level it has just been tested with nuclear single-gene phylogenies and the full Paraneoptera has never been studied with mitochondrial genomes. There is a broad acceptance that Paraneoptera is a monophyletic group of hemimetabolous insects, comprising the Hemiptera, Thysanoptera, and Psocodea, but the basal relationships within this group are quite controversial. The Condylognatha proposal (Hemiptera + Thysanoptera) was supported by several studies [83,97-102], although spermatological characters [2], fossil studies [103,104] and combined molecular and morphological data [4] suggested an alternative sistergroup relationship between Psocodea and Thysanoptera. Psocodea, however, is a fully accepted clade, even if the two orders included have been proposed to be mutually paraphyletic [96,105].

In our results, even using the AA-BI-CAT, which seemed to eliminate LBA artefacts for other clades, the basal Paraneoptera relationships were in contradiction with the generally accepted hypotheses. Psocodea was not monophyletic because Thysanoptera was recovered as the closest to Phthiraptera, and consequently Condylognatha is not supported. In fact, *Thrips imaginis* and the Phthiraptera genomes, were recovered as sister with high support in most of the methods tested. This result, although unexpected, cannot be readily dismissed as wrong and deserves more scrutiny (see for example [106]).

Eumetabola: Assessing the limits of the mitogenomic data

To try to understand what are the informative limits of the insect mitochondrial genomes, we raised the global divergence in our dataset by joining Paraneoptera and Holometabola genomes. With ML-AA and BI-DNA, all long branches grouped, a result obviously produced by LBA. Although resolution improved when the BI-AA-CAT was used, this method was not able to deal with the increased divergence and the result was not satisfactory (Figure 4). A tree with similar problems resulted when using the model CAT-BP, optimized to reduce the effects of compositional heterogeneity. Mainly, the hymenopterans remained within the long-branched cluster, although successfully including the short-branched Hymenoptera *Perga condei* with their relatives. Generally, although some signal was detected, this must be lower than the noise and considerable systematically erroneous relationships were recovered.

Given the observed inconsistencies when the divergence is increased in insects, we should question the utility of the Arthropoda mitogenomes to recover supraordinal phylogenetic information because of mutational saturation, at least with the current methodological offer.

Mitogenomic data have been used to successfully address several phylogenetic questions within mammals [107,108] and birds [53]. In both cases, however, relationships at the root level were not fully resolved, like the basal relationships between paleognaths and neognaths in birds, and Theria (marsupials plus placentals) versus Marsupionta (monotremes plus marsupials) hypotheses in mammals. In an ecdysozoan mitogenomics study testing the affinities of the three Panarthropoda phyla and the Mandibulata vs. Myriochelata hypothesis, Rota-Stabelli et al (2010) [60] also described difficulties caused by LBA. They obtained reasonable results only by removing rapidly evolving lineages and appyling the CAT model. Within Arthropoda, Nardi et al (2003) [22] presented an unexpected result based on a mitogenomic phylogeny: the paraphyly of the hexapods. They found crustaceans as sister to Insecta, and Collembola as sister to both. This result was discarded by Delsuc et al (2003) [23], who tried to avoid saturation and composition heterogeneity by recoding nucleotides as purines (R) and pyrimidines (Y), recovering then a monophyletic Hexapoda. Later, Cameron et al (2004) [21] performed a detailed battery of analysis including the major arthropod groups to test the hexapod monophyly. They removed hymenopteran and paraneopteran genomes from the analysis due to their extreme divergences. Even so, they could not obtain a conclusion

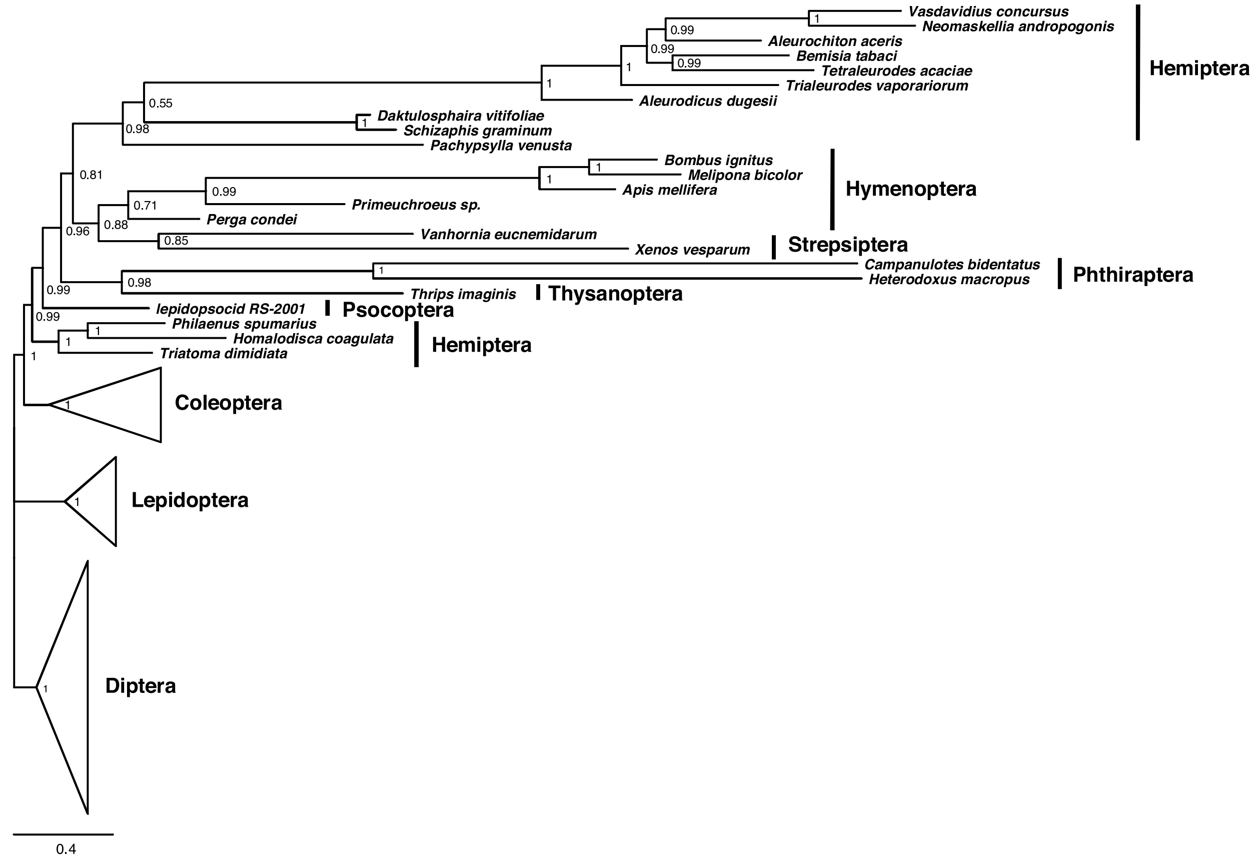


Figure 4 Eumetabola phylogeny. Eumetabola phylogeny obtained using BI-AA-CAT (3288 positions). Values at nodes show posterior probabilities and scale bar represents substitutions/site.

about the relationships due to the strong topological instability of the trees.

Cook et al (2005) [24] assessed the same question concluding that Crustacea and Hexapoda were mutually paraphyletic, although not including unstable lineages like Hymenoptera, the wallaby louse (*Heterodoxus macropus*) and others with long branches. Removing some important lineages, may strongly affect the general topology and the inferred evolutionary history. Carapelli et al (2007) [25] obtained the same conclusions by Cook et al (2005) [24] when including several additional genomes in the analysis and using a new model of amino acid replacement for Pancustracea, MtPan. They presented two large phylogenies based on DNA and protein alignments that supported a non-monophyletic Hexapoda, both obtaining a higher likelihood score under MtPan than when using MtArt or MtRev. However Carapelli et al did not test as many strategies and combinations like others did, in which case they would have probably arrived to similar contradicting conclusions. Some of the relationships within the Insecta they recovered were in strong disagreement with previous morphological and molecular evidence. For example: 1) They recovered a supported association between Strepsiptera and the Crustacean *Armillifer armillatus* (Pentastomida), two problematic yet clearly not related organisms sharing exceptional rates of evolution. 2) Diptera was included in the polyneopteran insect lineage when using BI-DNA and as an independent lineage from all the rest of the Insecta class when using BI-AA with MtPan model. 3) The positions of the orthropterans *Gryllotalpa orientalis* and *Locusta migratoria* remained unclear in their analysis. 4) The plecopteran *Pteronarcys* was recovered outside the polyneopterans, clustering with the Diptera. All of these cases were strongly supported by Bayesian posterior probabilities, but it is known that biases in deep phylogenies might increase supports of incorrect relationships. They attributed the non-monophyletic clade of Holometabola to a biased sampling, lacking orders like Mecoptera, Siphonaptera, Trichoptera or Neuroptera, but once more they removed the Hymenoptera from the analyses. Timmermans et al (2007) [26] re-evaluated the Collembola position using ribosomal protein gene sequences, which resulted in the supported monophyly of Hexapoda for all methodologies used (MP, ML and BI). They also found the inconsistency between nuclear and mitochondrial data when analyzing pancustracean relationships and clearly claimed that “caution is needed when applying mitochondrial markers in deep phylogeny”.

The limitations of the mitochondrial genome as phylogenetic marker were already pointed out by Curole and Kocher (1999) [109], when an increasing number of mitogenomes were sequenced and resulting phylogenies conflicted with morphological and nuclear hypotheses in the deep relationships of tetrapods and arthropods, as well as in mammals [110]. Within Insecta, more than one hundred mitogenomes are available now in GenBank/DDBJ/EMBL and they have been used to successfully resolve intra-ordinal relationships, such as in Diptera [28], Hymenoptera [29], Orthoptera [30] and Nepomorpha (Heteroptera) [31]. We report the difficulties to work on inter-ordinal relationships within Insecta, although showing that they can be generally avoided by using the BI-AA under the site-heterogeneous mixture model (CAT). However, we conclude that divergences in mitochondrial sequences above super-order levels represent an insurmountable problem for current methods. This result is at least valid for Arthropoda mitochondrial genomes, but difficult to extrapolate to other groups of organisms. We must remember some exceptional characteristics of the Insecta and Arthropoda in general, like high AT-content, the parasitic life-styles present in some groups or explosive radiation events in others. It is thus possible that a more relaxed evolutionary process in other metazoans allows for slightly deeper studies, and the limits of each dataset should be independently assessed.

Mitochondrial single genes reliability

Gene exclusion is one of the commonly used strategies to improve phylogenies and we tried to better understand the contribution of each mitochondrial gene to the phylogeny. Indeed, we found important variability in the phylogenetic signal of the different genes (Table 1). Five of the thirteen genes were especially informative in the topology resolution: *cox1*, *nad1*, *cytb*, *nad2* and *nad4*. On the contrary *atp6*, *atp8*, *cox2*, *cox3* and *nad4L* datasets produced the most different topologies. According to scale-factor values, *nad1*, *nad3*, *nad4*, *nad5* and *atp6* were the genes with a global divergence closest to the whole mitochondrial genome. *nad2*, *nad6*, *cox1*, *atp8* and *cytb* were the outliers in this case, giving the most deviated values. For both parameters, only *nad1* and *nad4* were among the best genes. Interestingly, some of the genes, for example *cox1*, performed very well regarding topology, but strongly deviated in divergence. The opposite applies to *atp6*. The unusually fast substitution rate of *cox1* compared to the mitochondrial mean (scale-factor = 2.0732) should be highlighted because this is the most common mitochondrial marker in single gene studies of insects and it is broadly used to infer molecular clocks in evolutionary time-based studies. According to this result, *cox1* seems to be a highly variable gene in insects, which makes it very suitable for the study of recent relationships and for DNA barcoding studies of this group of organisms.

Table 1 Scale-factor and Robinson-Foulds distances for individual mitochondrial genes

	Scale-factor	Robinson-Foulds
nad1	**1.1089**	**4**
nad2	0.7142	**5**
nad3	**0.9596**	7
nad4	**0.9771**	**6**
nad5	**0.9095**	8
nad6	0.7075	8
nad4L	0.7822	10
cox1	2.0732	**2**
cox2	0.8565	8
cox3	1.1510	8
atp6	**1.0279**	9
atp8	0.6826	11
cytb	1.3414	**4**

The best values for each parameter and the five best-scored genes for Robinson-Foulds are highlighted in bold.

Considering topology resolution as a priority in systematic studies, we selected the five best-scored genes for further comparisons with the whole genome. The phylogeny that resulted from their combined use reproduced a very similar topology to that of the entire dataset in several cases (Table 2). Thus, the resolution of the five gene combinations is comparable to that of a full genome, a result that could be explained by the inclusion of noise by the less informative genes. In conclusion, we suggest that the use of a selection of the most suitable genes is a valid (and simpler) strategy that produces results equivalent to the use of the entire genome. In order to apply this strategy in insect mitochondrial studies, we identify *cox1*, *nad1*, *cytb*, *nad2* and *nad4* as the best genes for topology, and *nad1*, *nad3*, *nad4*, *nad5* and *atp6* for branch lengths. We emphasize the importance of deciding what aspect of the mitogenome we want to estimate using a subset of genes, whether topology or branch length, because some of these genes (notoriously *cox1*) perform very well in one regard and poorly in the other.

Conclusions

Although several innovative phylogenetic methods have been developed to improve mitochondrial phylogenetic trees in some groups of organisms, results have been controversial in insects, leading to different conclusions that most often disagree with more generally accepted

Table 2 Scale-factor and Robinson-Foulds distance when comparing five concatenated genes versus whole genome in different datasets

	Scale-factor	Robinson-Foulds
Paraneoptera	1.19471	1
Paraneoptera (long-branched taxa excluded)	1.03562	1
Holometabola	1.26578	7
Holometabola (long-branched taxa excluded)	0.58196	2
Eumetabola	0.79679	13
Eumetabola (long-branched taxa excluded)	1.10733	4

relationships obtained from nuclear and morphological data. Thus, insects constitute a perfect model to test different methodologies and to better understand phylogenetic inference behaviour. Here we tested a battery of those strategies with three datasets of complete mitochondrial genomes of Insecta, including problematic taxa usually excluded from the analyses, and we compared the results with the current nuclear and morphological state of knowledge. The results suggested that the use of amino acid sequences instead of DNA is more appropriated at the inter-ordinal level and that the use of the site-heterogeneous mixture model (CAT) under a Bayesian framework, currently implemented in the software PhyloBayes, substantially avoids LBA artefacts. We show that inferring phylogenies above the super-order level constitutes the limit of the phylogenetic signal contained in insect mitochondrial genomes for currently available phylogenetic methods. For many of the relationships studied, we demonstrate for the first time that, with the proper methodology, mitochondrial data supports the most generally accepted hypotheses based on nuclear and morphological data. Thus, we confirm the non-monophyly of Homoptera within Paraneoptera, and recover Strepsiptera as a sister order to Coleoptera. In the basal splitting events in Holometabola we recover the Hymenoptera-Mecopterida association, and Coleoptera + Strepsiptera form a clade sister to the rest of Holometabola, which coincides with one of the two most accepted hypotheses. Recovered basal relationships in Paraneoptera differ from the currently accepted hypothesis in the position of Phthiraptera, which is recovered as sister to Thysanoptera, resulting in a paraphyletic Psocodea. By comparing single-gene to whole genome tree topologies, we select the five genes best performing for deep Insect phylogenetic inference. The combined used of these five genes (*cox1*, *nad1*, *cytb*, *nad2* and *nad4*) produces results comparable to those of mitogenomes, and we recommend the prioritary use of these markers in future studies.

Methods

Alignments

A total of 55 complete or almost complete Eumetabola mitochondrial genomes (17 of Paraneoptera and 38 of Holometabola) were downloaded from GenBank (Additional file 1: Table S1). Analyses were conducted using 3 datasets; 1) Holometabola, 2) Paraneoptera and 3) Eumetabola (Paraneoptera + Holometabola) in order to assess phylogenetic behaviour in a higher divergence level.

Every gene was translated to protein according to the arthropod mitochondrial genetic code and individually aligned using Mafft 5.861 [111]. To produce the DNA alignments, gaps generated in the protein alignment were transferred to the non-aligned DNA sequences using PutGaps software [112]. The resulting DNA and protein alignments for each gene were concatenated after removing problematic regions using Gblocks 0.91 [113] under a relaxed approach [15] with the next set of parameters: “Minimum Number Of Sequences For A Conserved Position” = 9, “Minimum Number Of Sequences For A Flank Position” = 13, “Maximum Number Of Contiguous Nonconserved Positions” = 8, “Minimum Lenght Of A Block” = 10, “Allowed Gap Positions” = “With Half”, and the kind of data was “by codons” for DNA and “Protein” for the aminoacids.

tRNA and rRNA sequences were individually aligned using ProbconsRNA 1.1 [114] and ambiguously aligned regions removed with Gblocks with the same parameters used for DNA. For the Paraneoptera dataset, both *tRNA-Leu* sequences from *Aleurodicus dugesii* were removed because they were extremely long in comparison to the rest and affected the alignment mechanism. For the Holometabola dataset, the large subunit ribosomal RNA sequence from *Anophophora glabripennis*, the *tRNA-Met* from *Ostrinia nubilalis* and *Ostrinia furnacalis*, and the *tRNA-Trp* from *Cysitomia duplonata* were unusually short and were not included. All these fragments were excluded from the Eumetabola dataset as well. Sequences were concatenated, and gaps were used instead of the removed RNAs and the few lacking coding genes.

Strategies for phylogenetic analysis

We tested several strategies for phylogenetic analyses on the three datasets. These differed in the phylogenetic algorithm, the treatment of saturation, and the use of different models of replacement: 1) Maximum likelihood on protein alignments under the MtRev and MtArt models (ML-AA); 2) Maximum likelihood on protein alignments under Empirical profile mixture models (20 and 60 profiles) (ML-AA-CAT) [38,115] 3) Bayesian inference on protein alignments under the MtRev model; 4) Bayesian inference on DNA alignments including only first and second codon positions for the 13 coding genes under the GTR+I+G model (BI-DNA); 5) Bayesian inference on

DNA alignments including first and second codon positions of the 13 coding genes, plus 22 tRNA and 2 rRNA [35] and under the GTR+I+G model; 6) Bayesian inference with a site specific rate model for all DNA + RNA positions [35] 7) Bayesian inference under the CAT model on DNA alignments including only first and second codon positions of the 13 coding genes; 8) Bayesian inference under the CAT model of protein alignments from the 13 coding genes (BI-AA-CAT) (Additional file 1: Table S2).

For maximum likelihood analyses, the software PhyML 2.4.4 [116] with the empirical MtRev model and six gamma rate categories was used. PhyML-CAT applying mixture models (C20 and C60) [115] was used when testing an alternative to empirical rate matrices in ML. For Bayesian inference, we used MrBayes v. 3.1.2 [117] and PhyloBayes 2.3 [38]. For MrBayes calculations in DNA alignments we used two partitions (first and second position of every codon), the GTR+I+G model, and four chains of 5.000.000 trees, sampling every 5000 generations. When including coding genes + tRNA + rRNA, sequences were partitioned in three independent partitions, one for each sequence type. For MrBayes analyses on protein alignments we used the MtRev model and four chains of 1.000.000 trees, sampling every 1000 generations, and applied a burn-in of 10% generations. For PhyloBayes analyses we used the site-heterogeneous mixture model CAT model for aminoacid sequences and the GTR-CAT model for the nucleotide sequences, and we run two independent chains of 5000 cycles, removing the first 1000 and sampling one point every five. For the site-specific rate model, characters were divided into six discrete rate categories using TreePuzzle [118] and partitioned in MrBayes from fastest to slowest, following a similar approach than in Kjer & Honeycutt [35]. Convergence of independent runs was checked with the software Tracer v1.4.

For the whole Eumetabola dataset, the CAT-BP model was tested, using the software nhPhylobayes v.023 [119,120]. This model is supposed to better account for amino-acid compositional heterogeneity, because it allows breakpoints along the branches of the phylogeny at which the amino acid composition can change. The number of components in the mixture were fixed to 120, according to the previous CAT-based phylogeny for Eumetabola. Four independent chains were run, and only two of them converged after highly demanding computation. Taking every tenth sampled tree, a 50% majority rule conseus tree was computed using the converged chains.

To statistically compare the CAT model with the stardard site-homogeneus models, cross validation statistics with PhyloBayes 3.3b were performed between the amino acid models (MtRev and CAT), as described in Philippe et al (2011) [121].

Mitochondrial single genes reliability

In order to explore the contribution of each individual gene to the concatenated tree, 13 single-gene phylogenies from the Paraneoptera dataset were reconstructed with BI-DNA excluding third codon positions. We scored each single-gene resulting phylogeny based on Robinson-Foulds distances and relative scale-factor values [122] using the complete mitochondrial tree as reference. The 5 best-scored genes were selected according to Robinson-Foulds distances and they were used to infer Paraneoptera, Holometabola and Eumetabola 5-gene phylogenies. Again, Robinson-Foulds distances and relative scale-factor values were calculated. In the same way, we also tested 5-gene performance when following a common practice in mitochondrial phylogenies of insects: the removal of rapidly evolving lineages with branch lengths deviating from the mean of the reference tree. To do that, taxa with a divergence to the root of the tree higher than 0.5 substitutions/position for Paraneoptera and Holometabola datasets and higher than 0.6 substitutions/position for the Eumetabola were removed. Thus, a total of 6 datasets were scored for the Robinson-Foulds distance and the scale-factor.

Additional material

Additional file 1: Additional Text, Figures and Tables. a) Table S1. List of mitochondrial genomes used in the study. b) Table S2. Number of characters in the final alignments for each phylogenetic reconstruction method tested. c) Simulations methods. d) Simulations results and discussion. d) Figure S1. Simulations. e) References.

Acknowledgements

We wish to thank José Castresana for his contribution and helpful ideas in designing the experiments and for comments on the manuscript, Cymon Cox for help with the nhPhylobayes analysis, and three anonymous reviewers for highly useful comments. Support for this research was provided by the Spanish Ministerio de Ciencia e Innovación project CGL2007-60516/BOS and CGL2010-21226/BOS; and predoctoral fellowships I3P-CPG-2006 and BES-2008-002054 to GT.

Author details

[1]Institut de Biologia Evolutiva (CSIC-UPF), Pg. Marítim de la Barceloneta 37, 08003 Barcelona, Spain. [2]Departament de Genètica i Microbiologia, Universitat Autònoma de Barcelona, Edifici C, 08193 Bellaterra, Spain.

Authors' contributions

GT designed the experiments and analyzed the data. GT and RV contributed in discussing and writing the paper. Both authors read and approved the final manuscript.

Received: 18 July 2011 Accepted: 27 October 2011
Published: 27 October 2011

References

1. Hennig W: *Die Stammesgeschichte der Insekten* Kramer, Frankfurt; 1969.
2. Kristensen NP: *Phylogeny of extant hexapods. The insects of Australia: A textbook for students and research workers* Melbourne University Press; 1991.

3. Kristensen NP: **Phylogeny of Endopterygote insects, the most successful lineage of living organisms.** *European Journal of Entomology* 1999, **96(3)**:237-253.
4. Whiting MF, Carpenter JC, Wheeler QD, Wheeler WC: **The Stresiptera problem: Phylogeny of the holometabolous insect orders inferred from 18S and 28S ribosomal DNA sequences and morphology.** *Systematic Biology* 1997, **46(1)**:1-68.
5. Wheeler WC, Whiting M, Wheeler QD, Carpenter JM: **The phylogeny of the extant hexapod orders.** *Cladistics* 2001, **17(4)**:403-404.
6. Kjer KM: **Aligned 18S and insect phylogeny.** *Systematic Biology* 2004, **53(3)**:506-514.
7. Ogden TH, Whiting MF, Wheeler WC: **Poor taxon sampling, poor character sampling, and non-repeatable analyses of a contrived dataset do not provide a more credible estimate of insect phylogeny: a reply to Kjer.** *Cladistics* 2005, **21(3)**:295-302.
8. Zdobnov EM, Bork P: **Quantification of insect genome divergence.** *Trends in Genetics* 2007, **23**:16-20.
9. Savard J, Tautz D, Richards S, Weinstock G, Gibbs R, Werren J, Tettelin H, M Lercher M: **Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of Holometabolous insects.** *Genome Research* 2006, **16(11)**:1334-1338.
10. Misof B, Niehuis O, Bischoff I, Rickert A, Erpenbeck D, Staniczek A: **Towards an 18S phylogeny of hexapods: Accounting for group-specific character covariance in optimized mixed nucleotide/doublet models.** *Zoology* 2007, **110(5)**:409-429.
11. Wiegmann BM, Trautwein MD, Kim JW, Cassel BK, Bertone MA, Winterton SL, Yeates DK: **Single-copy nuclear genes resolve the phylogeny of the holometabolous insects.** *BMC Biology* 2009, **7**:36.
12. Krauss V, Thümmler C, Georgi F, Lehmann J, Stadler PF, Eisenhardt C: **Near intron positions are reliable phylogenetic markers: an application to holometabolous insects.** *Mol Biol Evol* 2008, **25**:821-830.
13. Longhorn SJ, Pohl HW, Vogler AP: **Ribosomal protein genes of holometabolan insects reject the Halteria, instead revealing a close affinity of Strepsiptera with Coleoptera.** *Mol Phylogenet Evol* 2010, **55(3)**:846-859.
14. McKenna DD, Farrell BD: **9-genes reinforce the phylogeny of holometabola and yield alternate views on the phylogenetic placement of Strepsiptera.** *PloS One* 2010, **5(7)**:e11887.
15. Talavera G, Castresana J: **Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments.** *Systematic Biology* 2007, **56(4)**:564-577.
16. Hassanin A, Leger N, Deutsch J: **Evidence for multiple reversals of asymmetric mutational constraints during the evolution of the mitochondrial genome of Metazoa, and consequences for phylogenetic inferences.** *Systematic Biology* 2005, **54(2)**:277-298.
17. Fitch WM, Markowit E: **An improved method for determining codon variability in a gene and its application to rate of fixation of mutations in evolution.** *Biochemical Genetics* 1970, **4(5)**:579-93.
18. Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F: **Heterotachy and long-branch attraction in phylogenetics.** *BMC Evol Biol* 2005, **5**:50.
19. Pagel M, Meade A: **Modelling heterotachy in phylogenetic inference by reversible-jump Markov chain Monte Carlo.** *Philosophical Transactions of the Royal Society B-Biological Sciences* 2008, **363(1512)**:3955-3964.
20. Giribet G, Ribera C: **A review of arthropod phylogeny: New data based on ribosomal DNA sequences and direct character optimization.** *Cladistics* 2000, **16(2)**:204-231.
21. Cameron SL, Miller KB, D'Haese CA, Whiting MF, Barker SC: **Mitochondrial genome data alone are not enough to unambiguously resolve the relationships of Entognatha, Insecta and Crustacea sensu lato (Arthropoda).** *Cladistics* 2004, **20(6)**:534-557.
22. Nardi F, Spinsanti G, Boore JL, Carapelli A, Dallai R, Frati F: **Hexapod origins: Monophyletic or paraphyletic?** *Science* 2003, **299(5614)**:1887-1889.
23. Delsuc F, Phillips MJ, Penny D: **Comment on "Hexapod origins: Monophyletic or paraphyletic?".** *Science* 2003, **301(5639)**:1482.
24. Cook CE, Yue Q, Akam M: **Mitochondrial genomes suggest that hexapods and crustaceans are mutually paraphyletic.** *Proc Biol Sci* 2005, **272(1569)**:1295-1304.
25. Carapelli A, Lio P, Nardi F, van der Wath E, Frati F: **Phylogenetic analysis of mitochondrial protein coding genes confirms the reciprocal paraphyly of Hexapoda and Crustacea.** *BMC Evol Biol* 2007, **7(Suppl 2)**:S8.
26. Timmermans M, Roelofs D, Mariën J, van Straalen NM: **Revealing pancrustacean relationships: Phylogenetic analysis of ribosomal protein genes places Collembola (springtails) in a monophyletic Hexapoda and reinforces the discrepancy between mitochondrial and nuclear DNA markers.** *BMC Evolutionary Biology* 2008, **8**:83.
27. Boore JL, Collins TM, Stanton D, Daehler LL, Brown WM: **Deducing the pattern of arthropod phylogeny from mitochondrial-DNA rearrangements.** *Nature* 1995, **13**:163-165.
28. Cameron SL, Lambkin CL, Barker SC, Whiting MF: **A mitochondrial genome phylogeny of Diptera: whole genome sequence data accurately resolve relationships over broad timescales with high precision.** *Systematic Entomology* 2007, **32(1)**:40-59.
29. Cameron SL, Dowton M, Castro LR, Ruberu K, Whiting MF, Austin AD, Diement K, Stevens J: **Mitochondrial genome organization and phylogeny of two vespid wasps.** *Genome* 2008, **51(10)**:800-808.
30. Fenn J, Song H, Cameron SL, Whiting MF: **A preliminary mitochondrial genome phylogeny of Orthoptera (Insecta) and approaches to maximizing phylogenetic signal found within mitochondrial genome data.** *Molecular Phylogenetics and Evolution* 2008, **49(1)**:59-68.
31. Hua JM, Li M, Dong P, Cui Y, Xie Q, Bu W: **Phylogenetic analysis of the true water bugs (Insecta: Hemiptera: Heteroptera: Nepomorpha): evidence from mitochondrial genomes.** *BMC Evolutionary Biology* 2009, **9**:134.
32. Masta SE, Longhorn SJ, Boore JL: **Arachnid relationships based on mitochondrial genomes: Asymmetric nucleotide and amino acid bias affects phylogenetic analyses.** *Molecular Phylogenetics and Evolution* 2009, **50(1)**:117-128.
33. Burger G, Gray MW, Lang BF: **Mitochondrial genomes: anything goes.** *Trends in Genetics* 2003, **19(12)**:709-716.
34. Lin C, Danforth B: **How do insect nuclear and mitochondrial gene substitution patterns differ? Insights from Bayesian analyses of combined datasets.** *Mol Phylogenet Evol* 2004, **30(3)**:686-702.
35. Kjer KM, Honeycutt RL: **Site specific rates of mitochondrial genomes and the phylogeny of eutheria.** *BMC Evolutionary Biology* 2007, **7**:8.
36. Shao R, Dowton M, Murrell A, Barker SC: **Rates of gene rearrangement and nucleotide substitution are correlated in the mitochondrial genomes of insects.** *Mol Biol Evol* 2003, **20(10)**:1612-1619.
37. Castro LR, Austin AD, Dowton M: **Contrasting rates of mitochondrial molecular evolution in parasitic diptera and hymenoptera.** *Mol Biol Evol* 2002, **19(7)**:1100-1113.
38. Lartillot N, Philippe H: **A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process.** *Molecular Biology and Evolution* 2004, **21(6)**:1095-1109.
39. Kolaczkowski B, Thornton J: **Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous.** *Nature* 2004, **431**:980-984.
40. Rodriguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H: **Detecting and overcoming systematic errors in genome-scale phylogenies.** *Systematic Biology* 2007, **56(3)**:389-399.
41. Felsenstein J: **Cases in which parsimony or compatibility methods will be positively misleading.** *Systematic Zoology* 1978, **27(4)**:401-410.
42. Tarrío R, Rodriguez-Trelles F, Ayala F: **Shared nucleotide composition biases among species and their impact on phylogenetic reconstructions of the drosophilidae.** *Mol Biol Evol* 2001, **18(8)**:1464-1473.
43. Rosenberg M, Kumar S: **Heterogeneity of nucleotide frequencies among evolutionary lineages and phylogenetic inference.** *Mol Biol Evol* 2003, **20(4)**:610-621.
44. Rosenberg M, Kumar S: **Taxon sampling, bioinformatics, and phylogenomics.** *Systematic Biology* 2003, **52(1)**:119-124.
45. Brinkmann H, Van der Giezen M, Zhou Y, Poncelin de Raucourt G, Philippe H: **An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics.** *Systematic Biology* 2005, **54(5)**:743-757.
46. Philippe H, Delsuc F, Brinkmann H, Lartillot N: **Phylogenomics.** *Annual Review of Ecology, Evolution and Systematics* 2005, **36**:541-562.
47. Soria-Carrasco V, Valens-Vadell M, Pena A, Anton J, Amann R, Castresana J, Rossello-Mora R: **Phylogenetic position of Salinibacter ruber based on concatenated protein alignments.** *Syst Appl Microbiol* 2007, **30(3)**:171-179.
48. Abascal F, Posada D, Knight RD, Zardoya R: **Parallel evolution of the genetic code in arthropod mitochondrial genomes.** *PLoS Biology* 2006, **4**:711-718.

49. Abascal F, Posada D, Zardoya R: **MtArt: a new model on amino acid replacement for Arthropoda.** *Mol Biol Evol* 2007, **24**:1-5.
50. Ruiz-Trillo I, Riutort M, Littlewood DT, Herniou EA, Baguñà J: **Acoel flatworms: Earliest extant bilaterian metazoans, not members of Platyhelminthes.** *Science* 1999, **283(5409)**:1919-1923.
51. Brinkmann H, Philippe H: **Archaea sister group of bacteria? Indications from tree reconstruction artifacts in ancient phylogenies.** *Mol Biol Evol* 1999, **16(6)**:817-825.
52. Burleigh J, Mathews S: **Phylogenetic signal in nucleotide data from seed plants: Implications for resolving the seed plant tree of life.** *American Journal of Botany* 2004, **91**:1599-1613.
53. Harrison GL, McLenachan PA, Phillips MJ, Slack KE, Cooper A, Penny D: **Four new avian mitochondrial genomes help get to basic evolutionary questions in the late Cretaceous.** *Mol Biol Evol* 2004, **21(6)**:974-983.
54. Lartillot N, Brinkmann H, Philippe H: **Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model.** *BMC Evolutionary Biology* 2007, **7(Suppl 1)**:S4.
55. Roure B, Philippe H: **Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference.** *BMC Evolutionary Biology* 2011, **11**:17.
56. Lartillot N, Lepage T, Blanquart S: **PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating.** *Bioinformatics* 2009, **25(17)**:2286-2288.
57. Philippe H, Derelle R, Lopez P, Pick K, Borchiellini C, Boury-Esnault N, Vacelet J, Renard E, Houliston E, Quéinnec E, Da Silva C, Wincker P, Le Guyader H, Leys S, Jackson DJ, Schreiber F, Erpenbeck D, Morgenstern B, Wörheide G, Manuel M: **Phylogenomics Revives Traditional Views on Deep Animal Relationships.** *Current Biology* 2009, **19(8)**:706-712.
58. Lartillot N, Philippe H: **Improvement of molecular phylogenetic inference and the phylogeny of Bilateria.** *Philos Trans R Soc Lond B Biol Sci* 2008, **363**:1463-1472.
59. Sperling EA, Peterson KJ, Pisani D: **Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly or Eumetazoa.** *Mol Biol Evol* 2009, **26(10)**:2261-2274.
60. Rota-Stabelli O, Kayal E, Gleeson D, Daub J, Boore JL, Telford MJ, Pisani D, Blaxter M, Lavrov DV: **Ecdysozoan mitogenomics: evidence for a common origin of the legged invertebrates, the Panarthropoda.** *Genome Biology and Evolution* 2010, **2**:425-440.
61. Baurain D, Brinkmann H, Philippe H: **Lack of resolution in the animal phylogeny: Closely spaced cladogeneses or undetected systematic errors?** *Mol Biol Evol* 2007, **24**:6-9.
62. Philippe H, Brinkmann H, Martinez P, Riutort M, Baguña J: **Acoel flatworms are not platyhelmintes: evidence from phylogenomics.** *PLoS ONE* 2007, **2**: e717.
63. Delsuc F, Tsagkogeorga G, Lartillot N, Philippe H: **Additional molecular support for the new chordate phylogeny.** *Genesis* 2008, **46(11)**:592-604.
64. Carapelli A, Vannini L, Nardi F, Boore JL, Beani L, Dallai R, Frati F: **The mitochondrial genome of the entomophagous endoparasite Xenos vesparum (Insecta: Strepsiptera).** *Gene* 2006, **376(2)**:248-259.
65. Whiting MF: **Long-Branch Distraction and the Strepsiptera.** *Systematic Biology* 1998, **47(1)**:134-138.
66. Crowson RA: *The biology of the Coleoptera* London: Academic Press; 1981.
67. Kathirithamby J: **Review of the order Strepsiptera.** *Systematic Entomology* 1989, **14(1)**:41-92.
68. Kukalova-Peck J, Lawrence J: **Evolution of the hind wing in coleoptera.** *Canadian Entomologist* 1993, **125**:181-258.
69. Whiting MF: **Phylogeny of the holometabolous insect orders: molecular evidence.** *Zoologica Scripta* 2002, **31**:3-17.
70. Chalwatzis N, Baur A, Stetzer E, Kinzelbach R, Zimmermann F: **Strongly expanded 18S rRNA genes correlated with a peculiar morphology in the insect order of Strepsiptera.** *Zoologica* 1995, **98**:115-126.
71. Chalwatzis N, Hauf J, Van de Peer Y, Kinzelbach R, Zimmermann F: **18S ribosomal RNA genes of insects: Primary structure of the genes and molecular phylogeny of the Holometabola.** *Annals of the Entomological Society of America* 1996, **89**:788-803.
72. Hwang UW, Kim W, Tautz D, Friedrich M: **Molecular phylogenetics at the Felsenstein zone: Approaching the Strepsiptera problem using 5.8S and 28S rDNA sequences.** *Molecular Phylogenetics and Evolution* 1998, **9(3)**:470-480.
73. Carmean D, Crespi B: **Do long branches attract flies?** *Nature* 1995, **373**:666-666.
74. Huelsenbeck JP: **Systematic Bias in Phylogenetic Analysis: Is the Strepsiptera Problem Solved?** *Systematic Biology* 1998, **47(3)**:519-537.
75. Rokas A, Kathirithamby J, Holland PW: **Intron insertion as a phylogenetic character: the engrailed homeobox of Strepsiptera does not indicate affinity with Diptera.** *Insect Molecular Biology* 1999, **8(4)**:527-530.
76. Hayward DC, Trueman JW, Bastiani MJ, Ball EE: **The structure of the USP/RXR of Xenos pecki indicates that Strepsiptera are not closely related to Diptera.** *Development Genes and Evolution* 2005, **215(4)**:213-219.
77. Bonneton F, Brunet FG, Kathirithamby J, Laudet V: **The rapid divergence of the ecdysone receptor is a synapomorphy for Mecopterida that clarifies the Strepsiptera problem.** *Insect Molecular Biology* 2006, **15(3)**:351-362.
78. Bonneton F, Zelus D, Iwema T, Robinson-Rechavi M, Laudet V: **Rapid divergence of the ecdysone receptor in Diptera and Lepidoptera suggests coevolution between ECR and USP-RXR.** *Mol Biol Evol* 2003, **20(4)**:541-553.
79. Friedrich F, Beutel RG: **Goodbye Halteria? The thoracic morphology of Endopterygota (Insecta) and its phylogenetic implications.** *Cladistics* 2010, **26**:1-34.
80. Beutel RG, Friedrich F, Hörnschemeyer T, Phol H, Hünefeld F, Beckmann F, Meier R, Misof B, Whiting MF, Vilhelmsen L: **Morphological and molecular evidence converge upon a robust phylogeny of the megadiverse Holometabola.** *Cladistics* 2010, **26**:1-15.
81. Castro LR, Dowton M: **The position of the Hymenoptera within the Holometabola as inferred from the mitochondrial genome of Perga condei (Hymenoptera: Symphyta: Pergidae).** *Molecular Phylogenetics and Evolution* 2005, **34(3)**:469-479.
82. Castro LR, Dowton M: **Mitochondrial genomes in the Hymenoptera and their utility as phylogenetic markers.** *Systematic Entomology* 2007, **32(1)**:60-69.
83. Boudreaux HB: *Arthropod phylogeny with special reference to insects* New York: John Wiley & Sons; 1979.
84. Beutel R, Gorb S: **Ultrastructure of attachment specializations of hexapods, (Arthropoda): evolutionary patterns inferred from a revised ordinal phylogeny.** *Journal Of Zoological Systematics And Evolutionary Research* 2001, **39**:177-207.
85. Krenn HW: **Evidence from mouthpart structure on interordinal relationships in Endopterygota?** *Arthropod Systematics & Phylogeny* 2007, **65(1)**:7-14.
86. Whiting M: **Phylogeny of the Holometabolous Insects.** *Assembling the Tree of Life* Oxford University Press; 2004, 345-359.
87. Gullan PJ, Cranston PS: *The Insects: An Outline Of Entomology (3rd Edition)* Wiley-Blackwell; 2004.
88. Hornschemeyer T: **Phylogenetic significance of the wing-base of the Holometabola (Insecta).** *Zoologica Scripta* 2002, **31**:17-29.
89. Grimaldi D, Engel M: *Evolution of the insects* Cambrigde University Press; 2005.
90. Beutel RG, Pohl H: **Endopterygote systematics - where do we stand and what is the goal (Hexapoda, Arthropoda)?** *Systematic Entomology* 2006, **31(2)**:202-219.
91. Campbell BC, Steffen-Campbell JD, Sorensen JT, Gill RJ: **Paraphyly of Homoptera and Auchenorrhyncha inferred from 18S rDNA nucelotide sequences.** *Systematic Entomology* 1995, **20(3)**:175-194.
92. Sorensen JT, Campbell BC, Gill RJ, Steffen-Campbell JD: **Non-Monophyly of Auchenorrhyncha (Homoptera), Based Upon 18s rDNA Phylogeny - Eco-Evolutionary and Cladistic Implications within Pre-Heteropterodea Hemiptera (s.l) and a Proposal for New Monophyletic Suborders.** *Pan-Pacific Entomologist* 1995, **71(1)**:31-60.
93. Von Dohlen CD, Moran NA: **Molecular Phylogeny of the Homoptera: A Paraphyletic Taxon.** *Journal of Molecular Evolution* 1995, **41**:211-223.
94. Borner C: **Zur Systematik der Hexapoden.** *Zool Anz* 1904, **27**:511-533.
95. Yoshizawa K, Saigusa T: **Phylogenetic analysis of paraneopteran orders (Insecta: Neoptera) based on forewing base structure, with comments on monophyly of Auchenorrhyncha (Hemiptera).** *Systematic Entomology* 2001, **26(1)**:1-13.
96. Johnson KP, Yoshizawa K, Smith VS: **Multiple origins of parasitism in lice.** *Proc Biol Sci* 2004, **271(1550)**:1771-1776.
97. Kristensen NP: **The phylogeny of hexapod orders a critical review of recent accounts.** *Journal of Zoological Sytematics and Evolutionary Research* 1975, **13(1)**:1-44.
98. Kristensen NP: **Phylogeny of insect orders.** *Annual Review Of Entomology* 1981, **26**:135-157.

99. Kristensen NP: **The phylogeny of hexapod orders a critical review of recent accounts.** *Journal of Zoological Sytematics and Evolutionary Research* 1975, **13(1)**:1-44.
100. Kristensen NP: **Phylogeny of insect orders.** *Annual Review Of Entomology* 1981, **26**:135-157.
101. Hamilton K: **Morphology and evolution of the rhynchotan head (insecta, hemiptera, homoptera).** *Canadian Entomologist* 1981, **113(11)**:953-974.
102. Hennig W, Pont A, Schlee D: *Insect phylogeny* John Wiley & Sons; 1981.
103. Sharov AG: *Basic arthropodan stock with special reference to insects* Pergamon Press, New York; 1966.
104. Sharov AG: **The phylogenetic relations of the order Thysanoptera.** *Entomol Obozr* 1972, **51(4)**:854-858.
105. Murrell A, Barker SC: **Multiple origins of parasitism in lice: phylogenetic analysis of SSU rDNA indicates that the Phthiraptera and Psocoptera are not monophyletic.** *Parasitol Res* 2005, **97(4)**:274-80.
106. Shao R, Barker SC: **The Highly Rearranged Mitochondrial Genome of the Plague Thrips, *Thrips imaginis* (Insecta: Thysanoptera): Convergence of Two Novel Gene Boundaries and an Extraordinary Arrangement of rRNA Genes.** *Mol Biol Evol* 2003, **20(3)**:362-370.
107. Arnason U, Adegoke JA, Bodin K, Born EW, Esa YB, Gullberg A, Nilsson M, Short RV, Xu X, Janke A: **Mammalian mitogenomic relationships and the root of the eutherian tree.** *Proc Natl Acad Sci USA* 2002, **99(12)**:8151-8156.
108. Phillips MJ, Penny D: **The root of the mammalian tree inferred from whole mitochondrial genomes.** *Molecular Phylogenetics and Evolution* 2003, **28(2)**:171-185.
109. Curole J, Kocher T: **Mitogenomics: digging deeper with complete mitochondrial genomes.** *Trends Ecol Evol* 1999, **14(10)**:394-398.
110. Springer M, de Jong W: **Which mammalian supertree to bark up?** *Science* 2001, **291**:1709-1711.
111. Katoh K, Kuma K, Toh H, Miyata T: **MAFFT version 5: improvement in accuracy of multiple sequence alignment.** *Nucleic Acids Research* 2005, **33(2)**:511-518.
112. Fitzpatrick D, Pentony M: **PutGaps: DNA gapped file from Amino Acid alignment.**[http://bioinf.nuim.ie/software/putgaps].
113. Castresana J: **Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis.** *Mol Biol Evol* 2000, **17**:540-552.
114. Do CB, Mahabhashyam MS, Brudno M, Batzoglou S: **ProbCons: Probabilistic consistency-based multiple sequence alignment.** *Genome Research* 2005, **15(2)**:330-340.
115. Quant LS, Gascuel O, Lartillot N: **Empirical profile mixture models for phylogenetic reconstruction.** *Bioinformatics* 2008, **24**:2317-2323.
116. Guindon S, Gascuel O: **A simple, fasta and accurate algorithm to estimate phylogenies by maximum likelihood.** *Systematic Biology* 2003, **52(5)**:696-704.
117. Huelsenbeck JP, Ronquist F: **MRBAYES: Bayesian inference of phylogenetic trees.** *Bioinformatics* 2001, **17(8)**:754-755.
118. Schmidt HA, Strimmer K, Vingron M, Von Haeseler A: **TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing.** *Bioinformatics* 2002, **18**:502-504.
119. Blanquart S, Lartillot N: **A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneious sequence evolution.** *Mol Biol Evol* 2006, **23(11)**:2058-2071.
120. Blanquart S, Lartillot N: **A site- and time-heterogeneous model of amino acid replacement.** *Mol Biol Evol* 2008, **25(5)**:842-858.
121. Philippe H, Brinkmann H, Copley RR, Moroz LL, Nakano H, Poustka AJ, Wallberg A, Peterson KJ, Telford MJ: **Acoelomorph flatworms are deuterostomes related to Xenoturbella.** *Nature* 2011, **470**:255-258.
122. Soria-Carrasco V, Talavera G, Igea J, Castresana J: **The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees.** *Bioinformatics* 2007, **23**:2954-2956.

doi:10.1186/1471-2148-11-315
Cite this article as: Talavera and Vila: What is the phylogenetic signal limit from mitogenomes? The reconciliation between mitochondrial and nuclear data in the Insecta class phylogeny. *BMC Evolutionary Biology* 2011 **11**:315.

Additional file

What is the phylogenetic signal limit from mitogenomes? The reconciliation between mitochondrial and nuclear data in the Insecta Class phylogeny

Gerard Talavera & Roger Vila

Table S1 - List of mitochondrial genomes used in this study

Organism	GenBank Code	Taxonomical group
Holometabola		
Drosophila simulans	NC_005781.1	Diptera
Drosophila sechellia	NC_005780.1	Diptera
Drosophila mauritiana	NC_005779.1	Diptera
Drosophila melanogaster	NC_001709.1	Diptera
Drosophila yakuba	NC_001322.1	Diptera
Chrysomya putoria	AF352790.1	Diptera
Cochliomyia hominivorax	NC_002660.1	Diptera
Haematobia irritans	NC_007102.1	Diptera
Dermatobia hominis	NC_006378.1	Diptera
Bactrocera dorsalis	NC_008748.1	Diptera
Bactrocera oleae	NC_005333.1	Diptera
Ceratitis capitata	NC_000857.1	Diptera
Simosyrphus grandicornis	NC_008754.1	Diptera
Tricophtalma punctata	NC_008755.1	Diptera
Cydistomyia duplonata	NC_008756.1	Diptera
Anopheles gambiae	NC_002084.1	Diptera
Anopheles quadrimaculatus	NC_000875.1	Diptera
Aedes albopictus	NC_006817.1	Diptera
Bombyx mandarina	NC_003395.1	Lepidoptera
Bombyx mori	NC_002355.1	Lepidoptera
Antheraea pernyi	AY242996.1	Lepidoptera
Coreana raphaelis	DQ102703.1	Lepidoptera
Ostrinia furnacalis	NC_003368.1	Lepidoptera
Ostrinia nubilalis	NC_003367.1	Lepidoptera
Adoxophyes honmai	DQ073916.1	Lepidoptera
Bombus ignitus	DQ870926.1	Hymenoptera
Melipona bicolor	NC_004529.1	Hymenoptera
Apis mellifera	NC_001566.1	Hymenoptera
Vanhornia eucnemidarum	NC_008323.1	Hymenoptera
Primeuchroeus sp.	DQ302102.1 DQ302101.1	Hymenoptera
Perga condei	AY787816.1	Hymenoptera
Anoplophora glabripennis	NC_008221	Coleoptera
Crioceris duodecimpunctata	NC_003372.1	Coleoptera
Tribolium castaneum	NC_003081.1	Coleoptera
Rhagophthalmus lufengensis	DQ888607.1	Coleoptera
Rhagophthalmus ohbai	AB267275.1	Coleoptera
Pyrocoelia rufa	NC_003970.1	Coleoptera

Xenos vesparum	DQ364229.1	Strepsiptera
Paraneoptera		
Neomaskellia andropogonis	NC_006159.1	Hemiptera
Vasdavidius concursus	AY648941.2	Hemiptera
Aleurochiton aceris	NC_006160.1	Hemiptera
Bemisia tabaci	NC_006279.1	Hemiptera
Tetraleurodes acaciae	NC_006292.1	Hemiptera
Trialeurodes vaporariorum	NC_006280.1	Hemiptera
Aleurodicus dugesii	NC_005939.1	Hemiptera
Daktulosphaira vitifoliae	DQ021446.1	Hemiptera
Schizaphis graminum	NC_006158.1	Hemiptera
Pachypsylla venusta	AY278317.1	Hemiptera
Philaenus spumarius	AY630340.1	Hemiptera
Homalodisca coagulata	AY875213.1	Hemiptera
Triatoma dimidiata	NC_002609.1	Hemiptera
Heterodoxus macropus	NC_002651.1	Phthiraptera
Campanulotes bidentatus	NC_007884.1	Phthiraptera
Thrips imaginis	NC_004371.1	Thysanoptera
Lepidopsocid RS-2001	NC_004816.1	Psocoptera

Table S2 - Number of characters in the final alignments for each phylogenetic reconstruction method tested. Resulting trees shown in figures 2-4 are indicated.

	Paraneoptera	**Holometabola**	**Eumetabola**
ML (Protein)	2731 *(Fig.3)*	3501 *(Fig.2)*	3288
BI (Protein)	2731	3501	3288
BI (DNA 1st and 2nd position)	7010	7368	7232
BI (DNA 1st and 2nd position + RNA)	9536 *(Fig.3)*	10202 *(Fig.2)*	9548
BI (DNA + RNA) - Site specific rate model	13068	13889	-
BI (DNA 1st and 2nd position) - CAT model	7010	7368	7232
BI (Protein) - CAT model	2731 *(Fig.3)*	3501 *(Fig.2)*	3288 *(Fig.4)*

Simulations methods

We compared the efficiency of the protein-based phylogenetic reconstruction strategies using simulations. We performed simulated protein alignments with 1000 amino acid positions conducted along a reference phylogenetic tree with eight tips, with a global divergence equivalent to the Holometabola BI-DNA tree. In order to imitate possible LBA effects, two unrelated branches were forced to be six times longer than the average length of the rest (in the Holometabola BI-DNA tree the longest branches were four times bigger). One hundred simulations were performed using Seq-Gen [1] with six categories of rate heterogeneity (alfa = 0.872) and the MtRev evolutionary model. From these simulations, maximum likelihood trees with six categories of rate heterogeneity were inferred with Phyml 2.4.4, Bayesian inference with Mr.Bayes 3.1.2 and MtRev model, and PhyloBayes 2.3 under the CAT model. After that, we calculated the scale-factor, a relative value for comparing branch lengths between two trees, and the Robinson-Foulds distance, which calculates the topological differences between two trees, using Ktreedist 1.0 software [2] from each resulting tree versus the reference tree used to conduct the simulations.

Simulations results and discussion

Simulations were performed to confirm the ability of the CAT model to suppress the LBA bias compared to ML-AA and BI-DNA. We created two unrelated long branches in a tree with eight terminals with a similar divergence to the Holometabola dataset, and also exaggerated this divergence 2 and 3 times. The general tendency of the simulation test results was the same than the one observed in real data. For the three divergence levels explored, BI-AA-CAT produced the lowest percentage of trees grouping the two long branches as sister taxa. In divergence x1, BI-CAT did not group the long branches in any of the simulations (0%), while we obtained a 9% for ML-AA and a 10% for BI-DNA. For divergence x2, long branches were grouped together in 3% of the cases for BI-AA-CAT, 12% for ML-AA and 26% for BI-DNA. Finally for divergence x3, the values increased to 10% for BI-AA-CAT, 34% for ML-AA and 38 % for BI-DNA. These percentages do not evaluate intermediate LBA effects, where both branches might be closer than they should, but not strictly sisters. To evaluate the topological differences between the simulated trees and the reference topology, presumably a product of LBA, we calculated Robinson-Foulds distances and calculated the average of the 100 simulations for each divergence type and method. This revealed again the better performance of BI-AA-CAT, which obtained the lowest values, followed by ML-AA and BI-DNA respectively (Figure S1). A better performance of the amino acid sequences versus DNA was also reflected in these results, and their use together with a site-heterogeneous mixture model under a Bayesian framework is the suggested combination to avoid LBA artefacts.

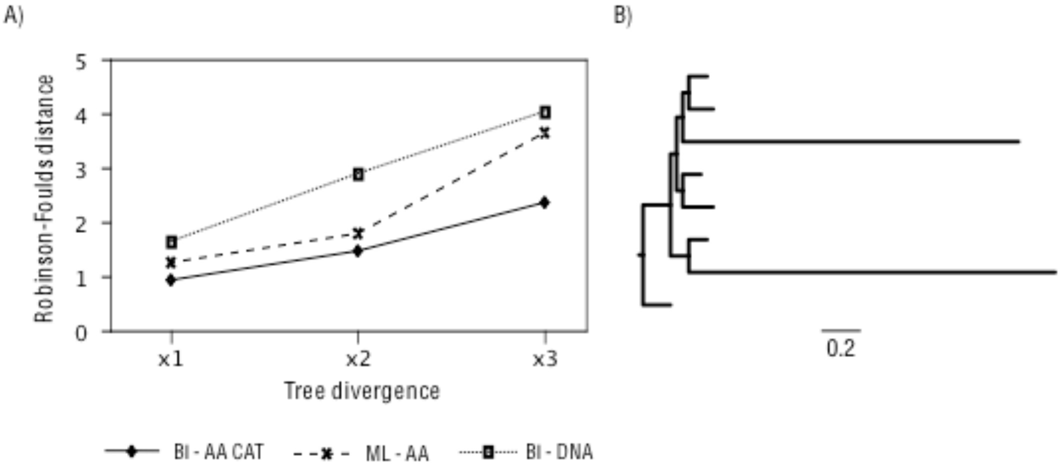


Figure S1 - Simulations

A) Average Robinson-Foulds distances relative to the reference tree calculated with ML - protein sequences (dotted line with squared symbols), BI - DNA excluding third codon positions (dashed line with cross symbols) and BI - protein with CAT model (solid line with diamonds) for three different tree divergences. B) Reference tree (divergence x1) used to conduct simulations. Scale bar represents 0.2 substitutions/site.

References:

1. Rambaut A, Grassly NC: **Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees.** *Comput Appl Biosci 1997,* **13**:235-238
2. Soria-Carrasco V, Talavera G, Igea J, Castresana J: **The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees.** *Bioinformatics 2007,* **23**:2954-2956

Chapter II

Talavera, G., Lukhtanov, V., Pierce, N.E., Vila, R. **2012**. Establishing criteria for higher-level classification using molecular data: the systematics of Polyommatus blue butterflies (Lepidoptera, Lycaenidae). *Cladistics.*

Published online. DOI: 10.1111/j.1096-0031.2012.00421.x

Cladistics

Cladistics (2012) 1–27

10.1111/j.1096-0031.2012.00421.x

Establishing criteria for higher-level classification using molecular data: the systematics of *Polyommatus* blue butterflies (Lepidoptera, Lycaenidae)

Gerard Talavera[a,b], Vladimir A. Lukhtanov[c,d], Naomi E. Pierce[e] and Roger Vila[a,]*

[a]Institut de Biologia Evolutiva (CSIC-UPF), Passeig Marítim de la Barceloneta, 37, 08003 Barcelona, Spain; [b]Departament de Genètica i Microbiologia, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain; [c]Department of Karyosystematics, Zoological Institute of Russian Academy of Science, Universitetskaya nab. 1, 199034 St Petersburg, Russia; [d]Department of Entomology, St Petersburg State University, Universitetskaya nab. 7/9, 199034 St Petersburg, Russia; [e]Department of Organismic and Evolutionary Biology and Museum of Comparative Zoology, Harvard University, 26 Oxford Street, Cambridge, MA 02138, USA

Accepted 11 June 2012

Abstract

Most taxonomists agree on the need to adapt current classifications to recognize monophyletic units. However, delineations between higher taxonomic units can be based on the relative ages of different lineages and/or the level of morphological differentiation. In this paper, we address these issues in considering the species-rich *Polyommatus* section, a group of butterflies whose taxonomy has been highly controversial. We propose a taxonomy-friendly, flexible temporal scheme for higher-level classification. Using molecular data from nine markers (6666 bp) for 104 representatives of the *Polyommatus* section, representing all but two of the 81 described genera/subgenera and five outgroups, we obtained a complete and well resolved phylogeny for this clade. We use this to revise the systematics of the *Polyommatus* blues, and to define criteria that best accommodate the described genera within a phylogenetic framework. First, we normalize the concept of section (*Polyommatus*) and propose the use of subtribe (Polyommatina) instead. To preserve taxonomic stability and traditionally recognized taxa, we designate an age interval (4–5 Myr) instead of a fixed minimum age to define genera. The application of these criteria results in the retention of 31 genera of the 81 formally described generic names, and necessitates the description of one new genus (***Rueckbeilia* gen. nov.**). We note that while classifications should be based on phylogenetic data, applying a rigid universal scheme is rarely feasible. Ideally, taxon age limits should be applied according to the particularities and pre-existing taxonomy of each group. We demonstrate that the concept of a morphological gap may be misleading at the genus level and can produce polyphyletic genera, and we propose that recognition of the existence of cryptic genera may be useful in taxonomy.

Despite current progress in morphological and molecular studies of ‘‘Blue’’ butterflies, subfamily Polyommatinae (Forster, 1936, 1938; Stempffer, 1937, Stempffer, 1967; Nabokov, 1945; Eliot, 1973; Als et al., 2004; Zhdanko, 2004; Stekolnikov and Kuznetzov, 2005; Wiemers et al., 2009; Stekolnikov, 2010), their higher-level systematics remain controversial. Eliot (1973) divided this subfamily into four tribes: Lycaenesthini, Candalidini, Niphandini and Polyommatini (Table 1). Among these tribes, the Polyommatini is the most diverse and arguably one of the most systematically difficult groups of butterflies, as stated by Eliot himself: ‘‘I have to admit complete failure in my efforts to subdivide it into natural groups, simply organizing it into 30 sections’’ (Eliot, 1973). His division of Polyommatini into sections has nevertheless been widely accepted by the scientific community (Hirowatari, 1992; Mattoni and Fiedler, 1993; Bálint and Johnson, 1994, 1995, 1997; Io, 1998; Pratt et al., 2006; Robbins and Duarte, 2006). Some entomologists prefer considering

*Corresponding author.
E-mail address: roger.vila@ibe.upf-csic.es

Table 1
Polyommatinae classification according to Eliot (1973)

Tribe	Section	Genera
Lycaenesthini		*Lycaenesthes* Moore, 1866; *Anthene* Doubleday, 1847; *Cupidesthes* Aurivillius, 1895; *Neurypexina* Bethune-Baker, 1910; *Neurellipes* Bethune-Baker 1910; *Monile* Ungemach, 1932; *Triclema* Karsch, 1893
Candalidini		*Candalides* Hübner, 1819; *Erina* Swainson, 1833 (= *Holochila* C. Felder, 1862); *Cyprotides* Tite, 1963; *Microscena* Tite, 1963; *Adaluma* Tindale, 1922; *Nesolycaena* Waterhouse & Lyell, 1905; *Zetona* Waterhouse, 1938; *Holochila* sensu auctt. nec C. Felder
Niphandini		*Niphanda* Moore, 1875
Polyommatini		
	Cupidopsis	*Cupidopsis* Karsch, 1895
	Una	*Una* de Nicéville, 1890; *Orthomiella* de Nicéville, 1890
	Petrelaea	*Petrelaea* Toxopeus, 1929; *Pseudonacaduba* Stempffer, 1943
	Nacaduba	*Nacaduba* Moore, 1881; *Prosotas* H. H. Druce, 1891; *Ionolyce* Toxopeus, 1929; *Catopyrops* Toxopeus, 1929; *Erysichton* Fruhstorfer, 1916; *Paraduba* Bethune-Baker, 1906; *Neolucia* Waterhouse & Turner, 1905; *Hypojamides* Riley, 1929
	Theclinesthes	*Theclinesthes* Röber, 1891; *Thaumaina* Bethune-Baker, 1908; *Utica* Hewitson, 1865, invalid, praeocc.
	Upolampes	*Upolampes* Bethune-Baker, 1908; *Caleta* Fruhstorfer, 1922; *Pycnophallium* Toxopeus, 1929; *Discolampa* Toxopeus, 1929 (= *Ethion* Shirozu & Saigusa, 1962); *Pistoria* Hemming, 1964 (= *Mambara* Bethune-Baker, 1908, praeocc.)
	Danis	*Danis* Fabricius, 1807 (= *Thysonotis* Hübner, 1819; *Hadothera* Billberg, 1820; *Damis* Boisduval, 1832); *Psychonotis* Toxopeus, 1930; *Epimastidia* H. H. Druce, 1891
	Jamides	*Jamides* Hübner, 1819; *Pepliphorus* Hübner, 1819 (= *Peplodyta* Toxopeus, 1929)
	Catochrysops	*Catochrysops* Boisduval, 1832; *Rysops* Eliot, 1973
	Lampides	*Lampides* Hübner, 1819 (= *Cosmolyce* Toxopeus, 1927; *Lampidella* Hemming, 1933)
	Callictita	*Callictita* Bethune-Baker, 1908
	Uranothauma	*Uranothauma* Butler, 1895
	Phlyaria	*Phylaria* Karsch, 1895
	Cacyreus	*Cacyreus* Butler, 1898 (= *Hyreus* Hübner, 1819, praeocc.); *Harpendyreus* Heron, 1909
	Leptotes	*Leptotes* Scudder, 1876; *Syntarucoides* Kaye, 1904; *Cyclyrius* Butler, 1897; *Syntarucus* Butler, 1900 (= *Langia* Tutt, 1906, praeocc.)
	Castalius	*Castalius* Hübner, 1819; *Tarucus* Moore, 1881
	Zintha	*Zintha* Eliot, 1973
	Zizeeria	*Zizeeria* Chapman, 1910; *Zizina* Chapman, 1910; *Pseudozizeeria* Beuret, 1955
	Famegana	*Famegana* Eliot, 1973
	Actizera	*Actizera* Chapman, 1910
	Zizula	*Zizula* Chapman, 1910
	Brephidium	*Brephidium* Scudder, 1876; *Oraidium* Bethune-Baker, 1914
	Everes	*Everes* Hübner, 1819 (= *Ununcula* van Eecke, 1915); *Cupido* Schrank, 1801 (= *Zizera* Moore, 1881); *Tiora* Evans, 1912; *Bothrinia* Chapman, 1909 (= *Bothria* Chapman, 1908, praeocc.); *Tongeia* Tutt, 1908; *Shijimia* Matsumura, 1919; *Talicada* Moore, 1881; *Binghamia* Tutt, 1908

Table 1
(*Contiuned*)

Tribe	Section	Genera
	Pithecops	*Pithecops* Horsfield, 1828; *Eupsychellus* Röber, 1891
	Azanus	*Azanus* Moore, 1881
	Eicochrysops	*Eicochrysops* Bethune-Baker, 1924
	Lycaenopsis	*Lycaenopsis* C. & R. Felder, 1865; *Neopithecops* Distant, 1884; *Parapithecops* Moore, 1884; *Megisba* Moore, 1881; *Pathalia* Moore, 1884; *Arletta* Hemming, 1935 (= *Moorea* Toxopeus, 1927, praeocc.); *Celastrina* Tutt, 1906; *Notarthrinus* Chapman, 1908; *Acytolepis* Toxopeus, 1927; *Oreolyce* Toxopeus, 1927; *Monodontides* Toxopeus, 1927; *Akasinula* Toxopeus, 1928; *Ptox* Toxopeus, 1928; *Udara* Toxopeus, 1928; *Rhinelephas* Toxopeus, 1928; *Uranobothria* Toxopeus, 1928; *Parelodina* Bethune-Baker, 1904; *Vaga* Zimmerman, 1958; *Papua* Röber, 1892, invalid, praeocc.; *Cyanirioides* Matsumura, 1919, invalid, praeocc.
	Glaucopsyche	*Glaucopsyche* Scudder, 1872; *Phaedrotes* Scudder, 1876; *Scolitantides* Hübner, 1819; *Apelles* Hemming, 1931; *Philotes* Scudder, 1876; *Turanana* Bethune-Baker, 1916 (= *Turania* Bethune-Baker, 1914, praeocc.); *Palaeophilotes* Forster, 1938; *Praephilotes* Forster, 1938; *Pseudophilotes* Beuret, 1955; *Shijimiaeoides* Beuret, 1955; *Sinia* Forster, 1949; *Iolana* Bethune-Baker, 1914; *Maculinea* van Eecke, 1915; *Caerulea* Forster, 1938; *Phengaris* Doherty, 1881
	Euchrysops	*Euchrysops* Butler, 1900; *Lepidochrysops* Hedicke, 1923 (= *Neochrysops* Bethune-Baker, 1923, praeocc.); *Thermoniphas* Karsch, 1895; *Oboronia* Karsch, 1893; *Athysanota* Karsch, 1895
	Polyommatus	*Polyommatus* Latreille, 1804; *Plebejus* Kluk, 1802; *Lycaeides* Hübner, 1819; *Cyaniris* Dalman, 1816; *Nomiades* Hübner, 1819; *Aricia* R. L., 1817 (= *Gynomorphia* Verity, 1929); *Pseudoaricia* Beuret, 1959; *Kretania* Beuret, 1959; *Ultraaricia* Beuret, 1959; *Agriades* Hübner, 1819; *Vacciniina* Tutt, 1909; *Albulina* Tutt, 1909; *Bryna* Evans, 1912; *Meleageria* Sagarra, 1925; *Agrodiaetus* Hübner, 1822 (= *Hirsutina* Tutt, 1909); *Lysandra* Hemming, 1933 (= *Uranops* Hemming, 1929, praeocc.); *Plebicula* Higgins, 1969; *Eumedonia* Forster, 1938; *Plebulina* Nabokov, 1944; *Icaricia* Nabokov, 1944; *Chilades* Moore, 1881; *Edales* Swinhoe, 1910; *Luthrodes* H. H. Druce, 1895; *Freyeria* Courvoisier, 1920; *Hemiargus* Hübner, 1818; *Itylos* Draudt, 1921; *Pseudochrysops* Nabokov, 1945; *Cyclargus* Nabokov, 1945; *Echinargus* Nabokov, 1945; *Pseudolucia* Nabokov, 1945; *Paralycaeides* Nabokov, 1945; *Nabokovia* Hemming, 1960 (= *Pseudothecla* Nabokov, 1945; praeocc.); *Parachilades* Nabokov, 1945

these sections, including the Polyommatus section, as tribes (Higgins, 1975; Zhdanko, 1983). Thus the *Polyommatus* section *sensu* Eliot, 1973 is equivalent to Polyommatini *sensu* Higgins, 1975.

The *Polyommatus* section is the most species-rich group within the blue butterflies, including about 460 species. It is generally cosmopolitan, but with most genera and species restricted to the Palearctic, Neotropical and Nearctic regions. Of a total of ca. 340–350 Palearctic species, ca. 130 belong to the monophyletic *Agrodiaetus*. About 20 species occur in North America (Opler and Warren, 2004) and at least 91 in the Neotropics (Lamas, 2004). Explosive chromosome evolution has evolved independently in at least three separate lineages, *Agrodaietus*, *Lysandra* and *Plebicula* (Kandul et al., 2004). Some lineages (e.g. *Polyommatus* s.s. and *Agrodiaetus*) have extremely high rates of diversification, resulting in numerous species in these lineages despite their young age (Kandul et al., 2004, 2007). In fact, *Agrodiaetus* displays one of the highest known diversification rates in the animal kingdom (Coyne and Orr, 2004). Homoploid hybrid speciation (considered to be rare in animals) has been hypothesized in the genus *Plebejus* (Gompert et al., 2006). The group displays an interesting pattern of wing colour evolution, including multiple independent cases of discoloration, a change in colour from blue to brown (Bálint and Johnson, 1997) and rapid colour changes that may reflect reinforcement (Lukhtanov et al., 2005) or ecological adaptation (Biro et al., 2003). Studies of the biology of these butterflies have focused on evolutionary processes (Krauss et al., 2004; Lukhtanov et al., 2005; Gompert et al., 2006; Kuhne and Schmitt, 2010; Lukhtanov, 2010), ecology (Vandewoestijne et al., 2008; Rusterholz and Erhardt, 2000), biogeography (Mensi et al., 1988; Schmitt et al., 2003; Schmitt, 2007; Vila et al., 2011), conservation (Brereton et al., 2008; Vila et al., 2010), cytogenetics (White, 1973; Lukhtanov and Dantchenko, 2002; Kandul et al., 2007; Vershinina and Lukhtanov, 2010), ecological physiology (Goverde et al., 2008), physiology and genetics of colour vision (Sison-Mangus et al., 2008), climate change (Carroll et al., 2009) and symbiosis (Pierce et al., 2002; Trager and Daniels, 2009).

A robust phylogenetic framework is fundamental for the advancement of these fields of research. Several modifications have been suggested to the tentative classification proposed by Eliot in 1973 (Bálint and Johnson, 1997; Zhdanko, 2004; Stekolnikov, 2010), but no comprehensive revision has been published so far.

The systematics of the section are especially problematic at the genus level. As many as 81 genera have been described within the section, but their morphological delineations are generally unclear and a wide array of taxonomic combinations are currently in use. Two extreme approaches exist: lumpers and splitters. The lumpers include the maximum number of species in one or a few genera. Examples include the monographs by Scott (1986) and by Gorbunov (2001), where nearly all the Holarctic species of the *Polyommatus* section are lumped into a single large genus (*Plebejus*). Splitters recognize numerous genera, with a genus described for every small species group. This approach has been a common practice for the *Polyommatus* section since the work of Forster (1938). The main consequence of the taxonomy of both lumpers and splitters is the same in one respect: they generate unstructured and uninformative classifications that do not reflect evolutionary relationships between the members of the section.

For example, some researchers divided the Holarctic species into three genera: *Chilades*, *Plebejus* and *Polyommatus* (Zhdanko, 1983; Hesselbarth et al., 1995), whereas others opted for four: *Chilades*, *Plebejus*, *Aricia* and *Polyommatus* (Kudrna, 2002). This created confusion as taxa of the *Aricia*, *Eumedonia*, *Albulina*, *Agriades* and *Vacciniina* species groups are sometimes included within the genus *Plebejus* (Hesselbarth et al., 1995) and sometimes within the genus *Polyommatus* (Zhdanko, 1983).

In the past 10 years, several molecular phylogenies have been published that focused on particular genera within the *Polyommatus* section (e.g. *Agrodiaetus*—Wiemers, 2003; Kandul et al., 2004, 2007; Vila et al., 2010; *Polyommatus*—Wiemers et al., 2010), or on more general issues such as biogeography and evolution (Schmitt et al., 2003; Krauss et al., 2004; Kuhne and Schmitt, 2010; Vila et al., 2011) and DNA barcoding (Wiemers and Fiedler, 2007; Lukhtanov et al., 2009). These studies were based on the analysis of limited numbers of molecular markers and most did not contain a representative collection of all the taxa of the *Polyommatus* section. Nevertheless, together these studies showed that most genera are young and closely related, explaining the controversial systematics of the group.

A recent seven-marker phylogeny was the first detailed hypothesis published for relationships in the *Polyommatus* section (Vila et al., 2011) with special attention to New World taxa. This study revealed that all the Neotropical genera—*Pseudolucia*, *Nabokovia*, *Eldoradina*, *Itylos*, *Paralycaeides*, *Hemiargus*, *Echinargus*, *Cyclargus* and *Pseudochrysops*—together formed a well supported monophyletic clade that is sister to the Old World and Nearctic taxa. The analyses showed that all Neotropical taxa belong to the *Polyommatus* section, thus the hypothesis that the Neotropical group is polyphyletic and that several taxa belong to other sections (Bálint and Johnson, 1994, 1995, 1997) was not supported. Vila et al. (2011) also determined that the *Everes* section is sister to the *Polyommatus* section. However, this study did not include a complete sampling for the Old World taxa.

We address here the analysis of phylogenetic relationships among worldwide taxa of the *Polyommatus* section. We use a combination of three mitochondrial genes and six nuclear markers to infer phylogenetic relationships between representatives of nearly all genera, subgenera and distinct species groups described within the section. We discuss principles of taxonomic classification above the species level (subgenus, genus, section and subtribe) and propose explicit criteria for defining genera in this group. We review the importance of molecular versus morphological data in evaluating our systematic hypothesis, and propose that the recognition of "cryptic genera" may be a useful concept in taxonomy. Finally, we rearrange the classification of the *Polyommatus* section and propose a new list of genera.

Materials and methods

Taxon sampling

We used 104 representatives of the *Polyommatus* section, including at least one representative of each described genus/subgenus for all but two genera that we were unable to collect (*Xinjiangia* Huang & Murayama, 1988 and *Grumiana* Zhdanko, 2004). Four representatives for the *Everes* and one for the *Leptotes* sections were used as outgroups. All specimens used in this study are listed in Table 2. The samples (bodies in ethanol and wings in glassine envelopes) are stored in the DNA and Tissues Collection of the Museum of Comparative Zoology (Harvard University, Cambridge, MA, USA).

DNA extraction and sequencing

Genomic DNA was extracted from a leg or from a piece of the abdomen of each specimen using the DNeasy^{TM} Tissue Kit (Qiagen Inc., Valencia, CA, USA) and following the manufacturer's protocols. Fragments from three mitochondrial genes—*cytochrome oxidase I* (*COI*) + *leu-tRNA* + *cytochrome oxidase II* (*COII*); and from six nuclear markers—*elongation factor-1 alpha* (*EF-1α*), *28S ribosome unit* (*28S*), *histone H3* (*H3*), *wingless* (*Wg*), *carbamoyl-phosphate synthetase 2/aspartate transcarbamylase/dihydroorotase* (*CAD*) and *internal transcribed spacer 2* (*ITS2*) were amplified by polymerase chain reaction and sequenced as described in Vila et al. (2011). The primers employed are shown in Table S1 (Appendix S1). The sequences obtained were submitted to GenBank under accession numbers JX093196–JX093497 (Table S2, Appendix S1).

Alignment

A molecular matrix was generated for each independent marker. All sequences were edited and aligned, together with those obtained in Vila et al. (2011), using Geneious 4.8.3 (Biomatters Ltd., 2009). *ITS2* sequences were aligned according to secondary structure using the *ITS2* Database Server (Koetschan et al., 2010), as described in Schultz and Wolf (2009). The HMM-Annotator tool (Keller et al., 2009) was used to delimit and crop the *ITS2* margins (*E*-value < 0.001, metazoan HMMs), preserving the proximal stems (25 nucleotides of 5.8S and 28S rDNA). The secondary structure of *ITS2* was predicted by custom homology modelling using the template structure of *Neolysandra coelestina* (MW99013) inferred by Wiemers et al. (2010), and at least 75% helix transfer was used (ITS2PAM50 matrix; gap costs: gap open 15, gap extension 2). For the outgroup taxa in *Everes* and *Leptotes* sections, the more closely related taxa *Chilades trochylus* MW99425 and *Tarucus theophrastus* MW02025 were used, respectively, as references for secondary structure prediction. For the few cases with incomplete proximal stem (3′ end), the short missing sequence was completed using the equivalent fragment from the template. These additions were necessary to obtain a correct alignment, and were removed for the posterior phylogenetic analysis. Sequences and secondary structures were aligned synchronously with 4SALE 1.5 (Seibel et al., 2006, 2008) using an *ITS2*-specific 12 × 12 scoring matrix.

Regions of the matrix lacking more than 50% of data, as well as ambiguously aligned regions, were removed using Gblocks ver. 0.96 under a relaxed criterion with the following parameters: −b2 = (50% + 1 of the sequences) −b3 = 3 −b4 = 5 −b5 = all (Castresana, 2000; Talavera and Castresana, 2007). This step was not applied to the *ITS2* alignment. The final combined alignment consisted of 6666 bp: 2172 bp of *COI* + *leu-tRNA* + *COII*, 1171 bp of *EF-1α*, 745 bp of *CAD*, 811 bp of *28S*, 370 bp of *Wg*, 1069 bp of *ITS2*, and 328 bp of *H3* (see Data S1).

Phylogenetic inference and dating

Maximum parsimony (MP), maximum likelihood (ML) and Bayesian inference (BI) were employed to estimate evolutionary relationships within Polyommatina. For MP analysis, the nine markers were concatenated in a single matrix and used as an input for the software PAUP ver. 4.0b10 (Swofford, 2000). Heuristic searches were performed with TBR branch swapping and 10 000 random taxon addition replicates, saving no more than 10 equally parsimonious trees per replicate. To estimate branch support on the recovered topology, nonparametric bootstrap values (Felsenstein, 1985) were assessed with PAUP ver. 4.0b10. One hundred bootstrap pseudoreplicates were obtained under a heuristic search with TBR branch swapping with 1000 random taxon addition replicates, saving no more than 10 equally parsimonious trees per replicate. Model-based approaches were conducted with BEAST ver. 1.6.0 (Drummond and Rambaut, 2007) for BI, and GARLI-PART ver. 0.97 (Zwickl, 2006) for ML. The data were partitioned by six markers, considering *COI* + *leu-tRNA* + *COII* a single evolutionary unit in the mitochondrial genome. jModeltest ver. 0.118 (Posada, 2008) was executed to select the best-fitting models for DNA substitution for each marker data set according to the Akaike information criterion (AIC). As a result, the HKY model was used for *H3*, the TN model for *CAD*, and a GTR model for the rest of the markers, in all cases with a gamma distribution (+G) and a proportion of invariants (+I) to account for heterogeneity in evolutionary rates among sites. The gamma distribution was estimated automatically from the data using six rate categories. Branch support was assessed by 100 bootstrap replicates for ML, and the

Table 2
Samples used in this study: taxon name, specimen label, sample accession number at MCZ and sample collection locality used in the analysis

Subtribe	Genus	Species & ssp.	Sample code	Locality
Polyommatina	*Agriades*	*glandon*	VL-05-Z994	Russia, Altai, Sailugem Range
Polyommatina	*Agriades*	*optilete optilete*	VL-01-B424	Russia, St. Petersburg, Tamengont
Polyommatina	*Agriades*	*optilete yukona*	JB-05-I879	Canada, Yukon, Dempster Hwy km 359
Polyommatina	*Agriades*	*orbitulus*	AD-03-B064	Russia, Altai, Aktash
Polyommatina	*Agriades*	*pheretiades*	NK-00-P690	Kazakhstan, Dzhambul reg., Kirgizski range
Polyommatina	*Agriades*	*podarce*	AS-92-Z130	USA, California, Leek Spring
Polyommatina	*Agriades*	*pyrenaicus dardanus*	AD-00-P259	Armenia, Gnishyk, Aiodzor Mts.
Polyommatina	*Alpherakya*	*sarta*	VL-02-X098	China, Xinjiang, Kuqa
Polyommatina	*Aricia*	*agestis*	NK-00-P712	Kazakhstan, Kayandy
Polyommatina	*Aricia*	*artaxerxes*	AD-02-W127	Russia, Primorski Krai Khanka Lake
Polyommatina	*Aricia*	*chinensis*	VL-05-Z997	Russia, Buryatia, Sosnovka, 900 m
Polyommatina	*Aricia*	*crassipuncta*	AD-00-P528	Armenia, Transcaucasus, Alibek Mt.
Polyommatina	*Aricia*	*nicias*	AD-03-B041	Russia, Altai, Aktash env.
Polyommatina	*Aricia*	*vandarbani*	VL-03-F745	Azerbaijan, Lerik, Talysh, 900–1000 m
Polyommatina	*Chilades*	*lajus*	DL-99-T242	Thailand, Prachuap Khiri Khan Province, Ampuh Thap Sakae
Polyommatina	*Cyaniris*	*semiargus belis*	AD-00-P369	Armenia, Zangezur mts., Akhtchi
Polyommatina	*Cyaniris*	*semiargus semiargus*	AD-00-P206	Russia, Low Volga, Volgograd reg., Kamytshinky
Polyommatina	*Cyclargus*	*ammon*	JE-01-C283	USA, Florida, Big Pine Key
Polyommatina	*Echinargus*	*isola*	AS-92-Z185	USA, California, Alpine, Carson River
Polyommatina	*Eldoradina*	*cyanea*	RV-05-M735	Peru, Lima, Oyón
Polyommatina	*Eumedonia*	*eumedon*	AD-03-B062	Russia, Altai, Aktash
Polyommatina	*Eumedonia*	*persephatta minshelkensis*	NK-00-P743	Kazakhstan, Shymkent reg., Karatau Mts.
Polyommatina	*Freyeria*	*putli*	RE-02-A007	Australia, Queensland, Trinity Beach
Polyommatina	*Freyeria*	*trochylus*	VL-01-L462	Turkey, Artvin, Kiliçkaya
Polyommatina	*Glabroculus*	*cyane*	VL-02-X159	Kazakhstan, Karaganda region, Aktchatau
Polyommatina	*Glabroculus*	*elvira*	NK-00-P793	Kazakhstan, Baltakul vlg.
Polyommatina	*Hemiargus*	*hanno bogotanus*	SR-03-K069	Colombia, Caldas, Chinchina
Polyommatina	*Hemiargus*	*hanno ceraunus*	MH-01-I001	Puerto Rico, Culebra Island, Flamenco Beach
Polyommatina	*Hemiargus*	*hanno gyas*	AS-92-Z255	USA, California, Los Angeles, Pyramid Lake
Polyommatina	*Hemiargus*	*hanno gyas*	DL-02-P801	USA, Arizona, Chiricahua Mts.
Polyommatina	*Hemiargus*	*huntingtoni*	RE-01-H234	Costa Rica, P.N. Santa Rosa, Guanacaste
Polyommatina	*Hemiargus*	*martha*	RV-04-I212	Peru, Huánuco
Polyommatina	*Hemiargus*	*ramon*	MFB-00-N223	Chile, Arica, Molino
Polyommatina	*Icaricia*	*acmon*	AS-92-Z184	USA, California, Alpine, Carson River
Polyommatina	*Icaricia*	*icarioides*	AS-92-Z065	USA, California, Nevada, Donner Pass
Polyommatina	*Icaricia*	*saepiolus*	AS-92-Z069	USA, California, Nevada, Donner Pass
Polyommatina	*Icaricia*	*shasta*	AS-92-Z465	USA, California, Nevada, Castle Peak
Polyommatina	*Itylos*	*huascarana*	RV-04-I403	Peru, Ancash, Pitec
Polyommatina	*Itylos*	*koa*	RV-03-V327	Peru, Junín, Huasahuasi
Polyommatina	*Itylos*	*mashenka*	MFB-00-N166	Peru, Junín
Polyommatina	*Itylos*	*sigal*	MFB-00-N220	Chile, Socoroma
Polyommatina	*Itylos*	*tintarrona*	RV-03-V182	Peru, Arequipa, Cañón del Colca
Polyommatina	*Itylos*	*titicaca*	MFB-00-N206	Chile, P.N. Lanca, Las Cuevas
Polyommatina	Kindermannia	*morgiana*	VL-02-X393	Iran, Kerman, Kuh-e-Lalizar Mts.
Polyommatina	*Kretania*	*alcedo*	VL-01-L319	Turkey, Erzurum Prov., Köprüköy
Polyommatina	*Kretania*	*eurypilus*	VL-01-L152	Turkey, Gümüshane Prov., 35 km SW Gümüshane, Dilekyolu
Polyommatina	*Kretania*	*eurypilus zamotajlovi*	SH-02-H006	Russia, Krasnodar Region, Abrau
Polyommatina	*Kretania*	*pylaon*	AD-00-P066	Russia, Volgograd, Kamyshinsky
Polyommatina	*Kretania*	*zephyrinus*	AD-00-P121	Armenia, Transcaucasus, Sevan, Shorzha
Polyommatina	*Luthrodes*	*cleotas*	CJM-07-J018	PNG, New Ireland Prov, Simberi Is.
Polyommatina	*Luthrodes*	*galba*	HU-08-D004	Cyprus, Ayios Nikolaos
Polyommatina	*Luthrodes*	*pandava*	MWT-93-A009	Malaysia, Kepong
Polyommatina	*Lysandra*	*bellargus*	AD-00-P129	Armenia, Transcaucasus, Amberd Valley, Aragatz Mt.
Polyommatina	*Lysandra*	*coridon borussia*	AD-00-P192	Russia, Tula region, Tatinki, 120 m
Polyommatina	*Lysandra*	*punctifera*	NK-02-A027	Morocco, High Atlas, Col-Tagh pass
Polyommatina	*Maurus*	*vogelii*	RVcoll09-X164	Morocco, Khenifra, S. Timahdite, Col du Zad
Polyommatina	*Nabokovia*	*cuzquenha*	RV-03-V234	Peru, Cuzco, Pisac
Polyommatina	*Nabokovia*	*faga*	MFB-00-N217	Chile, Socoroma
Polyommatina	*Neolysandra*	*coelestina alticola*	AD-00-P092	Armenia, Gegadyr, Gegamsky Mts.
Polyommatina	*Neolysandra*	*diana*	AD-00-P081	Armenia, Gegadyr, Gegamsky Mts., 1800m

Table 2
Samples used in this study: taxon name, specimen label, sample accession number at MCZ and sample collection locality used in the analysis

Subtribe	Genus	Species & ssp.	Sample code	Locality
Polyommatina	*Pamiria*	*chrysopis*	VL-05-Z998	Tajikistan, East Pamir, Sarykolski Range, Dunkeldyk Lake
Polyommatina	*Paralycaeides*	*inconspicua*	RV-03-V188	Peru, Arequipa, Cañón del Colca
Polyommatina	*Paralycaeides*	*vapa*	RV-03-V198	Peru, Puno, Chucuito
Polyommatina	*Patricius*	*lucifer*	VL-05-Z995	Russia, Altai, Chikhacheva Range, Sailugem Mt; 2300–2400 m
Polyommatina	*Plebejidea*	*loewii*	AD-00-P266	Armenia, Gnishyk, Aiodzor Mts.
Polyommatina	*Plebejus*	*anna*	AS-92-Z072	USA, California, Nevada, Donner Pass
Polyommatina	*Plebejus*	*argus*	NK-00-P135	Ukraine, Krim, Ai-Petri Mt.
Polyommatina	*Plebejus*	*argyrognomon*	AD-00-P560	Russia, Tula, Tatinki
Polyommatina	*Plebejus*	*idas armoricanella*	NK-00-P165	Russia, St. Petersburg, Luga
Polyommatina	*Plebejus*	*idas ferniensis*	NGK-02-C411	Canada, British Columbia, Castlegar
Polyommatina	*Plebejus*	*melissa*	AS-92-Z005	USA, California, Nevada, Verdi
Polyommatina	*Plebulina*	*emigdionis*	CCN-05-I856	USA, California, Kern, W. Onyx
Polyommatina	*Polyommatus*	*amandus*	NK-00-P596	Kazakhstan, Altai, Oktyabrsk
Polyommatina	*Polyommatus*	*amandus*	AD-00-P053	Russia, Volgograd region, Kamyshinsky
Polyommatina	*Polyommatus*	*amandus*	MAT-99-Q840	Spain, Pyrenees, Urús
Polyommatina	*Polyommatus*	*amandus amurensis*	AD-02-W109	Russia, Primorski Krai, S. Ussuri, Khanka Lake, Poganichnoye
Polyommatina	*Polyommatus*	*cornelia*	VL-01-L135	Turkey, Gümüshane Prov., 35 km SW Gümüshane, Dilekyolu
Polyommatina	*Polyommatus*	*damocles krymaeus*	NK-00-P103	Ukraine, Crimea, Kurortnoe
Polyommatina	*Polyommatus*	*damon damon*	MAT-99-Q841	Spain, Pyrenees, Urús
Polyommatina	*Polyommatus*	*daphnis*	NK-00-P108	Ukraine, Crimea, Kurortnoe
Polyommatina	*Polyommatus*	*dorylas armena*	AD-00-P312	Armenia, Gnishyk, Aiodzor Mts.
Polyommatina	*Polyommatus*	*erotides*	AD-03-B040	Kazakhstan, Tarbagatai Mts., Petrovskoe env.
Polyommatina	*Polyommatus*	*erschoffii*	AD-02-L274	Tajikistan, East Pamir, Sarykolski Range, Dunkeldyk Lake
Polyommatina	*Polyommatus*	*escheri*	MAT-99-Q838	Spain, Pyrenees, Urús
Polyommatina	*Polyommatus*	*glaucias*	AD-02-M278	Iran, Gorgan Prov., Shahkuh
Polyommatina	*Polyommatus*	*hunza*	VL-05-Z996	Tajikistan, East Pamir, Sarykolski Range, Dunkeldyk Lake
Polyommatina	*Polyommatus*	*icarus*	NK-00-P562	Kazakhstan, Altai, Oktyabrsk
Polyommatina	*Polyommatus*	*marcida*	AD-02-W258	Iran, Mazandaran, Geduk Pass and Veresk
Polyommatina	*Polyommatus*	*myrrha cinyraea*	AD-00-P389	Armenia, Zangezur Mts., Akhtchi
Polyommatina	*Polyommatus*	*nivescens*	MAT-99-Q904	Spain, Lleida, Rúbies
Polyommatina	*Polyommatus*	*ripartii budashkini*	NK-00-P859	Ukraine, Crimea, Karabi yaila
Polyommatina	*Polyommatus*	*stempfferi*	VL-02-X324	Iran, Esfahan, Khansar
Polyommatina	*Polyommatus*	*surakovi surakovi*	AD-00-P006	Armenia, Aiodzor mts., Gnishyk
Polyommatina	*Polyommatus*	*thersites*	MAT-99-Q947	France, Languedoc region, Mende
Polyommatina	*Polyommatus*	*thersites*	AD-00-P019	Armenia, Aiodzor Mts., Gnishyk, 1800 m
Polyommatina	*Polyommatus*	*venus*	NK-00-P810	Kazakhstan, Karzhantau vlg.
Polyommatina	*Pseudochrysops*	*bornoi*	MAC-04-Z114	Dominican Republic, Punta Cana
Polyommatina	*Pseudolucia*	*asafi*	RV-03-V020	Chile, Céspedes, Illapel
Polyommatina	*Pseudolucia*	*charlotte*	BD-02-B813	Chile, Temuco
Polyommatina	*Pseudolucia*	*chilensis*	MFB-00-N227	Chile, Farellones
Polyommatina	*Pseudolucia*	*sibylla*	RV-03-V112	Chile, Coquimbo, Río La Laguna
Polyommatina	*Pseudolucia*	*vera*	BD-02-B812	Chile, Temuco, Volcán Villarica
Polyommatina	*Rimisia*	*miris*	NK-00-P575	Kazakhstan, Altai, Oktyabrsk
Polyommatina	*Rueckbeilia*	*fergana*	NK-00-P777	Kazakhstan, Shymkent Reg., Karatau Mts., Turpan Pass
Cupidina	*Cupido*	*comyntas*	AS-92-Z312	USA, California, Davis
Cupidina	*Cupido*	*minimus*	AD-00-P540	Russia, Tula, Tatinki
Cupidina	*Talicada*	*nyseus*	JXM-99-T709	India, Karala, Trivandrum
Cupidina	*Tongeia*	*fischeri*	NK-00-P594	Kazakhstan, Altai, Oktyabrsk
Leptotina	*Leptotes*	*trigemmatus*	RV-03-V095	Chile, Coquimbo, Alcohuas

software SumTrees in the DendroPy phylogenetic Python library (Sukumaran and Holder, 2010) was used to generate a majority-rule bootstrap consensus tree.

BI with BEAST ver. 1.6.0 was used to estimate divergence times. Normally distributed tmrca priors including maximum and minimum ages from Vila et al. (2011) within the 95% HPD distribution were established on four well supported nodes, shown in Fig. 1. The resulting 95% HPD ranged from 1.5 to 3.3 Myr for node 1; from 5.5 to 13.1 Myr for node 2; from 8.4 to 16.8 Myr for node 3; and from 2.5 to 11.3 Myr for node 4. The uncorrelated relaxed clock (Drummond et al., 2006) and a constant population size under a coalescent model were established as priors. The rest of the settings and priors were set by default. Two

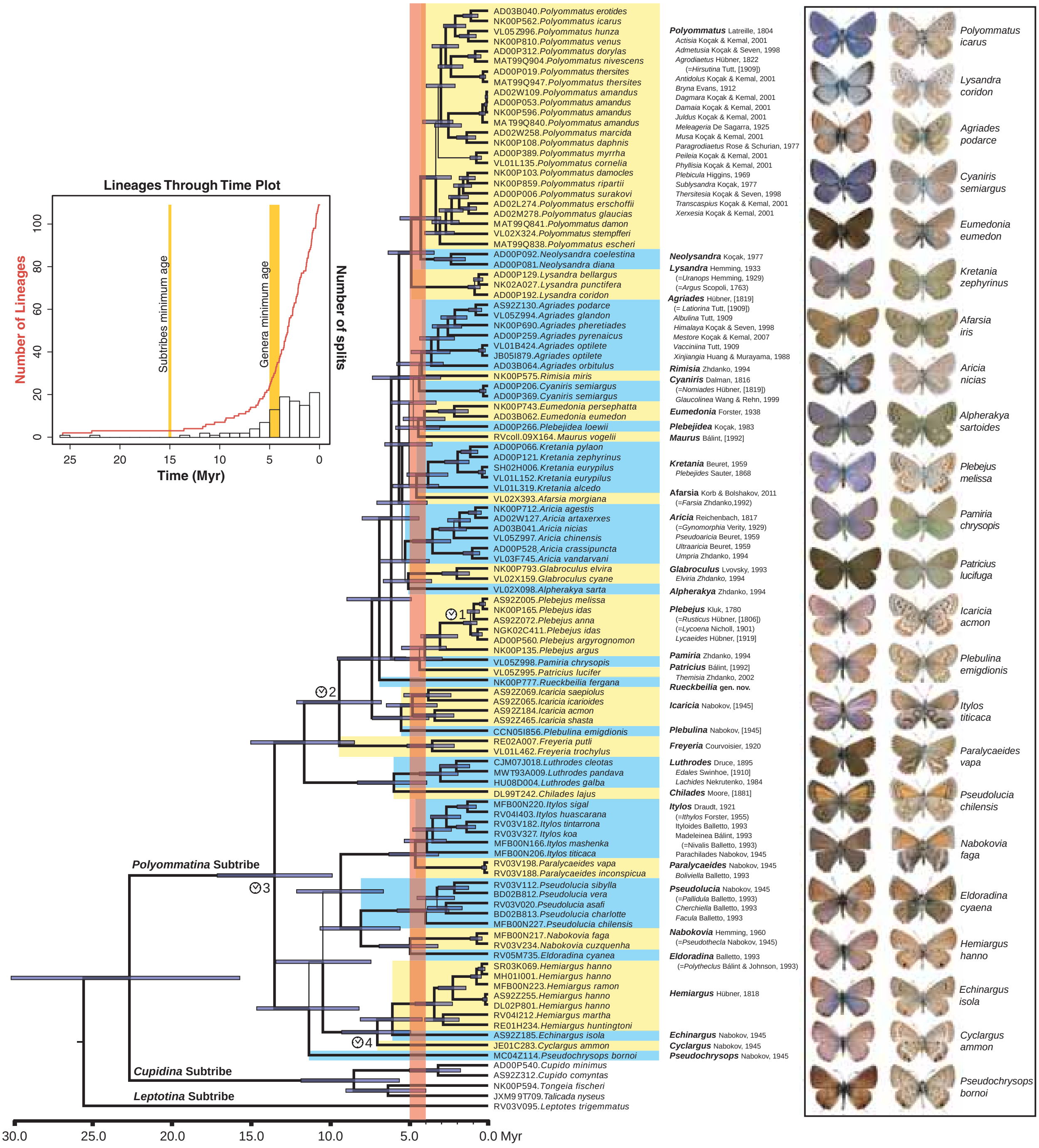

Lineages Through Time Plot
Number of Lineages
Number of splits
Subtribes minimum age
Genera minimum age
Time (Myr)
0
20
40
60
80
100
25
20
15
10
5
0
AD03B040.Polyommatus erotides
NK00P562.Polyommatus icarus
VL05Z996.Polyommatus hunza
NK00P810.Polyommatus venus
AD00P312.Polyommatus dorylas
MAT99Q904.Polyommatus nivescens
AD00P019.Polyommatus thersites
MAT99Q947.Polyommatus thersites
AD02W109.Polyommatus amandus
AD00P053.Polyommatus amandus
NK00P596.Polyommatus amandus
MA T99Q840.Polyommatus amandus
AD02W258.Polyommatus marcida
NK00P108.Polyommatus daphnis
AD00P389.Polyommatus myrrha
VL01L135.Polyommatus cornelia
NK00P103.Polyommatus damocles
NK00P859.Polyommatus ripartii
AD00P006.Polyommatus surakovi
AD02L274.Polyommatus erschoffii
AD02M278.Polyommatus glaucias
MAT99Q841.Polyommatus damon
VL02X324.Polyommatus stempfferi
MAT99Q838.Polyommatus escheri
AD00P092.Neolysandra coelestina
AD00P081.Neolysandra diana
AD00P129.Lysandra bellargus
NK02A027.Lysandra punctifera
AD00P192.Lysandra coridon
AS92Z130.Agriades podarce
VL05Z994.Agriades glandon
NK00P690.Agriades pheretiades
AD00P259.Agriades pyrenaicus
VL01B424.Agriades optilete
JB05I879.Agriades optilete
AD03B064.Agriades orbitulus
NK00P575.Rimisia miris
AD00P206.Cyaniris semiargus
AD00P369.Cyaniris semiargus
NK00P743.Eumedonia persephatta
AD03B062.Eumedonia eumedon
AD00P266.Plebejidea loewii
RVcoll.09X164.Maurus vogelii
AD00P066.Kretania pylaon
AD00P121.Kretania zephyrinus
SH02H006.Kretania eurypilus
VL01L152.Kretania eurypilus
VL01L319.Kretania alcedo
VL02X393.Afarsia morgiana
NK00P712.Aricia agestis
AD02W127.Aricia artaxerxes
AD03B041.Aricia nicias
VL05Z997.Aricia chinensis
AD00P528.Aricia crassipuncta
VL03F745.Aricia vandarvani
NK00P793.Glabroculus elvira
VL02X159.Glabroculus cyane
VL02X098.Alpherakya sarta
AS92Z005.Plebejus melissa
NK00P165.Plebejus idas
AS92Z072.Plebejus anna
NGK02C411.Plebejus idas
AD00P560.Plebejus argyrognomon
NK00P135.Plebejus argus
VL05Z998.Pamiria chrysopis
VL05Z995.Patricius lucifer
NK00P777.Rueckbeilia fergana
AS92Z069.Icaricia saepiolus
AS92Z065.Icaricia icarioides
AS92Z184.Icaricia acmon
AS92Z465.Icaricia shasta
CCN05I856.Plebulina emigdionis
RE02A007.Freyeria putli
VL01L462.Freyeria trochylus
CJM07J018.Luthrodes cleotas
MWT93A009.Luthrodes pandava
HU08D004.Luthrodes galba
DL99T242.Chilades lajus
MFB00N220.Itylos sigal
RV04I403.Itylos huascarana
RV03V182.Itylos tintarrona
RV03V327.Itylos koa
MFB00N166.Itylos mashenka
MFB00N206.Itylos titicaca
RV03V198.Paralycaeides vapa
RV03V188.Paralycaeides inconspicua
RV03V112.Pseudolucia sibylla
BD02B812.Pseudolucia vera
RV03V020.Pseudolucia asafi
BD02B813.Pseudolucia charlotte
MFB00N227.Pseudolucia chilensis
MFB00N217.Nabokovia faga
RV03V234.Nabokovia cuzquenha
RV05M735.Eldoradina cyanea
SR03K069.Hemiargus hanno
MH01I001.Hemiargus hanno
MFB00N223.Hemiargus ramon
AS92Z255.Hemiargus hanno
DL02P801.Hemiargus hanno
RV04I212.Hemiargus martha
RE01H234.Hemiargus huntingtoni
AS92Z185.Echinargus isola
JE01C283.Cyclargus ammon
MC04Z114.Pseudochrysops bornoi
AD00P540.Cupido minimus
AS92Z312.Cupido comyntas
NK00P594.Tongeia fischeri
JXM9 9T709.Talicada nyseus
RV03V095.Leptotes trigemmatus
1
2
3
4
Polyommatina Subtribe
Cupidina Subtribe
Leptotina Subtribe
30.0
25.0
20.0
15.0
10.0
5.0
0.0 Myr
Polyommatus Latreille, 1804
Actisia Koçak & Kemal, 2001
Admetusia Koçak & Seven, 1998
Agrodiaetus Hübner, 1822
(=Hirsutina Tutt, [1909])
Antidolus Koçak & Kemal, 2001
Bryna Evans, 1912
Dagmara Koçak & Kemal, 2001
Damaia Koçak & Kemal, 2001
Juldus Koçak & Kemal, 2001
Meleageria De Sagarra, 1925
Musa Koçak & Kemal, 2001
Paragrodiaetus Rose & Schurian, 1977
Peileia Koçak & Kemal, 2001
Phyllisia Koçak & Kemal, 2001
Plebicula Higgins, 1969
Sublysandra Koçak, 1977
Thersitesia Koçak & Seven, 1998
Transcaspius Koçak & Kemal, 2001
Xerxesia Koçak & Kemal, 2001
Neolysandra Koçak, 1977
Lysandra Hemming, 1933
(=Uranops Hemming, 1929)
(=Argus Scopoli, 1763)
Agriades Hübner, [1819]
(= Latiorina Tutt, [1909])
Albulina Tutt, 1909
Himalaya Koçak & Seven, 1998
Mestore Koçak & Kemal, 2007
Vacciniina Tutt, 1909
Xinjiangia Huang & Murayama, 1988
Rimisia Zhdanko, 1994
Cyaniris Dalman, 1816
(=Nomiades Hübner, [1819])
Glaucolinea Wang & Rehn, 1999
Eumedonia Forster, 1938
Plebejidea Koçak, 1983
Maurus Bálint, [1992]
Kretania Beuret, 1959
Plebejides Sauter, 1868
Afarsia Korb & Bolshakov, 2011
(=Farsia Zhdanko,1992)
Aricia Reichenbach, 1817
(=Gynomorphia Verity, 1929)
Pseudoaricia Beuret, 1959
Ultraaricia Beuret, 1959
Umpria Zhdanko, 1994
Glabroculus Lvovsky, 1993
Elviria Zhdanko, 1994
Alpherakya Zhdanko, 1994
Plebejus Kluk, 1780
(=Rusticus Hübner, [1806])
(=Lycoena Nicholl, 1901)
Lycaeides Hübner, [1919]
Pamiria Zhdanko, 1994
Patricius Bálint, [1992]
Themisia Zhdanko, 2002
Rueckbeilia gen. nov.
Icaricia Nabokov, [1945]
Plebulina Nabokov, [1945]
Freyeria Courvoisier, 1920
Luthrodes Druce, 1895
Edales Swinhoe, [1910]
Lachides Nekrutenko, 1984
Chilades Moore, [1881]
Itylos Draudt, 1921
(=Ithylos Forster, 1955)
Ityloides Balletto, 1993
Madeleinea Bálint, 1993
(=Nivalis Balletto, 1993)
Parachilades Nabokov, 1945
Paralycaeides Nabokov, 1945
Boliviella Balletto, 1993
Pseudolucia Nabokov, 1945
(=Pallidula Balletto, 1993)
Cherchiella Balletto, 1993
Facula Balletto, 1993
Nabokovia Hemming, 1960
(=Pseudothecla Nabokov, 1945)
Eldoradina Balletto, 1993
(=Polytheclus Bálint & Johnson, 1993)
Hemiargus Hübner, 1818
Echinargus Nabokov, 1945
Cyclargus Nabokov, 1945
Pseudochrysops Nabokov, 1945
Polyommatus icarus
Lysandra coridon
Agriades podarce
Cyaniris semiargus
Eumedonia eumedon
Kretania zephyrinus
Afarsia iris
Aricia nicias
Alpherakya sartoides
Plebejus melissa
Pamiria chrysopis
Patricius lucifuga
Icaricia acmon
Plebulina emigdionis
Itylos titicaca
Paralycaeides vapa
Pseudolucia chilensis
Nabokovia faga
Eldoradina cyaena
Hemiargus hanno
Echinargus isola
Cyclargus ammon
Pseudochrysops bornoi

Fig. 1. Bayesian chronogram for the newly proposed subtribe Polyommatina based on nine genes: *COI*, *leu-tRNA*, *COII*, *EF-1α*, *Wg*, *ITS2*, *CAD*, *28S* and *H3* (6666 bp). Thick lines indicate supported relationships (posterior probabilities ≥ 0.95); node bars show estimated divergence times uncertainty. Nearly all the extant genera are included in the phylogeny; representatives from the subtribes Cupidina and Leptotina were used as outgroups. Valid genus names are presented in bold. Subjective synonyms (that may yet be shown to represent valid subgenera with additional research) are shown after the valid names. Objective synonyms are indicated by "=". Normally distributed tmrca from inferred divergence times in Vila et al. (2011) were used as priors on the nodes 1–4. The phylogeny revealed unexpected relationships with respect to traditional classification. We rearranged the systematics of the group and proposed a new list of genera according to the following criteria: (i) taxa older than 5 Myr are considered genera; (ii) for taxa between 4 and 5 Myr we are conservative in the sense that we consider a clade to be a genus only if it has already been described, and do not consider it a genus if it has not; and (iii) taxa younger than 4 Myr are considered subgeneric. The 4–5-Myr time interval is highlighted in red. Applying these criteria resulted in the retention of 31 of the 81 genera formally described in the group, and necessitated the addition of one new genus. Minimum age thresholds used to define genera and subtribes are indicated in the lineage through time plot. The upper side and underside of representative adult specimens of the Polyommatina are shown on the right.

independent chains were run for 50 million generations each, sampling values every 1000 steps. A conservative burn-in of 500 000 generations was applied for each run after checking Markov chain Monte Carlo (MCMC) convergence through graphically monitoring likelihood values in Tracer ver. 1.5 (Rambaut and Drummond, 2007). Independent runs were combined in LogCombiner ver. 1.6.0 implemented in the software package BEAST and all parameters were analysed using the program Tracer to determine whether they had also reached stationarity. Tree topologies were assessed using TreeAnnotator ver. 1.6.0 in the BEAST package to generate a maximum clade credibility tree of all sampled trees with median node heights. Finally, FigTree ver. 1.2.2 (Rambaut, 2009) was used to visualize the consensus tree along with node ages, age deviations and node posterior probabilities.

Ancestral states reconstruction

Character evolution was reconstructed by estimating probabilities for ancestral character states with MESQUITE ver. 2.6 (Maddison and Maddison, 2007). Both MP and ML approaches were applied to the Bayesian tree for two discrete (absence or presence) morphological characters traditionally used to define the genus *Vacciniina*: (i) metallic marginal spots on the hind wing underside; and (ii) inner apical part of the valvae in the male genitalia with sclerotized ventral fold. A reduced phylogenetic tree excluding the basal Neotropical clade and outgroup was used.

Results and discussion

Higher-level systematics

The taxonomic system employed by Eliot (1973) grouped the genera in the rather unconventional category "section". This system is still widely used, and it coexists with several arrangements that use the more formal categories "tribe" and "subtribe". Since this study represents the first comprehensive revision of the group since Eliot, our goal is to normalize the systematics above the level of the genus. Our phylogeny (Fig. 1) shows that the *Polyommatus* section is monophyletic (see also Vila et al., 2011). We propose to use the term "Polyommatina subtribe" to replace Eliot's "*Polyommatus* section", and generally use the designation "subtribe" instead of "section" throughout. Thus Cupidina would be the sister to the Polyommatina, and Leptotina the sister to both. We estimate the ages of divergence for these subtribes to range between 22.8 and 25.7 Myr. In the lineages through time plot (Fig. 1), a relatively long period without diversification events, from 22.8 to 13.6 Ma, is observed. We have designated this period as a gap defining subtribes, and therefore consider subtribes to be those lineages older than 15 Myr. The three sections previously recognized by Eliot (1973) for the studied group fall within this definition of subtribe, as do most of the rest of sections in Polyommatini (Vila et al., 2011). In order to evaluate the four tribes within the subfamily Polyommatinae (e.g. Candalidini, Lycaenesthini, Niphandini and Polyommatini), an adequate threshold will need to be set for the tribal level using a more thorough phylogenetic analysis of the Lycaenidae that includes these taxa.

Genus concept

Since our aim is to establish a phylogenetically based classification system for the Polyommatina, criteria for delineating genera are important to establish. This is especially true given the wide array of taxonomic classifications that have been proposed for this group at the genus level, including drastic approaches that split the group into numerous nearly monotypic (consisting of a single species) genera (Forster, 1938; Zhdanko, 2004), or lumped all species into only a few genera (Zhdanko, 1983; Scott, 1986; Hesselbarth et al., 1995; Gorbunov, 2001; Kudrna, 2002).

Monophyly. One important criterion defining a genus is that it should be monophyletic. The majority of taxonomists currently believe that monophyly, in the narrow sense used by Hennig (Hennig, 1950, 1966; Envall, 2008; Hörandl and Stuessy, 2010) (= holophyly *sensu* Ashlock, 1971) is mandatory, at least for taxonomic categories above the species level (genus, family, etc.) (Schwenk, 1994; Groves, 2004). Paraphyletic taxa are incompatible with the principles of phylogenetic systematics (Schmidt-Lebuhn, 2011) and have relatively few defenders (Brummitt, 2003; Hörandl and Stuessy, 2010). Using paraphyletic groups in higher-level taxonomy poses serious problems as it can result in taxa that are neither mutually exclusive nor wholly inclusive of one another (Nelson et al., 2003). This gives rise to uncertainties and discrepancies in classifications. Thus avoiding paraphyletic groups and focusing on monophyletic entities *sensu* Hennig is the preferable option in practical terms. It is important to note, too, that the concept of monophyly applies to whole organisms. Trees inferred from single markers sometimes display paraphyletic relationships that reflect the evolutionary histories of individual genes rather than the species being studied. It is thus advisable to base taxonomic conclusions on multilocus analyses using the principle of character congruence as advocated by Kluge (1989) and Brower et al. (1996).

Still, the monophyly criterion alone is not enough to construct a taxonomic system. Nearly every phylogeny is a complicated structure consisting of numerous nested monophyletic lineages. The number of these nested clades is often much greater than the number of traditional taxonomic ranks. Therefore additional criteria need to be used to select which monophyletic lineages should be considered genera and which not, and similar criteria should be established for other ranks.

The morphological gap and the concept of cryptic genera. One criterion that can be used in defining a genus is the existence of a discontinuity in the distribution of morphological characters between one monophyletic group and another. The morphological gap (= morphological hiatus) seen between genera should be significantly larger than the gaps seen between species of the same genus. This criterion is widely used, but it is not ideal. First, it may be difficult to decide when a morphological gap is sufficient to separate genera (and it may be difficult to measure morphological gaps in the first place). Second, and most importantly, using this criterion can result in artificial taxonomic systems due to homoplasy. For example, the genus *Vacciniina* in its traditional conception includes three morphologically similar species: *V. optilete*, *V. alcedo* and *V. fergana* (Tuzov et al., 2000) (Fig. 2). However, our study demonstrates that these species represent three different evolutionary lineages that are not closely related (Fig. 1). In fact, we describe the new genus *Rueckbeilia* for the traditional species *V. fergana*, and include the species *V. alcedo* in the genus *Kretania* and the species *V. optilete* in the genus *Agriades*. Thus *Vacciniina sensu auctorum* represents three cryptic genera, i.e. three species clusters that cannot be separated from one another based on their morphological characters and, at the same time, cannot be lumped into a single genus as their combination would be polyphyletic. As a consequence, we suggest that the recognition of cryptic genera (Vilnet et al., 2007; Lucky and Sarnat, 2008) may be useful, in the same manner that the recognition of cryptic species is now widely used (Descimon and Mallet, 2009).

Cryptic genera are the consequence of unrecognized parallelisms in evolution of some morphological characters or of the long preservation of plesiomorphic states that are mistakenly considered synapomorphies; or of both processes acting simultaneously in different characters. For example, the species *V. optilete* seems to have independently evolved a wing pattern similar to those of *V. alcedo* and *V. fergana* (Fig. 2), whereas the "*Polyommatus*-like" structures of the male genitalia of these lineages (Stekolnikov, 2010) (Fig. 3) probably represent an ancestral condition that has been preserved for at least 6 million years (Fig. 4).

Age of lineage as a universal and unbiased criterion? Hennig (1966) proposed to synchronize taxonomic ranks universally according to geological ages. This would have the effect of making groups comparable and ranks definable. Since geological time is universal, the age of evolutionary lineages, generally estimated by the dating of nodes in phylogenetic trees, seems to be the only truly unbiased criterion by which taxonomic classifications above the level of biological species can be erected (Hennig, 1966). Avise and Johns (1999) devised a specific temporal-banding scheme to fit conventional Linnaean ranks. They proposed considering as genera those lineages that originated in the Pliocene (ca. 2–5 Ma); as subgenera the lineages above the level of species that

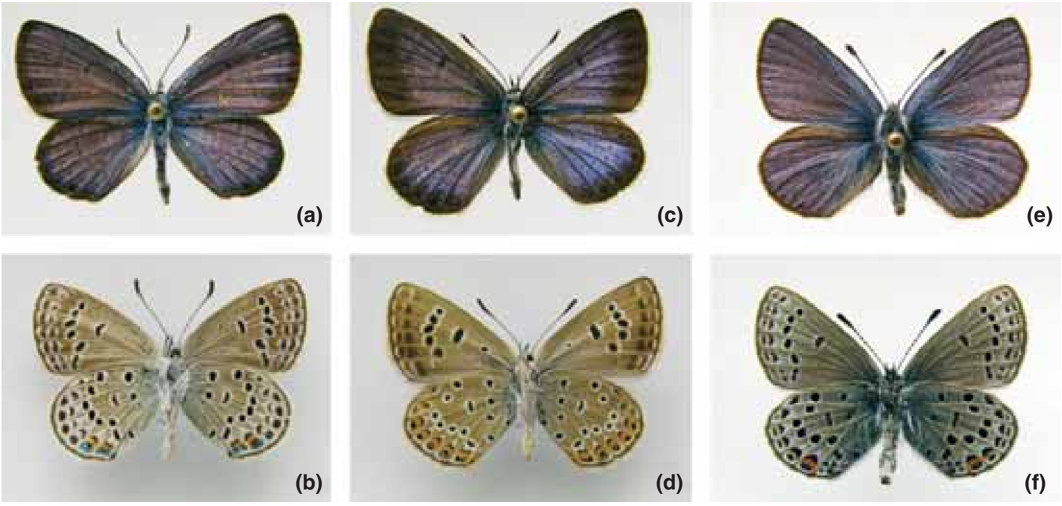


Fig. 2. Taxa representing three cryptic genera. (a,b) *Rueckbeilia fergana* (= "*Vacciniina*" *fergana*); (c,d) *Kretania alcedo* (= "Vacciniina" *alcedo*); (e,f) *Agriades optilete* (= "*Vacciniina*" *optilete*). These taxa were all considered species of the same genus (*Vacciniina*), although in fact they form three distinct genera according to the criteria described in this study. Despite their genetic differences, this artificial assemblage is strikingly convergent with respect to wing colour and pattern. They share the violet-blue colour of the wing upper side in males, and the similar wing underside with blue metallic scales that seems to have evolved independently at least twice.

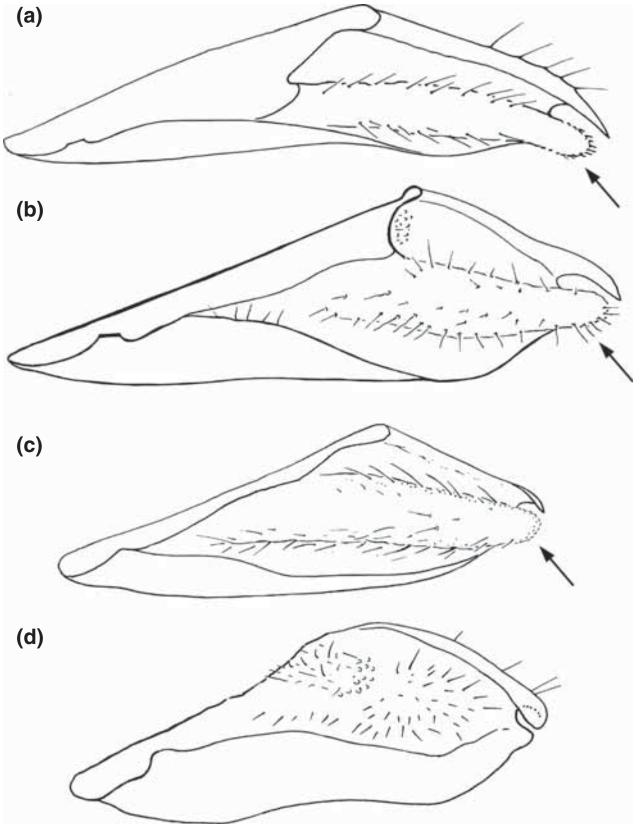


Fig. 3. Valva in the male genitalia. Inner part of valva with membranous ventral fold indicated by arrow (a–c) and without membranous ventral fold (d). (a) *Rueckbeilia fergana*; (b) *Kretania alcedo*; (c) *Agriades glandon*; (d) *Plebejus idas*. After Stekolnikov (2010) with modifications.

originated in the Pleistocene (0–2 Ma); and as tribes the lineages that originated in the Miocene (5–24 Ma).

However, this proposal has two main problems: it is not directly applicable to fossil organisms (Griffiths, 1973), and it would necessitate a major, even radical, rearrangement for current taxonomy. Acknowledging these difficulties, Avise and Mitchell (2007) launched the ‘‘timeclip proposal’’, which consists of labelling classic Linnaean taxa with timeclips that indicate their geological ages of origin. This could provide relevant additional information that could be updated easily without the need to alter taxonomy. Although the timeclip proposal is interesting, it still relies on a taxonomic system, and does not invalidate the need to establish true relationships within and between taxa and to decide how to determine taxonomic ranks.

We agree with the concept of relative ages, but we think this should be modified in at least two respects. First, the age thresholds must take into account the systematics and relative ages of different groups of organisms. Second, once the taxonomic ranks are established, diagnostic morphological characters should be explained or explored. Moreover, the rank of subtribe, which is especially useful in insect systematics, should be incorporated in the proposal of Avise and Johns (1999).

Stability and preservation of traditionally recognized taxa. The stability and preservation of traditionally recognized taxa must be taken into account in establishing classification guidelines (Godfray and Knapp, 2004). Indeed, stability is a concept that is positively valued by the International Commissions of Nomenclature, and that can, in some instances, take precedence over other principles. Applying a universal system of thresholds would result in taxonomic upheaval, mostly because at present there is deep discrepancy in the average age of the taxa accepted for different groups of organisms (Avise and Liu, 2011). In mammals, for example, many recognized genera are relatively young (3–5 Myr) (Castresana, 2001; Rowe et al., 2008; Abramson et al., 2009) with an estimated mean of 9.6 Myr (0.1–40) (Avise and Liu, 2011), whereas other groups may be relatively old, such as Decapoda, with an estimated mean of 60.2 Myr (16.8–135) (Avise and Liu, 2011) or Diptera (Drosophilidae, Chironomidae) with estimated means ranging from 30–40 Myr to more than 100 Myr (Avise and Johns, 1999; Cranston et al., 2010). Strong temporal banding heterogeneity among different organismal assemblages also occurs at higher taxonomic levels such as families and orders (Avise and Liu, 2011; Hedges and Kumar, 2009). Consequently, a universal system would require such a complete reorganization of the systematics of most groups of organisms that the overall effect would be deleterious to communication and understanding of taxonomic relationships.

Even if the most extreme cases, such as the relative ages of genera in Diptera or Decapoda, were to be modified to create a more balanced general classification, we propose that a temporal scheme should adapt to some degree to the particularities and pre-existing taxonomy of each group. Differences in the age thresholds might be necessarily pronounced in distantly related groups of taxa whose rates of diversification are likely to differ depending on intrinsic biological differences such as generation time and/or population size (e.g. Li, 1997), differences in the efficiency of DNA repair mechanisms (Britten, 1986), or differences in metabolic rate (Martin and Palumbi, 1993). The increased rate of nucleotide changes at several loci, including such usual phylogenetic markers as COI and CytB genes, can be affected in some phylogenetic lineages by positive selection due to their role in adaptation to specialized metabolic requirements (da Fonseca et al., 2008).

In the case of Polyommatina, the following thresholds provide a balanced classification that corresponds well with current evidence about relationships between groups: genera can be recognized as those lineages that originated in the late Miocene (older than 5 Myr), and subtribes those that originated in the early Miocene or

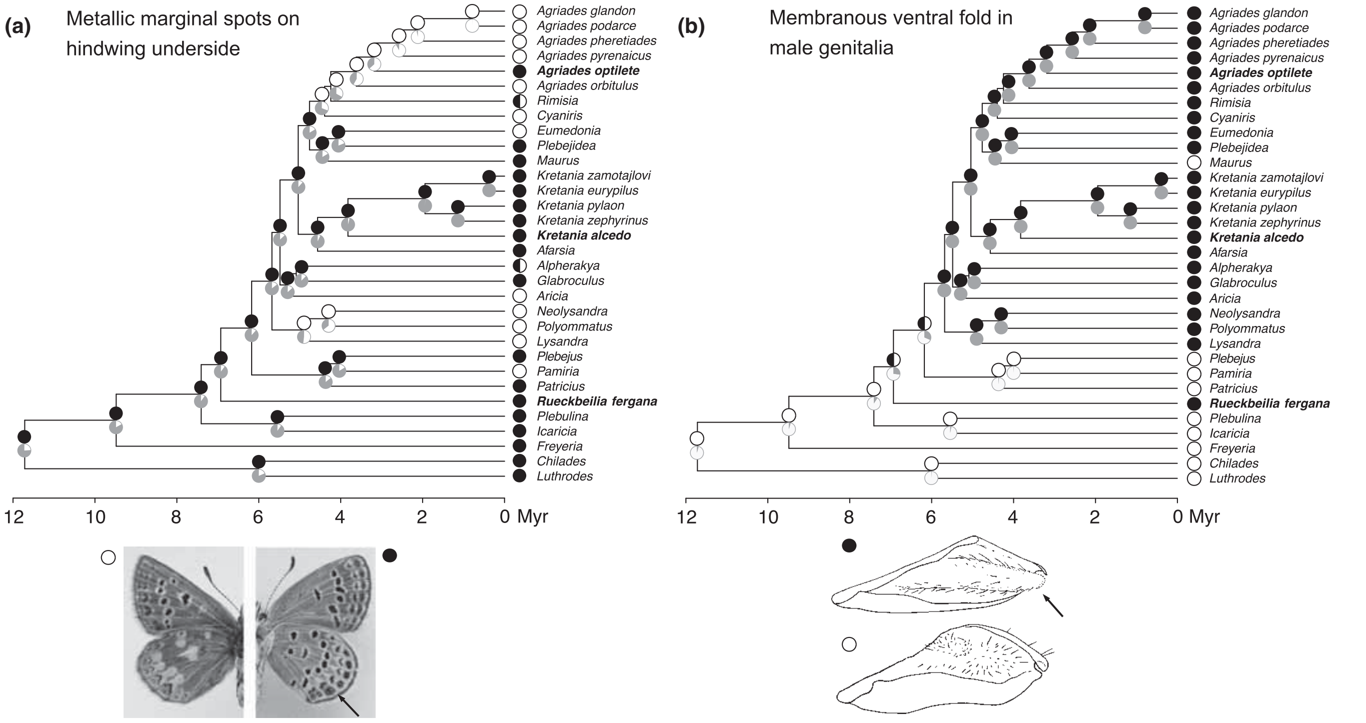


Fig. 4. Ancestral state reconstructions for two morphological characters traditionally defining the polyphyletic genus *Vacciniina* (taxa in bold). (a) Metallic marginal spots on the hind wing underside present (black circle) or absent (white circle). (b) Inner part of valvae in the male genitalia with membranous ventral fold (black circle) or without membranous ventral fold (white circle). Maximum parsimony (upper circles) and Maximum likelihood (lower grey circles) inferences are represented at nodes. Figures of genitalia are given after Stekolnikov (2010).

late Oligocene (older than 15 Myr). In the lineages through time plot, an increase in diversification can be seen starting at 5–4 Ma (Fig. 1), so we set the minimum age for genera at this point to avoid excessive splitting. This approach (plotting the number of lineages or branching events over time) is useful to illuminate diversification patterns in the group under study. Substantial changes in the rate of diversification mark key moments in the evolution of a group as a whole, and these are logical points to be used as age thresholds delimiting taxonomic ranks.

In our case, age thresholds were also selected so as to minimally affect the existing nomenclature and avoid the need for descriptions of new genera. A generic threshold of 3–4 Myr requires the creation of two new genera (the splitting of *Icaricia* and description of the new *Rueckbeilia*), while a 5–6 Myr threshold would have entailed a wide-scale synonymization (50 subjective synomyms) with excessive loss of phylogenetic information, and would have still required the description of *Rueckbeilia*. Wider thresholds would have also involved losing substantial input from the molecular data.

Accounting for uncertainty in age estimates. Additionally, any system of classification should recognize the uncertainty inherent in estimating evolutionary age given intrinsic errors associated with the methods of inference, especially when no paleontological material is available to calibrate a molecular clock. Absolute age is likely to vary depending on the analysis, and new information helping to calibrate the molecular clock, or additional methodological improvements, might affect age estimates. In contrast, relative ages among lineages are less affected by these factors because they do not depend on external information for tree calibration. The greater uncertainty in absolute age estimates compared with those based on relative ages is another reason to apply a temporal scheme specific to the group being studied, which could be adapted eventually to a different molecular substitution rate without major implications for the taxonomy of the group. A universal temporal scheme would suffer from taxonomic instability caused by uncertainty in absolute age, which is necessary when comparing taxa that are not closely related to each other. Divergent lineages sometimes display disparate molecular substitution rates, whereas closely related taxa tend to be more uniform in this regard (Martin and Palumbi, 1993; Li, 1997). The subtribe Polyommatina, a clade that evolved ca. 22.8 Ma, contains many taxa to be compared, but these are sufficiently evolutionarily and ecologically similar that they do not exhibit excessive variability in substitution rates among lineages.

In order to reduce the effect of the uncertainty in age estimates, and to avoid taxonomic instability because of small differences obtained using different phylogenetic analyses and/or novel calibration points, we propose using a time interval to set the limits of genus age, rather than a single date (e.g. 4.0–5.0 Myr for genus minimum age in our case). Thus lineages with a mean age within these intervals can be dealt with using this relatively conservative approach, as described below.

Importance of morphological diagnostic characters. Once the classification of a group is produced using the previously discussed criteria, the next critical step is to explore and explain the diagnostic morphological characters that define the proposed taxa. The exercise of integrating the molecular-based classification into a morphological framework has multiple benefits. It does not create a discontinuity with the previous morphology-based classifications; it avoids wasting the morphological data painstakingly gathered; and it allows for the reinterpretation of earlier work. It also facilitates the placement of extinct taxa and those that have not yet been sequenced, and overcomes the major drawback of a system based purely on molecular data.

Genera within the Polyommatina. In practice, we apply these criteria in the following manner.

1. We define as genus any lineage older than 5.0 Myr.
2. Between 4.0 and 5.0 Ma we are conservative, in the sense that we consider a clade to be a genus only if it has already been described, and do not consider it a genus if it has not.
3. We lump into another genus any lineage younger than 4.0 Myr.

Applying this taxonomy-friendly, flexible temporal scheme to the phylogeny and dating produced the division of the subtribe Polyommatina into 32 genera (Table 3). From this classification scheme, one new genus needs to be described and 39 names can be regarded as subjective synonyms or valid subgenera. The further designation of these 39 taxa as either subgenera or subjective synonyms requires additional data for all the species that each one represents, which is beyond of the scope of this paper.

For the 32 established genera, monophyly was statistically supported for the three phylogenetic methods used, with the sole exception of *Kretania*, where the phylogenetic position of *K. alcedo* was not resolved in the MP analysis (Table 3).

Composition and phylogenetic relationships of genera in the *Polyommatina*: before and after this study.

The nine-marker phylogeny revealed that the subtribe Polyommatina includes two major clades: the Neotropical clade (the genera *Pseudolucia*, *Nabokovia*, *Eldoradina*, *Itylos*, *Paralycaeides*, *Hemiargus*, *Echinargus*, *Cyclargus* and *Pseudochrysops*) and the non-Neotropical clade (the remaining genera).

The Neotropical clade. Relationships within the Neotropical clade have already been discussed in detail in a previous publication (Vila et al., 2011). Briefly, the Neotropical taxa are divided into four well supported clades. Two of these, probably sister clades, are formed by Andean, typically high-altitude taxa that occur south of Central Colombia. These are *Eldoradina* Balletto, 1993, *Nabokovia* Hemming, 1960 and *Pseudolucia* Nabokov, 1945 on one hand; and *Itylos* (= *Madeleinea* Bálint, 1993) and *Paralycaeides* Nabokov, 1945 on the other. The other two clades are formed by lowland taxa, including all the Caribbean representatives and species occurring north of Central Colombia, plus a few with more southern distributions. One clade is formed by *Cyclargus* Nabokov, 1945; *Echinargus* Nabokov, 1945 and *Hemiargus* Hübner, 1818; the other by *Pseudochrysops* Nabokov, 1945. The position of *Pseudochrysops* with respect to the other three clades is unresolved, probably due to its early divergence and very long branch.

The non-Neotropical clade. The non-Neotropical clade of the subtribe Polyommatina is strongly asymmetrical, with multiple nested lineages that are discussed below.

Chilades–Luthrodes clade. Within the non-Neotropical clade, *Chilades* Moore, [1881] (TS: *Papilio lajus* Stoll, [1780]) and *Luthrodes* Druce, 1895 (TS: *Polyommatus cleotas* Guérin-Méneville, [1831]) form a clade that is sister to the rest. The age of divergence between these two groups is 6.0 Myr, thus we consider them good genera despite the fact that most recent studies (Bridges, 1988) have lumped them together. Representatives of both *Chilades* and *Luthrodes* have a similar, most likely plesiomorphic, pattern on the wing underside with the presence of all the basic elements typical of the non-Neotropical Polyommatina. However, the male genitalia in *Luthrodes* are very distinct—clearly different from those in other genera—in the shape of the valvae, which are broad and trapeziform, and in the presence of a dorsal process at the distal end of the valvae that is markedly elongated and directed downwards (Bethune-Baker, 1913; Zhdanko, 1983, 2004; Stekolnikov and Kuznetzov, 2005). Within *Luthrodes*, the taxa *Edales* Swinhoe, [1910] (TS: *Lycaena pandava* Horsfield, [1829]) and *Lachides* Nekrutenko, 1987 (TS: *Lycaena galba* Lederer, 1855) are aged less than 4.0 Myr, and consequently should be considered subjective synonyms or subgenera of *Luthrodes*. However, this question may be better assessed after a study including additional species.

Bethune-Baker (1913) studied the male genitalia of *Chilades lajus* and showed that, unlike *Luthrodes*, the valvae are elongated and have a short dorsal process. In fact, the genital morphology of *Chilades* is more similar to that of *Freyeria* than to *Luthrodes*.

Table 3
Posterior probabilities and bootstrap values for monophyly in BI/ML/MP inferences, ages (mean and stdev), number of species, and larval food plant families for the 32 genera within the subtribe Polyommatina

Genus	Monophyly stability values	Genus age (Myr)	Number of species	Food plant
Polyommatus	100/100/100	4.3 [3.0–5.6]	183	Fabaceae
Neolysandra	100/100/100	4.3 [3.0–5.6]	6	Fabaceae
Lysandra	100/100/100	4.9 [3.4–6.4]	15	Fabaceae
Aricia	100/100/99	5.3 [3.7–6.9]	15	Geraniaceae
Glabroculus	100/100/100	5.1 [3.6–6.7]	2	Limoniaceae
Alpherakya	Single specimen	5.1 [3.6–6.7]	5	Crassulaceae
Agriades	100/73/81	4.2 [2.9–5.8]	19	Primulaceae, Saxifragaceae, Ericaceae Fabaceae
Rimisia	Single specimen	4.2 [2.9–5.8]	1	Fabaceae
Cyaniris	100/100/100	4.4 [3.0–5.7]	2	Fabaceae
Eumedonia	100/100/100	4.0 [2.7–5.4]	3	Geraniaceae
Plebejidea	Single specimen	4.0 [2.7–5.4]	2	Fabaceae
Maurus	Single specimen	4.4 [3.1–5.9]	1	Geraniaceae
Kretania	99/77/–	4.6 [3.1–6.1]	17	Fabaceae
Afarsia	Single specimen	4.6 [3.1–6.1]	9	Fabaceae
Plebejus	100/99/98	4.0 [2.7–5.5]	40	Fabaceae, Elaeagnaceae Empetraceae Ericaceae
Pamiria	Single specimen	4.0 [2.7–5.5]	7	Unknown
Patricius	Single specimen	4.4 [2.9–5.9]	7	Unknown
Rueckbeilia	Single specimen	6.9 [4.9–9.0]	2	Fabaceae
Icaricia	96/66/99	5.5 [3.8–7.4]	7	Polygonaceae Fabaceae
Plebulina	Single specimen	5.5 [3.8–7.4]	1	Chenopodiaceae
Freyeria	100/100/100	9.5 [6.8–12.2]	3	Boraginaceae, Phyllanthaceae, Fabaceae
Luthrodes	100/100/100	6.0 [3.9–8.3]	9	Fabaceae Cycadaceae
Chilades	Single specimen	6.0 [3.9–8.3]	ca. 12	Rutaceae Tiliaceae
Pseudolucia	100/100/98	8.1 [5.6–10.7]	46	Fabaceae Polygonaceae Portulacaceae Cuscutaceae
Nabokovia	100/100/100	5.0 [3.2–6.9]	3	Fabaceae
Eldoradina	Single specimen	5.0 [3.2–6.9]	2	Unknown
Itylos	99/74/76	4.6 [3.1–6.3]	24	Fabaceae
Paralycaeides	100/100/100	4.6 [3.1–6.3]	3	Fabaceae
Hemiargus	100/100/100	6.1 [4.2–8.1]	ca. 5	Fabaceae Cucurbitaceae Oxalidaceae
Echinargus	Single specimen	6.1 [4.2–8.1]	1	Fabaceae
Cyclargus	Single specimen	7.0 [4.9–9.3]	7	Asteraceae Fabaceae Malpighiaceae Sapindaceae
Pseudochrysops	Single specimen	11.4 [8.2–14.7]	1	Unknown

Freyeria clade. *Freyeria* Courvoisier, 1920 (TS: *Lycaena trochylus* Freyer, 1845) is frequently treated by modern authors as a subgenus of *Chilades* (Bálint and Johnson, 1997; Tolman and Lewington, 1997). Valvae in the male genitalia of *Freyeria* are elongated and have a short dorsal process (Zhdanko, 2004), and are generally similar to those of *Chilades*. However, molecular data demonstrate that *Freyeria* is not closely related to *Chilades* and represents a distinct clade that cannot possibly be subsumed within *Chilades* as it would result in a paraphyletic assemblage.

Our analysis includes one specimen of *Freyeria* from Turkey (*F. trochylus*) and one from Australia (*F. putli* (Kollar, [1844])). The taxon *F. putli* has until recently been considered a subspecies of *F. trochylus* (Common and Waterhouse, 1981; Parsons, 1999), but now most authors treat it as a good species (Bálint and Johnson, 1997; Braby, 2000). In our analysis, *F. trochylus* and *F. putli* appear as sister taxa, and we estimate that they diverged ca. 3.6 Ma. This is a surprisingly old divergence, and supports the recognition of *F. putli* as a distinct species.

Icaricia–Plebulina clade. In the original descriptions of the genera *Icaricia* Nabokov, [1945] (TS: *Lycaena icarioides* Boisduval, 1852) and *Plebulina* Nabokov, [1945] (TS: *Lycaena emigdionis* Grinnell, 1905), the author clearly indicated morphological characters that distinguished these genera from all other lycenids. In particular, Nabokov noted that *Plebulina* remarkably amalgamates the form of aedeagus similarly to *Plebejus*, with uncus, subunci, and valvae similar in shape to those found in *Albulina*. On the other hand, *Icaricia* remarkably combines a wing pattern similar to that of *Plebejus* with a shape of aedeagus similar to that found in *Aricia* (Nabokov, 1945). Since their description, however, the genera *Icaricia* and *Plebulina* generally have been treated as junior subjective synonyms, or as subgenera of either *Aricia* Reichenbach, 1817 or *Plebejus* Kluk, 1780 (Scott, 1986; Bálint and Johnson, 1997; Gorbunov, 2001; Brock and Kaufman, 2003; Opler and Warren, 2004). In all our analyses, the taxa within *Icaricia* and *Plebulina*, as well as the taxon *Lycaena saepiolus* Boisduval, 1852, form an exclusively Nearctic clade that is sister to all the rest of the Holarctic taxa. Such a topology in the phylogeny is unexpected given modern taxonomic treatments of these groups, and implies that *Icarica* and *Plebulina* cannot possibly be included within *Plebejus* or *Aricia*. This strongly supported result confirms that of Vila et al. (2011), who showed that this clade is the result of a relatively old colonization of the New World that occurred ca. 9.3 Ma. The age of divergence of *Icaricia* (including the taxon *I. saepiolus*) from the *Plebulina* lineage is 5.5 Myr. As a consequence, we maintain the monotypic *Plebulina* as a separate genus, a decision reinforced by the fact that *P. emigdionis* Grinnel, 1905 feeds on a different host-plant family (Chenopodiaceae) from the *Icaricia* taxa (Fabaceae and Polygonaceae), and by certain peculiarities of its larval morphology (Ballmer and Pratt, 1988). Interestingly, divergence ages within the *Icaricia* lineage are fairly old, reaching 4.8 Myr for the *I. acmon–I. shasta* versus *I. icarioides–I. saepiolus* split, which still falls within the 4.0–5.0 Myr genus timeframe. Since no separate genus name has ever been proposed for the *I. acmon–I. shasta* clade, we conservatively retain *Icaricia* as a single unit.

The genus Rueckbeilia (= *"Vacciniina" fergana* clade). The next well supported lineage found in our analysis is represented by a single species traditionally known as *Vacciniina fergana*. This species is recovered as sister to the rest of the Holarctic taxa, except for the *Icaricia–Plebulina* clade. This result is unexpected (but see Kandul et al., 2004; Lukhtanov et al., 2009) as the external morphology of *V. fergana* is extremely similar to *V. optilete*, the type-species of *Vacciniina*. This position of *V. fergana* in the phylogeny is strongly supported in all the analyses and thus cannot be considered an artifact. The deep divergence of the *V. fergana* lineage (6.9 Ma) indicates that it should be treated as a distinct genus, which we describe in the Appendix 1 under the name ***Rueckbeilia* gen. nov.** Interestingly, the isolated systematic position of *V. fergana* was not apparent in a detailed morphological study of this species (Stekolnikov, 2010). In fact, *V. fergana* exhibits a combination of primitive male genitalic characters that are found in some other taxa (Stekolnikov, 2010), and wing patterns that may represent a plesiomorphic condition in *Rueckbeilia*, *Glabroculus* and *Afarsia* + *Kretania*, but independently evolved in *Agriades optilete* (Fig. 4). A more detailed description of *Rueckbeilia* is given in the Appendix 1.

Patricius + (*Pamiria* + *Plebejus*) clade (lineage of *Plebejus sensu lato*). The grouping *Patricius* + (*Pamiria* + *Plebejus*) is recovered as a well supported clade in our phylogeny. This result is not trivial, as *Patricius* and *Pamiria* have usually been regarded as closely related to *Albulina* (Bálint and Johnson, 1997). However, the close relationship of *Patricius* (TS: *Lycaena lucifera* Staudinger, 1867), *Pamiria* (TS: *Lycaena chrysopis* Grum-Grshimailo, 1888) and *Plebejus* (TS: *Papilio argus* Linnaeus, 1758) had already been recognized by Zhdanko (2004), who noted that these genera shared similar *Plebejus*-like male genitalia. Within this clade, the genus *Patricius* is sister to the rest (divergence age 4.4 Myr), while *Pamiria* and *Plebejus* diverged 4.0 Ma.

In all our analyses, the studied species of *Lycaeides* Hübner [1819] (TS: *Papilio argyrognomon* Bergstrasser [1779], also includes *idas*, *melissa* and *anna*) form a clade that is sister to *Plebejus argus*, but its recent age (3.1 Myr) recommends the inclusion of *Lycaeides* within *Plebejus*. Noticeably, Nearctic *Lycaeides* representatives appear as polyphyletic, with unexpected, yet strongly supported, sister relationships between Old and New World taxa. This result is similar to that obtained independently by other researchers (Nice et al., 2005; Gompert et al., 2008; Vila et al., 2011) and deserves further analysis. A number of authors consider *Agriades*, *Alpherakya*, *Vacciniina*, *Plebejides* and *Plebejidea* as synonyms or subgenera of *Plebejus* (Bálint and Johnson, 1997; Gorbunov, 2001), but our results show that these taxa are more closely related to *Aricia* and *Polyommatus* than they are to *Plebejus*. Thus the prevalent use of *Plebejus* as a supergenus is not possible according to the recovered topology.

Polyommatus sensu lato clade. The rest of the Polyommatina taxa form a large clade consisting of 14 genera from *Alpherakya* to *Polyommatus* (Fig. 1). It corresponds to *Polyommatus sensu* Zhdanko, 1983 (but not to *Polyommatus sensu* Zhdanko, 2004) and can be defined by characters of male genitalia similar to those of *Polyommatus sensu stricto* (Zhdanko, 1983, 2004).

However, these genitalic characters may not constitute a true synapomorphy. Stekolnikov (2010) demonstrated a degree of heterogeneity in the male genitalia of this group, and a similar type of genitalia was found in *Rueckbeilia fergana*, which is not closely related. While the *Polyommatus sensu lato* clade is strongly supported, it is formed by several subclades that are older than 4 Myr. The evolutionary relationships among these supported subclades are in some cases unresolved, and we will discuss each in the following paragraphs.

The genus Alpherakya. *Alpherakya* Zhdanko, 1994 (TS: *Lycaena sarta* Alpheraky, 1881) is recovered as sister to *Glabroculus*, although this relationship is not well supported. It should also be noted that the wing patterns and food plants of these two taxa are different (Table 3). The morphology of this genus is characterized by a unique combination of traits that make its identification unmistakable. *Alpherakya* differs from all genera of the *Polyommatus sensu lato* clade, except for *Lysandra*, in having chequered wing fringes. It differs from *Lysandra* in having hairs on the eyes that are scarce and short, whereas in *Lysandra* they are long and dense. In male genitalia, the structure of the valvae is also diagnostic: valvae are comparatively short and broad, with a robust sclerotized inner fold, with a spade-shaped dorsal element in the apex and sclerotized ventral elements. *Alpherakya* can be separated from the taxa in the *Patricius* + (*Pamiria* + *Plebejus*) clade by the wide uncus (Zhdanko, 2004) and by the structure of valvae (Fig. 4). The larval food plants of the *Alpherakya* species are also peculiar: they feed on Crassulaceae (Zhdanko, 1997), whereas most species and genera of the subtribe Polyommatina are associated with Leguminosae or Geraniaceae. Bálint and Johnson (1997) considered *Alpherakya* as part of the genus *Plebejus*. However, our analysis, like the morphological analysis by Zhdanko (2004), does not support this hypothesis and demonstrates that these two genera are phylogenetically distant.

Glabroculus clade. Zhdanko (2004) synonymized *Glabroculus* Lvovsky, 1993 (TS: *Lycaena cyane* Eversmann, 1837) with *Plebejidea*, and considered *Elviria* (TS: *Lycaena elvira* Eversmann, 1854) to be a subgenus of *Plebejidea*. Bálint and Johnson (1997) considered *Glabroculus* (= *cyane*-group) as part of the genus *Polyommatus sensu lato*. Our data support none of these hypotheses. We show that neither *Plebejidea* nor *Polyommatus* is closely related to *Glabroculus*. Instead, *Glabroculus* appears as a sister to *Alpherakya*, although with low statistical support.

Morphologically, *Glabroculus* differs from *Polyommatus* by hairs on the eyes that are scarce and short (in *Polyommatus* they are long and dense) and by the presence of metallic marginal spots on the underside of hind wings. *Glabroculus* differs from the phylogenetically most closely related genus, *Alpherakya*, in having unchequered wing fringes. Moreover, the food plants of *Glabroculus* and *Alpherakya* are different (Table 3). The taxon *E. elvira* (the type-species of the nominal genus *Elviria*) was recovered as a sister to *G. cyane*, and the time of their divergence was estimated as ca. 2.0 Mya. Therefore *Elviria* can be considered a synonym of *Glabroculus*.

Aricia clade. The taxa representing *Aricia* (TS: *Papilio agestis* Denis & Schiffermüller, 1775) and *aratxerxes)*, *Umpria (*TS: *Lycaena chinensis* Murrey, 1874*)*, *Pseudoaricia* (TS: *Polyommatus nicias* Meigen, 1829) and *Ultraaricia* (TS: *Lycaena anteros* Freyer, 1839; includes the studied species *crassipuncta* and *vandarbani*) form a strongly supported clade. Since the divergences among them are younger than 4 Myr, the three latter taxa are subsumed within *Aricia*. The position of *Aricia* within the Polyommatini has been a subject of much discussion. Bálint and Johnson (1997) considered *Aricia* as closely related to the Neotropical taxon *Madeleinea*. Zhdanko (2004) also considered *Aricia* as one of the most basal within the *Polyommatus* section. In contrast, Stekolnikov (2010) found it to represent a young lineage closely related to *Polyommatus*. Our molecular data support the latter hypothesis, although the position of *Aricia* within the *Polyommatus sensu lato* clade is unresolved. Indeed, we recover *Aricia* as sister to *Alpherakya* + *Glabroculus*, but with low support.

Morphologically the genus is quite distinct. In the male genitalia, the aedeagus is lanceolate, with caulis developed, and entirely sclerotized, which is not observed in other taxa of the subtribe (Zhdanko, 2004). Among external characters, the naked eyes and absence of metallic spots on the underside of hindwings are characteristic, although they are not unique within the subtribe.

The genus Afarsia. (TS: *Cupido hyrcana* Lederer, 1869—an invalid name; the valid synonym is *Cupido morgiana* Kirby, 1871). The taxon *C. morgiana* was recognized as a distinctive entity by Zhdanko (1992, 2004) and Bálint and Johnson (1997), but its relationships with other taxa have never been properly documented. Bálint and Johnson (1997) placed it in the same group as *Patricius*, *Pamiria*, *Plebejidea*, *Vacciniina* and *Albulina*. In our reconstruction, it is recovered as sister to *Kretania*, but the support for this relationship is low. Its rather deep divergence (4.6 Myr) suggests that it should be treated as an independent genus. The genus name *Farsia* Zhdanko, 1992, for which *C. morgiana* is the type species, was preoccupied and the new name *Afarsia* Korb and Bolshakov, 2011 (= *Farsia* Zhdanko, 1992; nec *Farsia* Amsel, 1961) has recently been proposed as replacement (Korb and Bolshakov, 2011).

The morphology of the male genitalia of the genus *Afarsia* is similar to *Kretania sensu lato* (see below), but these two taxa are distinct in wing pattern: in *Afarsia* a discal spot on the fore wing upper side is always present and usually strongly enlarged, and one of the marginal metallic spots of the hind wing underside is enlarged. These characters of the wing pattern are also found in the genus *Albulina* (that was the reason why some authors placed *Afarsia* within or close to *Albulina*—see above). However, male genitalia in *Afarsia* are considerably different from those in *Albulina*, both in the structure of uncus, which is basally narrow with long slender arms, and in the shape of the valvae, which have a characteristically concave dorsal margin (Zhdanko, 2004).

Kretania clade. In all our analyses, the taxa within *Plebejides* (TS: *Lycaena pylaon* Fischer von Waldheim, 1832 and *P. zephyrinus*) and *Kretania sensu stricto* (TS: *Lycaena psylorita* Freyer, 1845, includes the studied species *K. eurypilus* and *K. zamotajlovi*), as well as the species *V. alcedo*, form a distinct, statistically well supported clade in ML and BI analyses that originated 4.6 Mya and should be considered as a genus. Within this genus, the species *V. alcedo* appears as sister to the rest, although the position of this taxon is unresolved in the MP analysis. The statistical support for the subclade *Kretania* s.s. + *Plebejides* is very high (100/100/100) and the time of divergence of this subclade is quite recent (ca. 1.9 Mya). The close relationship of *Kretania* s.s. and *Plebejides* was first suggested by Wiemers (2003) based on the molecular analysis of *COI* barcodes and nuclear *ITS2*. Interestingly, the close relationship between *V. alcedo*, *Kretania* s.s. and *Plebejides* has never been recognized by morphologists, who usually consider them as members of different, not closely related groups: *Plebejides* as a member of the *Plebejus* lineage (Zhdanko, 1983; Bálint and Johnson, 1997), *Kretania* as a member of the *Polyommatus* lineage (Bálint and Johnson, 1997), and the taxon *V. alcedo* as a species of *Vacciniina* (Bálint and Johnson, 1997). Nevertheless, these butterflies are fairly similar phenotypically. In fact, species of *Kretania* s.s. differ from *Plebejides* and *V. alcedo* largely in discoloured (brown) upper wings in males, but this is a labile character that has low value in genus-level taxonomy, as it seems to have evolved independently numerous times in the evolution of the Polyommatina (Bálint and Johnson, 1997; Lukhtanov et al., 2005). As a result, we propose the following new combinations: *Kretania alcedo* **comb. nov.**, *Kretania pylaon* **comb. nov.**, *Kretania zephyrinus* **comb. nov.**

The structure of the valvae in *Kretania sensu lato* (including *Plebejides* and the taxon *K. alcedo*) is typical of the genera *Polyommatus* or *Aricia* (Stekolnikov, 2010) (but not typical of the genus *Plebejus* as suggested by Zhdanko, 2004), the uncus is narrow (Zhdanko, 2004) and the wing pattern is extremely similar to that found in *Plebejus*. The combination of these morphological characters makes the genus *Kretania sensu lato* quite distinct.

The genus Maurus. The north African endemic species *Lycaena vogelii* Oberthür, 1920 has been included either within *Plebejus* or in the monotypic genus *Maurus* Bálint, [1992]. Our analysis recovers it as sister to the *Plebejidea–Eumedonia* clade with low support, but its age (4.4 Myr) is sufficient to maintain the genus *Maurus*. The morphology of the genitalia of *M. vogelii* has been described as close to that of *Plebejus* (Zhdanko, 2004). The external morphology of the genus is distinctive and can be recognized by the combination of chequered wing fringes and strongly enlarged discal spot on the fore wing upper side.

Plebejidea–Eumedonia clade. The genus *Plebejidea* (TS: *Lycaena loewii* Zeller, 1847) is usually considered to be close to *Glabroculus* (Tuzov et al., 2000; Zhdanko, 2004), *Polyommatus* (Bálint, 1991), or *Albulina* (Bálint and Johnson, 1997). Our data support none of these taxonomic hypotheses. Instead, in our reconstruction, *Plebejidea* appears as sister to *Eumedonia* with high statistical support. This result is unexpected, as representatives of *Plebejidea* and *Eumedonia* clearly differ in wing pattern and coloration and also in ecology: the species of *Eumedonia* inhabit humid biotopes and their larval food plants are species of Geraniaceae, whereas the species of *Plebejidea* inhabit very dry semi-desert biotopes and their larval food plants are xerophilous species of *Astragalus* (Fabaceae). The morphology of the male genitalia in *Plebejidea* is similar to that of *Glabroculus* (Zhdanko, 2004), but differs by a noticeable basal sclerotization of the subcostal groove of the valvae (Stekolnikov, 2010).

The genus Eumedonia. (TS: *Papilio eumedon* Esper, [1780]) has been considered to be close to *Aricia* (Bálint and Johnson, 1997; Tuzov et al., 2000) in part because they share the same larval food plants (Geraniaceae). However, our results do not support this close relationship, and differences in the structure of the uncus in the male genitalia (Zhdanko, 2004) also suggest that these genera are not closely related. In fact, the genus *Eumedonia* is morphologically quite distinct. It shares a similar form of the valvae in male genitalia with *Plebejidea*, the phylogenetically most closely related genus, as well as with the more distant *Polyommatus*, *Lysandra*, *Neolysandra*, *Aricia*, *Glabroculus* and *Alpherakya*, but differs from them in the narrow uncus and hairless eyes. The aedeagus in *Eumedonia* is comparatively slender and more pointed, resembling that in *Agriades* (Zhdanko, 2004), yet the wing patterns are very different between *Eumedonia* and *Agriades*.

The genus Cyaniris. The genus *Cyaniris* (TS: *Zephyrus argianus* Dalman, 1816, now regarded as a synonym of *Papilio semiargus* Rottemburg, 1775) is often considered to be close to *Polyommatus* s.s. (Hesselbarth et al., 1995; Bálint and Johnson, 1997), but this relationship was questioned on the basis of morphological (Zhdanko, 2004) and molecular analyses (Wiemers et al., 2010). Indeed, our data indicate that *Cyaniris* is not closely related to *Polyommatus* s.s. Instead, it forms a clade together with *Rimisia* and *Agriades sensu lato*, although the support for this relationship is not high. The age of divergence of the *Cyaniris* lineage (4.4 Myr) is sufficient to maintain it as an independent genus.

Cyaniris differs from *Polyommatus*, *Lysandra*, *Neolysandra*, *Aricia*, *Glabroculus*, *Alpherakya* and *Plebejidea* in having a narrow, nearly pointed uncus. It differs from other taxa that also have narrow uncus in the presence of hairs densely covering the eyes and by having a longer aedeagus (Zhdanko, 2004). Additionally, representatives of the genus have no marginal and submarginal pattern on the wing underside. The combination of these characters is characteristic for the genus *Cyaniris*.

The genus Rimisia. The monotypic Central Asian genus *Rimisia* (TS: *Lycaena miris* Staudinger, 1881) has been considered to be close to *Glabroculus* (Bálint and Johnson, 1997; Tuzov et al., 2000), with which it shares a similar pattern on the underside of the wings. This hypothesis is not supported by our data, since *Rimisia* is recovered as sister to *Agriades* with a divergence of more than 4 Myr. The genus *Rimisia* displays an unusual combination of morphological characters: valvae in the male genitalia similar to those of the species *Polyommatus icarus*, short and S-shaped aedeagus, naked eyes and peculiar female genitalia with small papillae anales (Zhdanko, 2004). *Rimisia miris* is considered to have no metallic marginal spots on the hind wings (Zhdanko, 2004), but our analysis of the morphology revealed that the species is variable with respect to this character and some specimens bear metallic scales on the marginal spots.

Agriades clade. According to our results, the genus *Agriades* (TS: *Papilio glandon* Prunner, 1798) originated 4.2 Mya and includes three monophyletic lineages that may be considered as subgenera: *Albulina* (*orbitulus*) (originated 3.6 Mya), *Vacciniina* s.s. (*optilete*) and *Agriades* s.s. (*glandon*, *pheretiades*, *podarce* and *pyrenaicus*) (the latter two split 3.2 Mya). These three taxa are often considered to be distinct genera (e.g. Higgins, 1975), and they indeed differ in their wing patterns (Fig. 5) and larval food plants (Table 3). The close relationship between *Albulina* and *Vacciniina* was recognized by Bálint and Johnson (1997). Our analysis strongly supports the grouping of *Agriades* s.s., *Albulina* and *Vacciniina* s.s. Within this group, *Agriades* s.s. and *Vacciniina* s.s. are sister taxa and *Albulina* is sister to the rest. As our study resulted in the fusion of the taxa *Agriades* s.s., *Albulina* and *Vacciniina* s.s in one genus, the following new combinations result: *Agriades optilete* **comb. nov,** *Agriades orbitulus* **comb. nov.**

Lysandra + (*Neolysandra* + *Polyommatus*) clade. This clade is recovered with a high support in our analysis, and it is estimated to have diverged ca. 5.7 Mya. Within this clade, three genera—*Lysandra*, *Neolysandra* and *Polyommatus*—are recognized in accordance with the criteria discussed above.

Lysandra clade. The genus *Lysandra* (TS: *Papilio coridon* Poda, 1761) is monophyletic and sister to the clade *Neolysandra* + *Polyommatus* with good support. The most characteristic morphological feature of the genus is the clearly chequered wing fringes. This character is not exclusive within the subtribe Polyommatina, and it is found in the distantly related genera *Alpherakya*, *Maurus* and *Grumiana*, as well as in some genera of the Neotropical clade. The hypothesis that *Lysandra* is a synonym of *Meleageria* (which includes the species *daphnis* and *marcida*) (Hesselbarth et al., 1995) is not supported by our phylogeny (see also Wiemers et al., 2010).

Neolysandra clade. In our analysis, the genus *Neolysandra* (TS: *Lycaena diana* Miller, 1912) emerges as a well supported lineage that is a sister to *Polyommatus*. Morphologically *Neolysandra* differs from other genera by the markedly wide and elliptical uncus. Moreover, it differs from the most similar genera *Lysandra* and *Polyommatus* in having short and scarce hairs covering the eyes and in displaying a reduced marginal and submarginal pattern on the wing underside (Zhdanko, 2004). In the molecular reconstruction made by Wiemers et al. (2010), *Neolysandra* was recovered as a polyphyletic taxon. Several reasons might explain this: the taxon sampling (the type species *N. diana* was not included), lack of resolution (the phylogeny was based on two relatively short sequences), and incomplete outgroup sampling (only the phylogenetically distant taxa *Cyaniris semiargus* and *Freyeria trochilus* were used to root the tree). What we consider *Neolysandra* (including the taxa *diana* and *coelestina*) corresponds to Wiemers' *Neolysandra* group I.

Polyommatus clade. In our analysis, the genus *Polyommatus* (TS: *Papilio icarus* Rottemburg, 1775) emerged as a distinct lineage about 4.3 Mya. It is composed of taxa sometimes included in the genera/subgenera *Actisia* Koçak & Kemal, 2001 (TS: *Lycaena actis* Herrich-Schäffer, 1851—a junior synonym, the valid synonym is *Lycaena atys* (Gerhard, 1851); *Admetusia* Koçak & Seven, 1998 (TS: *Papilio admetus* Esper, 1783); *Agrodiaetus* Hübner, 1822

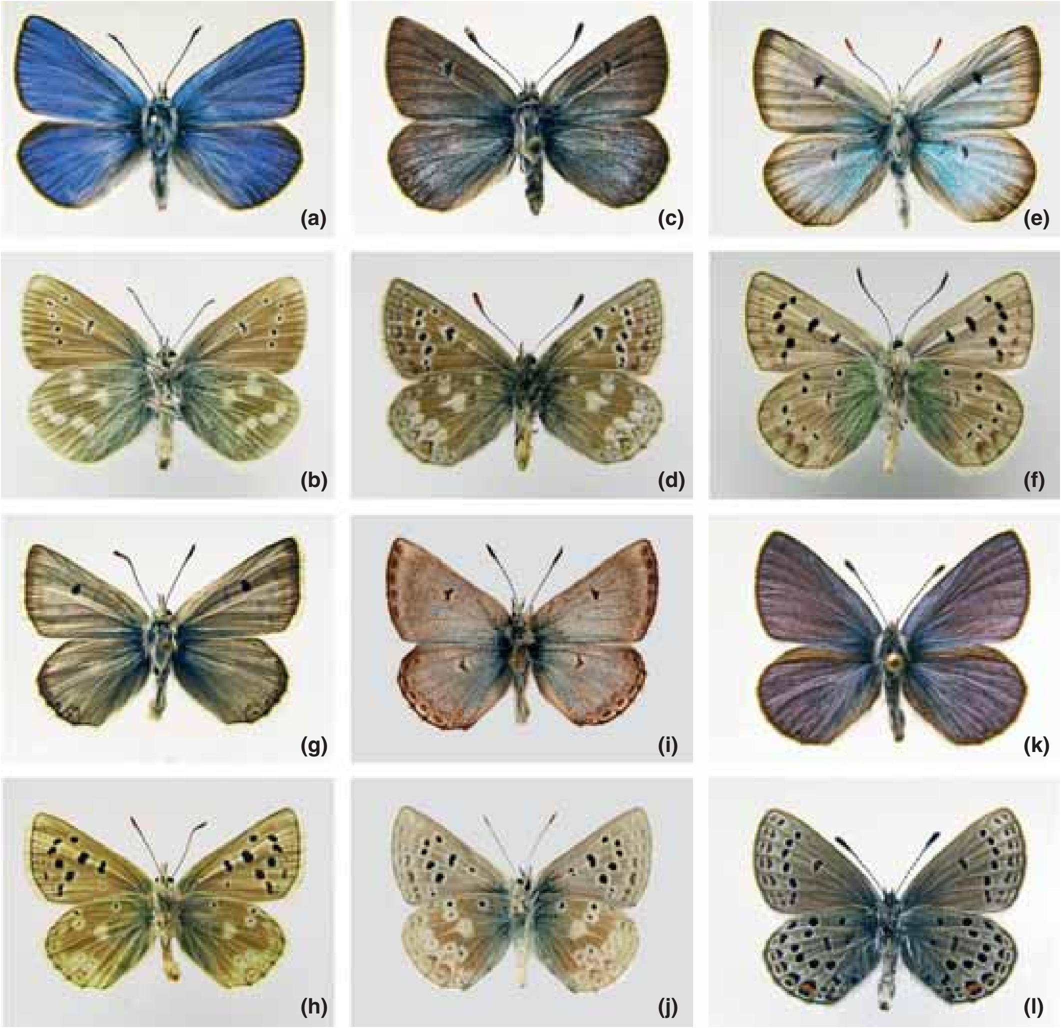


Fig. 5. Representative taxa of the genus *Agriades*. Similarly to other species-rich genera in the subtribe Polyommatina, despite their monophyly and genetic similarities, the genus *Agriades* is morphologically quite diverse with respect to both wing upper side and underside colours and patterns. (a,b) *Agriades orbitulus*; (c,d) *Agriades glandon*; (e,f) *Agriades pheretiades*; (g,h) *Agriades pyrenaicus*; (i,j) *Agriades podarce*; (k,l) *Agriades optilete*.

(= *Hirsutina* Tutt, [1909]) (TS: *Papilio damon* Denis & Schiffermüller, 1775); *Antidolus* Koçak & Kemal, 2001 (TS: *Papilio dolus* var. *antidolus* Rebel, 1901); *Bryna* Evans, 1912 (TS: *Lycaena stoliczkana* Felder & Felder, 1865); *Damaia* Koçak & Kemal, 2001 (TS: *Lycaena dama* Staudinger, 1892); *Meleageria* De Sagarra, 1925 (TS: *Papilio daphnis* Esper, 1778); *Musa* Koçak & Kemal, 2001 (TS: *Polyommatus musa* Koçak & Hosseinpour, 1996); *Paragrodiaetus* Rose & Schurian, 1977 (TS: *Lycaena glaucias* Lederer, 1870); *Peileia* Koçak & Kemal, 2001 (TS: *Polyommatus peilei* Bethune-Baker, 1921); *Phyllisia* Koçak & Kemal, 2001 (TS: *Papilio damon* var. *phyllis* Christoph, 1877); *Plebicula* Higgins, 1969 (TS: *Papilio argester* Bergträsser, 1779); *Polyommatus* Latreille, 1804 (TS: *Papilio icarus* Rottemburg, 1775); *Sublysandra* Koçak, 1977 (TS: *Lycaena candalus* Herrich-Schäffer, 1851); *Thersitesia* Koçak & Seven, 1998 (TS: *Lycaena thersites* Cantener, 1834); *Transcaspius* Koçak & Kemal, 2001 (TS: *Lycaena kindermanni* var. *transcaspica* Heyne, 1895); and *Xerxesia* Koçak & Kemal, 2001 (TS: *Lycaena damone* var. *xerxes* Staudinger, 1899). Several of these taxa are recovered as monophyletic, but no subclade is older than 4 Myr. Thus, according to our criteria, they should not be treated as genera. The composition and relationships obtained are notably similar to those obtained by Zhdanko (2004) based on a morphological analysis (e.g. *Lysandra* and *Neolysandra* are separate genera), but differ in some details (e.g. in the position of *Agrodiaetus*). Wiemers et al. (2010) specifically addressed relationships in this genus based on molecular data from two genetic markers and a different set of outgroup taxa. Deeper relationships are frequently not supported in their study and do not always match those obtained here. The most characteristic morphological features of the genus are the marked downward expansion of the ventral margin of the uncus and the presence of all the basic elements of the wing pattern (Zhdanko, 2004). *Polyommatus* differs from *Lysandra* in having white or grey (not chequered) fringes. It differs from *Neolysandra* in the presence of long hairs densely covering the eyes.

One of the subclades in our analysis is formed by the taxa traditionally included in *Agrodiaetus* (*P. damocles*, *P. ripartii*, *P. surakovi* and *P. damon*) and *Paragrodiaetus* (*P. glaucias* and *P. erschoffii*), thus our results confirm previous results showing that *Agrodiaetus* is a monophyletic entity that includes *Paragrodiaetus* (Kandul et al., 2004, 2007; Wiemers et al., 2010). Morphologically, the subgenus *Agrodiaetus* differs from other genera and subgenera of the subtribe Polyommatina in two autapomorphic characters of the male genitalia: distal extremity of aedeagus pronouncedly swollen (Zhdanko, 1983) and uncus markedly constricted dorsoventrally (Zhdanko, 2004). Our data also strongly support that the taxon *P. stempfferi* is sister to the *Agrodiaetus* clade, and that *P. escheri* is sister to the *P. stempfferi* + *Agrodiaetus* clade. The taxa *P. myrrha* and *P. cornelia*, representative of the taxon *Sublysandra*, form another subgroup of *Polyommatus* that is recovered with low support and with unresolved position. *Sublysandra* is usually considered to be a subgenus of *Polyommatus* (Bálint and Johnson, 1997; Zhdanko, 2004; Wiemers et al., 2010) and is morphologically similar to *Polyommatus* s.s. The subclade representing *Meleageria* (*P. daphnis* and *P. marcida*) is recovered with good support as sister to the species *P. amandus*. The close relationship between *P. amandus* and *P. daphnis* + *P. marcida* is surprising and has not been proposed previously.

The last supported subclade is formed by *Polyommatus* s.s. + (*Plebicula* + *Thersitesia*). The sister relationship of the taxa representing *Plebicula* (*P. dorylas* and *P. nivescesns*) and *Thersitesia* (*P. thersites*) was first recovered by Wiemers et al. (2010). *Polyommatus* s.s. was recovered as monophyletic with high support. Within this clade, the Central Asian species *P. hunza* and *P. venus* (which sometimes have been placed together in the genus *Bryna*) form a clade that is sister to the rest (*erotides* and *icarus*). This Central Asian subclade was also recovered by Wiemers et al. (2010).

Conclusion

A multilocus molecular phylogeny has clarified relationships within the Polyommatina, and molecular age estimates have helped to establish criteria specific for the higher-level taxonomy of this group. Each of the resulting clades that we designate to be a genus displays a distinguishing combination of morphological characters, but most of these characters are not unique to a single genus. The high evolutionary lability of many morphological characters traditionally used to infer relationships in this lineage of butterflies (metallic spots in the hind wing underside, blue versus brown male wing colour, shape of the valvae, membranous ventral fold in the inner part of valvae, marked discal spot on the fore wing upper side, number of segments in the antennal club, pilosity in the eyes, presence of small tails in the hind wing, etc.) is apparent, and explains why the taxonomy of the Polyommatina has been so controversial. Based on our phylogenetic results and the criteria outlined above, we propose the following systematic arrangement for the subtribe Polyommatina (in parentheses we list objective and subjective synonyms for the generic names, objective synonyms are indicated by the sign "="; in brackets we provide a tentative list of species for each genus in alphabetical order; likely synonyms for species are not included; species that were analysed in this study are highlighted in bold):

Subtribe **Polyommatina** *Swainson, 1827*

Genus ***Polyommatus*** Latreille, 1804 (*Actisia* Koçak & Kemal, 2001; *Admetusia* Koçak & Seven, 1998; *Agrodiaetus* Hübner, 1822 (= *Hirsutina* Tutt, [1909]); *Antidolus* Koçak & Kemal, 2001; *Bryna* Evans, 1912; *Dagmara* Koçak & Kemal, 2001; *Damaia* Koçak & Kemal, 2001; *Juldus* Koçak & Kemal, 2001; *Meleageria* De Sagarra, 1925; *Musa* Koçak & Kemal, 2001; *Paragrodiaetus* Rose & Schurian, 1977; *Peileia* Koçak & Kemal, 2001; *Phyllisia* Koçak & Kemal, 2001; *Plebicula* Higgins, 1969; *Sublysandra* Koçak, 1977; *Thersitesia* Koçak & Seven, 1998; *Transcaspius* Koçak & Kemal, 2001; *Xerxesia* Koçak & Kemal, 2001) [*P. abdon* Aistleitner & Aistleitner, 1994), *P. achaemenes* Skala, 2002, *P. actinides* (Staudinger, 1886), *P. admetus* (Esper, 1783), *P. aedon* (Christoph, 1877), *P. aereus* Eckweiler, 1998, *P. afghanicus* (Forster, 1973), *P. ahmadi* (Carbonell, 2001), *P. alcestis* Zerny, 1932, *P. aloisi* Bálint, 1998, *P. altivagans* (Forster, 1956), ***P. amandus* (Schneider, 1792)**, *P. amor* (Lang, 1884), *P. annamaria* Bálint, 1992, *P. anticarmon* (Koçak, 1983), *P. antidolus* (Rebel, 1901), *P. arasbarani* (Carbonell & Naderi, 2000), *P. ardschira* (Brandt, 1938), *P. ariana* (Moore, 1865), *P. aroaniensis* (Brown, 1976), *P. artvinensis* (Carbonell, 1997), *P. aserbeidschanus* (Forster, 1956), *P. atlanticus* (Elwes, 1906), *P. attalaensis* Carbonell, Borie & De Prins, 2004, *P. atys* (Gerhard, 1851), *P. avinovi Sthchetkin*, 1980, *P. baltazardi* (de Lesse, 1963), *P. baytopi* (de Lesse, 1959), *P. belovi* (Dantchenko & Lukhtanov, 2005), *P. bilgini* (Lukhtanov and Dantchenko, 2002), *P. bilucha* (Moore, 1884*)*, *P. birunii* Eckweiler & 10 Hagen, 1998, *P. bogra* Evans, 1932, *P. boisduvalii* (Herrich-Schäffer, 1843), *P. bollandi* Dumont, 1998, *P. buzulmavi* Carbonell, 1991, *P. caeruleus*

(Staudinger, 1871), *P. carmon* (Herrich-Schäffer, 1851, *P. celina* (Austaut, 1879), *P. charmeuxi* (Pagès, 1994), *P. cilicius* (Carbonell, 1998), *P. ciloicus* de Freina & Witt, 1983, **P. cornelia (Fryer, 1851)**, *P. cyaneus* (Staudinger, 1899), *P. dagestanicus* (Forster, 1960), *P. dagmara* (Grum-Grshimaïlo, 1888), *P. dama* (Staudinger, 1992), ***P. damocles*** **(Herrich-Schäffer, 1844)**, ***P. damon*** **(Denis & Schiffermüller, 1775)**, *P. damone* (Eversmann, 1841), *P. damonides* (Staudinger, 1899), *P. dantchenkoi* (Lukhtanov & Wiemers, 2003), ***P. daphnis*** **(Denis & Schiffermüller, 1775)**, *P. deebi* (Larsen, 1974), *P. demavendi* (Pfeiffer, 1938), *P. dizinensis* (Schurian, 1982), *P. dolus* (Hübner, 1823), ***P. dorylas*** **(Denis & Schiffermüller, 1775)**, *P. drunela* Swinhoe, 1910, *P. eckweileri* 10 Hagen, 1988, *P. ectabanensis* (de Lesse, 1964), *P. elbursicus* (Forster, 1956), *P. eleniae* Coutsis & De Prins, 2005, *P. erigone* (Grum-Grshimaïlo, 1890), *P. eriwanensis* (Forster, 1960), *P. ernesti* (Eckweiler, 1989), *P. eroides* (Frivaldszky, 1835), *P. eros* (Ochsenheimer, 1808), ***P. erotides*** **(Staudinger, 1892)**, ***P. erschoffii*** **(Lederer, 1869)**, ***P. escheri*** **(Hübner, 1823)**, *P. fabressei* (Oberthür, 1910), *P. faramarzi* Skala, 2001, *P. femininoides* (Eckweiler, 1987), *P. firdussii* (Forster, 1956), *P. florenciae* (Tytler, 1926), *P. forresti* Bálint, 1992, *P. frauvartianae* Bálint, 1997, *P. fulgens* (de Sagarra, 1925), ***P. glaucias*** **(Lederer, 1870)**, *P. golgus* (Hübner, 1813), *P. guezelmavi* Olivier, Puplesiene, van der Poorten, De Prins & Wiemers, 1999, *P. haigi* (Lukhtanov and Dantchenko, 2002), *P. hamadanensis* (de Lesse, 1959), *P. hopfferi* (Herrich-Schäffer, 1851), *P. huberti* (Carbonell, 1993), *P. humedasae* (Toso & Balletto, 1976), ***P. hunza*** **(Grum-Grshimaïlo, 1890)**, *P. icadius* **(Grum-Grshimaïlo, 1890)**, ***P. icarus*** **(Rottemburg, 1775)**, *P. interjectus* (de Lesse, 1960), *P. iphicarmon* Eckweiler & Rose, 1993, *P. iphidamon* (Staudinger, 1899), *P. iphigenia* (Herrich-Schäffer, 1847), *P. iphigenides* (Staudinger, 1886), *P. isauricoides* Graves, 1923, *P. ishkashimicus* Shchetkin, 1986, *P. juldusus* (Staudinger, 1886), *P. kamtshadalis* (Sheljuzhko, 1933), *P. karacetinae* (Lukhtanov and Dantchenko, 2002), *P. karatavicus* (Lukhtanov, 1990), *P. karindus* (Riley, 1921), *P. kendevani* (Forster, 1956), *P. khorasanensis* (Carbonell, 2001), *P. klausschuriani* 10 Hagen, 1999, *P. kurdistanicus* (Forster, 1961), *P. lama* (Grum-Grshimaïlo, 1891), *P. larseni* (Carbonell, 1994), *P. lukhtanovi* (Dantchenko, 2005), *P. luna* Eckweiler, 2002, *P. lycius* (Carbonell, 1996), *P. magnificus (Grum-Grshimaïlo, 1885)*, *P. maraschi* (Forster, 1956), ***P. marcida*** **(Lederer, 1870)**, *P. masulensis* 10 Hagen & Schurian, 2000, *P. mediator* (Dantchenko & Churkin, 2003), *P. melanius* (Staudinger, 1886), *P. menalcas* (Freyer, 1837), *P. menelaos* Brown, 1976, *P. meoticus* Zhdanko & Shchurov, 1998, *P. merhaba* De Prins, van der Poorten, Borie, van Oorschot, Riemis & Coenen, 1991, *P. mithridates* (Staudinger, 1878), *P. mofidii* (de Lesse, 1963), *P. morgani* (Le Cerf, 1909), *P. muellerae* Eckweiler, 1997, *P. muetingi* (Bálint, 1992), *P. musa* Koçak & Hosseinpour, 1996, ***P. myrrha*** **(Herrich-Schäffer, 1851)**, *P. nephohiptamenos* (Brown & Coutsis, 1978), *P. nepalensis* Forster, 1961, *P. ninae* (Forster, 1956), ***P. nivescens*** **(Keferstein, 1851)**, *P. nuksani* (Forster, 1937), *P. orphicus* (Kolev, 2005), *P. paulae* Wiemers & De Prins, 2004, *P. peilei* Bethune-Baker. 1921, *P. pfeifferi* (Brandt, 1938), *P. phyllides* (Staudinger, 1986), *P. phyllis* (Christoph, 1877), *P. pierceae* (Lukhtanov and Dantchenko, 2002), *P. pierinoi* Bálint, 1995, *P. poseidon* (Herrich-Schäffer, 1851), *P. poseidonides* (Staudinger, 1886), *P. posthumus* Christoph, 1877), *P. pseuderos* (Moore, 1879), *P. pulchella (Bernardi, 1951)*, *P. putnami* (Lukhtanov and Dantchenko, 2002), ***P. ripartii*** **(Freyer, 1830)**, *P. rjabovianus* (Koçak, 1980), *P. rovshani* (Dantchenko & Lukhtanov, 1994), *P. schuriani* (Rose, 1978), *P. sennanensis* (de Lesse, 1959), *P. sertavulensis* (Koçak, 1979), *P. shahkuhensis* (Lukhtanov & Shapoval, 2008), *P. shahrami* (Skala, 2001), *P. shamil* (Dantchenko, 2000), *P. shirkuhensis* 10 Hagen & Eckweiler, 2001, *P. sigberti* Olivier, van der Poorten, Puplesiene & De Prins, 2000, *P. sorkhensis* Eckweiler, 2003, ***P. stempfferi*** **(Brand, 1938)**, *P. stigmatifera* (Courvoisier, 1903), ***P. surakovi*** **(Dantchenko & Lukhtanov, 1994)**, *P. tankeri* (de Lesse, 1960), *P. tenhageni* Schurian & Eckweiler, 1999, *P. theresiae* Schurian, van Oorschot & van den Brink, 1992, ***P. thersites*** **(Cantener, 1834)**, *P. transcaspicus* (Heyne, 1895), *P. tshetverikovi* Nekrutenko, 1977, *P. tsvetajevi (Kurentzov, 1970)*, *P. turcicolus* (Koçak, 1977), *P. turcicus* (Koçak, 1977), *P. urmiaensis* (Schurian & 10 Hagen, 2003), *P. valiabadi* (Rose & Schurian, 1977), *P. vanensis* (de Lesse, 1958), *P. vaspurakani* (Lukhtanov & Dantchenko, 2003), ***P. venus*** **(Staudinger, 1886)**, *P. violetae* (Gómez-Bustillo, Expósito & Martínez, 1979), *P. vagneri* (Forster, 1956), *P. wiskotti* (Courvoisier, 1910), *P. yurinekrutenko* Koçak, 1996, *P. zapvadi* (Carbonell, 1993), *P. zarathustra* Eckweiler, 1997, *P. zardensis* Schurian & 10 Hagen, 2001]

Genus ***Neolysandra*** Koçak, 1977 [***N. coelestina*** **(Eversmann, 1843)**, *N. corona* (Verity, 1936), ***N. diana*** **(Miller, 1913)**, *N. ellisoni* (Pfeiffer, 1931), *N. fatima* Eckweiler & Schurian, 1980, *N. fereiduna* Skala, 2002].

Genus ***Lysandra*** Hemming, 1933 [(= *Uranops* Hemming, 1929); (= *Argus* Scopoli, 1763)] [*L. albicans* (Gerhard, 1851), *L. arzanovi* (Stradomsky & Shchurov, 2005, ***L. bellargus*** **(Rottemburg, 1775)**, *L. caelestissima* (Verity, 1921), ***L. coridon*** **(Poda, 1761)**, *L. corydonius* (Herrich-Schäffer, 1852), *L. dezina* de Freina & Witt, 1983, *L. gennargenti* Leigheb, 1987, *L. hispana* (Herrich-Schäffer, 1851), *L. melamarina* Dantchenko, 2000, *L. nufrellensis* Schurian, 1977, *L. ossmar* (Gerhard, 1851), ***L. punctifera*** **(Oberthür, 1876)**, *L. sheikh* Dantchenko, 2000, *L. syriaca* Tutt, 1910].

Genus ***Agriades*** Hübner, [1819] ((= *Latiorina* Tutt, [1909]); *Albulina* Tutt, 1909; *Himalaya* Koçak &

Seven, 1998; *Mestore* Koçak & Kemal, 2007; *Vacciniina* Tutt, 1909; *Xinjiangia* Huang & Murayama, 1988) [*A. amphirrhoe* (Oberthür, 1910), *A. arcaseia* (Fruhstorfer, 1916), *A. asiatica* (Elwes, 1882), *A. cassiope* Emmel & Emmel, 1998, *A. dis* (Grum-Grshimaïlo, 1891), ***A. glandon*** (**de Prunner, 1798**), *A. jaloka* (Moore, 1875), *A. janigena* (Riley, 1923), *A. kumukuleensis* (Huang & Murayama, 1988), *A. kurtjohnsoni* Bálint, 1997, *A. lehanus* (Moore, 1878), *A. luana* (Evans, 1915), *A. morsheadi* (Evans, 1923), ***A. optilete*** (**Knoch, 1781**), ***A. orbitulus*** (**de Prunner, 1798**), ***A. pheretiades*** (**Eversmann, 1843**), ***A. podarce*** (**Felder & Felder,1865),** ***A. pyrenaicus*** (**Boisduval, 1840**), *A. sikkima* (Moore, 1884)].

Genus ***Rimisia*** Zhdanko, 1994 [***R. miris*** (**Staudinger, 1881**)].

Genus ***Cyaniris*** Dalman, 1816 ((= *Nomiades* Hübner, [1819]); *Glaucolinea* Wang & Rehn, 1999) [*C. bellis* (Freyer, 1842), ***C. semiargus*** (**Rottemburg, 1775**)].

Genus ***Eumedonia*** Forster, 1938 [***E. eumedon*** (**Esper, 1780**), *E. kogistana* (Grum-Grshimaïlo, 1888), ***E. persephatta*** (**Alphéraky, 1881**)].

Genus ***Plebejidea*** Koçak, 1983 [*P. afshar* (Eckweiler, 1998), ***P. loewii*** (**Zeller, 1847**)].

Genus ***Maurus*** Bálint, [1992] [***M. vogelii*** (**Oberthür, 1920**)].

Genus ***Kretania*** Beuret, 1959 (*Plebejides* Sauter, 1868) [***K. alcedo*** (**Christoph, 1877**), *K. allardi* (Oberthür, 1874), *K. beani* (Bálint and Johnson, 1997), *K. csomai* (Bálint, 1992), ***K. eurypilus*** (**Freyer, 1851**), *K. hesperica* (Rambur, 1839), *K. iranica* (Forster, 1938), *K. martini* (Allard, 1867), *K. nicholli* (Elwes, 1901), *K. patriarcha* (Bálint, 1992), *K. philbyi* (Graves, 1925), *K. psylorita* (Freyer, 1845), ***K. pylaon*** (**Fischer von Waldheim, 1832**), *K. sephirus* (Frivaldszky, 1835), *K. trappi* (Verity, 1927), *K. usbeka* (Forster, 1939), ***K. zephyrinus*** (**Christoph, 1884**)].

Genus ***Afarsia*** Korb and Bolshakov, 2011 (= *Farsia* Zhdanko, 1992) [*A. antoninae* (Lukhtanov, 1999), *A. ashretha* (Evans, 1925), *A. hanna* (Evans, 1932), *A. iris* (Lang, 1884), *A. jurii* (Tshikolovets, 1997), ***A. morgiana (*****Kirby, 1871**), *A. omotoi* (Forster, 1972), *A. rutilans* (Staudinger, 1886), *A. sieversii* (Christoph, 1873)].

Genus ***Aricia*** Reichenbach, 1817 ((= *Gynomorphia* Verity, 1929); *Pseudoaricia* Beuret, 1959; *Ultraaricia* Beuret, 1959; *Umpria* Zhdanko, 1994) [***A. agestis*** (**Denis & Schiffermüller, 1775**), *A. anteros* (Freyer, 1838), ***A. artaxerxes*** (**Fabricius, 1793**), *A. bassoni* (Larsen, 1974), ***A. chinensis*** (**Murray, 1874**), *A. cramera* (Eschscholtz, 1821), ***A. crassipuncta*** (**Christoph, 1893**), *A. hyacinthus* (Herrich-Schäffer, 1847), *A. isaurica* (Staudinger, 1871), *A. montensis* (Verity, 1928), *A. morronensis* (Ribbe, 1910), ***A. nicias*** (**Meigen, 1829**), *A. teberdina* (Sheljuzhko, 1934), *A. torulensis* (Hesselbarth & Siepe, 1993), ***A. vandarbani*** (**Pfeiffer, 1937**)].

Genus ***Glabroculus*** Lvovsky, 1993 (*Elviria* Zhdanko, 1994) [***G. cyane*** (**Eversmann, 1837**), ***G. elvira*** (**Eversmann, 1854**).

Genus ***Alpherakya*** Zhdanko, 1994 [*A. bellona* (Grum-Grshimaïlo, 1888), *A. devanica* (Moore, 1875), *A. pilgram* (Bálint and Johnson, 1997), ***A. sarta*** (**Alphéraky, 1881**), *A. sartoides* (Swinhoe, 1910)].

Genus ***Plebejus*** Kluk, 1780 ((= *Rusticus* Hübner, [1806]); (= *Lycoena* Nicholl, 1901); *Lycaeides* Hübner, [1919]) [*P. aegina* (Grum-Grshimaïlo, 1891), *P. agnatus* (Staudinger, 1889), ***P. anna*** (**Edwards, 1861**), *P. argiva* (Staudinger, 1886), ***P. argus*** (**Linnaeus, 1758**), ***P. argyrognomon*** (**Bergsträsser, 1779**), *P. bergi* (Kusnezov, 1908), *P. caspicus* (Forster, 1936), *P. choltagi* (Zhdanko & Churkin, 2001), *P. christophi* (Staudinger, 1874), *P. cleobis* (Bremer, 1861), *P. dzhizaki* Zhdanko, 2000, *P. eversmanni* (Lang, 1884), *P. fridayi* Chermock, 1945, *P. fyodor* Hsu, Bálint & Johnson, 2000, *P. ganssuensis* (Grum-Grshimaïlo, 1891), ***P. idas*** (**Linnaeus, 1760**), *P. kwaja* (Evans, 1932), *P. lepidus* Zhdanko, 2000, *P. maidantagi* Zhdanko & Churkin, 2001, *P. maracandicus* (Erschoff, 1874), ***P. melissa*** (**Edwards, 1873**), *P. mongolicus* (Rühl, 1893), *P. noah* (Herz, 1900), *P. nushibi* Zhdanko, 2000, *P. planorum* (Alphéraky, 1881), *P. pseudaegon* (Butler, 1882), *P. qinghaiensis* (Murayama, 1992), *P. rogneda* (Grum-Grshimaïlo, 1990), *P. roxane* (Grum-Grshimaïlo, 1887), *P. samudra* (Moore, 1875), *P. samuelis* (Nabokov, 1844), *P. shuroabadicus* (Sthchetkin, 1963), *P. sinicus* (Forster, 1936), *P. subsolanus* (Eversmann, 1851), *P. tancrei* (Graeser, 1888), *P. tillo* Zhdanko & Churkin, 2001, *P. tomyris* (Grum-Grshimaïlo, 1890), *P. transcaucasicus* (Rebel, 1901), *P. uiguricus* Zhdanko, 2000].

Genus ***Pamiria*** Zhdanko, 1994 [***P. chrysopis*** (**Grum-Grshimaïlo, 1888**), *P. galathea* (Blanchard, 1844), *P. issa* (Zhdanko, 1992), *P. margo* Zhdanko, 2002, *P. metallica* (Felder & Felder, 1865), *P. omphisa* (Moore, 1875), *P. selma* (Koçak, 1996)].

Genus ***Patricius*** Bálint, [1992] (*Themisia* Zhdanko, 2002) [*P. felicis* (Oberthür, 1886), ***P. lucifer*** (**Staudinger, 1866**), *P. lucifugus* (Fruhstorfer, 1915), *P. lucina* (Grum-Grshimaïlo, 1902), *P. sagona Zhdanko*, 2002, *P. themis* (Grum-Grshimaïlo, 1891), *P. younghusbandi* (Elwes, 1906)].

Genus *Grumiana* Zhdanko, 2004 [*G. berezowskii* (Grum-Grshimaïlo, 1902) (not studied by us, morphologically close to *Plebejus* (Zhdanko, 2004)].

Genus ***Rueckbeilia*** gen. nov. [***R. fergana*** (**Staudinger, 1881**), *R. rosei* (Eckweiler, 1989)].

Genus ***Icaricia*** Nabokov, [1945] [***I. acmon*** (**Westwood, 1851**), *I. cotundra* Scott & Fisher, 2006, ***I. icarioides*** (**Boisduval, 1852**), *I. lupini* (Boisduval, 1869), *I. neurona* (Skinner, 1902)**,** ***I. saepiolus*** (**Boisduval, 1852**)**,** ***I. shasta*** (**Edwards, 1862**)].

Genus ***Plebulina*** Nabokov, [1945] [***P. emigdionis*** (**Grinnell, 1905**)].

Genus ***Freyeria*** Courvoisier, 1920 [*F. minuscule* (Aurivillius, 1909), ***F. putli*** **(Kollar, 1844),** ***F. trochylus*** **(Freyer, 1844)**].

Genus ***Luthrodes*** Druce, 1895 (*Edales* Swinhoe, [1910]; *Lachides* Nekrutenko, 1984) [*L. boopis* (Fruhstorfer, 1897), *L. buruana* (Holland, 1900), ***L. cleotas*** **(Guérin-Méneville, 1831)**, *L. contracta* (Butler, 1880), *L. ella* (Butler, 1881), ***L. galba*** **(Lederer, 1855)**, *L. mindora* (Felder & Felder, 1865), ***L. pandava*** **(Horsfield, 1829)**, *L. peripatria* (Hsu, 1989)].

Genus ***Chilades*** Moore, [1881] [*C. alberta* (Butler, 1901), *C. eleusis* (Demaison, 1888), *C. elicola* (Strand, 1911), *C. kedonga* (Grose-Smith, 1898), ***C. lajus*** **(Stoll, 1780)**, *C. naidina* (Butler, 1886), *C. parrhasius* (Fabricius, 1793), *C. sanctithomae* (Sharpe, 1893), *C. serrula* (Mabille, 1890)]. Species incertae sedis: *C. roemli* Kalis, 1933, *C. saga* (Grose-Smith, 1895), *C. yunnanensis* Watkins, 1927.

Genus ***Itylos*** Draudt, 1921 ((= *Ithylos* Forster, 1955); *Ityloides* Balletto, 1993; *Madeleinea* Bálint, 1993 (= *Nivalis* Balletto, 1993); *Parachilades* Nabokov, 1945) [*I. ardisensis* (Bálint & Lamas, 1997), *I. bella* (Bálint & Lamas, 1997), *I. cobaltana* (Bálint & Lamas, 1994), *I. colca* (Bálint & Lamas, 1997), *I. fumosus* (Balletto, 1993), *I. gradoslamasi* (Bálint, 1997), ***I. huascarana*** **(Bálint & Lamas, 1994),** ***I. koa*** **(Druce, 1876**), *I. lea* (Benyamini, Bálint and Johnson, 1995), *I. lolita* (Bálint, 1993), *I. ludicra* (Weymer, 1890), *I. malvasa* (Bálint & Pyrcz, 2000), ***I. mashenka*** **(Bálint, 1993)**, *I. mira* Bálint & Lamas, 1999, *I. moza* (Staudinger, 1894), *I. nodo* (Bálint and Johnson, 1995), *I. pacis* Draudt, 1921, *I. pasco* Bálint & Lamas, 1994, *I. pelorias* (Weymer, 1890), *I. pnin* Bálint, 1993, ***I. sigal*** **(Benyamini, Bálint and Johnson, 1995)**, ***I. tintarrona*** **(Bálint and Johnson, 1995)**, ***I. titicaca*** **(Weymer, 1890)**, *I. vokoban* (Bálint and Johnson, 1995)].

Genus ***Paralycaeides*** Nabokov, 1945 (*Boliviella* Balletto, 1993) [*P.* ***inconspicua*** **(Draudt, 1921)**, *P. shade* Bálint, 1993, ***P. vapa*** **(Staudinger, 1894)**].

Genus ***Pseudolucia*** Nabokov, 1945 ((= *Pallidula* Balletto, 1993); *Cherchiella* Balletto, 1993; *Facula* Balletto, 1992) [*P. andina* (Bartlett-Calvert, 1893), *P. annamaria* Bálint & Johnson, 1993, *P. arauco* Bálint, Benyamini & Johnson, 2001, *P. argentina* (Balletto, 1993), ***P. asafi*** **Benyamini, Bálint and Johnson, 1995**; *P. aureliana* Bálint & Johnson, 1993, *P. avishai* Benyamini, Bálint and Johnson, 1995; *P. barrigai* Benyamini & Bálint, 2011, *P. benyamini* Bálint and Johnson, 1995; ***P. charlotte*** **Bálint and Johnson, 1995**; ***P. chilensis*** **(Blanchard, 1852)**, *P. clarea* Bálint & Johnson, 1993, *P. collina* (Philippi, 1859), *P. dubi* Bálint, 2001, *P. faundezi* Benyamini & Bálint, 2011, *P. grata* (Köhler, 1934), *P. hazearum* Bálint & Johnson, 1993, *P. henyah* Bálint, Benyamini & Johnson, 2001, *P. humbert* Bálint and Johnson, 1995; *P. johnsoni* Benyamini & Bálint, 2011, *P. jujuyensis* Bálint, Eisele & Johnson, 2000, *P. kechico* Bálint, Benyamini & Johnson, 2001, *P. kinbote* Bálint & Johnson, 1993, *P. lanin* (Bálint & Johnson, 1993), *P. luzmaria* Benyamini & Bálint, 2011, *P. magellana* Benyamini, Bálint and Johnson, 1995; *P. munozae* Benyamini & Bálint, 2011, *P. neuqueniensis* Bálint and Johnson, 1995; *P. oligocyaena* (Ureta, 1956), *P. oraria* Bálint & Benyamini, 2001, *P. parana* Bálint, 1993, *P. patago* (Mabille, 1889), *P. penai* (Bálint & Johnson, 1993), *P. plumbea* (Butler, 1881), *P. scintilla* (Balletto, 1993), *P. shapiroi* Bálint and Johnson, 1995; ***P. sibylla*** **(Kirby, 1871)**, *P. sigal* Benyamini & Bálint, 2011, *P. talia* Bálint, Benyamini & Johnson, 1995, *P. tamara* Bálint and Johnson, 1995; *P. ugartei* Bálint & Benyamini, 2001, *P. valentina* Benyamini & Bálint, 2011, ***P. vera*** **Bálint & Johnson, 1993**, *P. whitakeri* Bálint and Johnson, 1995; *P. zina* Benyamini, Bálint and Johnson, 1995, *P. zoellneri* Benyamini & Bálint, 2011].

Genus ***Nabokovia*** Hemming, 1960 (= *Pseudothecla* Nabokov, 1945) [*N. ada* Bálint and Johnson, 1994, ***N. cuzquenha*** **Bálint & Lamas, 1997**, ***N. faga*** **(Dognin, 1895)**].

Genus ***Eldoradina*** Balletto, 1993 (= *Polytheclus* Bálint & Johnson, 1993) [***E. cyanea*** **(Balletto, 1993)**, *E. sylphis* Draudt, 1921].

Genus ***Hemiargus*** Hübner, 1818 [***H. huntingtoni*** **Rindge & Comstock, 1953**, ***H. martha*** **(Dognin, 1887)**, ***H. hanno*** **(Stoll, 1790)***, ***H. ramon*** **(Dognin, 1887)**].

Genus ***Echinargus*** Nabokov, 1945 [***E. isola*** **(Edwards, 1871)**].

Genus ***Cyclargus*** Nabokov, 1945 [***C. ammon*** **(Lucas, 1857)**, *C. dominicus* (Möschler, 1886), *C. kathleena* Johnson & Matusik, 1992, *C. oualiri* Brévignon, 2002, *C. shuturn* Johnson & Bálint, 1905, *C. sorpresus* Johnson & Matusik, 1992, *C. thomasi* (Clench, 1941)].

Genus ***Pseudochrysops*** Nabokov, 1945 [***P. bornoi*** **(Comstock & Huntington, 1943)**].

*This taxon seems to include at least two species but distributions and nomenclature are unclear.

Acknowledgements

We thank A.A. Stekolnikov for discussion of the manuscript and numerous comments. We thank A. Becerril, J. Beck, D. Benyamini, A.J. Berry, M.F. Braby, M.R. Canfield, M.A. Cornwall, N. Cottle, A.V. Dantchenko, R. Eastwood, J. Ebner, O. Gorbunov, M. Huben, E. John, N.P. Kandul, N.G. Kondla, G. Lamas, D.J. Lohman, J. Mathew, S.D.L. Miller, C.J. Muller, C.C. Nice, P.A. Opler, S. Ramírez, A. Shapiro, V. Shchurov, D.L. Stern, M.W. Tan, M.A. Travassos, A. Ugarte and A.D. Warren for providing material and field assistance. Support for this research was provided by the Spanish MICINN (project CGL2010-21226/BOS to G.T. and R.V. and predoctoral fellowship BES-2008-002054 to G.T.), by the Russian Foundation for Basic

Research (grants 12-04-00490, 11-04-00076, 11-04-00734 and 11-04-01119), by grant 16.518.11.7070 (Ministry of Education and Science of the Russian Federation), and by the programmes of the Presidium of Russian Academy of Science "Gene pools and genetic diversity" and "Origin of biosphere and evolution of geo-biological systems" to V.A.L.; grants from the Baker Foundation, the Green Memorial Fund of Harvard University and the Putnam Expeditionary Fund of the Museum of Comparative Zoology to N.E.P. and R.V. and NSF DEB-0447242 to N.E.P.

References

Abramson, N.I., Lebedev, V.S., Tesakov, A.S., Bannikova, A.A., 2009. Supraspecies relationships in the subfamily Arvicolinae (Rodentia, Cricetidae): an unexpected result of nuclear gene analysis. Mol. Biol. 43, 834–846.

Als, T.D., Vila, R., Kandul, N.P., Nash, D.R., Yen, S.-H., Hsu, Y.-F., Mignault, A.A., Boomsma, J.J., Pierce, N.E., 2004. The evolution of alternative parasitic life histories in Large Blue butterflies. Nature 432, 386–390.

Ashlock, P.D., 1971. Monophyly and associated terms. Syst. Zool. 20, 63–69.

Avise, J.C., Johns, G.C., 1999. Proposal for a standardized temporal scheme of biological classification for extant species. Proc. Natl Acad. Sci. USA 96, 7358–7363.

Avise, J.C., Liu, J.-X., 2011. On the temporal inconsistencies of Linnean taxonomic ranks. Biol. J. Linn. Soc. 102, 707–714.

Avise, J.C., Mitchell, D., 2007. Time to standardize taxonomies. Syst. Biol. 56, 130–133.

Bálint, Z., 1991. A xeromontane lycaenid butterfly species: *Plebejus pylaon* (Fischer von Waldheim, 1832) (Lepidoptera: Lycaenidae) and its relatives. Part I. Janus Pannonius Muz. Évk. 35, 33–69.

Bálint, Z., Johnson, K., 1994. Polyommatine lycaenids of the oreal biome in the Neotropics, Part I: The *Itylos* section (Lepidoptera: Lycaenidae, Polyommatini). Annls Hist.-Nat. Mus. Natn Hung. 86, 53–77.

Bálint, Z., Johnson, K., 1995. Neotropical polyommatine diversity and affinities, I. Relationships of the higher taxa (Lepidoptera: Lycaenidae). Acta Zool. Acad. Sci. Hung. 41, 211–235.

Bálint, Z., Johnson, K., 1997. Reformation of the *Polyommatus* section with taxonomic and biogeographic overview (Lepidoptera, Lycaenidae, Polyommatini). Neue Entomol. Nachr. 40, 1–68.

Ballmer, G.R., Pratt, G.F., 1988. A survey of the last instar larvae of the Lycaenidae (Lepidoptera) of California. J. Res. Lepid. 27, 1–81.

Bethune-Baker, G.T., 1913. Comparative notes on *Chilades galba*, Led., and *phiala*, Gr. Gr. Trans. Entomol. Soc. Lond. 61, 201–204.

Biomatters Ltd. 2009. Geneious ver. 4.8.3. Available at: http://www.geneious.com.

Biro, L.P., Bálint, Z., Kertesz, K., Vertesy, Z., Mark, G.I., Horvath, Z.E., Balazs, J., Mehn, D., Kiricsi, I., Lousse, V., Vigneron, J.P., 2003. Role of photonic-crystal-type structures in the thermal regulation of a Lycaenid butterfly sister species pair. Phys. Rev. E 67, 1–7.

Braby, M.F., 2000. Butterflies of Australia: Their Identification, Biology, and Distribution. CSIRO Publishing, Collingwood, Australia.

Brereton, T.M., Warren, M.S., Roy, D.B., Stewart, K., 2008. The changing status of the Chalkhill Blue butterfly *Polyommatus coridon* in the UK: the impacts of conservation policies and environmental factors. J. Insect Conserv. 12, 629–638.

Bridges, C.A. 1988. Catalogue of Family-group and Genus-group names (Lepidoptera: Rhopalocera). Self-published.

Britten, R.J., 1986. Rates of DNA-sequence evolution differ between taxonomic groups. Science 231, 1393–1398.

Brock, J.P., Kaufman, K., 2003. Butterflies of North America. Houghton Mifflin, New York, NY.

Brower, A.V.Z., DeSalle, R., Vogler, A.P., 1996. Gene trees, species trees and systematics: a cladistic perspective. Annu. Rev. Ecol. Syst. 27, 423–450.

Brummitt, R.K., 2003. Further dogged defense of paraphyletic taxa. Taxon 52, 803–804.

Carroll, M.J., Anderson, B.J., Brereton, T.M., Knight, S.J., Kudrna, O., Thomas, C.D., 2009. Climate change and translocations: the potential to re-establish two regionally-extinct butterfly species in Britain. Biol. Conserv. 142, 2114–2121.

Castresana, J., 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17, 540–552.

Castresana, J., 2001. Cytochrome *b* phylogeny and the taxonomy of great apes and mammals. Mol. Biol. Evol. 18, 465–471.

Common, I.F.B., Waterhouse, D.F., 1981. Butterflies of Australia. Angus and Robertson, Australia.

Coyne, J.A., Orr, A.H., 2004. Speciation. Sinauer, Sunderland.

Cranston, P.S., Hardy, N.B., Morse, G.E., Puslednik, L., McCluen, S.R., 2010. When molecules and morphology concur: the 'Gondwanan' midges (Diptera: Chironomidae). Syst. Entomol. 35, 636–648.

Descimon, H., Mallet, J., 2009. Bad species. In: Settele, J., Konvicka, M., Shreeve, T.G., Dennis, R., Van Dyck, H. (Eds.), Ecology of Butterflies in Europe. Cambridge University Press, Cambridge, UK, pp. 219–249.

Drummond, A.J., Rambaut, A., 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7, 214.

Drummond, A.J., Ho, S.Y.W., Phillips, M.J., Rambaut, A., 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4, e88.

Eliot, J.N., 1973. The higher classification of the Lycaenidae (Lepidoptera): a tentative arrangement. Bull. Br. Mus. Nat. Hist. 28, 371–505.

Envall, M., 2008. On the difference between mono-, holo-, and paraphyletic groups: a consistent distinction of process and pattern. Biol. J. Linn. Soc. 94, 217–220.

Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783–791.

da Fonseca, R.R., Johnson, W.E., O'Brien, S.J., Ramos, M.J., Antunes, A., 2008. The adaptive evolution of the mammalian mitochondrial genome. BMC Genomics 9, 119.

Forster, W., 1936. Beitrag zur Systematik des Tribus Lycaenini unter besonderer Berücksichtigung der *argyrognomon* – und der *argus*-Gruppe. Mitt. Münch. Ent. Ges. 26, 41–150.

Forster, W., 1938. Das System der paläarktischen Polyommatini (Lepidoptera, Lycaenidae). Mitt. Münch. Ent. Ges. 28, 97–118, 395.

Godfray, H.C.J., Knapp, S., 2004. Introduction. Taxonomy for the twenty-first century. Phil. Trans. R. Soc. Lond. B., Biol. Sci. 359, 559–569.

Gompert, Z., Fordyce, J.A., Forister, M.L., Shapiro, A.M., Nice, C.C., 2006. Homoploid hybrid speciation in an extreme habitat. Science 314, 1923–1925.

Gompert, Z., Fordyce, J.A., Forister, M.L., Nice, C.C., 2008. Recent colonization and radiation of North American Lycaeides (Plebejusx) inferred from mtDNA. Mol. Phylogenet. Evol. 48, 461–490.

Gorbunov, P.Y. 2001. The Butterflies of Russia (Lepidoptera: Hesperioidea and Papilionoidea): Classification, Genitalia, Keys for Identification. Thesis, Ekaterinburg.

Goverde, M., Bazin, A., Kéry, M., Shykoff, J.A., Erhardt, A., 2008. Positive effects of cyanogenic glycosides in food plants on larval development of the common blue butterfly. Oecologia 157, 409–418.

Griffiths, G.C.D., 1973. Some fundamental problems in biological classification. Syst. Biol. 22, 338–343.

Groves, C., 2004. The what, why and how of primate taxonomy. Int. J. Primatol. 25, 1105–1126.

Hedges, S.B., Kumar, S., 2009. The Timetree of Life. Oxford University Press, New York, NY.

Hennig, W., 1950. Grundzüge einer Theorie der phylogenetischen Systematik. Deutcher Zentralverlag, Berlin.

Hennig, W., 1966. Phylogenetic Systematics. University of Illinois Press, Urbana, IL.

Hesselbarth, G., Oorschot, H.V., Wagener, S. 1995. Die Tagfalter der Türkei unter Berücksichtigung der angrenzenden Länder. Author's edition, Bocholt, Germany.

Higgins, L.G., 1975. The Classification of European Butterflies. Collins, London, UK.

Hirowatari, T., 1992. A generic classification of the tribe Polyommatini of the Oriental and Australian regions (Lepidoptera, Lycaenidae, Polyommatinae). Bull. Univ. Osaka Prefect. Ser. B. 44, 1–102.

Hörandl, E., Stuessy, T.F., 2010. Paraphyletic groups as natural units of biological classification. Taxon 59, 1641–1653.

Io, C., 1998. The Classification and Identification of Chinese butterflies. Henan Science and Technology Publishing House, Zhengzhou, China.

Kandul, N.P., Lukhtanov, V.A., Dantchenko, A.V., Coleman, J.W.S., Sekercioglu, C.H., Haig, D., Pierce, N.E., 2004. Phylogeny of *Agrodiaetus* Hübner 1822 (Lepidoptera: Lycaenidae) inferred from mtDNA sequences of COI and COII and nuclear sequences of EF1-α: karyotype diversification and species radiation. Syst. Biol. 53, 278–298.

Kandul, N.P., Lukhtanov, V.A., Pierce, N.E., 2007. Karyotypic diversity and speciation in *Agrodiaetus* butterflies. Evolution 61, 546–559.

Keller, A., Schleicher, T., Schultz, J., Müller, T., Dandekar, T., Wolf, M., 2009. 5.8S-28S rRNA interaction and HMM-based ITS2 annotation. Gene 430, 50–57.

Kluge, A.G., 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). Syst. Zool. 38, 7–25.

Koetschan, C., Förster, F., Keller, A., Schleicher, T., Ruderisch, B., Schwarz, R., Müller, T., Wolf, M., Schultz, J., 2010. The ITS2 Database III – sequences and structures for phylogeny. Nucl. Acids Res. 38, 275–279.

Korb, S.K., Bolshakov, L.V., 2011. A catalogue of butterflies (Lepidoptera: Papilionoformes) of the former USSR, reformatted and updated. Eversmannia (Suppl. 2), 1–24.

Krauss, J., Schmitt, T., Seitz, A., Steffan-Dewenter, I., Tscharntke, T., 2004. Effects of habitat fragmentation on the genetic structure of the monophagous butterfly *Polyommatus coridon* along its northern range margin. Mol. Ecol. 13, 311–320.

Kudrna, O., 2002. The distribution Atlas of European butterflies. Oedippus 20, 1–342.

Kuhne, G., Schmitt, T., 2010. Genotype shifts along one generation of the blue butterfly *Polyommatus coridon* without changes in allele frequencies. Ann. Zool. Fenn. 47, 278–286.

Lamas, G., 2004. Checklist: Part 4A. Hesperioidea – Papilionoidea. In: Heppner, J.B. (Ed.), Atlas of Neotropical Lepidoptera. Association for Tropical Lepidoptera, Scientific Publishers, Gainesville, pp. 1–439.

Li, W., 1997. Molecular Evolution. Sinauer, Sunderland.

Lucky, A., Sarnat, E.M., 2008. New species of *Lordomyrma* (Hymenoptera: Formicidae) from southeast Asia and Fiji. Zootaxa 1681, 37–46.

Lukhtanov, V.A., 2010. Dobzhansky's rule and reinforcement of pre-zygotic reproductive isolation in zones of secondary contact. Zh. Obshch. Biol. 71, 372–385.

Lukhtanov, V.A., Dantchenko, A.V., 2002. Principles of highly ordered metaphase I bivalent arrangement in spermatocytes of *Agrodiaetus* (Lepidoptera). Chrom. Res. 10, 5–20.

Lukhtanov, V.A., Kandul, N.P., Plotkin, J.B., Dantchenko, A.V., Haig, D., Pierce, N.E., 2005. Reinforcement of pre-zygotic isolation and karyotype evolution in *Agrodiaetus* butterflies. Nature 436, 385–389.

Lukhtanov, V.A., Sourakov, A., Zakharov, E.V., Hebert, P.D.N., 2009. DNA barcoding Central Asian butterflies: increasing geographical dimension does not significantly reduce the success of species identification. Mol. Ecol. Res. 9, 1302–1310.

Maddison, W.P., Maddison, D.R. 2007. MESQUITE: a modular system for evolutionary analysis. Available at: http://mesquiteproject.org.

Martin, A.P., Palumbi, S.R., 1993. Body size, metabolic-rate, generation time, and the molecular clock. Proc. Natl Acad. Sci. USA 90, 4087–4091.

Mattoni, R.H.T., Fiedler, K., 1993. On the functional foreleg tarsus in *Caerulea* males (Lepidoptera: Lycaenidae: Polyommatini). J. Res. Lepid. 30, 289–296.

Mensi, P., Lattes, A., Salvidio, S., Balletto, E., 1988. Taxonomy, evolutionary biology and biogeography of South West European *Polyommatus coridon* (Lepidoptera: Lycaenidae). Zool. J. Linn. Soc. 93, 259–271.

Nabokov, V., 1944. Notes on the morphology of the genus *Lycaeides* (Lycaenidae, Lepidoptera). Psyche 50, 104–138.

Nabokov, V., 1945. Notes on Neotropical Plebejinae (Lycaenidae). Psyche 51, 1–61.

Nelson, C., Murphy, D.J., Ladiges, P.Y., 2003. Brummitt on paraphyly: a response. Taxon 52, 295–298.

Nice, C.C., Anthony, N., Gelembiuk, G., Raterman, D., ffrench-Constant, R., 2005. The history and geography of diversification within the butterfly genus *Lycaeides* in North America. Mol. Ecol. 14, 1741–1754.

Opler, P.A., Warren, A.D., 2004. Butterflies of North America. 2. Scientific Names List for Butterfly Species of North America, North of Mexico. Contributions of the C. P. Gillette Museum of Arthropod Diversity, Colorado State University, Colorado.

Parsons, M., 1999. The Butterflies of Papua New Guinea. Their Systematics and Biology. Academic Press, San Diego.

Pierce, N.E., Braby, M.F., Heath, A., Lohman, D.J., Mathew, J., Rand, D.B., Travassos, M.A., 2002. The ecology and evolution of ant association in the Lycaenidae (Lepidoptera). Ann. Rev. Entomol. 47, 733–771.

Posada, D., 2008. jModelTest: phylogenetic model averaging. Mol. Biol. Evol. 25, 1253–1256.

Pratt, G.F., Wright, D.M., Ballmer, G.R., 2006. Allozyme phylogeny of North American blues (Lepidoptera : Lycaenidae : Polyommatini). Pan-Pac. Entomol. 82, 283–295.

Rambaut, A. 2009. FigTree ver. 1.2.2. Available at: http://beast.bio.ed.ac.uk/FigTree.

Rambaut, A., Drummond, A.J. 2007. Tracer ver. 1.5. Available at: http://beast.bio.ed.ac.uk/Tracer.

Robbins, R.K., Duarte, M., 2006. Systematic placement of *Lycaena cogina* Schaus (Lepidoptera: Lycaenidae: Polyommatinae), a biogeographically disjunct New World species. Proc. Entomol. Soc. Wash. 108, 226–236.

Rowe, K.C., Reno, M.L., Richmond, D.M., Adkins, R.M., Steppan, S.J., 2008. Pliocene colonization and adaptive radiations in Australia and New Guinea (Sahul): multilocus systematics of the old endemic rodents (Muroidea: Murinae). Mol. Phyl. Evol. 47, 84–101.

Rusterholz, H.P., Erhardt, A., 2000. Can nectar properties explain sex-specific flower preferences in the Adonis Blue butterfly *Lysandra bellargus*? Ecol. Entomol. 25, 81–90.

Schmidt-Lebuhn, A.N., 2011. Fallacies and false premises – a critical assessment of the arguments for the recognition of paraphyletic taxa in botany. Cladistics 28, 174–187.

Schmitt, T., 2007. Molecular biogeography of Europe: Pleistocene cycles and postglacial trends. Front. Zool. 4, 11.

Schmitt, T., Giessl, A., Seitz, A., 2003. Did *Polyommatus icarus* (Lepidoptera: Lycaenidae) have distinct glacial refugia in southern Europe? Evidence from population genetics. Biol. J. Linn. Soc. 80, 529–538.

Schultz, J., Wolf, M., 2009. ITS2 Sequence-structure analysis in phylogenetics: a how-to manual for molecular systematics. Mol. Phyl. Evol. 52, 520–523.

Schwenk, K., 1994. Comparative biology and the importance of cladistic classification: a case study from the sensory biology of squamate reptiles. Biol. J. Linn. Soc. 52, 69–82.

Scott, J.A., 1986. The Butterflies of North America. Stanford University Press, Stanford, CA.

Seibel, P.N., Müller, T., Dandekar, T., Schultz, J., Wolf, M., 2006. 4SALE – a tool for synchronous RNA sequence and secondary structure alignment and editing. BMC Bioinformatics 7, 498.

Seibel, P.N., Müller, T., Dandekar, T., Wolf, M., 2008. Synchronous visual analysis and editing of RNA sequence and secondary structure alignments using 4SALE. BMC Res. Notes 1, 91.

Sison-Mangus, M.P., Briscoe, A.D., Zaccardi, G., Knüttel, H., Kelber, A., 2008. The lycaenid butterfly *Polyommatus icarus* uses a duplicated blue opsin to see green. J. Exp. Biol. 211, 361–369.

Stekolnikov, A.A., 2010. Evolution of the skeleton and musculature of the male genitalia in the family Lycaenidae (Lepidoptera): II. Infratribe Polyommatina Swainson, 1827. Entomol. Obozr. 89, 561–587.

Stekolnikov, A.A., Kuznetzov, V.I., 2005. Evolution of skeleton and musculature of the male genitalia in the family Lycaenidae (Lepidoptera): I. The Cupido, Glaucopsyche, Lycaenopsis, and Itylos sections. Entomol. Obozr. 84, 738–760.

Stempffer, H., 1937. Contribution a l'etude des Plebeiinae palearctiques. Bull. Soc. Entomol. Fr. 42, 211–218, 296–301.

Stempffer, H., 1967. The genera of the African Lycaenidae (Lepidoptera: Rhopalocera). Bull. Br. Mus. Nat. Hist. (Ent.) Suppl. 10, 1–322.

Sukumaran, J., Holder, M.T., 2010. DendroPy: a python library for phylogenetic computing. Bioinformatics 26, 1569–1571.

Swofford, D.L., 2000. PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods). Sinauer Associates, Sunderland, MA.

Talavera, G., Castresana, J., 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. 56, 564–577.

Tolman, T., Lewington, R., 1997. Butterflies of Britain and Europe, Collins Field Guide. HarperCollins, London.

Trager, M.D., Daniels, J.C., 2009. Ant tending of Miami blue butterfly larvae (Lepidoptera: Lycaenidae): partner diversity and effects on larval performance. Flor. Entomol. 92, 474–482.

Tuzov, V.K., Bogdanov, P.V., Churkin, S.V., Dantchenko, A.V., Devyatkin, A.L., Murzin, V.S., Samodurov, G.D., Zhdanko, A.B., 2000. Guide to the Butterflies of the Russia and Adjacent Territories (Lepidoptera, Rhopalocera). Vol. 2. Pensoft, Sofia-Moscow.

Vandewoestijne, S., Schtickzelle, N., Baguette, M., 2008. Positive correlation between genetic diversity and fitness in a large, well-connected metapopulation. BMC Biol. 6, 46.

Vershinina, A.O., Lukhtanov, V.A., 2010. Geographical distribution of the cryptic species *Agrodiaetus alcestis alcestis*, *A. alcestis karacetinae* and *A. demavendi* (Lepidoptera, Lycaenidae) revealed by cytogenetic analysis. Comp. Cytogenet. 4, 1–11.

Vila, R., Lukhtanov, V.A., Talavera, G., Gil, T.F., Pierce, N.E., 2010. How common are dot-like distribution ranges? Taxonomical oversplitting in Western European *Agrodiaetus* (Lepidoptera, Lycaenidae) revealed by chromosomal and molecular markers. Biol. J. Linn. Soc. 101, 130–154.

Vila, R., Bell, C.D., Macniven, R., Goldman-Huertas, B., Ree, R.H., Marshall, C.R., Bálint, Z., Johnson, K., Benyamini, D., Pierce, N.E., 2011. Phylogeny and palaeoecology of *Polyommatus* blue butterflies show Beringia was a climate-regulated gateway to the New World. Proc. R. Soc. Lond. B 278, 2737–2744.

Vilnet, A.A., Milyutina, I.A., Konstantinova, N.A., Ignatov, M.S., Troitsky, A.V., 2007. Phylogeny of the genus *Lophozia* (Dumort.) Dumort. s. str. inferred from nuclear and chloroplast sequences ITS1-2 and TRNL-F. Russ. J. Genet. 43, 1306–1313.

White, M.J.D., 1973. Animal Cytology and Evolution. Cambridge University Press, Cambridge, UK.

Wiemers, M. 2003. Chromosome differentiation and the radiation of the butterfly subgenus *Agrodiaetus* (Lepidoptera: Lycaenidae: *Polyommatus*) – a molecular phylogenetic approach. PhD thesis, University of Bonn. Available at: http://hss.ulb.uni-bonn.de/2003/0278/0278.htm.

Wiemers, M., Fiedler, K., 2007. Does the DNA barcoding gap exist? A case study in blue butterflies (Lepidoptera: Lycaenidae). Front. Zool. 4, 8.

Wiemers, M., Keller, A., Wolf, M., 2009. ITS2 secondary structure improves phylogeny estimation in a radiation of blue butterflies of the subgenus *Agrodiaetus* (Lepidoptera: Lycaenidae: *Polyommatus*). BMC Evol. Biol. 9, 300.

Wiemers, M., Stradomsky, B.V., Vodolazhsky, D.I., 2010. A molecular phylogeny of *Polyommatus* s. str. and *Plebicula* based on mitochondrial COI and nuclear ITS2 sequences (Lepidoptera: Lycaenidae). Eur. J. Entomol. 107, 325–336.

Zhdanko, A.B., 1983. A key to the lycaenid genera (Lepidoptera, Lycaenidae) of the USSR based on the characters of the male genitalia. Entomol. Obozr. 62, 131–152.

Zhdanko, A.B., 1992. *Farsia* subgen. n. – a new subgenus of the genus *Vacciniina* (Lycaenidae, Lepidoptera) from the Middle Asia. Zool. Zh. 71, 151–154.

Zhdanko, A.B., 1997. Lycaenid foodplants in Kazakhstan and Middle Asia. Atalanta 28, 97–110.

Zhdanko, A.B., 2004. A revision of the supraspecific taxa of the lycaenid tribe Polyommatini (Lepidoptera, Lycaenidae). Entomol. Obozr. 83, 645–663.

Zwickl, D.J. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. PhD thesis, University of Texas at Austin. Available at: https://www.nescent.org/wg_garli/Main_Page.

Supporting information

Additional Supporting information may be found in the online version of this article:

Appendix S1. Includes Tables S1–S3 and associated references.

Data S1. Talavera_et_al_2011.nex.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Appendix 1

Description of the new genus Rueckbeilia

Rueckbeilia Lukhtanov, Talavera, Pierce & Vila, gen. nov.

Type species *Lycaena fergana* Staudinger, 1881 (Stett. Ent. Z. p. 262) (type species originally described as "*Lyc.[aena] Loewii* Z. var.? *Fergana* Stgr.").

The name is feminine in gender.

Description

Head with whitish scales. Antennae approximately half as long as fore wing costa, with alternating black and white dots. The antennal club consists of 14 or 15 segments. Dorsal side of the club black with white end, ventral side reddish brown. Second segment of labial palpus white with blackish brush; third segment black. Eyes without hairs, bordered with snow-white scales. Length of fore wing 13–15 mm. Wing colour sexually dimorphic. Male upper side (Fig. 2a) violet-blue; black margin of the wings narrow (0.5–1 mm); veins slightly darkened distally; black discal spot of the fore wing small or unclear. Inner part of cilia dark grey, outer part white. Male underside (Fig. 2b) greyish brown with black spots encircled by white; hindwing with two to four orange submarginal spots; three to fourblack marginal spots near tornus with blue metallic scales. Female upper side brown with two to three orange submarginal spots on hindwing and with cilia as in the male. Female underside almost the same as in the male, but slightly darker.

Male genitalia

Uncus divided into two sclerotized lobes. Gnatos situated at their bases, in the form of sclerotized hooks. Juxta with two long narrow branches. Aedeagus straight and relatively short. Valvae (Fig. 3a) narrow, with a strongly convex and setose longitudinal membranous fold on the ventral wall. The costal margin of the valvae is bent medially, so that a membranous subcostal groove is formed between this margin and the longitudinal fold. Sacculus extends along the entire ventral margin of the valvae. The musculature of male genitalia has been investigated by Stekolnikov (2011), who has found that (i) the transversal intravalval musculature consists of a single undifferentiated muscle, (ii) the fixed insertion site of the intravalval muscle expands over the entire surface of the sacculus, and (iii) the fibres diverge in a radial pattern and attach to both the articular and the costal margin of the valvae.

Female genitalia

Ovipositor rather short. Anterior apophyses three times shorter than posterior ones. Antevaginal plate large, with two sclerotized lobes, and forming a deep concavity with membranous proboscis. Proboscis with a small, strongly sclerotized plate on the top and connected with ductus bursae. Bursa membranous, without signum.

Diagnosis

The external morphology of *Rueckbeilia* is most similar to the genera *Kretania* (especially *K. alcedo*) and *Agriades* (especially *A. optilete*). All these taxa share a similar wing pattern that seems to have evolved independently several times, and a possibly plesiomorphic structure of the male valvae with a well developed membranous median fold. However, *Rueckbeilia* represents a distinct monophyletic entity on the basis of molecular characters. It is not closely related to *Kretania* or *Agriades*, and can be distinguished from these and from other genera by using molecular markers from *COI*, *COII*, *EF-1α*, *Wg*, *ITS2*, *CAD*, *28S*, and *H3* (Table S3, Appendix S1). The mitochondrial diagnostic characters are in the following positions in *COI* + *COII* mtDNA: guanine (G) in position 1801 and thymine (T) in position 2139. For the nuclear marker *Wg*, diagnostic characters are in the following positions: adenine (A) in 217 and G in 222. For the nuclear marker *EF-1α*, diagnostic characters are A in position 295 and T in position 304. For the nuclear marker *CAD*, G in position 413 is a diagnostic character. For the nuclear marker *28S*, diagnostic characters are in the following positions: G in 284 and T in 586. For the nuclear marker *ITS2*, diagnostic characters are in the following positions: cytosine (C) in 12 and 1026 (positions refer to the alignment provided as Supplementary Table S3). Except for these fixed molecular differences that distinguish the genus *Rueckbeilia* from all other genera of the subtribe Polyommatina, there are numerous positions that differentiate the genus *Rueckbeilia* from particular genera (Table S2, Appendix S1). Although these characters are not genus-specific, they constitute unique combinations that can be used for diagnostics.

Distribution

Uzbekistan, Tajikistan, Kyrgyzstan, Kazakhstan, and Northwest China. Records for East Iran and Turkmenistan require verification.

Etymology

The name is given in honour of Eugen Rückbeil and his two sons, Georg and Wasily (second half of the 19th century–beginning of the 20th century, exact years unknown), famous Russian collectors of German origin who explored the butterfly fauna of Central Asia and West China.

Note. In addition to *R. fergana*, we provisionally include in the genus *Rueckbeilia* the phenotypically similar (but genetically still unstudied) taxon *Rueckbeilia rosei* (Eckweiler, 1989) **comb. nov.**, a species known from East Turkey and Iran and traditionally considered within the genus *Vacciniina*.

SUPPLEMENTARY DOCUMENTATION

Establishing criteria for higher level classification using molecular data: the systematics of *Polyommatus* blue butterflies (Lepidoptera, Lycaenidae)

Gerard Talavera[a,b], Vladimir A. Lukhtanov[c,d], Naomi E. Pierce[e] and Roger Vila[a]

[a]*Institut de Biologia Evolutiva (CSIC-UPF), Passeig Marítim de la Barceloneta, 37, 08003 Barcelona, Spain*

[b]*Departament de Genètica i Microbiologia, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain*

[c]*Department of Karyosystematics, Zoological Institute of Russian Academy of Science, Universitetskaya nab. 1, 199034 St Petersburg, Russia*

[d]*Department of Entomology, St Petersburg State University, Universitetskaya nab. 7/9, 199034 St Petersburg, Russia*

[e]*Department of Organismic and Evolutionary Biology and Museum of Comparative Zoology, Harvard University, 26 Oxford Street, Cambridge, Massachusetts 02138, USA*

Supplementary Table S1. Primer sequences. mt: mitochondrial, n: nuclear. T = thymine, A = adenine, G = guanine, C = cytosine, K = G+T, W = A+T, M = A+C, Y = C+T, R = A+G, S = G+C, V = G+A+C, I = Inosine, N = A+C+G+T.

Primer location	**Primer name**	**Direction**	**Sequence (5' to 3')**
mt *COI*	LCO1490[1]	forward	GGTCAACAAATCATAAAGATATTGG
mt *COI*	Ron[2,3]	forward	GGATCACCTGATATAGCATTCCC
mt *COI*	Nancy[3]	reverse	CCCGGTAAAATTAAAATATAAACTTC
mt *COI*	Tonya[3]	forward	GAAGTTTATATTTTAATTTTACCGGG
mt *COI*	Hobbes[3]	reverse	AAATGTTGNGGRAAAAATGTTA
mt *COI*	TN2126[4]	forward	TTGAYCCTGCAGGTGGWGGAG
mt *COII*	George[3,5]	forward	ATACCTCGACGTTATTCAGA
mt *COII*	Phyllis[3,5]	reverse	GTAATAGCIGGTAARATAGTTCA
mt *COII*	Strom[3,5]	forward	TAATTTGAACTATYTTACCIGC
mt *COII*	Eva[3,5]	reverse	GAGACCATTACTTGCTTTCAGTCATCT
mt *COII*	JL3146[4]	forward	GAGTTTCACCTTTAATAGAACA
mt *COII*	B-tLys[2]	reverse	GTTTAAGAGACCAGTACTTG
mt *COII*	JL2532[4]	forward	ACAGTAGGAGGATTAACAGGAG
n *CAD*	CAD787F[6]	forward	GGDGTNACNACNGCNTGYTTYGARCC
n *CAD*	CADFa[7]	forward	GDATGGTYGATGAAAATGTTAA
n *CAD*	CADRa[7]	reverse	CTCATRTCGTAATCYGTRCT
n *EF-1α*	ef135[8,9]	forward	CAAATGYGGTGGTATYGACAAACG
n *EF-1α*	ef684[8,9]	reverse	TCCTTRCGCTCCACSTGCCAYCC
n *EF-1α*	ef531[8,9]	forward	TACAGYGAGCSCCGTTTYGAGGA
n *EF-1α*	ef929[8,9]	reverse	GCCTCTTGGAGAGCTTCGTGGTG
n *EF-1α*	ef51.9[8,9]	forward	CARGACGTATACAAAATCGG
n *EF-1α*	efrcM4R[8,9]	reverse	ACAGCVACKGTYTGYCTCATRTC
n *H3*	H3F[10]	forward	ATGGCTCGTACCAAGCAGACVGC
n *H3*	H3R[10]	reverse	ATATCCTTRGGCATRATRGTGAC
n *ITS-2*	ITS-3[11]	forward	GCATCGATGAAGAACGCAGC
n *ITS-2*	ITS-4[11]	reverse	TCCTCCGCTTATTGATATGC
n *wg*	LepWg1[12]	forward	GARTGYAARTGYCAYGGYATGTCTGG
n *wg*	LepWg2E[7]	reverse	ACNACGAACATGGTCTGCGT
n *wg*	Wg1n[13]	forward	CGGAGATGCGMCAGGARTGC
n *wg*	Wg2n[13]	reverse	CTTTTTCCGTSCGACACAGYTTGC
n *28S*	S3660[14]	forward	GAGAGTTMAASAGTACGTGAAAC
n *28S*	A335[14]	reverse	TCGGARGGAACCAGCTACTA

[1] Folmer, O., Black, M., Hoeh, W., Lutz, R., & Vrijenhoek, R.C. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol. Marine Biol. Biotech.* 3, 294-299.

[2] Simon, C., Frati, F., Beckebach, A., Crespi, B., Liu, H. & Flook, P. 1994. Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. *Annals of the Entomological Society of America* 87(6), 651-701.

[3] Monteiro, A. & Pierce, N.E. 2001. Phylogeny of *Bicyclus* (Lepidoptera: Nymphalidae) inferred from COI, COII, and EF-1alpha gene sequences. *Molecular Phylogenetics and Evolution* 18, 264-281.

[4] Canfield M.R., Greene E., Moreau C.S., Chen N., & Pierce N.E. 2008. Exploring phenotypic plasticity and biogeography in emerald moths: A phylogeny of the genus *Nemoria* (Lepidoptera: Geometridae). *Molecular Phylogenetics and Evolution* 49(2), 477-87.

[5] Brower, A.V.Z. 1994. Phylogeny of *Heliconius* butterflies inferred from mitochondrial DNA sequences (Lepidoptera: Nymphalidae). *Molecular Phylogenetics and Evolution* 3(2), 159-174.

[6] Moulton, J.K. & Wiegmann, B.M. 2004. Evolution and phylogenetic utility of cad (rudimentary) among Mesozoic-aged eremoneuran Diptera (Insecta). *Molecular Phylogenetics and Evolution* 31, 363-378.

[7] Vila, R., Bell, C.D., Macniven, R., Goldman-Huertas, B., Ree, R.H., Marshall, C.R., Bálint, Z., Johnson, K., Benyamini, D., & Pierce, N.E. 2011. Phylogeny and palaeoecology of *Polyommatus* blue butterflies show Beringia was a climate-regulated gateway to the New World. Proceedings of the Royal Society B 278(1719), 2737-2744.

[8] Cho, S., Mitchell, A., Regier, J.C., Mitter, C., Poole, R.W., Friedlander, T.P., & Zhao, S. 1995. A highly conserved nuclear gene for low-level phylogenetics: elongation factor-1alpha recovers morphology-based tree for heliothine moths. *Mol. Biol. Evol.* 12 (4), 650-656.

[9] Kandul, N.P., Lukhtanov, V.A., Dantchenko, A.V., Coleman, J.W.S., Sekercioglu, C.H., Haig, D. & Pierce, N.E. 2004. Phylogeny of *Agrodiaetus* Hübner 1822 (Lepidoptera: Lycaenidae) inferred from mtDNA sequences of *COI* and *COII*, and nuclear sequences of *EF1-α*: karyotype diversification and species radiation. *Systematic Biology* 53 (2), 278-298.

[10] Colgan, D.J., McLauchlan, A., Wilson, G.D.F., Livingston, S.P., Edgecombe, G.D., Macaranas, J., Cassis G., & Gray, M.R. 1998. Histone H3 and U2 snRNA DNA sequences and arthropod molecular evolution. *Australian Journal of Zoology* 46, 419-437.

[11] White, T.J., Bruns, S., Lee, S., & Taylor, J. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics in *PCR protocols: a guide to methods and applications*, edited by M.A. Innis, Gelfandm D.H., J.J. Snisky, & T. J. White. Academic Press, New York, pp. 315-322.

[12] Brower, A.V.Z. & DeSalle, R. 1998. Patterns of mitochondrial versus nuclear DNA sequence divergence among nymphalid butterflies: the utility of wingless as a source of characters for phylogenetic inference. *Insect Molecular Biology* 7(1), 73-82.

[13] Designed by Ada Kalizewska (Harvard University, Cambridge, MA, USA).

[14] Sequeira, A.S., Normark, B.B., & Farrell, B. 2000. Evolutionary assembly of the conifer fauna: Distinguishing ancient from recent associations in bark beetles. *Proceedings of the Royal Entomological Society (London) B* 267, 2359-2366.

Supplementary Table S2. GenBank accession codes. Sequences obtained in this work range from JX093196 to JX093497.

Specimen Code	Taxon	*COI + COII*	*EF-1a*	*wg*	*CAD*	*ITS2*	*28S*	*H3*
VL02X393	*Afarsia morgiana*	JX093487	JX093302		JX093267	JX093394	JX093228	JX093344
VL05Z994	*Agriades glandon*	GQ128942				GQ129038		GQ128729
VL01B424	*Agriades optilete*	GQ129011	GQ128699	GQ128910	GQ128630	GQ129110	GQ128521	GQ128803
JB05I879	*Agriades optilete*	GQ129012	GQ128700	JX093444	GQ128631	GQ129111	GQ128522	GQ128804
AD03B064	*Agriades orbitulus*	GQ128945	GQ128634	GQ128842	GQ128560	GQ129043	GQ128450	GQ128734
NK00P690	*Agriades pheretiades*	JX093466	JX093284	GQ128838	GQ128556	GQ129039	GQ128446	GQ128730
AS92Z130	*Agriades podarce*	GQ128943	GQ128632	GQ128839	GQ128557	GQ129040	GQ128447	GQ128731
AD00P259	*Agriades pyrenaicus araraticus*	GQ128944	GQ128633	GQ128840	GQ128558	GQ129041	GQ128448	GQ128732
VL02X098	*Alpherakya sarta*	JX093451	JX093311	JX093446	JX093268	JX093402	JX093224	JX093353
NK00P712	*Aricia agestis*	AY496801	AY496824	GQ128843	GQ128561	GQ129044	GQ128451	GQ128735
AD02W127	*Aricia artaxerxes*	JX093482	JX093308	JX093415	JX093255	JX093372	JX093202	JX093318
VL05Z997	*Aricia chinensis*	JX093476						JX093364
AD00P528	*Aricia crassipuncta*	JX093459	JX093304	JX093414		JX093376	JX093201	JX093317
AD03B041	*Aricia nicias*	GQ128992	GQ128681	GQ128892	GQ128612	GQ129092	GQ128502	GQ128785
VL03F745	*Aricia vandarvani*	JX093480	JX093306			JX093395		JX093347
DL99T242	*Chilades lajus*	GQ128946	GQ128635	GQ128844	GQ128562		GQ128452	GQ128736
AS92Z312	*Cupido comyntas*	GQ128954	GQ128643	GQ128852	GQ128571	GQ129053	GQ128461	GQ128745
AD00P540	*Cupido minimus*	GQ128947	GQ128636	GQ128845	GQ128563	GQ129045	GQ128453	GQ128737
AD00P369	*Cyaniris semiargus*	JX093483		JX093413	JX093260	JX093367	JX093217	JX093339
AD00P206	*Cyaniris semiargus*	JX093491	JX093301	JX093412	JX093259	JX093371	JX093235	JX093338
JE01C283	*Cyclargus ammon*	GQ128948	GQ128637	GQ128846	GQ128564	GQ129046	GQ128454	GQ128738
AS92Z185	*Echinargus isola*	DQ018914	DQ018914	DQ018885	GQ128566	GQ129048	GQ128456	GQ128740
RV05M735	*Eldoradina cyanea*	GQ128952	GQ128641	GQ128850	GQ128569	GQ129051	GQ128459	GQ128743
AD03B062	*Eumedonia eumedon*	GQ128953	GQ128642	GQ128851	GQ128570	GQ129052	GQ128460	GQ128744
NK00P743	*Eumedonia persephatta*	JX093492	JX093279	JX093438	JX093269	JX093398	JX093233	JX093354
RE02A007	*Freyeria putli*	GQ128956	GQ128645	GQ128854	GQ128573	GQ129055	GQ128463	GQ128747
VL01L462	*Freyeria trochylus*	GQ128955	GQ128644	GQ128853	GQ128572	GQ129054	GQ128462	GQ128746
VL02X159	*Glabroculus cyane*	JX093489	JX093283	JX093434	JX093275	JX093387	JX093221	JX093356
NK00P793	*Glabroculus elvira*	JX093456	JX093295	JX093435	JX093271	JX093399	JX093229	JX093355
MH01I001	*Hemiargus hanno*	GQ128960	GQ128649	GQ128858	GQ128577	GQ129059	GQ128467	GQ128751
SR03K069	*Hemiargus hanno bogotanus*	GQ128957	GQ128646	GQ128855	GQ128574	GQ129056	GQ128464	GQ128748
DL02P801	*Hemiargus hanno gyas*	GQ128959	GQ128648	GQ128857	GQ128576	GQ129058	GQ128466	GQ128750
AS92Z255	*Hemiargus hanno gyas*	GQ128958	GQ128647	GQ128856	GQ128575	GQ129057	GQ128465	GQ128749
RE01H234	*Hemiargus huntingtoni*	GQ128949	GQ128638	GQ128847	GQ128565	GQ129047	GQ128455	GQ128739
RV04I212	*Hemiargus martha*	GQ128950	GQ128639	GQ128848	GQ128567	GQ129049	GQ128457	GQ128741
MFB00N223	*Hemiargus ramon*	GQ128961	GQ128650	GQ128859	GQ128578	GQ129060	GQ128468	GQ128752
AS92Z184	*Icaricia acmon*	GQ128962	GQ128651	GQ128860	GQ128579	GQ129061	GQ128469	GQ128753

AS92Z065	*Icaricia icarioides*	GQ128963	GQ128652	GQ128861	GQ128580	GQ129062	GQ128470	GQ128754
AS92Z069	*Icaricia saepiolus*	GQ128966	GQ128655		GQ128583	GQ129065	GQ128473	
AS92Z465	*Icaricia shasta*	GQ128967	GQ128656	GQ128864	GQ128584	GQ129066	GQ128474	GQ128757
RV04I403	*Itylos huascarana*	GQ128978	GQ128667	GQ128876	GQ128596	GQ129078	GQ128486	GQ128769
RV03V327	*Itylos koa*	GQ128979	GQ128668	GQ128877	GQ128597	GQ129079	GQ128487	GQ128770
MFB00N166	*Itylos mashenka*	GQ128969	GQ128658	GQ128866	GQ128586	GQ129068	GQ128476	GQ128759
MFB00N220	*Itylos sigal*	GQ128983	GQ128672	GQ128881	GQ128601	GQ129083	GQ128491	GQ128774
RV03V182	*Itylos tintarrona*	GQ128984	GQ128673	GQ128882	GQ128602	GQ129084	GQ128492	GQ128775
MFB00N206	*Itylos titicaca*	GQ128970	GQ128659	GQ128867	GQ128587	GQ129069	GQ128477	GQ128760
VL01L319	*Kretania alcedo*	JX093488	JX093281	JX093439	JX093273	JX093383	JX093196	JX093359
VL01L152	*Kretania eurypilus*	JX093457	JX093298	JX093433	JX093265	JX093389	JX093218	JX093345
SH02H006	*Kretania eurypilus zamotajlovi*	JX093458	JX093297	JX093432	JX093264	JX093384	JX093198	JX093343
AD00P066	*Kretania pylaon*	GQ128990	GQ128679	GQ128888	GQ128608	GQ129089	GQ128498	GQ128781
AD00P121	*Kretania zephyrinus*	JX093464	JX093285	JX093441	JX093261	JX093378		JX093346
RV03V095	*Leptotes trigemmatus*	JX093474	GQ128660	GQ128868	GQ128588	GQ129070	GQ128478	GQ128761
CJM07J018	*Luthrodes cleotas*	JX093484			JX093276	JX093365		JX093360
HU08D004	*Luthrodes galba*	JX093471	JX093305			JX093406		
MWT93A009	*Luthrodes pandava*	GQ128951	GQ128640	GQ128849	GQ128568	GQ129050	GQ128458	GQ128742
AD00P129	*Lysandra bellargus*	JX093472	JX093299	JX093410	JX093262	JX093380	JX093225	JX093340
AD00P192	*Lysandra coridon*	JX093495	JX093300			JX093377	JX093227	JX093342
NK02A027	*Lysandra punctifera*	JX093494	JX093282	JX093411	JX093263	JX093391	JX093226	JX093341
09X164	*Maurus vogelii*	JX093485	JX093316		JX093278	JX093382	JX093230	JX093361
RV03V234	*Nabokovia cuzquenha*	GQ128985	GQ128674	GQ128883	GQ128603	GQ129085	GQ128493	GQ128776
MFB00N217	*Nabokovia faga*	GQ128986	GQ128675	GQ128884	GQ128604	GQ129086	GQ128494	GQ128777
AD00P092	*Neolysandra coelestina*	JX093490	JX093303	JX093417	JX093258	JX093379	JX093223	JX093337
AD00P081	*Neolysandra diana*	JX093479	JX093309	JX093416	JX093256	JX093370	JX093203	JX093363
VL05Z998	*Pamiria chrysopis*	JX093469	JX093312	JX093447		JX093388		JX093348
RV03V188	*Paralycaeides inconspicua*	GQ128987	GQ128676	GQ128885	GQ128605	GQ129087	GQ128495	GQ128778
RV03V198	*Paralycaeides vapa*	GQ128988	GQ128677	GQ128886	GQ128606		GQ128496	GQ128779
VL05Z995	*Patricius lucifer*	JX093475		JX093443	JX093266	JX093386		JX093357
AD00P266	*Plebejidea loewii*	GQ128989	GQ128678	GQ128887	GQ128607	GQ129088	GQ128497	GQ128780
AS92Z072	*Plebejus anna*	GQ128972	GQ128661	GQ128869	GQ128589	GQ129071	GQ128479	GQ128762
NK00P135	*Plebejus argus*	JX093468	AY496828	GQ128889	GQ128609	JX093400	GQ128499	GQ128782
AD00P560	*Plebejus argyrognomon*	JX093467	AY496827	GQ128870	GQ128590	GQ129072	GQ128480	GQ128763
NK00P165	*Plebejus idas*	GQ128974	GQ128663	GQ128872	GQ128592	GQ129074	GQ128482	GQ128765
NGK02C411	*Plebejus idas arapraetextus*	GQ128973	GQ128662	GQ128871	GQ128591	GQ129073	GQ128481	GQ128764
AS92Z005	*Plebejus melissa*	GQ128975	GQ128664	GQ128873	GQ128593	GQ129075	GQ128483	GQ128766
CCN05I856	*Plebulina emigdionis*	GQ128991	GQ128680	GQ128890	GQ128610	GQ129090	GQ128500	GQ128783
AD00P053	*Polyommatus amandus*	JX093452	JX093315	JX093418	JX093277	JX093373	JX093204	JX093321
NK00P596	*Polyommatus amandus*	JX093481	JX093280	JX093431	JX093247	JX093404		JX093323
MAT99Q840	*Polyommatus amandus*	JX093453	JX093293	JX093449	JX093246	JX093408	JX093205	JX093325
AD02W109	*Polyommatus amandus*	JX093454	JX093310	JX093419	JX093245	JX093366	JX093215	JX093322
VL01L135	*Polyommatus cornelia*	JX093486	JX093294	JX093450	JX093244	JX093390	JX093208	JX093327
NK00P103	*Polyommatus damocles krymaeus*	AY496727	AY496839	JX093426	JX093240	HM210178	JX093232	JX093331
MAT99Q841	*Polyommatus damon*	AY496732	AY496835	GQ128841	GQ128559	GQ129042	GQ128449	GQ128733

NK00P108	*Polyommatus daphnis*	JX093461	JX093290	JX093424	JX093252	JX093403	JX093212	JX093329
AD00P312	*Polyommatus dorylas*	AY496813	AY496831	JX093420	JX093248	JX093375	JX093219	JX093336
AD03B040	*Polyommatus erotides*	JX093470	JX093307	JX093428	JX093253	JX093374	JX093209	JX093351
AD02L274	*Polyommatus erschoffii*	AY496743	JX093287		JX093236	HM210182	JX093206	JX093332
MAT99Q838	*Polyommatus escheri*	JX093465	JX093296	JX093425	JX093241	JX093407	JX093197	JX093350
AD02M278	*Polyommatus glaucias*	JX093478	JX093288	JX093442	JX093239	JX093381	JX093207	JX093330
VL05Z996	*Polyommatus hunza*	JX093463			JX093257	JX093385		JX093335
NK00P562	*Polyommatus icarus*	JX093497	AY496846	GQ128891	GQ128611	GQ129091	GQ128501	GQ128784
AD02W258	*Polyommatus marcida*	JX093460	JX093289	JX093422	JX093272	JX093368	JX093211	JX093328
AD00P389	*Polyommatus myrrha cinyraea*	JX093473	AY496844	JX093430	JX093243	JX093369	JX093234	JX093326
MAT99Q904	*Polyommatus nivescens*	JX093496	JX093286	JX093421	JX093249	JX093409	JX093222	JX093333
NK00P859	*Polyommatus ripartii*	AY496779	AY496834	JX093429	JX093237	HM210195	JX093214	JX093362
VL02X324	*Polyommatus stempfferi*	AY954000	JX093314	JX093427	JX093242	JX093393	JX093216	JX093324
AD00P006	*Polyommatus surakovi surakovi*	AY496792	AY496837	JX093440	JX093238	HM210199	JX093210	JX093349
AD00P019	*Polyommatus thersites*	JX093455	JX093291	JX093436	JX093250	JX093392	JX093199	JX093319
MAT99Q947	*Polyommatus thersites*	AY496818	JX093292	JX093437	JX093251	JX093401	JX093200	JX093320
NK00P810	*Polyommatus venus*	JX093462	JX093313	JX093423	JX093254	JX093405	JX093220	JX093334
MC04Z114	*Pseudochrysops bornoi*	GQ128994	GQ128683	GQ128894	GQ128614	GQ129094	GQ128504	GQ128787
RV03V020	*Pseudolucia asafi*	GQ128997	GQ128686	GQ128897	GQ128617	GQ129097	GQ128507	GQ128790
BD02B813	*Pseudolucia charlotte*	GQ128998	GQ128687	GQ128898	GQ128618	GQ129098	GQ128508	GQ128791
MFB00N227	*Pseudolucia chilensis*	GQ128999	GQ128688	GQ128899	GQ128619		GQ128509	GQ128792
RV03V112	*Pseudolucia sibylla*	GQ129006	GQ128694	GQ128905	GQ128625	GQ129105	GQ128516	GQ128798
BD02B812	*Pseudolucia vera*	GQ129008	GQ128696	GQ128907	GQ128627	GQ129107	GQ128518	GQ128800
NK00P575	*Rimisia miris*	JX093493	AY496848	JX093445	JX093270	JX093396	JX093213	JX093352
NK00P777	*Rueckbeilia fergana*	JX093477	AY496850	JX093448	JX093274	JX093397	JX093231	JX093358
JXM99T709	*Talicada nyseus*	GQ129009	GQ128697	GQ128908	GQ128628	GQ129108	GQ128519	GQ128801
NK00P594	*Tongeia fischeri*	GQ129010	GQ128698	GQ128909	GQ128629	GQ129109	GQ128520	GQ128802

Supplementary Table S3. Diagnostic molecular characters (nucleotides) for the genus *Rueckbeilia*. Positions with fixed differences between *Rueckbeilia* (highlighted in yellow) and other genera of the subtribe Polyommatina. Base positions correspond to those in the molecular matrix provided as a Supplementary File.

	COI+tRNA-Leu+COII																									**Wg**		**EF1-alpha**													**CAD**			**28S**		**H3**	**ITS2**	
	68	145	320	451	628	748	940	1000	1009	1022	1030	1087	1111	1156	1291	1294	1551	1626	1635	1671	1801	1906	1916	2033	2139	217	222	58	295	304	370	400	412	450	790	823	865	955	1051	1123	136	184	413	284	586	79	12	1026
Polyommatus	A/T	T/C	A/G	A	A/G	A/G	A	A	A/G	A	T/C	T/C	A	A/C/T	A/T	A/T	T	T/C	A	A/G	A	T	A	A/T	A	G	A	C	G	G	C	C	G	C/T	G	C	C	C	A	C	C/T	A	A	A	G	A	T	T
Neolysandra	A	T	A	A	T/C	A	A	A	A	A	T	T	A	A	A	A	T	T	A	A	A	T/C	A	A	A	G	A	C	G	G	C	C	G	C	G	C	C	C	A	C	C	A	A	A	G	A	T	T
Lysandra	A	T	A	A	A	A	A	A	A	A/G	T	T	A	A	A	A	T	T	A	A	A	T	A	A	A	G	A	C/T	G	G	C	C	G	C	G	C	C	C	A	C	C	A	A	A	G	A	T	T
Aricia	A	T	A	A	A	A	A	A	A	A	T	T/C	A	A	A	A/T	T	T	A	A	A	T	A	A	A	G	A	C	G	G	C	C	G/T	C/T	G	C	C	C/T	A	C/T	C	A	A	A	G	A	T	T
Glabroculus	A	T	A	A	A	A	A	A	A/G	A	T	T	A	A	A	A	T	T	A	A	A	T	A	A	A	G	A	C	G	G	C	C	G	C	G	C	C	C	A	C	C	A	A	A	G	A	T	T
Alpherakya	A	T	A	A	A	A	A	A	A	A	T	T	A	T	A	A	T	T	A	A	A	T	A	A	A	G	A	C	G	G	C	C	G	C	G	C	C	C	A	C	C	A	A	A	G	A	T	T
Agriades	A	T	A	A	A/T	A	A/G	A	A	A	T/C	T	A	A	A	A	T	T/C	A	A	A	T	A	A	A	G	A	C	G	G	C	C	G	C	G	C	C	C	A/T	C/T	C	A	A	A	G	A	T	T
Rimisia	N	N	A	A	A	A	A	A	G	A	T	T	A	A	A	A	T	T	A	A	A	T	A	A	A	G	A	C	G	G	C	C	G	C	G	T	C	C	A	C	C	A	A	A	G	A	T	T
Cyaniris	A	T	A	C	A	A	A	A	G	A	C	T	A	A	A	A	T	C	A	A	A	T	A	A	A	G	A	C	G	G	C	C	G/A	C	G	C	C	C	A	C	C	A	A	A	G	A	T	T
Eumedonia	A	T	A	A/T	A	A	A	A	A	A	T	T	A	A	A	A	T	T	A	A	A	T	A/G	A	A	G	A	C	G	G	C	C	G	C	G	C	C	C	A	C	C	A	A	A	G	A	T	T
Plebejidea	A	T	A	A	A	A	A	A	A	A	T	T	A	A	T	A	T	T	A	A	A	T	A	A	A	G	A	C	G	G	C	C	G	C	G	C	C	C	A	C	C	A	A	A	G	A	T	T
Maurus	A	T	A	C	A	A	A	A	A	A	T	T	A	A	A	A	T	T	A	A	A	T	A	A	A	N	N	N	N	N	N	N	N	N	G	C	C	C	A	T	N	A	A	A	G	A	T	T
Kretania	A	C	A	A	A	A	A	A	A	A	T	T/C	A	A	A	A	T	T	A	A	A	T	A	A	A	G	A	C	G	G	C	C	G	C/T	G	C	C	C	A	C	C	A	A	A	G	A	T	T
Afarsia	A	T	A	A	A	A	A	A	A	A	T	T	A	A	T	A	T	T	A	A	A	T	A	A	A	N	N	C	G	G	C	C	G	C	G	C	C	C	N	C	C	A	A	A	T	A	T	T
Plebejus	A	T	A	A/G	A	A	A	A	A	A	T/C	T	A	A/G	A	A	T/C	T/C	A	A	A	T	A	A	A	G	A	C	G	G	C	C	G	C	G	C	C	C	A	C	C	A	A	A	G	A	T	T
Pamiria	A	T	A	A	A	A	A	A	A	A	T	T	A	A	A	A	T	T	A	A	A	T	A	T	A	G	A	C	G	G	C	C	G	C	G	C	T	C	A	C	N	N	N	N	N	N	T	T

Patricius	A	T	A	A	A	A	A	A	A	A	T	T	A	A	A	A	T	T	A	A	A	T	A	A	A	G	A	N	N	N	N	N	N	N	N	N	N	N	N	N	C	A	A	N	N	N	T	T
Rueckbeilia	T	C	T	T	T	T	T	G	G	G	C	C	T	T	T	T	A	C	G	G	G	C	G	T	T	A	G	T	A	T	A	T	C	T	A	T	T	T	G	T	T	G	G	G	T	T	C	C
Icaricia	A	T	A	A/T/C	A	A/G	A/T	A	A	A	T	T	A	A	A	A	T/C	T	A/G	A	A	T	A	A	A	G	A	C	G	G	C	C	G	C	G	C	C	C	A	C	C	A	A	A	G	A	T	T
Plebulina	A	T	A	A	A	G	A	A	A	A	T	T	A	A	A	A	T	T	A	A	A	T	A	A	A	G	A	C	G	G	C	C	G	C	G	C	C	C	A	C	C	A	A	A	G	A	T	T
Freyeria	A	T	A	A	A	A	A	C/T	A	A	T	T	A	A/T	A	A/T	T	T	A	A	A	T	A	A	A	G	A	C	G	G	C	C	C	C	G	C	C	C/T	A	C	C/T	A	A	A	G	A/T	T	T
Luthrodes	A	T	A	A	A	A	A/T	A/T	A	A	A	T/C	A	A	A/C/T	A	A	T	A	A	A	T	A	A	A	G	A	C	G	G	C	C	G	C	G	C	C	C	A	C	C	A	A	A	G	A	T	T
Chilades	A	T	A	A	A	A	A	T	A	A	T	T	T	T	A	A	T	T	A	A	A	T	A	A	A	G	A	C	G	G	C	C	G	C	G	C	C	C	A	C	C	A	A	A	G	A	T	T
Pseudolucia	A	T	A	A	A	A/G	A	A	A	A	T	T	A/T	A	A	A	A	A	A	A	A	T	A	A	A	G	A	C	G	G	C/A	C	G	C	G	C	C	C/T	A/T/G	C/T	C	A/G	A	A	G	A/G	T	T
Nabokovia	A	T	A	A	A	A	A	A	A	A	T	T	A	A	A	A	T	T	A	A	A	T	A	A	A	G	A	C	G	G	C	C	G	C	G/A	C	C	C	A	C	T	A	A	A	G	A	T	T
Eldoradina	A	T	A	A	A	A	A	A	A	A	T	T	A	A	A	A	T	T	A	A	A	T	A	A	A	G	A	C	G	G	C	C	G	C	G	C	C	C	A	T	T	A	A	A	G	A	T	T
Itylos	A	T	A	A	A	A	A/G	A	A/G	A	T	T	A	A/T	A	A/T	T	T/C	A	A/G	A	T/C	A	A	A	G	A	C	G	G	C	C	G	C	G	C	C	C/T	A	C	C/T	A	A	A	G	A	T	T
Paralycaeides	A	T	A	A	A	G	A	A	A	A	T	T	A	A	A	A	T	T	A	A	A	T	A	A	A	G	A	C	G	G	C	C	G	C	G	C	C	C	A	C	C	A	A	A	G	A	T	T
Hemiargus	N	N	A	A	T	A	A	C/T	A	A/G	T	T/C	A	A	T	A/T	T	T	A	A	A	T	A	A	A	G	A	C	G	G	C	C	G	C	G	C	C	C	A	C/T	C	A	A	A	G	A	T	T
Echinargus	A	T	A	A/T	A	A	A	T	T	T	T	T	A	A	A	A	T/A	T	A	A	A	T	A	A	A	G	A	C	G	G	C	C	G	C	G/A	C/T	C	C	A	C	C	A	A	A	G	A	T	T
Cyclargus	A	T	A	T	A	G	A	A	A	A	T	C	A	A	A	A	T	T	A	A	A	T	A	A	A	G	A	C	G	G	C	C	G	C	G	C	C	C	A	C	C	A	A	A	G	A	T	T
Pseudochrysops	A	T	A	A	A	A	A	T	A	A	A	T	A	A	A	A	A	T	A	A	A	T	A	A	A	G	A	C	G	G	C	T	G	T	G	C	T	C	A	C	C	A	A	A	G	A	T	T
Cupido	A	T	A	A	A	A	A	A	A	A	T	T	A	T/C	A	T	T	T	A	A	A	T/C	A	A	A	G	A	C	G	G	G	G	G	C/T	G	C	C	C	A/C	C	C	G	A	A	A	C	T	T
Tongeia	A	T	A	A	A	A	A	A	T	A	T	T	T	A	A	T	T	T	A	A	A	T	A	A	A	G	G	C	G	G	G	G	G	C	A	C	C	C	T	C	C	G	A	A	G	A	T	T
Talicada	A	T	T	A	A	A	A	A	A	A	T	T	A	N	A	T	T	T	A	A	A	C	A	A	A	G	A	C	G	G	A	A	G	C	G	C	T	C	T	C	C	G	A	A	G	A	T	T
Leptotes	A	T	A	A	A	A	A	A	A	A	T	T	A	A	A	T	C	T	A	A	A	T	A	A	A	G	A	C	G	G	C	C	A	C	G	C	C	C	C	T	C	C	A	A	G	C	T	T

Chapter III

Talavera, G., Dincă, V., Vila, R. Factors affecting biodiversity assessments with the GMYC model: insights from a DNA-barcoding survey of butterfly fauna. *In prep.*

Factors affecting biodiversity assessments with the GMYC model: insights from a DNA-barcoding survey of butterfly fauna

Gerard Talavera[a,b], Vlad Dincă[a,c] and Roger Vila[a,*]

[a]Institut de Biologia Evolutiva (CSIC-UPF), Passeig Marítim de la Barceloneta, 37, 08003 Barcelona, Spain

[b]Departament de Genètica i Microbiologia, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain

[c]Department of Zoology, Stockholm University, 106 91, Stockholm, Sweden.

[*]Corresponding author: roger.vila@csic.es

ABSTRACT

DNA sequencing of living organisms is an increasingly common activity that is promoting a new boost of taxonomical work incorporating molecular data. The use of the generalized mixed Yule-coalescent (GMYC) model has become one of the most popular approaches for species delimitation in groups with uncertain taxonomy based on single-locus data. This method essentially classifies branches in a tree as intra- or interspecific by maximizing the overall likelihood. Despite the fact that GMYC is intensely used, its performance has not been thoroughly assessed with empirical data on a solid taxonomic framework. We here examine an array of factors affecting GMYC resolution and provide success rates using a taxonomically well resolved DNA barcoding dataset of butterflies representing the fauna from an entire country (Romania). The dataset is ambiguity free (634 bp for all sequences), comprehensive (176 species), intensely and homogeneously sampled (1303 samples across the country covering all main biotopes), and taxonomy was independently assessed based on both external and internal morphology, including linear and geometric morphometry when needed. We focused on three aspects potentially shaping the branching pattern of the tree and likely to be determinant in GMYC results: 1) tree reconstruction method, 2) taxon sampling coverage/taxon level, and 3) geographic sampling intensity/geographic scale. We show that approximately 80% of the species studied are recovered as genetic clusters by GMYC. A 10% of the unrecognized species are not monophyletic and these failures can be attributed to intrinsic properties of the data. Our results demonstrate that this method is remarkably stable under a wide array of circumstances, including high singleton presence (up to 95%), taxon level (as low as between three to five species), and presence of gaps in intraspecific sampling coverage (removal of all intermediate haplotypes). As a conclusion, we provide a list of recommendations to maximize efficiency in GMYC analyses that will be of interest to researchers planning to use this method.

INTRODUCTION

DNA techniques have provided hope to accelerate global biodiversity exploration by massively sequencing living organisms. Two alternative strategies are generally followed: massive single-locus sequencing projects allow the fast discovery of unknown diversity, while comprehensive genomic approaches enable a finer way to study already known organisms. Although it is generally agreed that multiple genetic markers allowing coalescence-based studies are optimal for studying species boundaries when combined to independent sources of evidence (morphological, behavioural, karyological, etc.) (Will and Rubinoff, 2004; Rubinoff and Holland, 2005; Will *et al*, 2005; Knowles and Carstens, 2007; Carstens and Knowles, 2007), the single-locus modus operandi still allows broader surveying and has flourished in recent years mainly by the hand of DNA barcoding (Hebert *et al*, 2003). Although DNA barcoding was not conceived as a tool to delimit species, it established the base for derived DNA-based methods for species delimitation (Davis and Nixon, 1992; Brower, 1999; Pons *et al*, 2006; Cummings *et al*, 2008; Masters *et al*, 2011). The use of the generalized mixed Yule-coalescent (GMYC) model (Pons *et al*, 2006) has become one of the most popular approaches for species delimitation based on phylogenetic data. This method does not require any previous information about species and has been specifically developed for single-locus data, making it particularly suitable for phylogenetic community ecology studies or to explore groups with uncertain taxonomy. The GMYC method classifies branches in a gene tree as intra- or interspecific by maximizing the likelihood of a GMYC evolution model. Branching events between species are modelled with a Yule model, i.e. assuming a constant speciation rate and no extinction (Nee *et al*, 1994; Barraclough and Nee, 2001) and branching events within species are modelled using a neutral coalescent process (Hudson, 1990). Both intra- and interspecific branching models are generalized to allow qualitative departures from the basic models. Thus, short terminal branches will correspond to genetic clusters (species) separated by generally longer internal branches.

While GMYC species delimitation has been broadly used in a variety of organisms, the performance of the method has not yet been accurately tested in a solid taxonomic framework. Most of the work done to evaluate GMYC performance has been based on simulations. Papadopoulou *et al* (2008) and Lohse (2009) contributed with interesting discussions regarding the influence of the sampling scheme, Esselstyn *et al* (2012) assessed the significance of different coalescent parameters on individual gene trees and Reid and Carstens (2012) tested the effect of tree depth (i.e. length of the interspecific branches as a factor of the effective population size), the importance of allele sampling within species, and the relevance of the DNA fragment length. Although some insights have been obtained from empirical studies (e.g. Pons *et al*, 2011; Esselstyn *et al*, 2012), the nature and magnitude of many potential biases have not been evaluated in real cases. Most importantly, most study systems used were not adequate for various reasons and the results were inconclusive. Ideally, a dataset to test species delimitation performance needs to be extensive (including a substantial number of species), homogeneous (sampling intensity and range should be as similar as possible across taxa), and taxonomy needs to be resolved and assessed independently (to avoid circular reasoning). Thus, we decided to specifically address this question using a highly suitable empirical dataset. This consisted of a complete and fairly homogeneous DNA barcode library for the butterfly fauna on an entire country (Romania) (Dincă *et al*, 2011a). This country includes about one-third of the European butterfly fauna (182 confirmed species) that is representative for temperate Europe and displays high eco-region richness. Since the European butterflies are arguably the best studied invertebrates in the world, the detailed morphological study of all 1387 samples (180 species) in Dincă *et al* (2011a), including external and internal, linear and geometric morphometry when necessary, makes the taxonomy highly precise for this particular dataset. We focused on three aspects potentially shaping the branching pattern of the tree and likely to be determinant in GMYC results: 1) tree reconstruction method, 2) taxon sampling coverage/taxon level, and 3) geographic sampling intensity/geographic scale.

First, tree reconstruction algorithms based on evolutionary models are subject to variability affecting relative branch lengths estimations. A further degree of variability can be obtained by applying different molecular clocks to get ultrametric trees. GMYC users typically choose less time-consuming programs like RaxML (Stamatakis, 2006) plus a subsequent branch length transformation with PATHd8 (Britton *et al*, 2006, 2007) or r8s (Sanderson, 2003). Software directly producing ultrametric trees (e.g. BEAST) avoids intermediate steps, but is very computationally demanding for large datasets, thus limiting applicability (Puillandre *et al*, 2012).

Second, variable rates of genetic changes among species might affect species delimitation, especially when the method is applied at wide taxon level. GMYC has been indistinctly used at several taxonomical levels, from an entire phylum (Barraclough *et al*, 2009), or a superfamily (Monaghan *et al*, 2009), to species complexes (Fontaneto *et al*, 2007a; Papadopoulou *et al*, 2009a; Gebiola *et al*, 2012). However, it is to be expect that pronounced differences in tree shape depending on the outgroup taxa included may result in a different behaviour of the GMYC model.

Lastly, the selection of sampling across territory has been discussed as a limitation for the GMYC approach (Lohse, 2009; Papadopoulou *et al*, 2009b; Pons *et al*, 2011; Reid and Carstens, 2012). While it has been reported, based on simulations, that a very complete representation of the population genetic structure within the study region for each taxon might be essential to avoid artifactual delimitations mainly in terms of oversplitting (Lohse, 2009), others argued that in real datasets this is not such a major concern (Papadopoulou *et al*, 2009b; Reid and Carstens, 2012). On the other side, increasing the geographic scale of the study could result in oversplitting as well because of a dilution of the sampling intensity and homogeneity across space (Bergsten *et al*, 2012).

All these factors may strongly affect the branching pattern of the trees, and due to the plasticity of candidate datasets (type of organisms, geographical range, sampling intensity, etc.) they can create strong imbalances between the information included in the portion of the tree under a Yule process and the portion under a coalescent process (Figure 1). Since all variables mentioned above can act in synergy and affect GMYC delimitations, a proper empirical dataset to test

GMYC is one where the three issues could be assessed together. A critical requirement is that a good previous taxonomic knowledge exists in order to contrast GMYC delimitations with detailed morphology-based determinations. The taxon level must be deep enough to display a substantial level of cladogenesis and the species richness high enough to obtain statistically valuable results. Similarly, geographical area needs to be wide enough to comprise ecosystem diversity. On the other hand, intraspecific taxon coverage needs to be sufficient in terms of populations studied and fairly homogeneous across species.

Based on the Romanian butterfly dataset, we study the GMYC accuracy and the possible reasons for failure under a range of different phylogenetic approaches. We also evaluate consequences from different factors that in theory could create imbalances between the Yule and coalescent portions of the trees (Figure 1), such as sampling limitations or taxonomical depth. With the evidence obtained, we provide new insights into the proper conditions for the efficient application of the GMYC method.

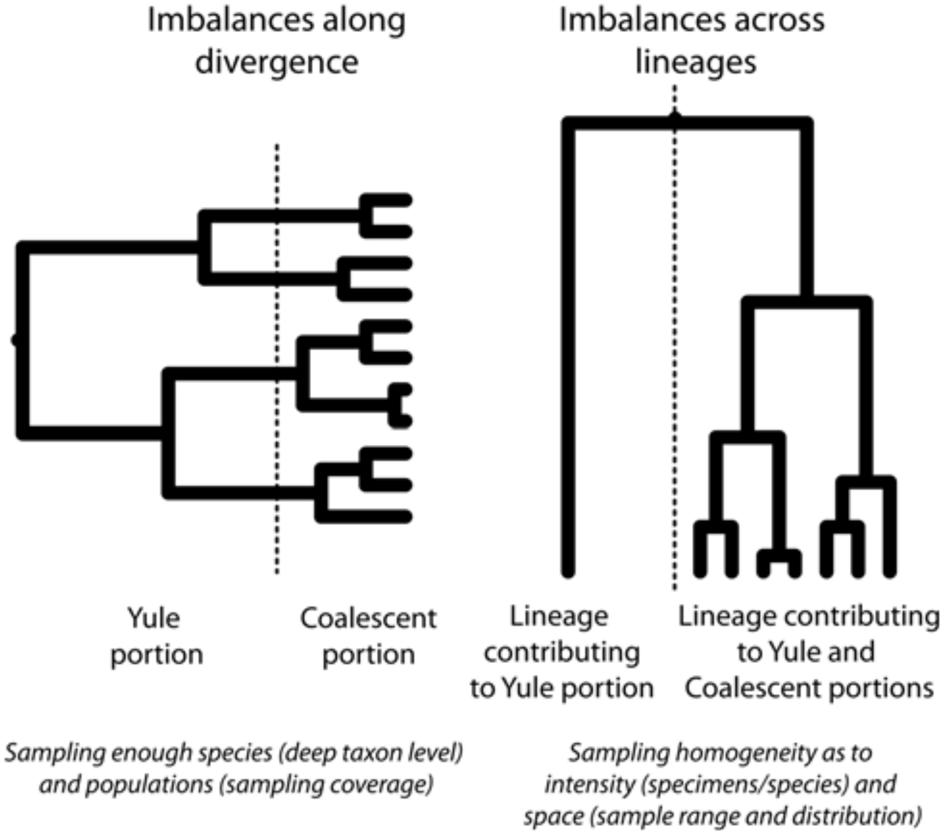

Imbalances along divergence
Imbalances across lineages
Yule portion
Coalescent portion
Lineage contributing to Yule portion
Lineage contributing to Yule and Coalescent portions
Sampling enough species (deep taxon level) and populations (sampling coverage)
Sampling homogeneity as to intensity (specimens/species) and space (sample range and distribution)

Figure 1. Two types of imbalances between the Yule and coalescent portions in the tree shape that may negatively affect GMYC performance. A) Both portions should be well represented in the tree (sufficient species included and sufficient intraspecific variability). B) Differences across lineages should be minimized (sufficient sampling intensity and as homogeneous as possible).

MATERIALS AND METHODS

GMYC performance was evaluated in three different ways: 1) effects caused by branch length optimization to obtain ultrametric trees, 2) by the Yule-coalescent branching proportionalities derived from the taxon level explored for a specific area in combination with 3) the intensity on the taxon coverage applied for specific taxa.

1) Dataset characteristics

Our analyses started with the 1387 DNA barcodes dataset of 180 species of Romanian butterflies from Dincă *et al* (2011a). Sequence length was filtered to avoid the presence of missing data resulting in an ambiguity free, 634 base pair alignment, of 1303 specimens representing 176 species. This dataset included 97% of the Romanian butterfly species sampled from numerous sites of the Romanian territory of ca. 237 thousand square kilometres. The sampling covered many different types of habitats and distant locations in order to improve the assessment of intraspecific variability. Overall, the average number of specimens per species was of 7.4. Only four species in the dataset (*Allancastria cerisyi, Polyommatus amandus, Nymphalis l-album* and *Limenitis reducta*) were represented by one specimen, while the maximum number of specimens reached 23 in the case of *Pyrgus armoricanus*. The 1303 samples were collected from 134 localities across Romania's territory, with an average of 5.2 sites per species (ranging between one and 20). Repeated haplotypes were removed using Collapse 1.2 (Posada, 2004) to produce a final matrix of 495 haplotypes representing 176 species.

2) Phylogenetic approaches used

Ultrametric phylogenetic trees were obtained following different strategies under a relaxed lognormal coalescent clock and a uniform clock. For a uniform clock, branch lengths were normalized using penalized likelihood (PL) with cross validation (CV) in r8s (Sanderson, 2003), and d8 and MPL (Britton *et al*, 2002) algorithms in PATHd8 (Britton *et al*, 2006, 2007). Maximum likelihood (ML) trees were generated with Garli 1.0 (Zwickl, 2006) except for the ones applying Patdh8, which were obtained with RAxML 7.0.3 (Stamatakis, 2006) in order to test the fastest combination of methods. Also, a strict clock was applied to bayesian inference in BEAST 1.5.3 (Drummond and Rambaut, 2007). For a relaxed clock, ChronoPL function in Ape library (Paradis *et al*, 2004) implemented in R with a lambda = 0.5 (which allows rates to vary among branches) and CV was applied to ML trees, and BEAST 1.5.3 was used to get trees under an uncorrelated lognormal relaxed clock (Drummond *et al*, 2006). Coalescent and Yule tree priors were also evaluated in the bayesian trees. As a result, a total of eight tree reconstruction procedures were followed for further GMYC evaluation. For BEAST inference, four gamma rate categories were selected and a randomly generated initial tree was used. Two independent chains were run for a variable number of generations (between 10 and 50 million) depending on the dataset. Values were sampled every 10% of the run length and convergence was inspected in Tracer v.1.5 (Rambaut and Drummond, 2007). The substitution model used for both maximum likelihood and bayesian approaches was GTR+I+G according to the AIC criterion, obtained from jModeltest (Posada, 2008) outputs.

Trees coming from penalized likelihood in r8s were not fully dichotomous and the R function “multi2di” in the Ape library dividing multichotomies in order of appearance in the tree (random=false) was used. For ML trees, some internal branches with 0 values were found, and in that case they were substituted by 0.000001 before normalizing branch lengths.

In order to evaluate the capability of GMYC to detect species boundaries from our COI barcoding dataset, the species delimitation tool (Pons *et al*, 2006; Fontaneto *et al*, 2007b; Monaghan *et al*, 2009) was conducted in the SPLITS package (Ezard and

Fujisawa, 2009) implemented in R statistical software. Single and multiple-threshold options were optimized for all the resulting trees. The function automatically outputs a likelihood ratio test (LRT) between the null and GMYC models, a number of ML clusters and entities with their respective intervals of confidence and the threshold time (for the single option) or multiple threshold times (for the multiple option) where jumps between the Yule and coalescent portions are found. LRTs between single and multiple threshold options were performed to select the model that better fits the data. Unless otherwise stated all the analyses were conducted on haplotype trees with no repeated haplotypes.

3) Subclade tests

The 495-haplotypes dataset was progressively partitioned in order to assess consequences of reducing the taxonomic level (i.e. the internal Yule-coalescent branching pattern proportionality). Thus, phylogenetic trees from subsets containing species grouped by five butterfly families were also evaluated for the four bayesian approaches previously examined, corresponding to 141 haplotypes for Lycaenidae + Riodinidae, 233 for Nymphalidae, 52 for Pieridae, 52 for Hesperiidae and 17 for Papilionidae. The resulting clustering and entity delimitations were compared with the full Rophalocera phylogenies. Trees resulting from a strict clock and a coalescent constant size prior were used for further analysis.

Next, all possible subclades were generated from each family tree and the significance of the GMYC versus the null model for each partition was evaluated. We obtained 65 possible tree partitions with GMYC fit significance and for them the program performance was evaluated under the single-threshold approach. Finally, we compared the estimated number of GMYC entities obtained for the full family tree to those obtained for each of the subclades.

4) Taxon coverage tests

Taxon coverage effects were studied following three approaches. First, a sampling reduction (or coalescent portion reduction) for each of the 18 species with intra-specific genetic distances > 0.5% that were correctly recovered as entities by

GMYC was done by selecting only the two most genetically distant specimens in the dataset and eliminating the rest. These individuals were often not the most geographically distant. Second, we forced an unbalanced sampling by enlarging the dataset with samples alien to the study area (Romania) independently for the species *Polyommatus icarus* and *Papilio machaon* (coalescent portion increase). A total of 99 extra samples of *P. icarus* and 18 of *P. machaon* where incorporated to the analyses (Supplementary Table S1). Most samples of *P. icarus* originated from the dataset used by Dincă *et al* (2011b) (and see references therein) to which we added four samples from Hausmann *et al* (2011). This dataset covered a very large geographic area, ranging from Iberia to Asian Russia and from Israel to Finland. Almost all samples of *P. machaon* represent original data, with the exception of two specimens taken from Hausmann *et al* (2011). Our dataset included 16 samples from Spain and two from Germany (Supplementary Table S1). Subsequently, the total number of sequences per species was randomly reduced to 50%, 20% and 10% (10 replicates each) on both species and new tree inferences were done with BEAST (coalescent prior + strict clock) for all the resulting datasets. Lastly, to simulate possible effects derived from low intraspecific richness, we progressively reduced species representations to singletons in 40%, 50%, 70%, 90%, 95% and 100% of the 176 species. The original dataset already had 48 (27%) species with a single haplotype and, as a control we also evaluated the dataset excluding every singleton (128 species). To avoid stochasticity, ten replicates were done per each percentage by randomly selecting the species to be reduced to singletons.

RESULTS

1) Phylogenetic trees performance

When analysing the entire dataset, the number of resulting GMYC entities was similar (from 181 to 188) for the different phylogenetic combinations explored (Figure 2A), which also displayed comparable values of confidence intervals (CI) (Supplementary Table S2). The number of morphospecies clades that exactly matched a GMYC entity (correctly identified species) ranged between 140 and 143 from a total of 176, and the percentages of identification success ranged between 79.54% and 81.25% (Figure 2A). The tree performed with RaxML and normalized with PATHd8 (with both algorithms d8 and MPL) reported the best match between GMYC entities and morphospecies. One exception was the PL with CV from a ML tree using the ChronoPL function in R, which resulted in unexpectedly low values for the obtained number of entities (148), the correctly identified morphospecies (98), and the percentage of success (55.68%).

When assessing the family trees performance, the four previously used bayesian models were also evaluated. Almost identical tendencies were detected between models using a coalescent or a Yule prior, and between a relaxed or strict molecular clock (Supplementary Table S3) (Figure 2B). As inferred in the full dataset, the percentage of GMYC success was close to 80% for all the families, with the exception of the Papilionidae (100% in this case).

GMYC performance by families was also evaluated using sequence trees instead of haplotype trees. Indeed, the choice of using a BEAST genealogy instead of a phylogenetic tree as GMYC input allows the use of the entire datasets without the requisite to discard repeated haplotypes. Such a test was done by using BEAST genealogy reconstruction with a coalescent prior and a strict molecular clock, and almost identical values to the previous approach were obtained (Figure 2C) (Supplementary Table S4).

In all the approaches, the GMYC single threshold fit the data better than the multiple threshold option (Supplementary Table S2-S4).

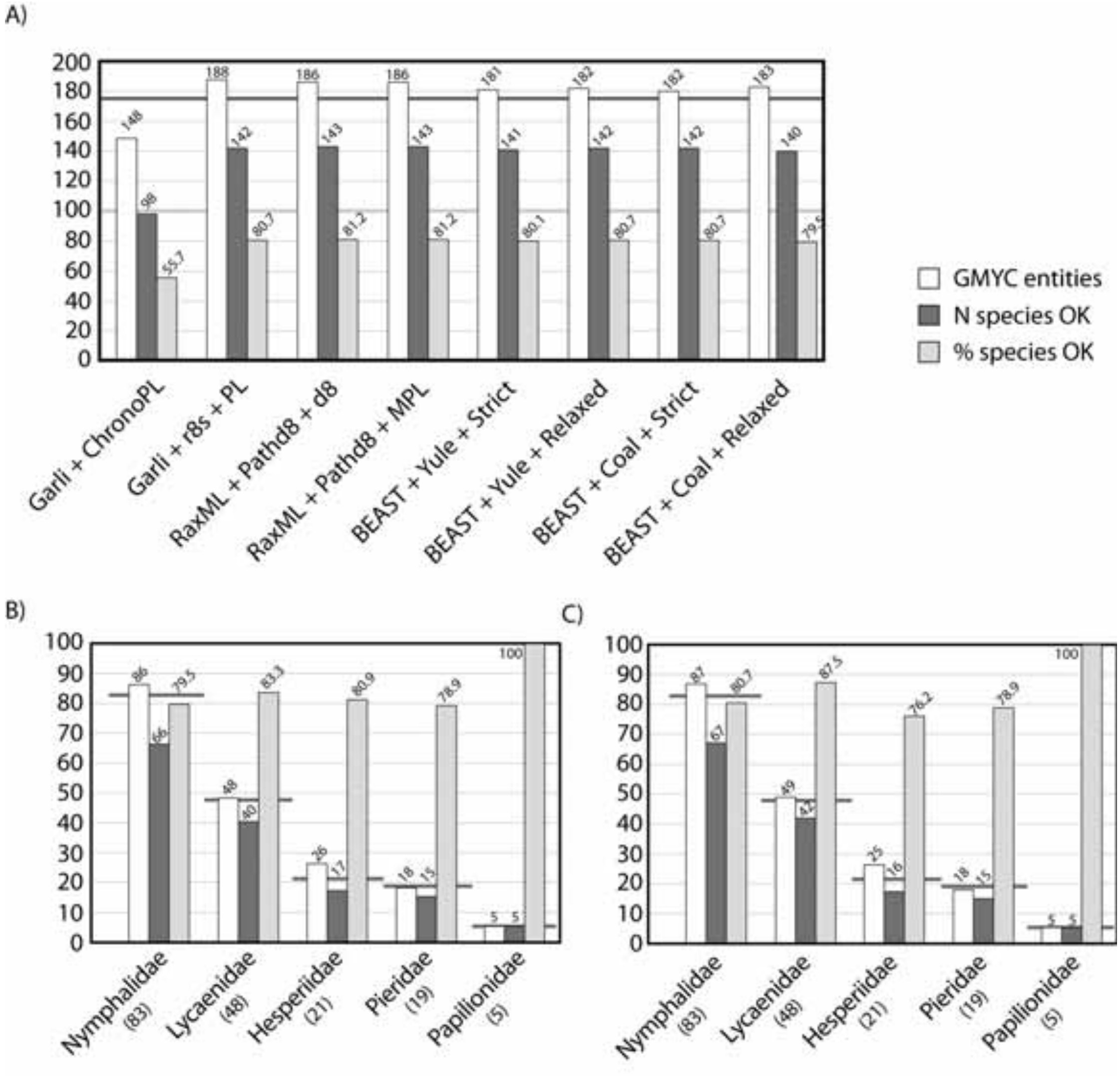


Figure 2. A) Number of GMYC entities, number and percentage of correct GMYC hits for the eight different phylogenetic procedures tested on a complete haplotype tree. Same parameters for the five butterfly families analysed independently using B) haplotype trees and C) trees with all the sequences (including identical ones), both performed with BEAST with a coalescent tree prior and a strict molecular clock. Horizontal grey lines represent the correct number of species.

2) Sensitivity to the Yule-coalescent branching proportionalities

Testing GMYC significance. The GMYC method needs to detect switches from Yule to coalescent accumulation of lineages in the data. As a first step, the method performs a likelihood ratio test (LRT) to evaluate the Yule-coalescent equilibrium compared to the null model for a given dataset, a requisite for a correct usage of the tool. Thus, a significant threshold detects a switch between the Yule and the coalescent process in the tree.

The combination of taxon level and taxon coverage (intensity of sampling for a specific taxon and/or area) will influence the Yule-coalescent balance in the tree. Thus, from a species complex to an entire superorder and/or from a regional park to an entire continent, GMYC can be used under a wide parameter space of possible combinations. If any of those requisites is inappropriate, the GMYC model will not fit the tree significantly better than the null model. We performed analyses progressively reducing Yule and coalescent portions of the tree, paying special attention to the limits of the significant level of the GMYC model. The reduction of the Yule portion was assessed by analysing all 298 possible subclades within family trees. A total of 65 clades (21.8%) produced significance for GMYC (likelihood ratio test *P*-value < 0.05). While results were slightly variable according to the taxa involved, a clear pattern of GMYC significance decrease with reduced number of species and depth of the subclades was observed (Figure 3, Supplementary Figure S1). No subclade with less than three species was significant, and none with less than six species was strongly significant (*P* < 0.001).

Regarding the reduction of the coalescent proportion by increasing the percentage of singletons (species represented by a single haplotype), GMYC was significant up to a 95% of singletons (Figure 5B). At this percentage LRT *P*-values abruptly increased and only six of the ten cases analysed achieved GMYC significance. Such a value suggests that a surprisingly low proportion of species with population structure are necessary for GMYC significance.

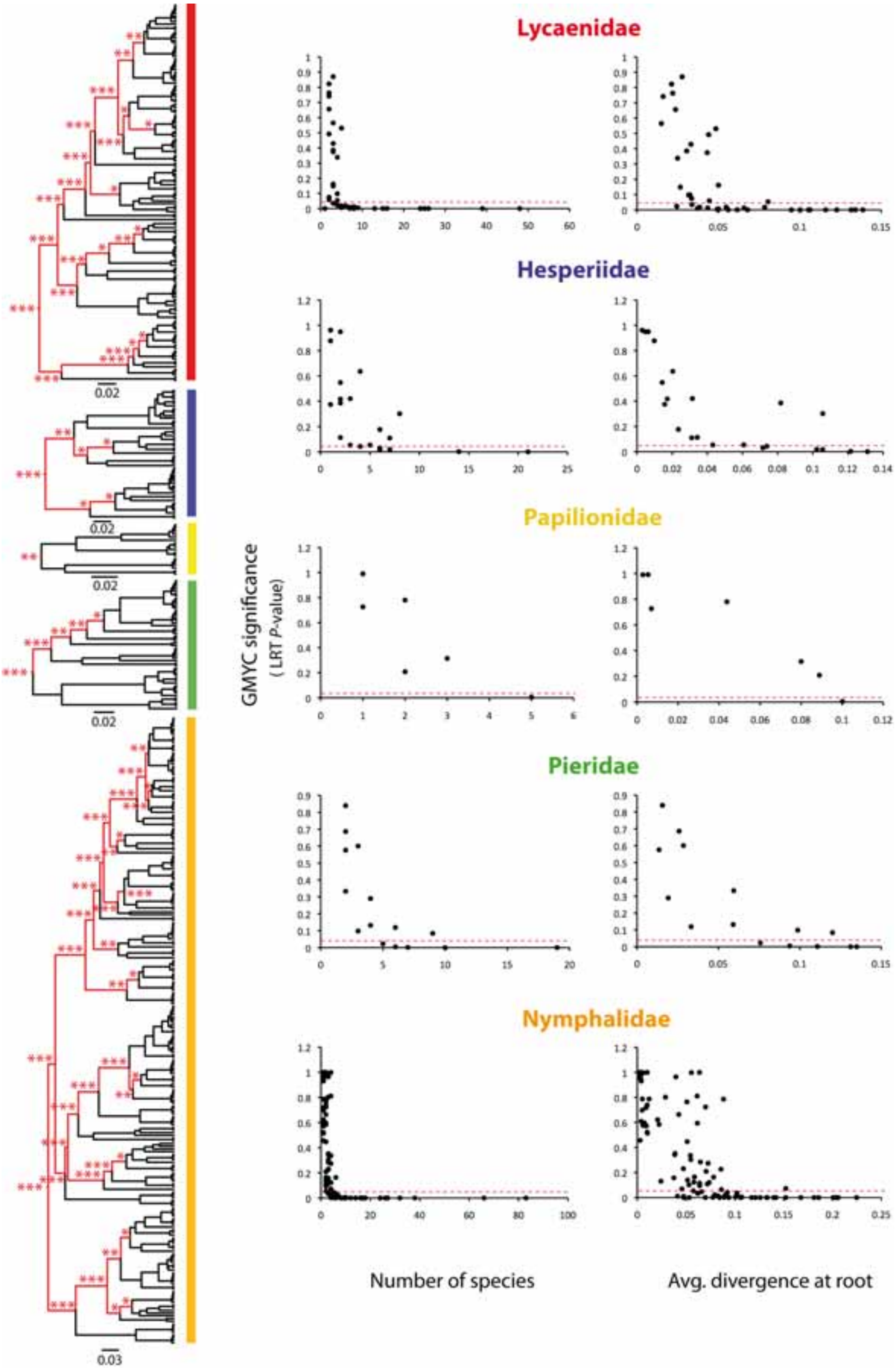

Lycaenidae
Hesperiidae
Papilionidae
Pieridae
Nymphalidae
GMYC significance
(LRT P-value)
Number of species
Avg. divergence at root
0.02
0.02
0.02
0.02
0.03
*
**
***
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.2
0
10
20
30
40
50
60
0.05
0.1
0.15
5
15
25
0.02
0.04
0.06
0.08
0.12
0.14
2
3
4
6
80
100
0.2
0.25

Figure 3. Importance of taxon level for GMYC significance. GMYC significance was tested for all possible subclades within family haplotype trees (left column). Nodes are labelled according to the level of GMYC significance (no symbol = $P > 0.05$, * = $P < 0.05$, ** = $P < 0.01$, *** = $P < 0.001$) obtained for the subclade that they define. The black part of the trees represents the subclades where differences between Yule and coalescent portions are not significant and are thus not suitable for being analysed independently under the GMYC model. Plots show GMYC significance (likelihood ratio test *P*-value) vs. the number of species included in the subclades (central column) and vs. the subclade tree depth (average divergence from root to tips) (right column).

GMYC performance at lower taxon levels. All the GMYC significant subclades in the family trees were identified (Figure 3) (Supplementary Figure S1) and GMYC performance assessed (Table 1). Of the total 65 significant subclades, only 12 (18.5%) presented discordances in the inferred number of GMYC entities when compared to the same clade in the entire tree. The *Charcharodus* clade (Hesperiidae) and the *Brenthis* + *Argynnis* clade (Nymphalidae) were those most affected by the reduction of Yule portion of the tree. The observed differences, though, generally involved one or very few taxa, they were not necessarily related to the level of GMYC significance, and no clear tendency of splitting or lumping was observed. Thus, we can state that, within the boundaries imposed by GMYC significance, taxon level did not have a capital influence on species recognition.

Reduction of intermediate intra-specific divergences. The removal of intermediate haplotypes and posterior tree reconstruction for the taxa with highest intra-specific divergence (*p*-distance > 0.5%) among those that were correctly recovered as single entities did not cause any changes to GMYC estimations (Table 2). This result is remarkable since these species were selected among those most close to the GMYC threshold and indicates that intermediate haplotypes for particular taxa within a deep tree do not influence GMYC performance.

Table 1. Comparison between the GMYC number of entities for the full family tree and each GMYC-significant subclade showing any difference. Clade numbers are shown in Supplementary Figure S1 and extended information for all subclades in Supplementary Table S6. Levels of significances are (*) significant, P < 0.05; (**) very significant, P < 0.01; (***) extremely significant, P < 0.001.

Family	Clade Number	GMYC significance	N. entities full tree	N. entities subclade tree	CI	Comment
Nymphalidae	39	**	6	5	3-6	*Lasiommata maera* oversplitted (two entities) in the full tree
Nymphalidae	40	***	36	37	36-38	*Coenonympha arcania* and *C. leander* recovered as conspecific in the full tree
Nymphalidae	46	**	4	5	4-5	*Coenonympha arcania* and *C. leander* recovered as conspecific in the full tree
Nymphalidae	58	***	10	9	8-10	*Melitaea aurelia* oversplitted (two entities) in the full tree
Nymphalidae	89	*	7	6	5-8	*Argynnis pandora* oversplitted (two entities) in the full tree
Nymphalidae	92	**	12	9	4-15	*Argynnis pandora* oversplitted (two entities) in the full tree *Brenthis daphne* oversplitted (two entities) in the full tree *Brenthis daphne* and *B. ino* recovered as conspecific in the subclade tree
Nymphalidae	99	***	19	21	15-22	*Argynnis aglaja* oversplitted (two entities) in the subclade tree *Brenthis ino* oversplitted (two entities) in the subclade tree
Nymphalidae	103	***	21	22	18-24	*Argynnis aglaja* oversplitted (two entities)in the subclade tree
Lycaenidae	10	*	2	3	3-3	*Plebejus argus* oversplitted (two entities)in the subclade tree
Hesperiidae	11	*	6	5	5-6	*Thymelicus sylvestris* oversplitted in two entities in the subclade tree and in three entities in the full tree
Hesperiidae	20	*	8	5	4-8	*Carcharodus alceae* oversplitted (two entities) in the full tree *Carcharodus floccifera* oversplitted (two entities) in the full tree *Carcharodus floccifera* and *C. orientalis* recovered as conspecific in the subclade tree
Hesperiidae	21	*	9	6	5-9	*Carcharodus alceae* oversplitted (two entities) in the full tree *Carcharodus floccifera* oversplitted (two entities) in the full tree *Carcharodus floccifera* and *C. orientalis* recovered as conspecific in the subclade tree

Table 2. Among the species that were recovered as an entity, those displaying the highest intra-specific divergence (p-distance > 0.5%) were used to test for the importance of intermediate haplotypes. Trees and GMYC were re-analysed retaining only the two most divergent haplotypes, thus simulating gaps in specimen sampling that could produce intraspecific splits. None of the taxa analysed showed splitting when only the most extreme haplotypes were considered.

Family	Species	*p*-distance	Splitting?
Lycaenidae	*Calloprhys rubi*	0.008	no
	Cyaniris semiargus	0.008	no
	Eumedonia eumedon	0.008	no
	Maculinea nausithous	0.008	no
	Polyommatus icarus	0.011	no
	Polyommatus thersites	0.011	no
Nymphalidae	*Melitaea cinxia*	0.009	no
	Aglais aglaja	0.009	no
	Coenonympha glycerion	0.005	no
	Brenthis hecate	0.005	no
	Erebia melas	0.008	no
	Melitaea trivia	0.008	no
	Brenthis ino	0.008	no
	Arethusana arethusa	0.008	no
Pieridae	*Pieris brassicae*	0.008	no
	Pieris rapae	0.008	no
Papilionidae	*Papilio machaon*	0.016	no
	Zerynthia polyxena	0.009	no

Sampling geographical range expansion. Intraspecific divergences (coalescent portion) were progressively increased by expanding the sampling area for the species *Polyommatus icarus* (Lycaenidae) and *Papilio machaon* (Papilionidae). Thus, an unbalanced situation of the sampling area was induced between those species and the rest of the taxa, which only included Romanian specimens. GMYC results were different in the two cases. For *P. icarus*, a maximum of five GMYC entities were recovered when a total of 110 specimens (44 haplotypes) from a very large geographical area were used (Supplementary Table S1). The number of GMYC entities progressively decreased when sequences were randomly removed in proportions of 50%, 20% and 10% (Figure 4). In all cases, higher numbers of GMYC entities were estimated in comparison to the original dataset that included only samples from Romania (when only a single GMYC entity was recovered). Such effect was not detected in the case of *P. machaon*, where a single entity was recovered in all the analyses. However, for this species (25 specimens added, 16 haplotypes) the area covered and the increase of maximum intraspecific divergence were smaller than in the case of *P. icarus*.

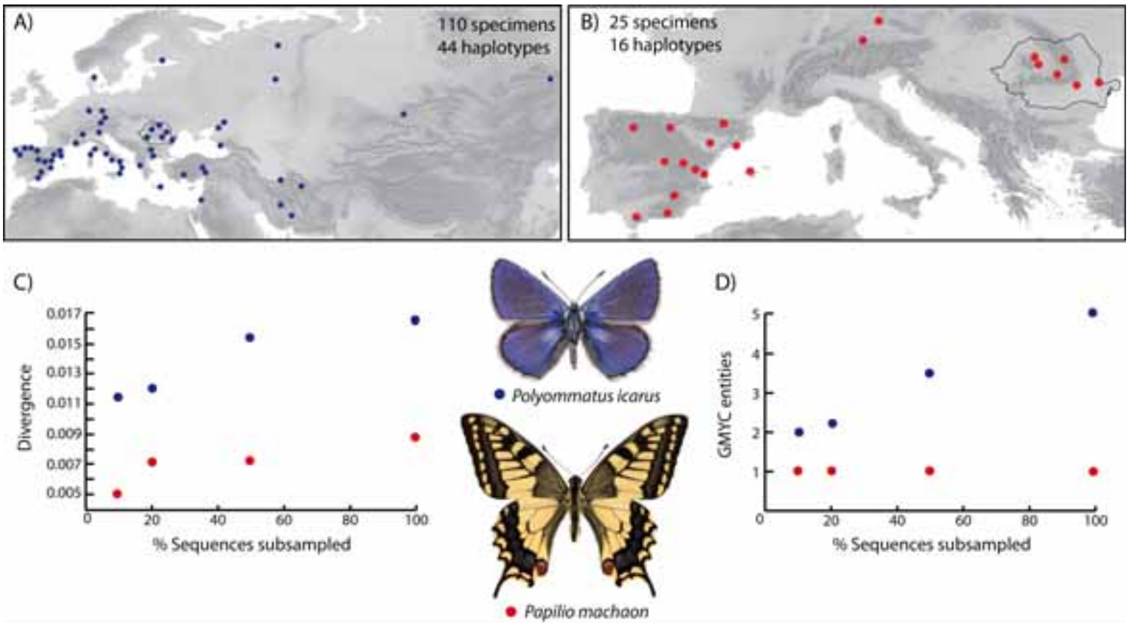


Figure 4. Sampling geographical range expansion for *Polyommatus icarus* (A) and *Papilio machaon* (B). These species were correctly recovered as single entities when only Romania was sampled. An unbalanced geographic sampling produced GMYC splits for *P. icarus*, that were progressively reduced when using different levels of random subsampling. A more limited expansion for *P. machaon* does not produce splitting. Plots show maximum intraspecific divergences (C) and number of GMYC entities (D) vs. percentage of sequences subsampled.

Increasing the percentage of species with unique haplotype. The original dataset contained 48 species represented by a unique haplotype (or singletons) from a total of 176 species (27% of singletons). An increase of the singletons percentage supposed a slightly progressive improvement of the percentage of GMYC success (percentage of species correctly recognised as entities) (Figure 5A). Note, however, that the challenge for the method was increasingly simpler as singletons reduced the possibilities of splitting. Indeed a higher success was accomplished by a reduction of the number of GMYC entities, many of which were previously erroneous splits. A tree without singletons (excluding the original 48 singletons) was also evaluated, resulting in a GMYC performance (83%) similar to that of the initial dataset.

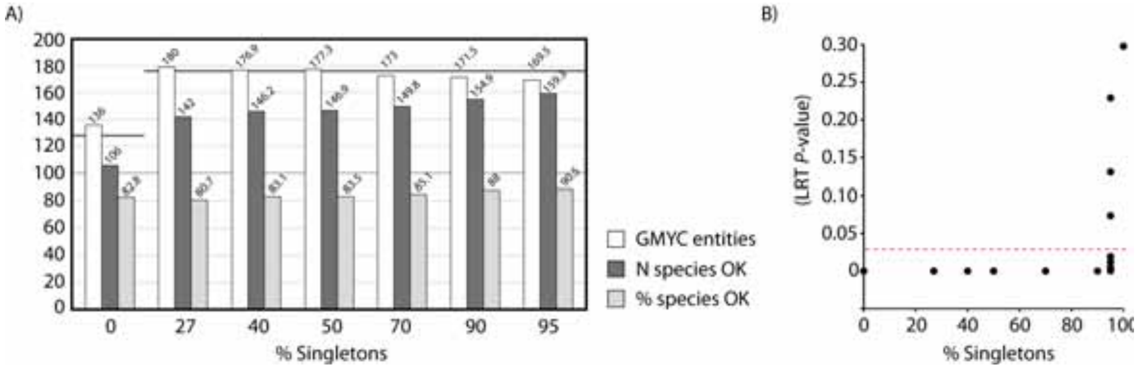


Figure 5. A) Number of GMYC entities, number of correct hits and percentage for different percentage (0%, 27%, 40%, 50%, 70%, 90% and 95%) of singletons (species represented by a single haplotype). Horizontal grey lines represent the correct number of species. B) GMYC significance (likelihood ratio test *P*-value) vs. percentage of singletons.

DISCUSSION

1) What is the real rate of success of GMYC?

The percentage of failure was very similar for all approaches used, excluding ChronoPL, and ranged between 20.45% and 19.31% (34 and 36 species respectively). Generally, all methods had problems in delimiting the same set of species and the same effects for each taxon were recurrently observed across methods. The 36 cases of failure can be grouped in three main categories: 16 cases of lumping species pairs, 14 cases of oversplitting (into two entities in all cases except in one case that is divided in three) and six cases of paraphyly (Supplementary Table S5). To assess the real performance of the GMYC method, one should take into account that not all these cases represent failures of the method *per se*, but may be due to various factors such as limited resolution of the data (e.g. incomplete lineage sorting of the marker), quality of the phylogenetic reconstruction, or incomplete taxonomical framework (e.g. cryptic species or taxonomical oversplitting).

For the 16 cases of “lumping”, five species pairs (*Colias crocea* - *C. erate*, *Apatura ilia* - *A. metis*, *Coenonympha tullia* - *C. rhodopensis*, *Erebia ligea* - *E. euryale*, *Hipparchia fagi - H. syriaca*) were recovered as conspecific because they displayed very low levels of genetic divergence (less than 0.5% in all cases) and lacked reciprocal monophyly. Moreover, at least one of these pairs (*Colias crocea - C. erate*) is strongly suspected of hybridization or historical introgression as natural hybrids have been recorded (Annex 4 of Dincă *et al,* 2011a and references therein). In addition, some of these species may require taxonomic revision and several authors have generated debates whether they represent species or subspecies (e.g. *Coenonympha tullia - C. rhodopensis*) (Annex 4 of Dincă *et al*, 2011a and references therein). The remaining three species pairs were each monophyletic, but formed closely related pairs displaying low levels of genetic divergence. However, they seem to display constant morphological differences and their status as distinct species is rather well established. A certain GMYC lumping tendency has been proven by simulation tests and attributed to a poor ability of the model to identify incomplete lineage sorting or clades undergoing rapid radiation (Esselstyn *et al*, 2012; Reid and Carstens, 2012), a phenomenon that may eventually blur the boundary between tokogeny and phylogeny among incipient species, and that can be more pronounced when a thorough regional sampling is achieved. In our dataset, no such tendency was observed since cases of lumping and splitting were similar in number, and actually a good number of the cases of lumping documented coincided with non-reciprocal monophyly.

Among the 14 cases of oversplitting (~8%), six displayed levels of maximum intraspecific divergence equal or higher than 2% (*Lasiommata maera* (2%), *Melitaea aurelia* (2.2%), *Thymelicus sylvestris* (2.7%), *Pyrgus armoricanus* (2.8%), *Hipparchia semele* (3.5%) and *Boloria euphrosyne* (5%)) and have already been highlighted as potential cases of cryptic species deserving further studies (Dincă *et al*, 2011a). Indeed, it has been claimed that GMYC is a tool to identify potential cryptic species when morphospecies are divided in two or more GMYC entities (Fontaneto *et al*, 2009; Ceccarelli *et al*, 2012). However, an artifactual oversplitting tendency has also been suspected in some instances (Lohse, 2009; Hendrich *et al*, 2011; Crawford *et al*, 2010; Monaghan *et al*, 2009; Esselstyn *et al*, 2012), and cases of oversplitting may represent either real failures of GMYC or gaps in our taxonomical knowledge. Despite the European butterfly fauna has been studied in depth, several unexpected discoveries of cryptic species have been recently reported, such as *Zerynthia cassandra* (Nazari and Sperling, 2007; Dapporto, 2010), *Polyommatus celina* (Dincă *et al*, 2011b) and *Leptidea juvernica* (Dincă *et al*, 2011c). While it is not impossible that a few of the 14 cases of splitting may actually represent cryptic species, there is currently no relevant morphological, ecological or additional molecular data supporting this hypothesis, and we consider them as failures of the GMYC method due to pronounced intraspecific genetic variability.

Finally, two of the three species pairs displaying paraphyly constitute suspected cases of introgression (*Lysandra coridon* - *L. bellargus* and *P. napi* - *P. bryoniae*). Additionally, *P. bryoniae* and *C. orientalis* are taxa whose taxonomic status has been debated and may still require further studies (see Annex 4 of Dincă *et al*, 2011a and references therein). For these reasons, these three species pairs should not be treated as failures of the GMYC method. All in all, it is reasonable to consider that only 20 out of 36 cases may represent failures of the GMYC. This means that the actual performance of the GMYC would rise to 88.63% (BEAST coalescent relaxed), 89.2% (BEAST Yule strict) and 89.77% (BEAST coalescent strict and Yule relaxed) (Supplementary Table S5).

These percentages of success suggest that the GMYC approach can represent a powerful tool in assessing biodiversity based exclusively on DNA sequence data and in the absence of taxonomical knowledge. Any final value, though, may be greatly dependent on other parameters that affecting the internal branching pattern of the tree, which will be basically shaped by the attributes of each specific dataset sampling, as discussed below.

2) Effects of phylogenetic reconstruction on GMYC performance

2.1- Relevance of the ultrametric tree obtention method

Optimizing branch lengths in a phylogenetic tree to convert it to ultrametric is required for the GMYC method, but this process can be computationally demanding for large-scale datasets. As a consequence, the fastest methods to obtain ultrametric trees are most frequently used, as for example the algorithms implemented in the software PATHd8 (Ceccarelli *et al*, 2012; Crawford *et al*, 2010; Hendrich *et al*, 2010; Kergoat *et al*, 2011; Nekola *et al*, 2009; Astrin *et al*, 2012). In our study, BEAST inferences were the most computationally demanding, followed by Garli ML inferences + PL in r8s. The fastest inferences were the ones using Garli ML + ChronoPL in R and RaxML + PATHd8 algorithms. We report very poor GMYC resolution for the tree from the ChronoPL function, but success rates from PATHd8 resulting trees were comparable to those of BEAST genealogies and even slightly higher, probably a stochastic effect. Thus, we can conclude that all the studied methods performed similarly, and only the ChronoPL function is clearly unsuitable to perform GMYC and should be avoided. The RaxML + PATHd8 algorithm would be the preferred choice among the ones tested if time or computation capacity are constrains, as is frequently the case.

Some studies recommend BEAST as input for GMYC and show that the performance of the coalescent tree prior is better than that of the Yule prior (Monaghan *et al*, 2009) at least when the analysis is based on the COI marker (Ceccarelli *et al*, 2012). Given that the GMYC uses coalescence as a null model, the coalescent tree prior is considered a more adequate option, and appears to fit better the majority of the datasets in model comparisons (Monaghan *et al*, 2009). Regarding the molecular clock models, Monaghan *et al* (2009) observed in their dataset that a relaxed molecular clock with Yule prior resulted in a greater number of GMYC entities than other methods (strict or coalescent). We observe very small differences in the different BEAST options, and we cannot incline in favour of any of them since all fit within the estimated intervals of confidence (Supplementary Table S2-S4). As a general rule, our findings suggest that GMYC is much more sensitive to sampling effects than to the tree reconstruction method, as will be discussed below.

2.2- Haplotype collapsing or all specimens?

Collapsing sequences to haplotypes is a common approach before running GMYC, as the model cannot handle polytomies and zero-length terminal branches. Thus the removal of identical haplotypes is necessary when the input tree is inferred using a classic phylogenetic reconstruction method. However, when a genealogy-based inference approach is employed, such as BEAST, identical sequences will be treated as different alleles coalescing back to their most recent common ancestor, which will insert non-zero branch lengths also providing relevant information to the GMYC calculations. For our dataset, the GMYC performance of genealogies obtained by using the models implemented in the package BEAST was not found to be different to that of the collapsed datasets (Figure 2C and Supplementary Table S4), and for that reason, the use of less demanding methods would be preferred, since large-scale biodiversity surveys may include vast numbers of specimens and the inference of BEAST genealogies will exponentially increase computational time when not removing duplicate haplotypes.

3) Effects of dataset characteristics on GMYC performance

Imbalances between the Yule and coalescent portions of the tree may be crucial for the performance of the method and will depend on each specific dataset. Several factors can shape the internal branching pattern favouring one or the other portion. For example, an increased number of deep coalescent events in the gene tree, large population sizes or fast speciation rates will produce a shift of the estimated threshold to higher divergences, potentially leading to GMYC underestimations (Esselstyn *et al*, 2012). Nevertheless, our empirical results show that both GMYC significance and success rate are remarkably stable over a wide array of conditions altering the dataset characteristics and, consequently, tree shape.

3.1- Taxon level effects

Scientists' interests in biodiversity range from detailed studies of species-groups across their entire distribution, to large-scale biodiversity surveys of all the taxa occurring in a specific area, such as in phylogenetic community ecology. That is, from clade sampling (studying the diversity within a clade) to area sampling (studying the local/regional fauna or flora). For both scenarios, the structure of the Yule portion of the tree will play an important role. For example it may be different to work within a radiation including a set of closely related taxa than at higher taxonomic levels within a specific area where sister taxa can be present or not. The inclusion of sister taxa will have a lot of relevance narrowing the inter-specific divergences, which may shape the tree as a continuum pattern of splits with less chances of finding switches in the likelihood of GMYC calculations, a situation potentially leading to underestimations. On the other hand, deep taxon level surveys have the risk of including strong heterogeneity in speciation rates or in intraspecific population structure across taxa, for example.

Our dataset is in an intermediate position between the two extremes discussed. It represents a rather diverse superfamily survey for a substantial geographical area and includes both deeply diverged lineages and extremely recent speciation events, of which the 16 lumped species pairs are a good example. Genera include a number of species present in the study region ranging from one to 14.

When progressively reducing the size of the target trees in our dataset, we show that a surprisingly low number of species is required to produce GMYC significance (a minimum of between three and five depending on the taxa). Not only this, but only a small degree of variability in GMYC success was documented depending on the taxon level analysed, usually representing a single case. We mostly observe single changes in the resulting number of GMYC entities without a clear trend to either splitting or lumping, nor to increasing or decreasing success (Table 1), a result that might be even attributable to stochasticity. Worth noting, a wide applicability of the method with regards to taxon level does not imply that all datasets are adequate, and some studies applying GMYC to datasets including species complexes with a very small number of closely related species are usual (e.g. Fontaneto *et al*, 2007a; Papadopoulou *et al*, 2009a; Gebiola *et al*, 2012), but could likely be problematic.

3.2- Impact of singleton percentage

Any survey might be subject to collecting single specimens representing rare species. However, many singletons in a dataset may be due to multiple individuals sharing a single haplotype. Our dataset illustrates this: 48 species (27%) display a single haplotype, but only four of these are represented by a single specimen. Thus, a proportion of lineages will be singletons according to the intensity and success of the collecting work, but also to the characteristics of the taxa within the area studied. The higher the number of singletons, the lower the coalescent portion in our tree providing information to the GMYC model. Indeed, having a low coalescent portion has been understood as detrimental for GMYC resolution (e.g. Lim *et al*, 2011). However, some studies found a good GMYC performance with proportions of singletons close to 60% (Monaghan *et al*, 2009; Ceccarelli *et al*, 2012), and Reid and Carstens, (2012) concluded, based on simulations, that GMYC works efficiently with singletons while other taxa better represented allow to calibrate the divergence threshold. Our results point to that direction and shows that up to 95% of singletons the method is operational and the success rate does not decrease. Actually, performance increases because turning taxa into singletons avoids the tendency to oversplitting that some of them display. Obviously, the higher the singletons present, the less biologically meaningful is the result.

3.3- Is the inclusion of outgroup taxa relevant?

Users sometimes add and sometimes exclude outgroup taxa from their datasets, and a consensus about their relevance does not seem to be established. Adding outgroup taxa to a low taxon level phylogeny will increase the Yule portion of the tree and thus in most cases will help balancing the Yule-coalescent equilibrium. For example, Powell *et al* (2011) added an outgroup to their tree to improve the power of the algorithm to estimate parameters associated with the Yule portion of the model because the number of early branches was relatively low.

4) Sampling scheme relevance

4.1- Does a continuous geographic sampling matter?

An ideal biodiversity survey should cover as many populations as possible representing a continuous geographic sampling in order to better assess biodiversity for one specific area of study. Lohse (2009) suggested that undersampling of population demes may oversplit entities in GMYC analysis when <20% of demes are sampled. Papadopoulou *et al* (2009b) replied arguing that geographical coherence of demes may or may not exist in real data (Avise, 2009), and Reid and Carstens (2012) showed that such a demic sampling might not be a common problem for real datasets. Our results indicate that intermediate haplotypes of a given species are basically irrelevant, in the sense that their removal does not produce a split in that species. The result is even stronger if we take into account that the test was performed with the taxa that displayed intraspecific divergences nearest to the GMYC divergence threshold. This suggests that gaps in intraspecific sampling are rather unimportant as long as they don't affect the maximum divergence. Thus, a continuous and homogeneous sampling is likely less important than a sampling directed to obtain the most diverged populations. This could be achieved by trying to cover the most distant or isolated populations, the widest variety of habitats, and the phenotypically differentiated populations. This result contradicts the suggestion that geographical gaps for one species may theoretically lead to oversplitting (Lohse, 2009). The implications are important especially because the viability to collect complete series of demes will decrease when increasing the geographic area surveyed. It is worth noting that intermediate haplotypes in a single species might not influence the exact divergence of the Yule-coalescent threshold estimated by GMYC, but their general presence or absence across the taxa in the tree may do so.

A suitable number of samples per species cannot be simply a function of the size of the region studied, since the genetic richness and structure within the surveyed area will be important too. Moreover, the characteristics of the fauna studied may be relevant as well. Since butterflies are a group with relatively high dispersal rates that can maintain certain levels of gene flow at different geographic scales, one may ask, for example, what happens in organisms with higher levels of population isolation or specialization, or displaying low population sizes. In this context, plotting the number of haplotypes versus the number of specimens or populations for each species as in Figure 5 may help evaluating the sampling effort with respect to the area and the characteristics of our taxa. For our dataset, it can be seen that above ca. 10 specimens or eight populations per species the number of haplotypes are very variable depending on the species, and no clear correlation is retained. This suggests that saturation is being reached at least for some species. Given that we have shown that maximum intraspecific divergence is more important than sheer haplotype numbers, accumulation curves showing how this parameter grows with sampling effort for each species might help planning and monitoring the sampling strategy.

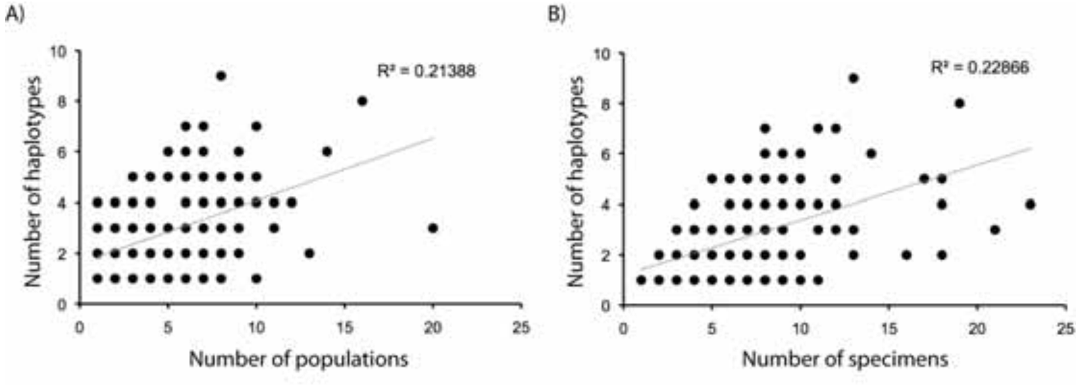


Figure 6. Plots showing the number of haplotypes vs. the number of populations (A) and specimens (B) for each of the species in the dataset. Note that above ca. ten specimens or eight populations per species the correlation is lost because of saturation.

4.2- Geographic range balance among organisms surveyed

The fact that all the specimens included in a tree originate from the same surveyed area can be crucial for a reliable result because an unbalanced situation from both the Yule and the coalescent portions might introduce a bias in the tree. We created a highly unbalanced area sampling for two species with respect to the rest by enlarging their area of study. It resulted in multiple splitting for *P. icarus* and, by contrast, *P. machaon* did not experience any change. Indeed, sampling expansion in the case of *P. icarus* was much more important than that of *P. machaon* in terms of area and specimens, and was reflected in a higher increase in maximum intraspecific divergence. Moreover, there is an important number of species of the *Polyommatus* group in Romania (*amandus*, *thersites*, *daphnis*, *dorylas*, *coridon* and *bellargus*), which represents a less flexible Yule portion to be displaced in detriment to the coalescent portion. On the other hand, only five species of Papilionidae are present. In addition, no pair represents closely related sister species, and the closest species to *P. machaon* belong to different genera (*Zerynthia* and *Allancastria*). This situation represents an enormous Yule depth that is difficult to counteract by increasing the coalescent portion. In conclusion, area sampling imbalances across taxa might indeed result in splitting when a sufficient increase of instraspecific divergence occurs with respect to the general tree shape.

5) General usefulness of GMYC

The GMYC is a stable, objective and biologically meaningful methodology for biodiversity exploration with single-locus data. We obtain a general rate of success about the 80%, which would need to be corrected to 90% if we don't account for intrinsic limitations of the dataset. GMYC can be described as a method to objectively find the most suitable divergence percentage to define species for a given tree. Thus the best possible performance will be that produced by the threshold minimizing the sum of lumping plus splitting cases. In our case, we find a set of 36 cases that are repeatedly wrongly categorized by GMYC trials. These include 16 cases of lumping species pairs, 14 cases of oversplitting and 6 cases of non-monophyly. Thus, we see that the result could hardly be improved by a single threshold because it is already situated in a midpoint between lumping and splitting. Indeed, the best possible GMYC result is determined in each dataset by the number of cases of non-monophyly plus the degree of overlap between intra- and interspecific divergences. We thus consider this result to be almost the best possible given the characteristics of our dataset, which vindicates the validity of the method. This limitation could be in theory overcome by a multiple-threshold model. However, we find that the multiple-threshold model never fits our dataset significantly better than the single-threshold one. This could be explained either because the cases of lumping and splitting are scattered and mixed in the tree (and thus too many thresholds would be necessary) or because some limitations in the multiple-threshold model exist that should be improved. Indeed our cases of lumping and splitting are scattered along the tree and both tendencies occur in closely related taxa. For example, the subfamily Satyrinae in general and the genus *Hipparchia* in particular include cases of both lumping and splitting. This phenomenon is expected to be common in many datasets and may be the main reason for the limited applicability of the multiple-threshold model.

As the developers claim, this is a method for evaluating biodiversity from unknown areas and/or living groups, although users have applied the method as a criterion to take specific decisions, which highlighted the limits of GMYC. It is important to keep in mind what we can know and we cannot know when applying the GMYC on single-locus based phylogenetic trees. Based on the results here presented and previously published tests showing that performance is good for the vast majority of species and that the method is stable in a wide array of conditions, GMYC promises to be greatly useful to estimate potential species numbers and highlight candidate cryptic species, although it should not be used as an ultimate criterion for species description by itself since a percentage of error always exists.

CONCLUSIONS

Our results demonstrate that GMYC performance is little affected within a good part of the conditions tested, including the phylogenetic methods used to obtain the ultrametric tree, the taxon levels explored, and the quality of coverage for each taxon. Generally speaking, except for instance in the case of the ChronoPL method, this space of relative stability is approximately delimited by GMYC significance values. We report a high performance rate for GMYC with empirical data, which is close to 80% when taking into account the dataset as a whole, and close to 90% when not accounting for failures derived from data limitations such as non-monophyletic lineages. These results suggest that the GMYC is a generally suitable tool for most studies, including biodiversity assessments, phylogenetic community ecology, and species complexes provided that a minimal set of conditions are fulfilled. A synopsis of the factors tested, results obtained and practical recommendations are summarized in Table 3.

Table 3. Summary of factors tested, results obtained and practical recommendations for biodiversity studies employing the GMYC model.

Factors tested	Results	Practical recommendations for GMYC studies
1. Phylogenetic methods		
1.1. Tree inference method	Similar results for all ML and Bayesian methods tested.	Not necessary to test several options.
1.1. Ultrametric tree obtention method after ML	PATHd8 and r8s produce similar results. ChronoPL works substantially worse.	The usage of rapid algorithms as PATHd8 is recommended to accelerate computing. Do not use ChronoPL.
1.2. Haplotype vs. sequence trees	No substantial differences.	It is possible to use haplotype trees in order to accelerate computing.
2. Taxon level		
2.1. Superfamily to family reduction	No substantial differences. A substantial Yule proportion is present in both cases.	Higher taxonomic levels can be equally operative, at least for taxa with a good number of species.
2.2. Reduction of clades to the minimal GMYC significance	Low Yule portion levels destabilize GMYC when less than 3 to 5 species are included (depending on the taxa). Within GMYC significance, variation of performance is small.	If only interested in one or few species, add outgroup taxa. Avoid studies with extremely few species and always test for GMYC significance.
3. Sampling coverage		
3.1. Reduction of haplotypes intermediate to the most divergence in a species	No induction of splits at our geographical range. The most critical for a given species is the maximum intraspecific divergence, not so much the presence of intermediate haplotypes.	Sampling should be designed not to miss extreme haplotypes: try to cover the most distant or isolated populations, the widest variety of habitats, and those populations phenotypically differentiated. Plotting maximum intraspecific divergences accumulation curves for each species may help monitoring the sampling strategy.
3.2. Unbalanced geographical range	Species may split when new geographically distant haplotypes increase intraspecific divergence.	The extension of sampling area needs to be uniform over all species for comparable results.
3.3. Percentage of singletons	A high singleton percentage can be accommodated by GMYC (up to 95%). Threshold and percentages not widely affected within GMYC significance.	A high percentage of singletons may not affect the percentage of correctly identified species, but may critically reduce the meaningfulness of the analysis.

ACKNOWLEDGEMENTS

We want to thank Anna Papadopoulou for fruitful discussions and suggestions. Support for this research was provided by the Spanish MICINN (project CGL2010-21226/BOS and predoctoral fellowship BES-2008-002054 to G.T.).

REFERENCES

Astrin, J.J., Stüben, P.E., Misof, B., Wägele, J.W., Gimnich, F., Raupach, M.J., Ahrens, D. **2012**. Exploring diversity in cryptorhynchine weevils (Coleoptera) using distance-, character- and tree-based species delineation. *Molecular Phylogenetics and Evolution* 63:1-14.

Avise, J.C. **2009**. Phylogeography: retrospect and prospect. *Journal of Biogeography* 36:3–15.

Barraclough, T.G., Nee, S. **2001**. Phylogenetics and speciation. *Trends in Ecology and Evolution* 16(7):391-399.

Barraclough, T.G., Hughes, M., Ashford-Hodges, N., Fujisawa, T. **2009**. Inferring evolutionary significant units of bacterial diversity from broad environmetal surveys of single-locus data. *Biology Letters* 5(3):425-428.

Bergsten, J., Bilton, D.T., Fujisawa, T., Elliott, M., Monaghan, M.T., Balke, M., Hendrich, L., Geijer, J., Herrmann, J., Foster, G.N., Ribera, I., Nilsson, A.N., Barraclough, T.G., Vogler, A. **2012.** The effect of geographical scale of sampling on DNA barcoding. *Systematic Biology* 61(5):851-869.

Britton, T., Oxelman, B., Vinnersten, A., Bremer, K. **2002**. Phylogenetic dating with confidence intervals using mean path lengths. *Molecular Phylogenetics and Evolution* 24:58-65.

Britton, T., Anderson, C., Jacquet, D., Lundqvist, S., Bremer K. **2006**. PATHd8 - a program for phylogenetic dating of large trees without a molecular clock. Avalilable at www.math.su.se/PATHd8/

Britton, T., Anderson, C., Jacquet, D., Lundqvist, S., Bremer, K. **2007**. Estimating divergence times in large phylogenetic trees. *Systematic Biology*, 56:741-752

Brower, A.V.Z. **1999**. Delimitation of phylogenetic species with DNA sequences: A critique of Davis and Nixon's population aggregation analysis. *Systematic Biology* 48: 199-213.

Carstens, B.C., Knowles, L.L. **2007**. Estimating species phylogeny from gene-tree probabilities despite incomplete lineage-sorting: an example from Melanoplus grasshoppers. *Systematic Biology* 56:400–411.

Ceccarelli, F.S., Sharkey, M.J., Zaldívar-Riverón, A. **2012**. Species identification in the taxonomically neglected, hihgly diverse, neotropical parasitoid wasp genus Notiospathius (Braconidae: Doryctinae) based on an integrative molecular and morphological approach. *Molecular Phylogenetics and Evolution* 62:485-495.

Crawford, A.J., Lips, K.R., Bermingham, E. **2010**. Epidemic disease decimates amphibian abundance, species diversity, and evolutionary history in the highlands of central Panama. *Proceedings of the National Academy of Sciences USA* 107(31):13777-13782.

Cummings, M.P., Neel, M.C., Shaw, K.L. **2008**. A genealogical approach to quantifying lineage divergence. *Evolution* 62(9):2411-2422.

Dapporto, L. **2010**. Speciation in Mediterranean refugia and post-glacial expansion of *Zerynthia polyxena* (Lepidoptera, Papilionidae). *Journal of Zoological Systematics and Evolutionary Research* 48:229–237.

Davis, J.I., Nixon, K.C. **1992**. Populations, genetic variation, and the delimitation of phylogenetic species. *Systematic Biology* 41:421–435.

Dincă, V., Zakharov, E., Hebert, P., Vila, R., **2011a**. Complete DNA barcode reference library for a country's butterfly fauna reveals high performance for temperate Europe. *Proceedings of the Royal Society B: Biological Sciences* 278, 347–355.

Dincă, V., Dapporto, L., Vila, R. **2011b**. A combined genetic-morphometric analysis unravels the complex biogeographical history of Polyommatus icarus and Polyommatus celina common blue butterflies. *Molecular Ecology* 20(18):3921-35.

Dincă, V., Lukhtanov, V.A., Talavera, G., Vila, R. **2011c**. Unexpected layers of cryptic diversity in wood white Leptidea butterflies. *Nature Communications* 2:234.

Drummond, A.J., Ho, S.Y.W., Phillips, M.J., Rambaut, A. **2006**. Relaxed Phylogenetics and Dating with Confidence. *PLoS Biology* 4(5):e88.

Drummond, A.J., Rambaut, A. **2007**. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* 7:214.

Esselstyn, J.A., Evans, B.J., Sedlock, J.L., Khan, F.A.A., Heaney, L.R. **2012**. Single-locus species delimitation: a test of the mixed Yule-coalescent model, with an empirical application to Philippine round-leag bats. *Proceedings of the Royal Society B: Biological Sciences* doi:10.1098/rspb.2012.0705

Ezard, T., Fujisawa, T., Barraclough, T. **2009**. Splits: Species' Limits by Threshold Statistics. R package version 1.0 [http://R-Forge.R-project.org/projects/splits/]

Fontaneto, D., Boschetti, C., Ricci, C. **2007a**. Cryptic diversification in ancient asexuals: evidence from the bdelloid rotifer *Philodina flaviceps*. *Journal of Evolutionary Biology* 21(2):580-587.

Fontaneto, D., Herniou, E.A., Boschetti, C., Caprioli, M., Melone, G., Ricci, C., Barraclough, T.G. **2007b**. Independently evolving species in asexual bdelloid rotifers. *PLoS Biology* 5:914–921.

Fontaneto, D., Kaya, M., Herniou, E.A., Barraclough, T.G. **2009**. Extreme levels of hidden diversity in microscopic animals (Rotifera) revealed by DNA taxonomy. *Molecular Phylogenetics and Evolution* 53(1):182-189.

Gebiola, M., Gómez-Zurita, J., Monti, M.M., Navone, P., Bernardo, U. **2012**. Integration of molecular, ecological, morphological and endosymbiont data for species delimitation within the Pnigalio soemius complex (Hymenoptera:Eulophidae). *Molecular Ecology* 21(5):1190-1208.

Hausmann, A., Haszprunar, H., Segerer, A.H., Speidel, W., Behounek, G., Hebert, P.D.N. **2011**. Now DNA-barcoded: the butterflies and larger moths of Germany. Spixiana 34:47–58.

Hebert P.D.N., Cywinska A., Ball S.L. **2003**. Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences* 270:313–321.

Hendrich, L., Pons, J., Ribera, I., Balke, M. **2010**. Mitochondrial *Cox1* Sequence data reliably uncover patterns of insect diversity but suffer from high lineage-idiosyncratic error rates. *PLoS One* 5(12):e14448.

Hudson, R.R. **1990**. Gene genealogies and coalescent process. In *Oxford Surveys in Evolutionary Biology*. Futuyma, D.J., Antonivics, J., [Eds]. Oxford University Press.

Kergoat, G.J., Le Ru, B.P., Genson, G., Cruaud, C., Couloux, A., Delobel, A. **2011**. Phylogenetics, species boundaries and timing of resource tracking in a highly specialized group of seed beetles (Coleoptera: Chrysomelidae: Bruchinae). *Molecular Phylogenetics and Evolution* 59:746-760.

Knowles, L.L., Carstens, B.C. **2007**. Delimiting species without monophyletic gene trees. *Systematic Biology* 56:887–895.

Lohse, K. **2009**. Can mtDNA barcodes be used to delimit species? A response to Pons *et al*. (2006). *Systematic Biology* 8(4):439-442.

Masters, B.C., Fan, V., Ross, H.A. **2011**. Species Delimitation - a Geneious plugin for the exploration of species boundaries. *Molecular Ecology Resources* 11:154-7

Monaghan, M.T., Wild, R., Elliot, M., Fujisawa, T., Balke, M., Inward, D.J.G., Lees D.C., Ranaivosolo, R., Eggleton, P., Barraclough, T.G., Vogler, A.P. **2009**. Accelerated species inventory on Madagascar using coalescent-based models of species delineation. *Systematic Biology* 58(3):298-311

Nazari, V., Sperling, F. **2007**. Mitochondrial DNA divergence and phylogeography in western Palaearctic Parnassiinae (Lepidoptera: Papilionidae): How many species are there? *Insect Systematics and Evolution* 38:121–138.

Nee, S., May, R.M. & Harvey, P.H. **1994**. The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 344(1309):305-311.

Nekola, J.C., Coles, B.F., Bergthorsson, U. **2009**. Evolutionary pattern and process within the Vertigo gouldii (Mollusca: Pulmonata, Pupillidae) group of minute North American land snails. *Molecular Phylogenetics and Evolution* 53:1010-1024.

Paradis, E., Claude, J., Strimmer, K. **2004**. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20: 289–290.

Papadopoulou, A., Bergsten, J., Fujisawa, T., Monaghan, M.T., Barraclough, T.G., Vogler, A.P. **2008**. Speciation and DNA barcodes: testing the effects of dispersal on the formation of discrete sequence clusters. *Philosophical Transactions of the Royal Society B*: *Biological Sciences* 363:2987–2996.

Papadopoulou, A., Anastasiou, I., Keskin, B., Vogler, A.P. **2009a**. Comparative phylogeography of tenebrionid beetles in the Aegean archipelago: the effect of dispersal ability and habitat preference. *Molecular Ecology* 18:2503–2517.

Papadopoulou, A., Monaghan, M.T., Barraclough, T.G., Vogler, A.P. **2009b**. Sampling error does not invalidate the Yule-Coalescent model for species delimitation. A response to Lohse (2009). *Systematic Biology* 58(4):442-444.

Pons, J., Barraclough, T., Gomez-Zurita, J., Cardoso, A., Duran, D., Hazell, S., Kamoun, S., Sumlin, W., Vogler, A. **2006**. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology* 55:595–610.

Pons J., Fujisawa T., Claridge E.M., Anthony Savill R., Barraclough T.G., Vogler A.P. **2011**. Deep mtDNA subdivision within Linnean species in an endemic radiation of tiger beetles from New Zealand (genus Neocicindela). *Molecular Phylogenetics and Evolution* 59:251-262.

Posada, D. **2004**. Collapse: Describing haplotypes from sequence alignments. [http://darwin.uvigo.es/software/collapse.html].

Posada, D. **2008**. jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution* 25:1253–1256.

Powell, J.R., Monaghan, M.T., OPik, M., Rillig, M.C. **2011**. Evolutionary criteria outperform operational approaches in producing ecologically relevant fungal species inventories. *Molecular Ecology* 20(3):655–666.

Puillandre, N., Lambert, A., Brouillet, S., Achaz, G. **2012**. ABGD, Automatic barcode gap discovery for primary species delimitation. *Molecular Ecology* 21(8):1864-77.

Rambaut, A., Drummond, A.J. **2007**. Tracer v1.4, Available from http://beast.bio.ed.ac.uk/Tracer

Reid, N.M., Carstens, B.C. **2012**. Phylogenetic estimation error can decrease the accuracy of species delimitation: a Bayesian implementation of the general mixed Yule-coalescent model. *BMC Evolutionary Biology* 12:196.

Rubinoff, D., Holland, B.S. **2005**. Between two extremes: mitochondrial DNA is neither the panacea nor the nemesis of phylogenetic and taxonomic inference. *Systematic Biology* 54:952–961.

Sanderson, M.J. **2003**. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19: 301–302.

Stamatakis, A. **2006**. RAxML-VI-HPC: Maximum Likelihood-based Phylogenetic Analyses with Thousands of Taxa and Mixed Models. *Bioinformatics* 22(21):2688–2690.

Will, K.W., Rubinoff, D. **2004**. Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics* 20:47–55.

Will, K.W., Mishler B.D., Wheeler Q.D. **2005**. The perils of DNA barcoding and the need for integrative taxonomy. *Systematic Biology* 54:844–851.

Zwickl, D.J. **2006**. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. PhD thesis, University of Texas at Austin. Available at https://www.nescent.org/wg_garli/Main_Page.

SUPPLEMENTARY INFORMATION

Factors affecting biodiversity assessments with the GMYC model: insights from a DNA-barcoding survey of butterfly fauna

Gerard Talavera[a,b], Vlad Dincă[a,c] and Roger Vila[a,*]

[a]Institut de Biologia Evolutiva (CSIC-UPF), Passeig Marítim de la Barceloneta, 37, 08003 Barcelona, Spain

[b]Departament de Genètica i Microbiologia, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain

[c]Department of Zoology, Stockholm University, 106 91, Stockholm, Sweden.

[*]Corresponding author: roger.vila@csic.es

Supplementary Table S1. List of *Polyommatus icarus* and *Papilio machaon* sequences used in this study apart from the Romanian dataset in Dincă et al (2011a). *P. icarus* sequences lacking GenBank accession numbers share haplotypes with deposited sequences (see Dincă et al 2011b).

Sample ID / GenBank accession number	Taxon	Locality data	Country
FJ428817	*P. icarus*	Iskr, Pasarel	Bulgaria
09V959 / JN084708	*P. icarus*	N Pirin Mt., Banderitsa nut-vihren hut road	Bulgaria
09V960 / JN084697	*P. icarus*	N Pirin Mt., Banderitsa nut-vihren hut road	Bulgaria
08L303 / JN084695	*P. icarus*	Hammer Bakker	Denmark
07C710 / JN084694	*P. icarus*	Uusimaa Loviisa Harmaakallio	Finland
09X260 / JN084698	*P. icarus*	Bédoine, Provence	France
09T515 / JN084701	*P. icarus*	Bonifacio, Corsica	France
08P209	*P. icarus*	Courbassil, Languedoc-Roussillon	France
09V243	*P. icarus*	Domaine de la chasse de Puits de Rians, Provence	France
09X210	*P. icarus*	Oraison, Provence	France
09X222 / JN084688	*P. icarus*	Sault, Provence	France
07C552 / JN084693	*P. icarus*	Hungen Steinheim, Hessen	Germany
GU655005	*P. icarus*	Muenchen, Obermenzing, Bavaria	Germany
GU688449	*P. icarus*	Koenigsdorf TOEL, Bavaria	Germany
HM391783	*P. icarus*	Wiesenfelden, Hoehenberg, Bavaria	Germany
HQ563555	*P. icarus*	Toepen, Fattigsmuehle, Bavaria	Germany
150308PP22 / JN084709	*P. icarus*	Bahía de Almyros, Almyros, Crete	Greece
FJ428819	*P. icarus*	Dodoni, near Igoumenista	Greece
150308PP10 / JN084691	*P. icarus*	Montes de Ornos, Crete	Greece
150308PP62 / JN084692	*P. icarus*	Montes de Ornos, Crete	Greece
150308PP67	*P. icarus*	Montes de Ornos, Crete	Greece
AY556866	*P. icarus*	Mt. Falakro	Greece
AY556927	*P. icarus*	Hajiabad, Golestan	Iran
EU597142	*P. icarus*	N. Torbat-e-Heydariyeh, Khorasan	Iran
FJ428823	*P. icarus*	Nir Sannich Abad, Yazd	Iran
FJ428824	*P. icarus*	S. Lalehzar, Kerman	Iran
FJ428826	*P. icarus*	Mount Tavor	Israel
LD-292	*P. icarus*	Aspromonte, Santuario Polsi, Calabria	Italy
07D884 / JN084705	*P. icarus*	Carpineto Romano, Monti Lepini, Roma	Italy
07D864	*P. icarus*	Filettino, Monti Simbruini, Frosinone, Lazio	Italy
09T572	*P. icarus*	Monte Perone Elba island	Italy
07E062 / JN084702	*P. icarus*	Monte Sirino, Potenza, Basilicata	Italy
EU597139	*P. icarus*	near Trento	Italy
07D812 / JN084700	*P. icarus*	Ostia Antica, Roma	Italy
07E207 / JN084703	*P. icarus*	Ozein-Visyes, Cogne Valley, Valle d'Aosta	Italy
07E051	*P. icarus*	P.N. La Sila, Calabria	Italy
09T559 / JN084706	*P. icarus*	Salerno, Capri island	Italy
09T560	*P. icarus*	Salerno, Capri island	Italy
07D899	*P. icarus*	Serra del Prete, Mont Pollino, Calabria	Italy
FJ663998	*P. icarus*	Oktyabrsky v.	Kazakhstan
FJ663999	*P. icarus*	Oktyabrsky v.	Kazakhstan
08J150	*P. icarus*	Ponte Pedrinha, Beça, Boticas	Portugal
08J127	*P. icarus*	Seculca, Beça, Boticas	Portugal
08J199	*P. icarus*	Serra do Larouco, Padornelos, Montalegre	Portugal

Sample ID / GenBank accession number	Taxon	Locality data	Country
FJ428801	*P. icarus*	Azov, Rostov-on-Don	Russia
FJ428802	*P. icarus*	Azov, Rostov-on-Don	Russia
GQ885173	*P. icarus*	Belokalitvensky District	Russia
FJ428803	*P. icarus*	Dugino, Azov, Rostov-on-Don	Russia
FJ428820	*P. icarus*	Dugino, Azov, Rostov-on-Don	Russia
FJ428822	*P. icarus*	Dugino, Azov, Rostov-on-Don	Russia
FJ428818	*P. icarus*	Nov. Chara, Chita	Russia
FJ428821	*P. icarus*	Sochi, Krasnodar	Russia
EU597140	*P. icarus*	Ukhta, Komi Republic	Russia
EU597141	*P. icarus*	Ukhta, Komi Republic	Russia
FJ428825	*P. icarus*	Zav’yalovsky, Udmurtia	Russia
07F062	*P. icarus*	10 km S. of Morella, Castellón	Spain
09X521	*P. icarus*	2 km NW El Vallecillo, Teruel	Spain
08P416 / JN084704	*P. icarus*	Burón, Picos de Europa, León	Spain
08H434	*P. icarus*	Campo Real, Madrid	Spain
06G433	*P. icarus*	Can Rovira, El Brull, Barcelona	Spain
08H410	*P. icarus*	Cantoblanco, Madrid	Spain
08J960 / JN084689	*P. icarus*	Fuente Albergue de San Francisco, Güejar Sierra	Spain
08J962	*P. icarus*	Fuente Albergue de San Francisco, Güejar Sierra	Spain
08H592	*P. icarus*	Gualda, La Alcarria, Guadalajara	Spain
06K697	*P. icarus*	Guillimona, La Sagra, Huéscar, Granada	Spain
06K698	*P. icarus*	Guillimona, La Sagra, Huéscar, Granada	Spain
08L473	*P. icarus*	Ivars de Noguera, Lleida	Spain
09V396	*P. icarus*	La Tancada, Delta de l'Ebre, Tarragona	Spain
08L281 / JN084687	*P. icarus*	Los Collados - Benimaurell, Vall de Laguart, Alicante	Spain
08H255	*P. icarus*	Mequinenza, Huesca	Spain
07C623	*P. icarus*	Meranges, Girona	Spain
08J968	*P. icarus*	near Veleta, Monachil, Sierra Nevada, Granada	Spain
08R458	*P. icarus*	Padrón, Picarana-Rio Tinto, Galicia	Spain
08L092	*P. icarus*	Ruta de la Sima, El Vallecillo, Teruel	Spain
08R025	*P. icarus*	Sant Vicent, Lliria, Valencia	Spain
08M674	*P. icarus*	Serra Major, Morera de Montsant, Tarragona	Spain
06G507	*P. icarus*	Sierra de Alcubierre, Aragón	Spain
09V446	*P. icarus*	San Juan (Sierra Nevada), Granada	Spain
09V486 / JN084707	*P. icarus*	Sierra de la Sagra, Granada	Spain
09V488	*P. icarus*	Sierra de la Sagra, Granada	Spain
09V902 / JN084699	*P. icarus*	Sierra de la Sagra, Granada	Spain
09V905	*P. icarus*	Sierra de la Sagra, Granada	Spain
09V907	*P. icarus*	Sierra de la Sagra, Granada	Spain
09V941	*P. icarus*	Sierra de la Sagra, Granada	Spain
08P076	*P. icarus*	Torrent de Ridolaina, Cerdanya, Girona	Spain
08R488	*P. icarus*	Tui, Rio Louro-Magdalena, Galicia	Spain
07C463	*P. icarus*	UAB, Bellaterra, Barcelona	Spain
08H232	*P. icarus*	UAB, Bellaterra, Barcelona	Spain
AY556949	*P. icarus*	Ubierna, Burgos	Spain
08P653	*P. icarus*	Valle (Valle-Zurea), Lena, Asturias	Spain
08J899	*P. icarus*	Vega de Espinareda, León	Spain
08L276	*P. icarus*	Vistabella del Maestrazgo, Castellón	Spain
07C177 / JN084696	*P. icarus*	13 km N. of Saimbeily, Adana	Turkey
07C179	*P. icarus*	13 km N. of Saimbeily, Adana	Turkey
07F274	*P. icarus*	13 km N. of Saimbeily, Adana	Turkey

Sample ID / GenBank accession number	Taxon	Locality data	Country
AY556994	*P. icarus*	Dedegol Gecidi, Isparta	Turkey
07F178	*P. icarus*	Kiskaçli, Kayseri	Turkey
07C260	*P. icarus*	Lake Tuzla, Kayseri	Turkey
07F225	*P. icarus*	Ski resort, Erciyes Mountain, Kayseri	Turkey
08J361 / GU669689	*P. machaon*	Vallgrassa, Parc Natural del Garraf, Barcelona	Spain
08R265 / GU669683	*P. machaon*	Vilamós, Vall d'Arán, Lleida	Spain
08J346 / GU669682	*P. machaon*	Cantallops, Alt Empordà, Girona	Spain
08H454 / GU676652	*P. machaon*	Campo Real, Madrid	Spain
08J326 / GU676539	*P. machaon*	Barranco de Valcuerna, Candasnos, Aragón	Spain
08J745 / HM901252	*P. machaon*	Casa de Camineros, Tuéjar, Com. Valenciana	Spain
311007WR55 / GU675945	*P. machaon*	Montecañada, Paterna, Com. Valenciana	Spain
08H627 / GU676601	*P. machaon*	El Gastor, Cádiz, Andalucía	Spain
09V419 / HM901399	*P. machaon*	San Juan (Sierra Nevada), Andalucía	Spain
09T105 / HM901370	*P. machaon*	Uña, Cuenca, Castilla - La Mancha	Spain
270708ZB68 / GU675943	*P. machaon*	Huelga-Utrera, Pontones, Castilla - La Mancha	Spain
08J830 / GU675872	*P. machaon*	Mansilla de las Mulas, Castilla y León	Spain
08R325 / GU675856	*P. machaon*	Pic Tomir, Mallorca	Spain
08P251 / GU676002	*P. machaon*	Santa Maria, Mallorca	Spain
08P286 / GU676000	*P. machaon*	Genova, Mallorca	Spain
08P267 / GU675999	*P. machaon*	S'Albufera, Mallorca	Spain
JF415720	*P. machaon*	Bavaria, Oberbayern, Diessen	Germany
GU707119	*P. machaon*	Bavaria, Oberpfalz, Regensburg, Irlbach bei Wenzenbach	Germany

Supplementary Table S2. GMYC results from the different phylogenetic approaches tested.

	lk null model	lk GMYC	lk ratio	LR test	number ML clusters	CI	Number ML entities	CI	threshold time
SINGLE									
Yule_Strict	3934.455	**4129.876**	390.8424	0***	129	126-131	181	178-186	-0.008415945
Yule_Relaxed	3928.651	**4013.804**	170.3067	0***	131	129-135	182	179-191	-0.01380598
Coal_Strict	3939.987	**4138.193**	396.4108	0***	128	127-130	180	179-188	-0.008624911
Coal_Relaxed	3958.194	**4127.415**	338.4417	0***	129	127-130	183	178-189	-0.008218051
ML PL CV r8s	1235.403	**1470.540**	470.2732	0***	124	124-125	188	183-191	-1.451404
ML PL CV chronopl	4159.176	**4183.56**	48.76785	1.46e-10***	124	116-130	148	135-164	-0.01126325
ML d8 Pathd8	1150.136	**1358.805**	417.339	0***	129	127-129	186	182-189	-2.145605
ML MPL Pathd8	762.7698	**970.0913**	414.6431	0***	129	126-129	186	180-189	-4.700002
MULTIPLE									
Yule_Strict	3934.455	4131.71	394.5106	0***	129	128-130	192	191-200	-0.01135483 -0.006699273 -0.004591989 -0.002971353
Yule_Relaxed	3928.651	4020.096	182.8906	0***	130	130-138	195	188-206	-0.06608312 -0.01391933 -0.007289458 -0.005235013
Coal_Strict	3939.987	4141.827	403.6793	0***	133	133-136	191	189-195	-0.00996776 -0.006435187 -0.004538289 -0.002499381
Coal_Relaxed	3958.194	4131.155	345.9221	0***	131	130-134	194	190-196	-0.01067919 -0.006410255 -0.003729548 -0.002040071
ML PL CV r8s	1235.403	1475.165	479.5247	0***	129	126-129	194	190-194	-1.923096 -1.451404 -1.236565 -1.002014 -0.987953 -0.746125 -0.712135 -0.527453 -0.527453
ML PL CV chronopl	4159.176	4193.28	68.20856	1.116107e-11***	113	103-115	148	121-156	-0.03808485 -0.01641639 -0.009144724 -0.006581627 -0.004474474 -0.003295786
ML d8 Pathd8	1150.136	1363.427	426.5832	0***	129	129-131	197	196-205	-3.06515 -1.597791 -1.369536 -1.239102 -0.760852
ML MPL Pathd8	762.7698	989.5435	453.5474	0***	127	124-128	191	160-191	-52.66667 -40.4 -34.66667 -32.84211 -27.42857 -13.6 -3.142858 -2 -2 -1.800001 -1.000001 -1 -1 -0.999999 -0.666668

Supplementary Table S3. GMYC results for the 5 family trees from different phylogenetic approaches tested.

Nymphalidae	lk null model	lk GMYC	lk ratio	LR test	number ML clusters	CI	Number ML entities	CI	threshold time
SINGLE									
Yule_Strict	1693.975	**1797.557**	207.1633	0***	59	58-61	86	84-89	-0.007881439
Yule_Relaxed	1706.067	**1770.221**	128.3070	0***	61	59-63	86	82-90	-0.009885226
Coal_Strict	1695.889	**1804.698**	217.6187	0***	60	59-62	85	84-88	-0.007379892
Coal_Relaxed	1707.212	**1790.444**	166.4639	0***	59	58-61	86	84-89	-0.007835904
MULTIPLE									
Yule_Strict	1693.975	1799.134	210.3169	0***	57	57-60	90	82-97	-0.01100496 -0.006518004 -0.004171957 -0.003030275
Yule_Relaxed	1706.067	1771.770	131.4063	0***	61	59-62	93	90-101	-0.01240216 -0.00694334 -0.004011274
Coal_Strict	1695.889	1805.386	218.9937	0***	57	57-60	90	89-95	-0.01012431 -0.004640785 -0.002748426
Coal_Relaxed	1707.212	1790.890	167.3564	0***	59	59-60	94	94-99	-0.007835904 -0.003429501

Lycaenidae	lk null model	lk GMYC	lk ratio	LR test	number ML clusters	CI	Number ML entities	CI	threshold time
SINGLE									
Yule_Strict	1007.234	**1050.765**	87.06108	0***	37	37-38	48	47-51	-0.007580533
Yule_Relaxed	1011.613	**1044.066**	64.90507	5.251355e-14***	37	37-39	48	47-51	-0.008193093
Coal_Strict	1009.095	**1054.238**	90.28693	0***	37	37-38	48	48-51	-0.006980703
Coal_Relaxed	1010.825	**1050.084**	78.518	1.110223e-16***	37	37-39	48	48-52	-0.007139122
MULTIPLE									
Yule_Strict	1007.234	1051.242	88.0151	0***	38	37-38	52	48-60	-0.007580533 -0.003546308 -0.002418972
Yule_Relaxed	1011.613	1045.296	67.36502	3.618217e-13***	38	36-39	51	47-62	-0.008193093 -0.005034371 -0.002723347
Coal_Strict	1009.095	1055.571	92.9523	0***	39	37-41	53	48-59	-0.006980703 -0.004167267 -0.00208586
Coal_Relaxed	1010.825	1051.612	81.57525	4.440892e-16***	41	37-42	53	48-61	-0.007139122 -0.00399817 -0.002268585

Hesperiidae	lk null model	lk GMYC	lk ratio	LR test	number ML clusters	CI	Number ML entities	CI	threshold time
SINGLE									
Yule_Strict	304.9764	**317.4483**	24.94384	1.586367e-05***	15	14-15	26	24-27	-0.003534026
Yule_Relaxed	304.9306	**317.3763**	24.8915	1.626840e-05***	15	14-15	26	24-27	-0.003628842
Coal_Strict	304.5646	**318.0188**	26.90846	6.153331e-06***	15	14-15	26	24-27	-0.003113034
Coal_Relaxed	304.4982	**317.7212**	26.44603	7.69157e-06***	15	14-15	26	24-27	-0.003148664
MULTIPLE									
Yule_Strict	304.9764	317.5517	25.15055	4.692192e-05***	15	15-15	27	25-27	-0.003534026 -0.0017913
Yule_Relaxed	304.9306	317.4491	25.03705	4.945407e-05***	15	15-15	27	25-27	-0.003628842 -0.001845086
Coal_Strict	304.5646	318.1572	27.18524	1.823616e-05***	15	15-15	27	24-27	-0.003113034 -0.001567430
Coal_Relaxed	304.4982	317.8812	26.76588	2.216722e-05***	15	14-16	27	24-27	-0.003148664 -0.001570409

Pieridae	lk null model	lk GMYC	lk ratio	LR test	number ML clusters	CI	Number ML entities	CI	threshold time
SINGLE									
Yule_Strict	316.5903	**329.9467**	26.71291	6.762331e-06***	13	13-14	18	18-19	-0.006079802
Yule_Relaxed	316.5355	**329.4494**	25.82774	1.036318e-05***		13-14	18	18-19	-0.006302636
Coal_Strict	318.9399	**332.5433**	27.20691	5.327666e-06***	13	13-14	18	18-19	-0.005451966
Coal_Relaxed	318.8055	**332.0226**	26.43425	7.735432e-06***	13	13-14	18	18-19	-0.005536053
MULTIPLE									
Yule_Strict	316.5903	330.0786	26.97658	2.009677e-05***	13	12-14	19	16-20	-0.006079802 -0.002094096
Yule_Relaxed	316.5355	329.6252	26.17937	2.911539e-05***	13	12-14	19	16-20	-0.006302636 -0.00215506
Coal_Strict	318.9399	332.6425	27.40533	1.645906e-05***	14	12-14	19	16-21	-0.005451966 -0.002681625
Coal_Relaxed	318.8055	332.1749	26.73891	2.244702e-05***	14	12-14	19	16-21	-0.005536053 -0.00268956

Papilionidae	lk null model	lk GMYC	lk ratio	LR test	number ML clusters	CI	Number ML entities	CI	threshold time
SINGLE									
Yule_Strict	83.4571	**89.93471**	12.95523	0.00473442**	4	4-4	5	5-5	-0.007473451
Yule_Relaxed	83.04935	**86.4619**	6.825112	0.07768594n.s.	4	2-5	5	2-6	-0.01082893
Coal_Strict	84.15425	**90.64417**	12.97984	0.004680402**	4	4-4	5	5-5	-0.007046265
Coal_Relaxed	84.21109	**89.3125**	10.20282	0.01691851*	4	3-5	5	3-6	-0.007822454
MULTIPLE									
Yule_Strict	83.4571	90.33898	13.76378	0.008088414**	4	4-4	7	5-7	-0.007473451 -0.001693895
Yule_Relaxed	83.04935	86.84839	7.598086	0.1074611n.s.	4	4-5	7	6-8	-0.01082893 -0.002470347
Coal_Strict	84.15425	91.16514	14.02178	0.007225881**	4	4-4	7	6-7	-0.007046265 -0.001445077
Coal_Relaxed	84.21109	91.09236	13.76253	0.008092815**	3	3-4	5	5-6	-0.07306488 -0.0008719775

Supplementary Table S4. GMYC results for the 5 family genealogies including all surveyed specimens and using BEAST with a coalescent prior and a strict molecular clock.

Nymphalidae	lk null model	lk GMYC	lk ratio	LR test	number ML clusters	CI	Number ML entities	CI	threshold time
SINGLE									
Coal_Strict	5983.871	**6141.304**	314.8668	0***	78	75-80	87	84-89	-0.005069584
MULTIPLE									
Coal_Strict	5983.871	6144.126	320.5111	0***	87	77-87	99	86-99	-0.009179678 -0.005069584 -0.003207249 -0.002075409

Lycaenidae	lk null model	lk GMYC	lk ratio	LR test	number ML clusters	CI	Number ML entities	CI	threshold time
SINGLE									
Coal_Strict	3495.44	**3542.685**	94.48992	0***	46	44-47	49	46-52	-0.00646989
MULTIPLE									
Coal_Strict	3495.44	3543.69	96.50027	0***	53	45-53	59	48-59	-0.00646989 -0.002224977 -0.000801447

Hesperiidae	lk null model	lk GMYC	lk ratio	LR test	number ML clusters	CI	Number ML entities	CI	threshold time
SINGLE									
Coal_Strict	1406.335	**1435.158**	57.64748	1.869283e-12***	24	23-24	25	24-27	-0.003536388
MULTIPLE									
Coal_Strict	1406.335	1438.363	64.0571	1.758038e-12***	35	26-35	42	29-42	-0.002793438 -0.001485183 -0.000411419

Pieridae	lk null model	lk GMYC	lk ratio	LR test	number ML clusters	CI	Number ML entities	CI	threshold time
SINGLE									
Coal_Strict	1545.134	**1569.078**	47.88962	2.247822e-10***	18	18-19	18	18-20	-0.005019711
MULTIPLE									
Coal_Strict	1545.134	1570.916	51.5654	6.623458e-10***	25	18-25	26	18-26	-0.005019711 -0.003255062 -0.001794941

Papilionidae	lk null model	lk GMYC	lk ratio	LR test	number ML clusters	CI	Number ML entities	CI	threshold time
SINGLE									
Coal_Strict	209.8431	**217.1453**	14.60445	0.002187856**	4	4-5	5	5-6	-0.006500339
MULTIPLE									
Coal_Strict	209.8431	217.2163	14.74638	0.005257206**	4	3-4	6	6-7	-0.006500339 -0.002385335

Supplementary Table S5. Cases of GMYC failure and their cause. The numbers in parentheses indicate the number of GMYC entities recovered through oversplitting.

Genus	Species	Coalescent – Relaxed Clock	Coalescent – Strict Clock	Yule – Relaxed Clock	Yule - Strict Clock
Papilio	*machaon*	oversplitting (2)	oversplitting (2)	oversplitting (2)	oversplitting (2)
Colias	*erate*	conspecific (*C. crocea*)	conspecific (*C. crocea*)	conspecific (*C. crocea*)	conspecific (*C. crocea*)
Colias	*crocea*	conspecific (*C. erate*)	conspecific (*C. erate*)	conspecific (*C. erate*)	conspecific (*C. erate*)
Pieris	*napi*	paraphyly (*P. bryoniae*)	paraphyly (*P. bryoniae*)	paraphyly (*P. bryoniae*)	paraphyly (*P. bryoniae*)
Pieris	*bryoniae*	paraphyly (*P. napi*)	paraphyly (*P. napi*)	paraphyly (*P. napi*)	paraphyly (*P. napi*)
Boloria	*euphrosyne*	oversplitting (2)	oversplitting (2)	oversplitting (2)	oversplitting (2)
Brenthis	*daphne*	oversplitting (2)	oversplitting (2)	oversplitting (2)	oversplitting (2)
Apatura	*ilia*	conspecific (*A. metis*)	conspecific (*A. metis*)	conspecific (*A. metis*)	conspecific (*A. metis*)
Apatura	*metis*	conspecific (*A. ilia*)	conspecific (*A. ilia*)	conspecific (*A. ilia*)	conspecific (*A. ilia*)
Argynnis	*aglaja*	oversplitting (2)			
Argynnis	*pandora*	oversplitting (2)	oversplitting (2)	oversplitting (2)	oversplitting (2)
Melitaea	*athalia*	oversplitting (2)		oversplitting (2)	oversplitting (2)
Melitaea	*aurelia*	oversplitting (2)	oversplitting (2)	oversplitting (2)	oversplitting (2)
Coenonympha	*arcania*	conspecific (*C. leander*)	conspecific (*C. leander*)	conspecific (*C. leander*)	conspecific (*C. leander*)
Coenonympha	*leander*	conspecific (*C. arcania*)	conspecific (*C. arcania*)	conspecific (*C. arcania*)	conspecific (*C. arcania*)
Coenonympha	*tullia*	conspecific (*C. rhodopensis*)	conspecific (*C. rhodopensis*)	conspecific (*C. rhodopensis*)	conspecific (*C. rhodopensis*)
Coenonympha	*rhodopensis*	conspecific (*C. tullia*)	conspecific (*C. tullia*)	conspecific (*C. tullia*)	conspecific (*C. tullia*)
Erebia	*ligea*	conspecific (*E. euryale*)	conspecific (*E. euryale*)	conspecific (*E. euryale*)	conspecific (*E. euryale*)
Erebia	*euryale*	conspecific (*E. ligea*)	conspecific (*E. ligea*)	conspecific (*E. ligea*)	conspecific (*E. ligea*)
Hipparchia	*fagi*	conspecific (*H. syriaca*)	conspecific (*H. syriaca*)	conspecific (*H. syriaca*)	conspecific (*H. syriaca*)
Hipparchia	*syriaca*	conspecific (*H. fagi*)	conspecific (*H. fagi*)	conspecific (*H. fagi*)	conspecific (*H. fagi*)
Hipparchia	*semele*	oversplitting (2)	oversplitting (2)	oversplitting (2)	oversplitting (2)
Lasiommata	*maera*	oversplitting (2)	oversplitting (2)		oversplitting (2)
Cupido	*alcetas*	conspecific (*C. decolorata*)	conspecific (*C. decolorata*)	conspecific (*C. decolorata*)	conspecific (*C. decolorata*)
Cupido	*decolorata*	conspecific (*C. alcetas*)	conspecific (*C. alcetas*)	conspecific (*C. alcetas*)	conspecific (*C. alcetas*)

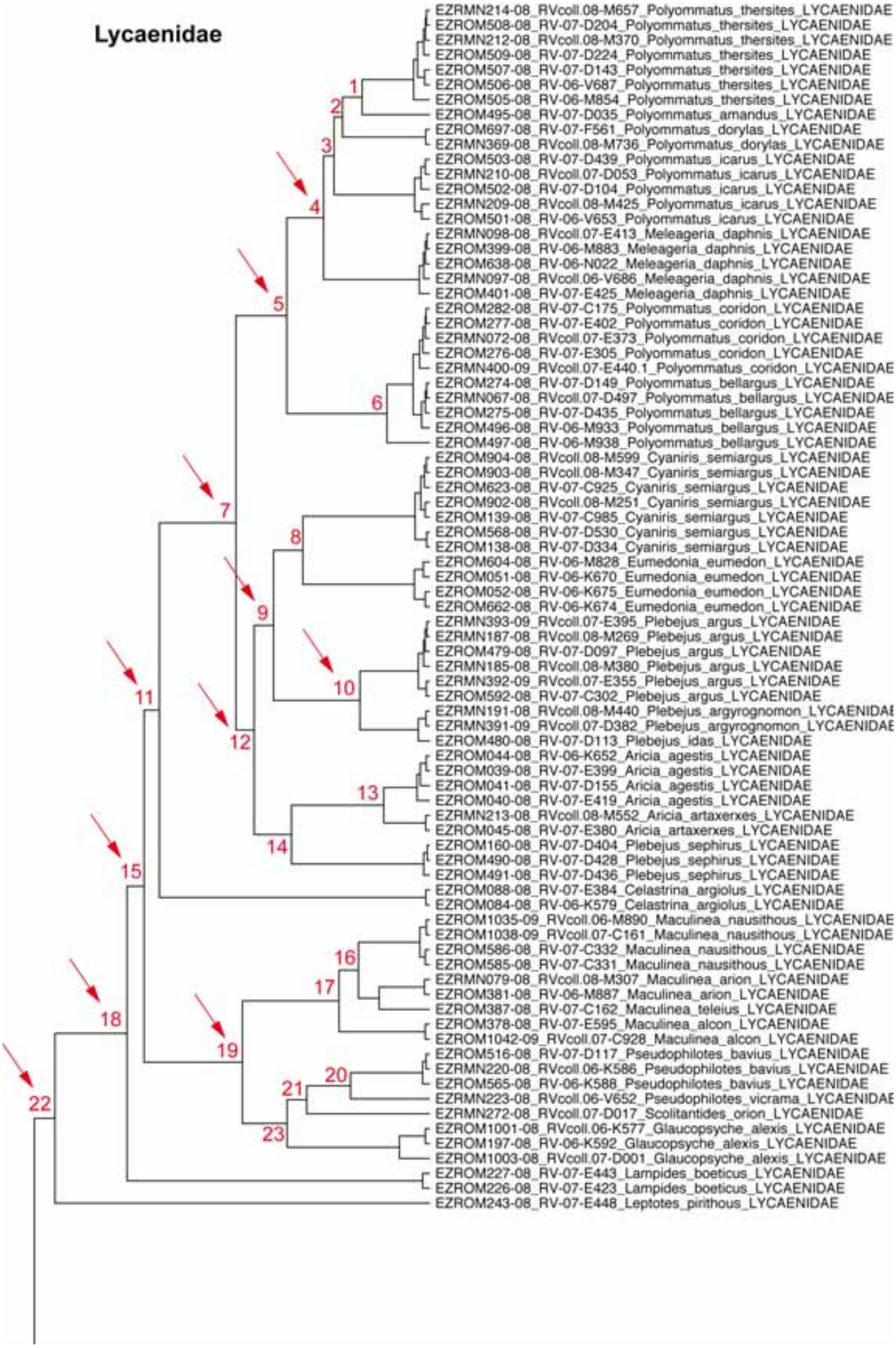
Genus	**Species**	**Coalescent – Relaxed Clock**	**Coalescent – Strict Clock**	**Yule – Relaxed Clock**	**Yule - Strict Clock**
Cupido	*osiris*	oversplitting (2)	oversplitting (2)	oversplitting (2)	oversplitting (2)
Glaucopsyche	*alexis*	oversplitting (2)	oversplitting (2)	oversplitting (2)	oversplitting (2)
Lysandra	*bellargus*	paraphyly (*L. coridon*)	paraphyly (*L. coridon*)	paraphyly (*L. coridon*)	paraphyly (*L. coridon*)
Lysandra	*coridon*	paraphyly (*L. bellargus*)	paraphyly (*L. bellargus*	paraphyly (*L. bellargus*)	paraphyly (*L. bellargus*)
Plebejus	*argyrognomon*	conspecific (*P. idas*)	conspecific (*P. idas*)	conspecific (*P. idas*)	conspecific (*P. idas*)
Plebejus	*idas*	conspecific (*P. argyrognomon*)	conspecific (*P. argyrognomon*)	conspecific (*P. argyrognomon*)	conspecific (*P. argyrognomon*)
Carcharodus	*alceae*	oversplitting (2)	oversplitting (2)	oversplitting (2)	oversplitting (2)
Carcharodus	*orientalis*	paraphyly (*C. flocciferus*)	paraphyly (*C. flocciferus*)	paraphyly (*C. flocciferus*)	paraphyly (*C. flocciferus*)
Carcharodus	*flocciferus*	paraphyly (*C. orientalis*)	paraphyly (*C. orientalis*)	paraphyly (*C. orientalis*)	paraphyly (*C. orientalis*)
Pyrgus	*armoricanus*	oversplitting (2)	oversplitting (2)	oversplitting (2)	oversplitting (2)
Thymelicus	*sylvestris*	oversplitting (3)	oversplitting (3)	oversplitting (3)	oversplitting (3)
	General failure	**36 taxa (20,45%)**	**34 taxa (19,31%)**	**34 taxa (19,31%)**	**35 taxa (19,89)**
	Only GMYC failure	**20 taxa (11,36%)**	**18 taxa (10,23%)**	**18 taxa (10,23%)**	**19 taxa (10,80%)**

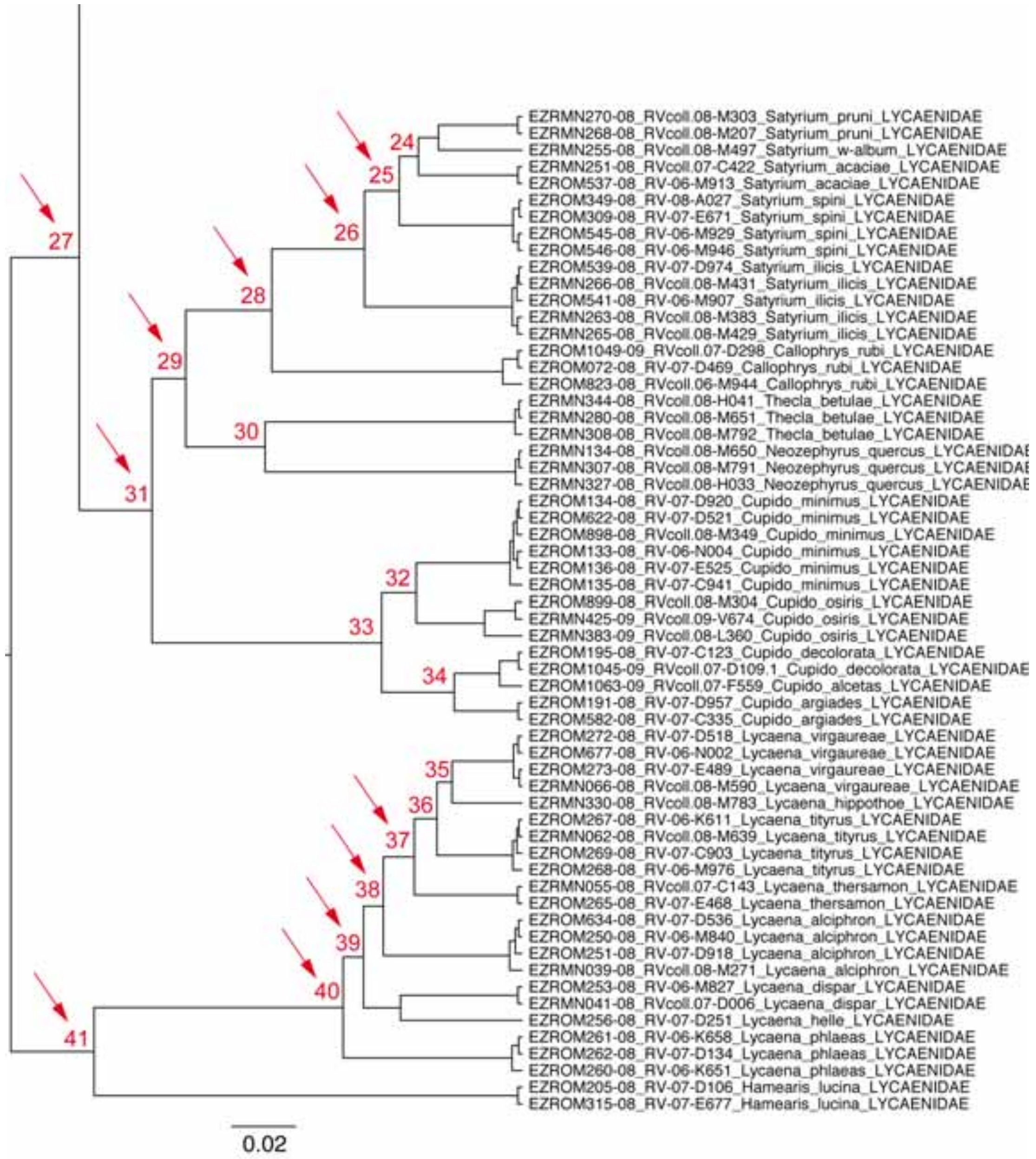
Supplementary Table S6. Comparison between GMYC number of entities for the full family tree and each significant subtree. Clade numbers correspond to the ones shown in Supplementary Figure 1. Levels of significances are (*): Significant, 0.01 to 0.05; (**): Very significant, 0.001 to 0.01; (***) Extremely significant, <0.001.

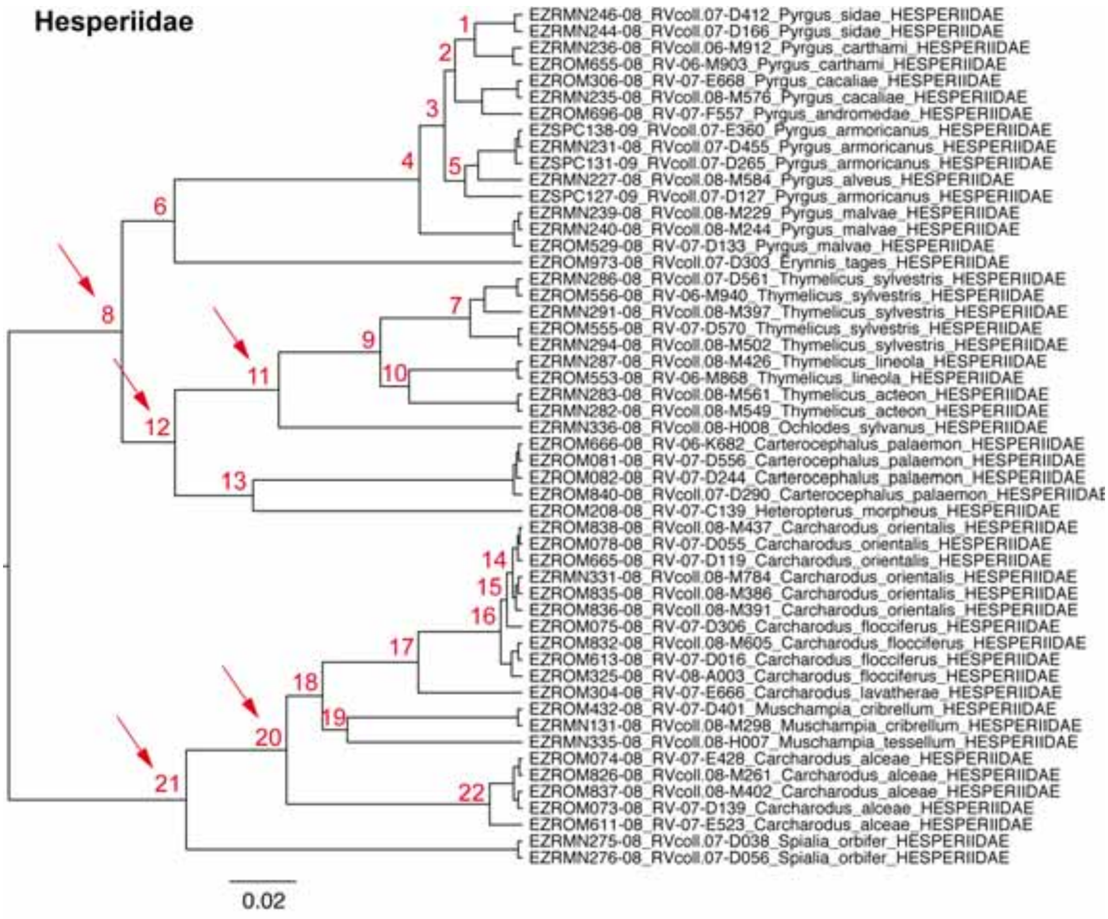
Family	Clade Number	Level of GMYC significance	Number of entities for full tree	Number of entities for clade tree	CI	Comment
Pieridae	3	*	5	5	5-10	
Pieridae	5	**	6	6	6-7	
Pieridae	6	**	7	7	6-7	
Pieridae	7	***	10	10	10-10	
Nymphalidae	5	**	5	5	4-6	
Nymphalidae	9	***	11	11	11-12	
Nymphalidae	11	*	4	4	4-4	
Nymphalidae	12	***	13	13	13-14	
Nymphalidae	13	***	6	6	6-6	
Nymphalidae	17	***	17	17	17-18	
Nymphalidae	19	*	3	3	3-3	
Nymphalidae	21	**	4	4	4-4	
Nymphalidae	22	***	25	25	25-26	
Nymphalidae	29	***	26	26	26-27	
Nymphalidae	30	***	6	6	6-6	
Nymphalidae	32	***	8	8	8-8	
Nymphalidae	34	***	32	32	32-33	
Nymphalidae	**39**	******	**6**	**5**	**3-6**	*Lasiommata maera* oversplitted (two entities) in the full tree
Nymphalidae	**40**	*******	**36**	**37**	**36-38**	*Coenonympha arcania* and *C. leander* recovered as conspecific in the full tree
Nymphalidae	44	*	3	3	3-9	
Nymphalidae	**46**	******	**4**	**5**	**4-5**	*Coenonympha arcania* and *C. leander* recovered as conspecific in the full tree
Nymphalidae	56	***	64	64	63-66	
Nymphalidae	**58**	*******	**10**	**9**	**8-10**	*Melitaea aurelia* oversplitted (two entities) in the full tree
Nymphalidae	61	*	3	3	3-3	
Nymphalidae	62	**	4	4	4-4	
Nymphalidae	63	***	12	12	10-12	
Nymphalidae	67	***	16	16	14-16	
Nymphalidae	74	***	26	26	25-27	
Nymphalidae	78	*	7	7	7-8	
Nymphalidae	79	***	9	9	9-9	
Nymphalidae	80	***	28	28	27-29	
Nymphalidae	82	***	10	10	10-10	

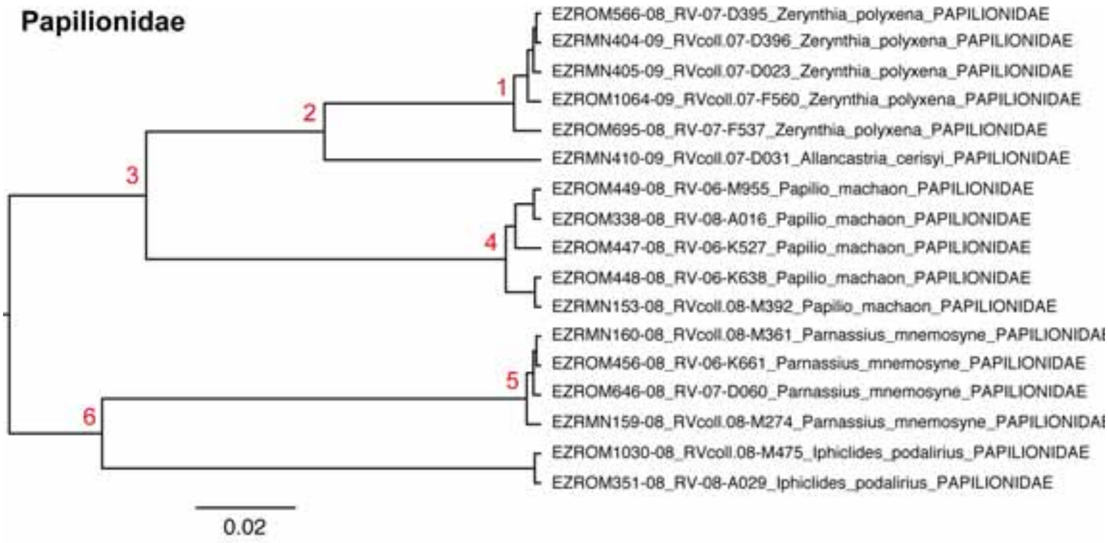
Nymphalidae	**89**	*****	**7**	**6**	**5-8**	*Argynnis pandora* oversplitted (two entities) in the full tree
Nymphalidae	**92**	******	**12**	**9**	**4-15**	*Argynnis pandora* oversplitted (two entities) in the full tree *Brenthis daphne* oversplitted (two entities) in the full tree *Brenthis daphne* and *B. ino* recovered as conspecific in the subclade tree
Nymphalidae	**99**	*******	**19**	**21**	**15-22**	*Argynnis aglaja* oversplitted (two entities) in the subclade tree *Brenthis ino* oversplitted (two entities) in the subclade tree
Nymphalidae	101	*	5	5	5-6	
Nymphalidae	**103**	*******	**21**	**22**	**18-24**	*Argynnis aglaja* oversplitted (two entities)in the subclade tree
Nymphalidae	105	*	7	7	5-8	
Lycaenidae	4	**	5	5	5-6	
Lycaenidae	5	**	7	7	7-8	
Lycaenidae	7	***	14	14	14-15	
Lycaenidae	9	*	4	4	4-5	
Lycaenidae	**10**	*****	**2**	**3**	**3-3**	*Plebejus argus* oversplitted (two entities)in the subclade tree
Lycaenidae	11	***	15	15	15-16	
Lycaenidae	12	***	7	7	7-8	
Lycaenidae	15	***	24	24	24-26	
Lycaenidae	18	***	25	25	25-27	
Lycaenidae	19	*	9	9	8-10	
Lycaenidae	22	***	26	26	26-28	
Lycaenidae	25	*	4	4	4-4	
Lycaenidae	26	**	5	5	5-6	
Lycaenidae	27	***	39	39	39-41	
Lycaenidae	28	*	6	6	6-8	
Lycaenidae	29	***	8	8	8-9	
Lycaenidae	31	***	13	13	13-16	
Lycaenidae	37	*	4	4	4-5	
Lycaenidae	38	*	5	5	5-6	
Lycaenidae	39	***	7	7	7-8	
Lycaenidae	40	***	8	8	8-9	
Lycaenidae	41	***	9	9	9-10	
Hesperiidae	8	**	17	17	3-18	
Hesperiidae	**11**	*****	**6**	**5**	**5-6**	*Thymelicus sylvestris* oversplitted in two entities in the subclade tree and in three entities in the full tree
Hesperiidae	12	*	8	8	7-9	
Hesperiidae	**20**	*****	**8**	**5**	**4-8**	*Carcharodus alceae* oversplitted (two entities) in the full tree *Carcharodus floccifera* oversplitted (two entities) in the full tree *Carcharodus floccifera* and *C. orientalis* recovered as conspecific in the subclade tree
Hesperiidae	**21**	*****	**9**	**6**	**5-9**	*Carcharodus alceae* oversplitted (two entities) in the full tree *Carcharodus floccifera* oversplitted (two entities) in the full tree *Carcharodus floccifera* and *C. orientalis* recovered as conspecific in the subclade tree

Supplementary Figure S1. Family trees with numbers assigned to nodes and arrows showing a significant pattern to be analyzed by GMYC.

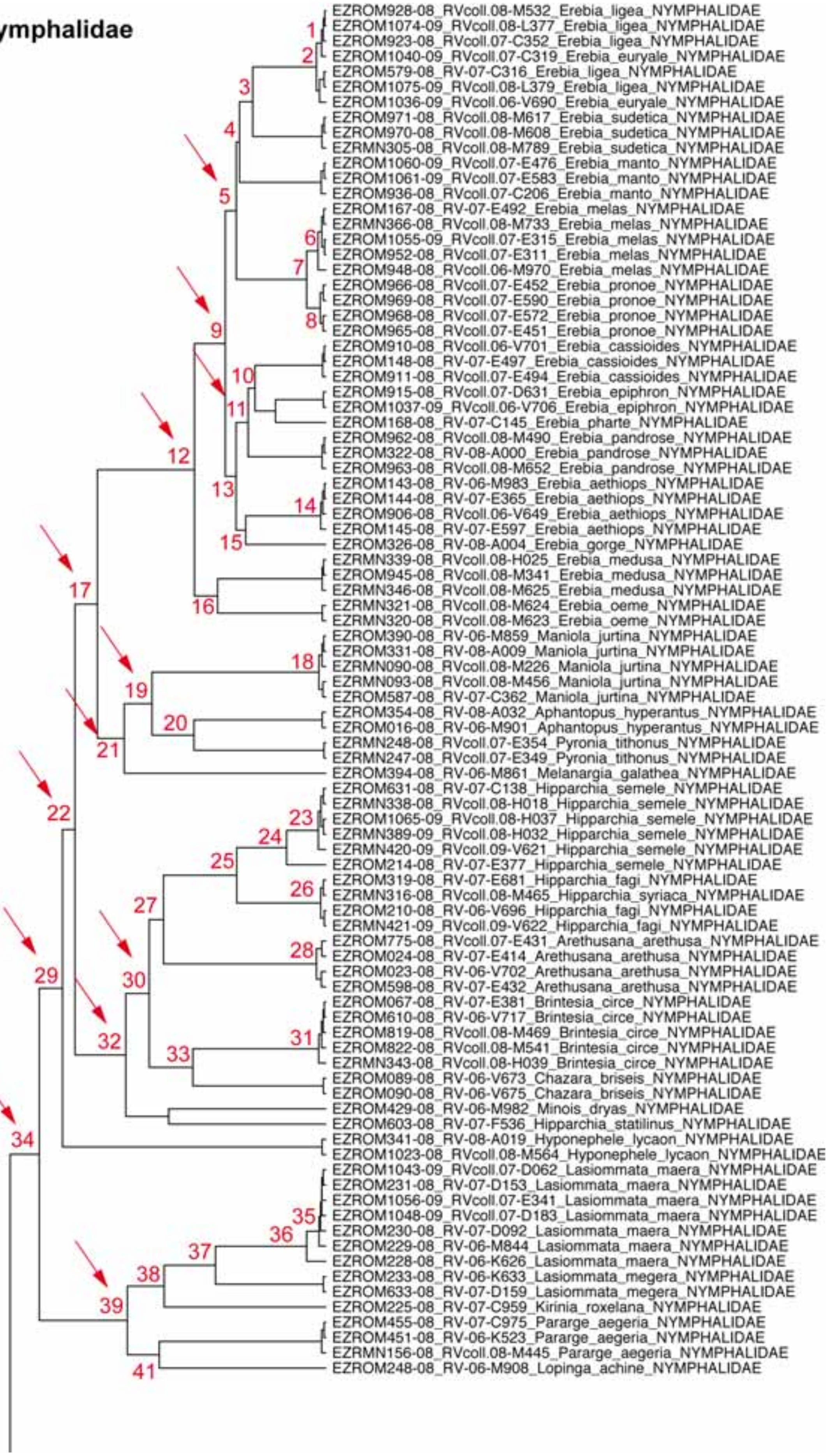

Lycaenidae
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
EZRMN214-08_RVcoll.08-M657_Polyommatus_thersites_LYCAENIDAE
EZROM508-08_RV-07-D204_Polyommatus_thersites_LYCAENIDAE
EZRMN212-08_RVcoll.08-M370_Polyommatus_thersites_LYCAENIDAE
EZROM509-08_RV-07-D224_Polyommatus_thersites_LYCAENIDAE
EZROM507-08_RV-07-D143_Polyommatus_thersites_LYCAENIDAE
EZROM506-08_RV-06-V687_Polyommatus_thersites_LYCAENIDAE
EZROM505-08_RV-06-M854_Polyommatus_thersites_LYCAENIDAE
EZROM495-08_RV-07-D035_Polyommatus_amandus_LYCAENIDAE
EZROM697-08_RV-07-F561_Polyommatus_dorylas_LYCAENIDAE
EZRMN369-08_RVcoll.08-M736_Polyommatus_dorylas_LYCAENIDAE
EZROM503-08_RV-07-D439_Polyommatus_icarus_LYCAENIDAE
EZRMN210-08_RVcoll.07-D053_Polyommatus_icarus_LYCAENIDAE
EZROM502-08_RV-07-D104_Polyommatus_icarus_LYCAENIDAE
EZRMN209-08_RVcoll.08-M425_Polyommatus_icarus_LYCAENIDAE
EZROM501-08_RV-06-V653_Polyommatus_icarus_LYCAENIDAE
EZRMN098-08_RVcoll.07-E413_Meleageria_daphnis_LYCAENIDAE
EZROM399-08_RV-06-M883_Meleageria_daphnis_LYCAENIDAE
EZROM638-08_RV-06-N022_Meleageria_daphnis_LYCAENIDAE
EZRMN097-08_RVcoll.06-V686_Meleageria_daphnis_LYCAENIDAE
EZROM401-08_RV-07-E425_Meleageria_daphnis_LYCAENIDAE
EZROM282-08_RV-07-C175_Polyommatus_coridon_LYCAENIDAE
EZROM277-08_RV-07-E402_Polyommatus_coridon_LYCAENIDAE
EZRMN072-08_RVcoll.07-E373_Polyommatus_coridon_LYCAENIDAE
EZROM276-08_RV-07-E305_Polyommatus_coridon_LYCAENIDAE
EZRMN400-09_RVcoll.07-E440.1_Polyommatus_coridon_LYCAENIDAE
EZROM274-08_RV-07-D149_Polyommatus_bellargus_LYCAENIDAE
EZRMN067-08_RVcoll.07-D497_Polyommatus_bellargus_LYCAENIDAE
EZROM275-08_RV-07-D435_Polyommatus_bellargus_LYCAENIDAE
EZROM496-08_RV-06-M933_Polyommatus_bellargus_LYCAENIDAE
EZROM497-08_RV-06-M938_Polyommatus_bellargus_LYCAENIDAE
EZROM904-08_RVcoll.08-M599_Cyaniris_semiargus_LYCAENIDAE
EZROM903-08_RVcoll.08-M347_Cyaniris_semiargus_LYCAENIDAE
EZROM623-08_RV-07-C925_Cyaniris_semiargus_LYCAENIDAE
EZROM902-08_RVcoll.08-M251_Cyaniris_semiargus_LYCAENIDAE
EZROM139-08_RV-07-C985_Cyaniris_semiargus_LYCAENIDAE
EZROM568-08_RV-07-D530_Cyaniris_semiargus_LYCAENIDAE
EZROM138-08_RV-07-D334_Cyaniris_semiargus_LYCAENIDAE
EZROM604-08_RV-06-M828_Eumedonia_eumedon_LYCAENIDAE
EZROM051-08_RV-06-K670_Eumedonia_eumedon_LYCAENIDAE
EZROM052-08_RV-06-K675_Eumedonia_eumedon_LYCAENIDAE
EZROM662-08_RV-06-K674_Eumedonia_eumedon_LYCAENIDAE
EZRMN393-09_RVcoll.07-E395_Plebejus_argus_LYCAENIDAE
EZRMN187-08_RVcoll.08-M269_Plebejus_argus_LYCAENIDAE
EZROM479-08_RV-07-D097_Plebejus_argus_LYCAENIDAE
EZRMN185-08_RVcoll.08-M380_Plebejus_argus_LYCAENIDAE
EZRMN392-09_RVcoll.07-E355_Plebejus_argus_LYCAENIDAE
EZROM592-08_RV-07-C302_Plebejus_argus_LYCAENIDAE
EZRMN191-08_RVcoll.08-M440_Plebejus_argyrognomon_LYCAENIDAE
EZRMN391-09_RVcoll.07-D382_Plebejus_argyrognomon_LYCAENIDAE
EZROM480-08_RV-07-D113_Plebejus_idas_LYCAENIDAE
EZROM044-08_RV-06-K652_Aricia_agestis_LYCAENIDAE
EZROM039-08_RV-07-E399_Aricia_agestis_LYCAENIDAE
EZROM041-08_RV-07-D155_Aricia_agestis_LYCAENIDAE
EZROM040-08_RV-07-E419_Aricia_agestis_LYCAENIDAE
EZRMN213-08_RVcoll.08-M552_Aricia_artaxerxes_LYCAENIDAE
EZROM045-08_RV-07-E380_Aricia_artaxerxes_LYCAENIDAE
EZROM160-08_RV-07-D404_Plebejus_sephirus_LYCAENIDAE
EZROM490-08_RV-07-D428_Plebejus_sephirus_LYCAENIDAE
EZROM491-08_RV-07-D436_Plebejus_sephirus_LYCAENIDAE
EZROM088-08_RV-07-E384_Celastrina_argiolus_LYCAENIDAE
EZROM084-08_RV-06-K579_Celastrina_argiolus_LYCAENIDAE
EZROM1035-09_RVcoll.06-M890_Maculinea_nausithous_LYCAENIDAE
EZROM1038-09_RVcoll.07-C161_Maculinea_nausithous_LYCAENIDAE
EZROM586-08_RV-07-C332_Maculinea_nausithous_LYCAENIDAE
EZROM585-08_RV-07-C331_Maculinea_nausithous_LYCAENIDAE
EZRMN079-08_RVcoll.08-M307_Maculinea_arion_LYCAENIDAE
EZROM381-08_RV-06-M887_Maculinea_arion_LYCAENIDAE
EZROM387-08_RV-07-C162_Maculinea_teleius_LYCAENIDAE
EZROM378-08_RV-07-E595_Maculinea_alcon_LYCAENIDAE
EZROM1042-09_RVcoll.07-C928_Maculinea_alcon_LYCAENIDAE
EZROM516-08_RV-07-D117_Pseudophilotes_bavius_LYCAENIDAE
EZRMN220-08_RVcoll.06-K586_Pseudophilotes_bavius_LYCAENIDAE
EZROM565-08_RV-06-K588_Pseudophilotes_bavius_LYCAENIDAE
EZRMN223-08_RVcoll.06-V652_Pseudophilotes_vicrama_LYCAENIDAE
EZRMN272-08_RVcoll.07-D017_Scolitantides_orion_LYCAENIDAE
EZROM1001-08_RVcoll.06-K577_Glaucopsyche_alexis_LYCAENIDAE
EZROM197-08_RV-06-K592_Glaucopsyche_alexis_LYCAENIDAE
EZROM1003-08_RVcoll.07-D001_Glaucopsyche_alexis_LYCAENIDAE
EZROM227-08_RV-07-E443_Lampides_boeticus_LYCAENIDAE
EZROM226-08_RV-07-E423_Lampides_boeticus_LYCAENIDAE
EZROM243-08_RV-07-E448_Leptotes_pirithous_LYCAENIDAE

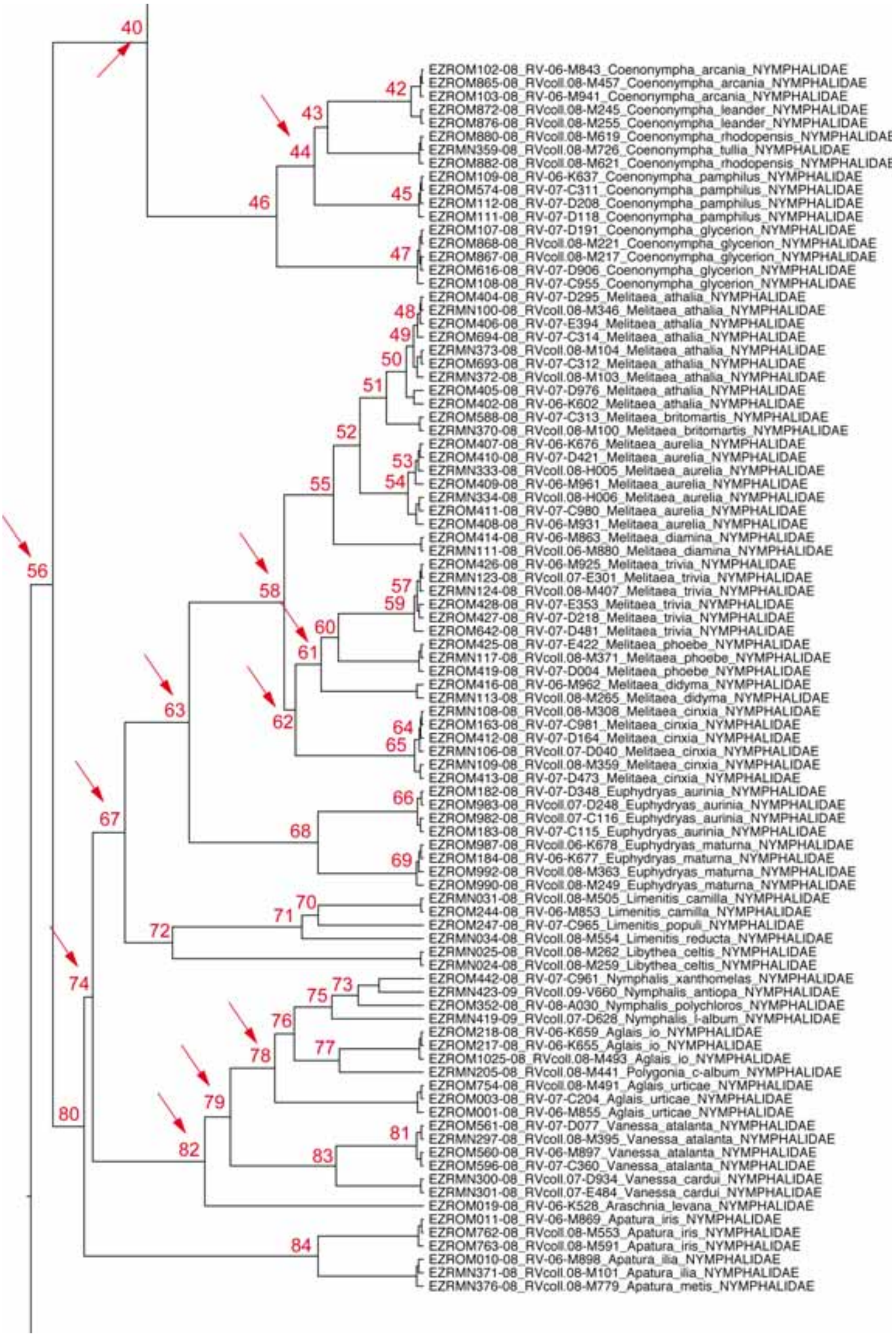

24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
EZRMN270-08_RVcoll.08-M303_Satyrium_pruni_LYCAENIDAE
EZRMN268-08_RVcoll.08-M207_Satyrium_pruni_LYCAENIDAE
EZRMN255-08_RVcoll.08-M497_Satyrium_w-album_LYCAENIDAE
EZRMN251-08_RVcoll.07-C422_Satyrium_acaciae_LYCAENIDAE
EZROM537-08_RV-06-M913_Satyrium_acaciae_LYCAENIDAE
EZROM349-08_RV-08-A027_Satyrium_spini_LYCAENIDAE
EZROM309-08_RV-07-E671_Satyrium_spini_LYCAENIDAE
EZROM545-08_RV-06-M929_Satyrium_spini_LYCAENIDAE
EZROM546-08_RV-06-M946_Satyrium_spini_LYCAENIDAE
EZROM539-08_RV-07-D974_Satyrium_ilicis_LYCAENIDAE
EZRMN266-08_RVcoll.08-M431_Satyrium_ilicis_LYCAENIDAE
EZROM541-08_RV-06-M907_Satyrium_ilicis_LYCAENIDAE
EZRMN263-08_RVcoll.08-M383_Satyrium_ilicis_LYCAENIDAE
EZRMN265-08_RVcoll.08-M429_Satyrium_ilicis_LYCAENIDAE
EZROM1049-09_RVcoll.07-D298_Callophrys_rubi_LYCAENIDAE
EZROM072-08_RV-07-D469_Callophrys_rubi_LYCAENIDAE
EZROM823-08_RVcoll.06-M944_Callophrys_rubi_LYCAENIDAE
EZRMN344-08_RVcoll.08-H041_Thecla_betulae_LYCAENIDAE
EZRMN280-08_RVcoll.08-M651_Thecla_betulae_LYCAENIDAE
EZRMN308-08_RVcoll.08-M792_Thecla_betulae_LYCAENIDAE
EZRMN134-08_RVcoll.08-M650_Neozephyrus_quercus_LYCAENIDAE
EZRMN307-08_RVcoll.08-M791_Neozephyrus_quercus_LYCAENIDAE
EZRMN327-08_RVcoll.08-H033_Neozephyrus_quercus_LYCAENIDAE
EZROM134-08_RV-07-D920_Cupido_minimus_LYCAENIDAE
EZROM622-08_RV-07-D521_Cupido_minimus_LYCAENIDAE
EZROM898-08_RVcoll.08-M349_Cupido_minimus_LYCAENIDAE
EZROM133-08_RV-06-N004_Cupido_minimus_LYCAENIDAE
EZROM136-08_RV-07-E525_Cupido_minimus_LYCAENIDAE
EZROM135-08_RV-07-C941_Cupido_minimus_LYCAENIDAE
EZROM899-08_RVcoll.08-M304_Cupido_osiris_LYCAENIDAE
EZRMN425-09_RVcoll.09-V674_Cupido_osiris_LYCAENIDAE
EZRMN383-09_RVcoll.08-L360_Cupido_osiris_LYCAENIDAE
EZROM195-08_RV-07-C123_Cupido_decolorata_LYCAENIDAE
EZROM1045-09_RVcoll.07-D109.1_Cupido_decolorata_LYCAENIDAE
EZROM1063-09_RVcoll.07-F559_Cupido_alcetas_LYCAENIDAE
EZROM191-08_RV-07-D957_Cupido_argiades_LYCAENIDAE
EZROM582-08_RV-07-C335_Cupido_argiades_LYCAENIDAE
EZROM272-08_RV-07-D518_Lycaena_virgaureae_LYCAENIDAE
EZROM677-08_RV-06-N002_Lycaena_virgaureae_LYCAENIDAE
EZROM273-08_RV-07-E489_Lycaena_virgaureae_LYCAENIDAE
EZRMN066-08_RVcoll.08-M590_Lycaena_virgaureae_LYCAENIDAE
EZRMN330-08_RVcoll.08-M783_Lycaena_hippothoe_LYCAENIDAE
EZROM267-08_RV-06-K611_Lycaena_tityrus_LYCAENIDAE
EZRMN062-08_RVcoll.08-M639_Lycaena_tityrus_LYCAENIDAE
EZROM269-08_RV-07-C903_Lycaena_tityrus_LYCAENIDAE
EZROM268-08_RV-06-M976_Lycaena_tityrus_LYCAENIDAE
EZRMN055-08_RVcoll.07-C143_Lycaena_thersamon_LYCAENIDAE
EZROM265-08_RV-07-E468_Lycaena_thersamon_LYCAENIDAE
EZROM634-08_RV-07-D536_Lycaena_alciphron_LYCAENIDAE
EZROM250-08_RV-06-M840_Lycaena_alciphron_LYCAENIDAE
EZROM251-08_RV-07-D918_Lycaena_alciphron_LYCAENIDAE
EZRMN039-08_RVcoll.08-M271_Lycaena_alciphron_LYCAENIDAE
EZROM253-08_RV-06-M827_Lycaena_dispar_LYCAENIDAE
EZRMN041-08_RVcoll.07-D006_Lycaena_dispar_LYCAENIDAE
EZROM256-08_RV-07-D251_Lycaena_helle_LYCAENIDAE
EZROM261-08_RV-06-K658_Lycaena_phlaeas_LYCAENIDAE
EZROM262-08_RV-07-D134_Lycaena_phlaeas_LYCAENIDAE
EZROM260-08_RV-06-K651_Lycaena_phlaeas_LYCAENIDAE
EZROM205-08_RV-07-D106_Hamearis_lucina_LYCAENIDAE
EZROM315-08_RV-07-E677_Hamearis_lucina_LYCAENIDAE
0.02

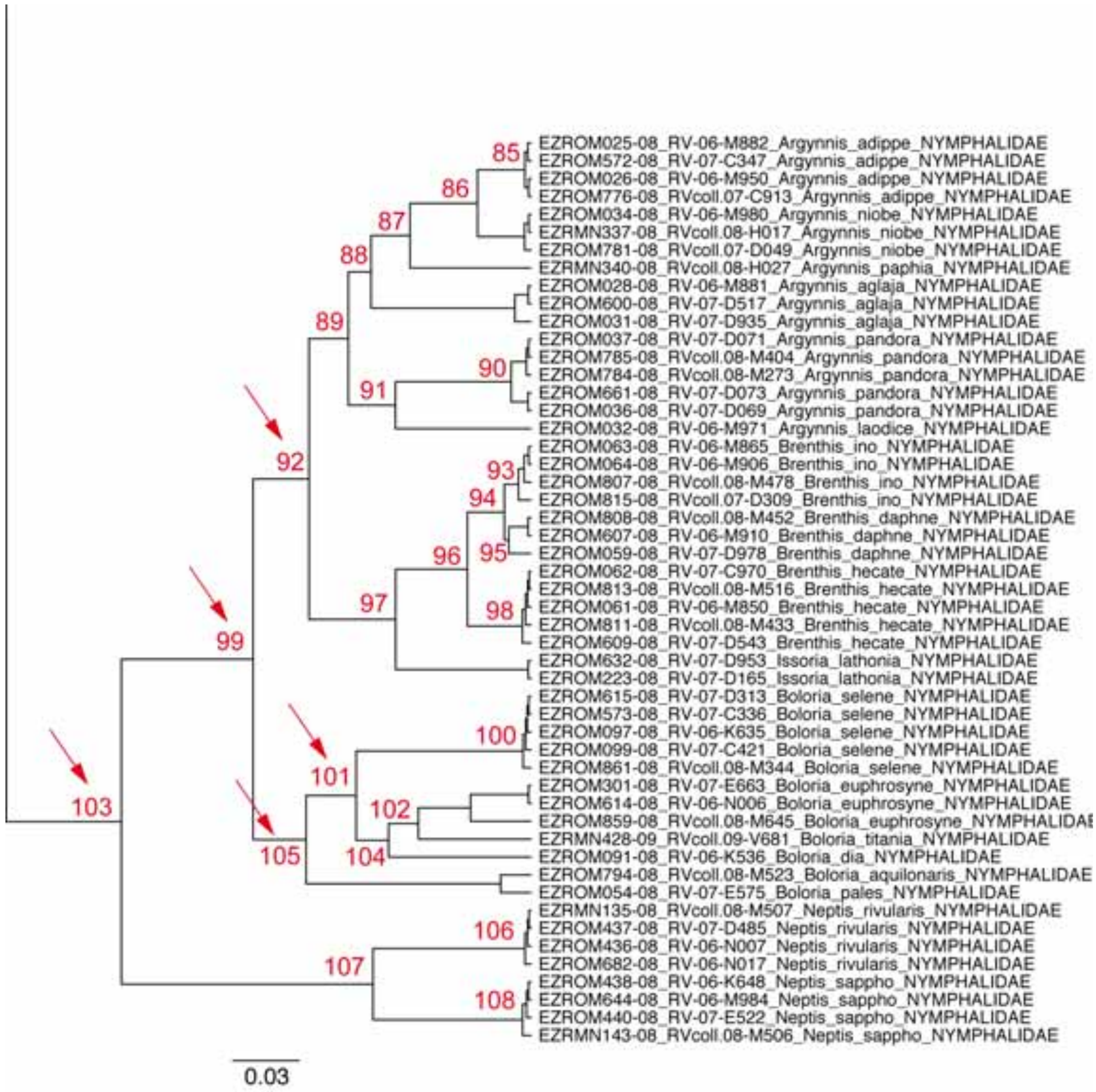

Hesperiidae
EZRMN246-08_RVcoll.07-D412_Pyrgus_sidae_HESPERIIDAE
EZRMN244-08_RVcoll.07-D166_Pyrgus_sidae_HESPERIIDAE
EZRMN236-08_RVcoll.06-M912_Pyrgus_carthami_HESPERIIDAE
EZROM655-08_RV-06-M903_Pyrgus_carthami_HESPERIIDAE
EZROM306-08_RV-07-E668_Pyrgus_cacaliae_HESPERIIDAE
EZRMN235-08_RVcoll.08-M576_Pyrgus_cacaliae_HESPERIIDAE
EZROM696-08_RV-07-F557_Pyrgus_andromedae_HESPERIIDAE
EZSPC138-09_RVcoll.07-E360_Pyrgus_armoricanus_HESPERIIDAE
EZRMN231-08_RVcoll.07-D455_Pyrgus_armoricanus_HESPERIIDAE
EZSPC131-09_RVcoll.07-D265_Pyrgus_armoricanus_HESPERIIDAE
EZRMN227-08_RVcoll.08-M584_Pyrgus_alveus_HESPERIIDAE
EZSPC127-09_RVcoll.07-D127_Pyrgus_armoricanus_HESPERIIDAE
EZRMN239-08_RVcoll.08-M229_Pyrgus_malvae_HESPERIIDAE
EZRMN240-08_RVcoll.08-M244_Pyrgus_malvae_HESPERIIDAE
EZROM529-08_RV-07-D133_Pyrgus_malvae_HESPERIIDAE
EZROM973-08_RVcoll.07-D303_Erynnis_tages_HESPERIIDAE
EZRMN286-08_RVcoll.07-D561_Thymelicus_sylvestris_HESPERIIDAE
EZROM556-08_RV-06-M940_Thymelicus_sylvestris_HESPERIIDAE
EZRMN291-08_RVcoll.08-M397_Thymelicus_sylvestris_HESPERIIDAE
EZROM555-08_RV-07-D570_Thymelicus_sylvestris_HESPERIIDAE
EZRMN294-08_RVcoll.08-M502_Thymelicus_sylvestris_HESPERIIDAE
EZRMN287-08_RVcoll.08-M426_Thymelicus_lineola_HESPERIIDAE
EZROM553-08_RV-06-M868_Thymelicus_lineola_HESPERIIDAE
EZRMN283-08_RVcoll.08-M561_Thymelicus_acteon_HESPERIIDAE
EZRMN282-08_RVcoll.08-M549_Thymelicus_acteon_HESPERIIDAE
EZRMN336-08_RVcoll.08-H008_Ochlodes_sylvanus_HESPERIIDAE
EZROM666-08_RV-06-K682_Carterocephalus_palaemon_HESPERIIDAE
EZROM081-08_RV-07-D556_Carterocephalus_palaemon_HESPERIIDAE
EZROM082-08_RV-07-D244_Carterocephalus_palaemon_HESPERIIDAE
EZROM840-08_RVcoll.07-D290_Carterocephalus_palaemon_HESPERIIDAE
EZROM208-08_RV-07-C139_Heteropterus_morpheus_HESPERIIDAE
EZROM838-08_RVcoll.08-M437_Carcharodus_orientalis_HESPERIIDAE
EZROM078-08_RV-07-D055_Carcharodus_orientalis_HESPERIIDAE
EZROM665-08_RV-07-D119_Carcharodus_orientalis_HESPERIIDAE
EZRMN331-08_RVcoll.08-M784_Carcharodus_orientalis_HESPERIIDAE
EZROM835-08_RVcoll.08-M386_Carcharodus_orientalis_HESPERIIDAE
EZROM836-08_RVcoll.08-M391_Carcharodus_orientalis_HESPERIIDAE
EZROM075-08_RV-07-D306_Carcharodus_flocciferus_HESPERIIDAE
EZROM832-08_RVcoll.08-M605_Carcharodus_flocciferus_HESPERIIDAE
EZROM613-08_RV-07-D016_Carcharodus_flocciferus_HESPERIIDAE
EZROM325-08_RV-08-A003_Carcharodus_flocciferus_HESPERIIDAE
EZROM304-08_RV-07-E666_Carcharodus_lavatherae_HESPERIIDAE
EZROM432-08_RV-07-D401_Muschampia_cribrellum_HESPERIIDAE
EZRMN131-08_RVcoll.08-M298_Muschampia_cribrellum_HESPERIIDAE
EZRMN335-08_RVcoll.08-H007_Muschampia_tessellum_HESPERIIDAE
EZROM074-08_RV-07-E428_Carcharodus_alceae_HESPERIIDAE
EZROM826-08_RVcoll.08-M261_Carcharodus_alceae_HESPERIIDAE
EZROM837-08_RVcoll.08-M402_Carcharodus_alceae_HESPERIIDAE
EZROM073-08_RV-07-D139_Carcharodus_alceae_HESPERIIDAE
EZROM611-08_RV-07-E523_Carcharodus_alceae_HESPERIIDAE
EZRMN275-08_RVcoll.07-D038_Spialia_orbifer_HESPERIIDAE
EZRMN276-08_RVcoll.07-D056_Spialia_orbifer_HESPERIIDAE
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
0.02

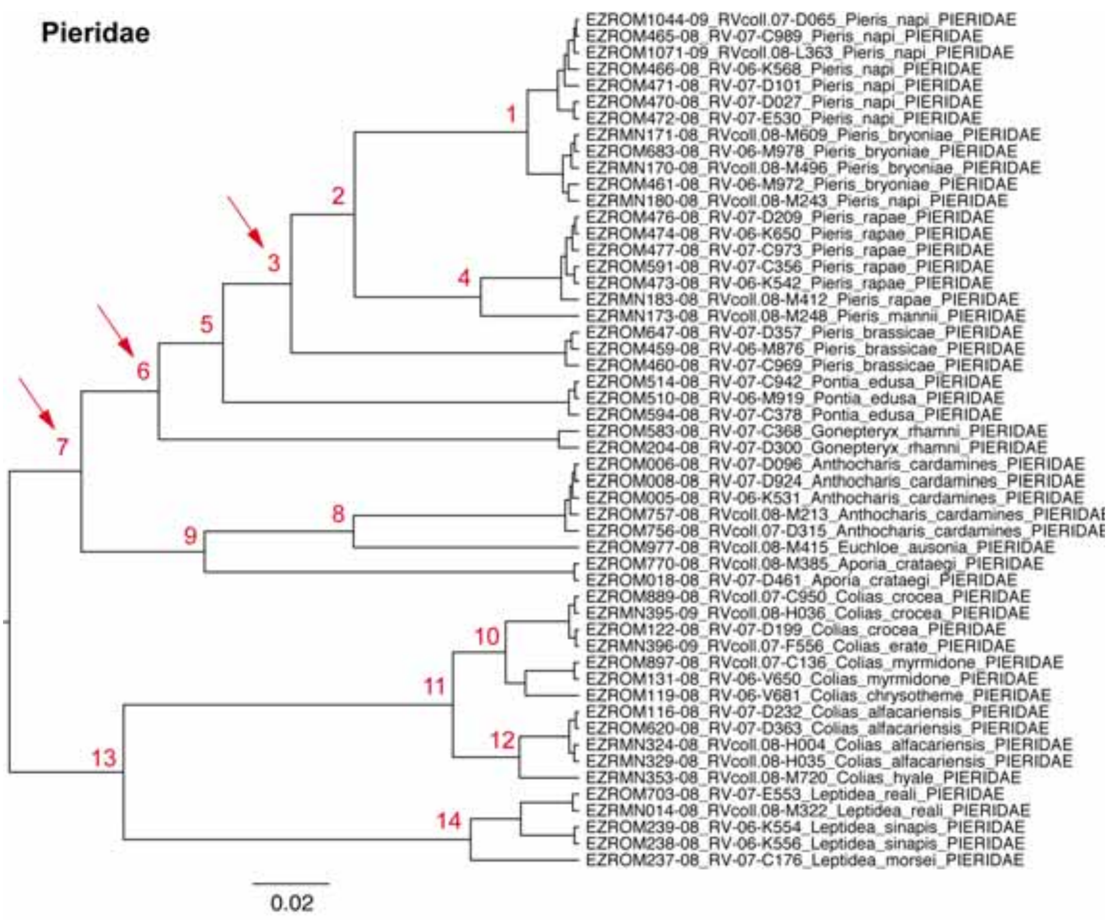

Papilionidae
EZROM566-08_RV-07-D395_Zerynthia_polyxena_PAPILIONIDAE
EZRMN404-09_RVcoll.07-D396_Zerynthia_polyxena_PAPILIONIDAE
EZRMN405-09_RVcoll.07-D023_Zerynthia_polyxena_PAPILIONIDAE
EZROM1064-09_RVcoll.07-F560_Zerynthia_polyxena_PAPILIONIDAE
EZROM695-08_RV-07-F537_Zerynthia_polyxena_PAPILIONIDAE
EZRMN410-09_RVcoll.07-D031_Allancastria_cerisyi_PAPILIONIDAE
EZROM449-08_RV-06-M955_Papilio_machaon_PAPILIONIDAE
EZROM338-08_RV-08-A016_Papilio_machaon_PAPILIONIDAE
EZROM447-08_RV-06-K527_Papilio_machaon_PAPILIONIDAE
EZROM448-08_RV-06-K638_Papilio_machaon_PAPILIONIDAE
EZRMN153-08_RVcoll.08-M392_Papilio_machaon_PAPILIONIDAE
EZRMN160-08_RVcoll.08-M361_Parnassius_mnemosyne_PAPILIONIDAE
EZROM456-08_RV-06-K661_Parnassius_mnemosyne_PAPILIONIDAE
EZROM646-08_RV-07-D060_Parnassius_mnemosyne_PAPILIONIDAE
EZRMN159-08_RVcoll.08-M274_Parnassius_mnemosyne_PAPILIONIDAE
EZROM1030-08_RVcoll.08-M475_Iphiclides_podalirius_PAPILIONIDAE
EZROM351-08_RV-08-A029_Iphiclides_podalirius_PAPILIONIDAE
1
2
3
4
5
6
0.02

Nymphalidae


EZROM928-08_RVcoll.08-M532_Erebia_ligea_NYMPHALIDAE
EZROM1074-09_RVcoll.08-L377_Erebia_ligea_NYMPHALIDAE
EZROM923-08_RVcoll.07-C352_Erebia_ligea_NYMPHALIDAE
EZROM1040-09_RVcoll.07-C319_Erebia_euryale_NYMPHALIDAE
EZROM579-08_RV-07-C316_Erebia_ligea_NYMPHALIDAE
EZROM1075-09_RVcoll.08-L379_Erebia_ligea_NYMPHALIDAE
EZROM1036-09_RVcoll.06-V690_Erebia_euryale_NYMPHALIDAE
EZROM971-08_RVcoll.08-M617_Erebia_sudetica_NYMPHALIDAE
EZROM970-08_RVcoll.08-M608_Erebia_sudetica_NYMPHALIDAE
EZRMN305-08_RVcoll.08-M789_Erebia_sudetica_NYMPHALIDAE
EZROM1060-09_RVcoll.07-E476_Erebia_manto_NYMPHALIDAE
EZROM1061-09_RVcoll.07-E583_Erebia_manto_NYMPHALIDAE
EZROM936-08_RVcoll.07-C206_Erebia_manto_NYMPHALIDAE
EZROM167-08_RV-07-E492_Erebia_melas_NYMPHALIDAE
EZRMN366-08_RVcoll.08-M733_Erebia_melas_NYMPHALIDAE
EZROM1055-09_RVcoll.07-E315_Erebia_melas_NYMPHALIDAE
EZROM952-08_RVcoll.07-E311_Erebia_melas_NYMPHALIDAE
EZROM948-08_RVcoll.06-M970_Erebia_melas_NYMPHALIDAE
EZROM966-08_RVcoll.07-E452_Erebia_pronoe_NYMPHALIDAE
EZROM969-08_RVcoll.07-E590_Erebia_pronoe_NYMPHALIDAE
EZROM968-08_RVcoll.07-E572_Erebia_pronoe_NYMPHALIDAE
EZROM965-08_RVcoll.07-E451_Erebia_pronoe_NYMPHALIDAE
EZROM910-08_RVcoll.06-V701_Erebia_cassioides_NYMPHALIDAE
EZROM148-08_RV-07-E497_Erebia_cassioides_NYMPHALIDAE
EZROM911-08_RVcoll.07-E494_Erebia_cassioides_NYMPHALIDAE
EZROM915-08_RVcoll.07-D631_Erebia_epiphron_NYMPHALIDAE
EZROM1037-09_RVcoll.06-V706_Erebia_epiphron_NYMPHALIDAE
EZROM168-08_RV-07-C145_Erebia_pharte_NYMPHALIDAE
EZROM962-08_RVcoll.08-M490_Erebia_pandrose_NYMPHALIDAE
EZROM322-08_RV-08-A000_Erebia_pandrose_NYMPHALIDAE
EZROM963-08_RVcoll.08-M652_Erebia_pandrose_NYMPHALIDAE
EZROM143-08_RV-06-M983_Erebia_aethiops_NYMPHALIDAE
EZROM144-08_RV-07-E365_Erebia_aethiops_NYMPHALIDAE
EZROM906-08_RVcoll.06-V649_Erebia_aethiops_NYMPHALIDAE
EZROM145-08_RV-07-E597_Erebia_aethiops_NYMPHALIDAE
EZROM326-08_RV-08-A004_Erebia_gorge_NYMPHALIDAE
EZRMN339-08_RVcoll.08-H025_Erebia_medusa_NYMPHALIDAE
EZROM945-08_RVcoll.08-M341_Erebia_medusa_NYMPHALIDAE
EZRMN346-08_RVcoll.08-M625_Erebia_medusa_NYMPHALIDAE
EZRMN321-08_RVcoll.08-M624_Erebia_oeme_NYMPHALIDAE
EZRMN320-08_RVcoll.08-M623_Erebia_oeme_NYMPHALIDAE
EZROM390-08_RV-06-M859_Maniola_jurtina_NYMPHALIDAE
EZROM331-08_RV-08-A009_Maniola_jurtina_NYMPHALIDAE
EZRMN090-08_RVcoll.08-M226_Maniola_jurtina_NYMPHALIDAE
EZRMN093-08_RVcoll.08-M456_Maniola_jurtina_NYMPHALIDAE
EZROM587-08_RV-07-C362_Maniola_jurtina_NYMPHALIDAE
EZROM354-08_RV-08-A032_Aphantopus_hyperantus_NYMPHALIDAE
EZROM016-08_RV-06-M901_Aphantopus_hyperantus_NYMPHALIDAE
EZRMN248-08_RVcoll.07-E354_Pyronia_tithonus_NYMPHALIDAE
EZRMN247-08_RVcoll.07-E349_Pyronia_tithonus_NYMPHALIDAE
EZROM394-08_RV-06-M861_Melanargia_galathea_NYMPHALIDAE
EZROM631-08_RV-07-C138_Hipparchia_semele_NYMPHALIDAE
EZRMN338-08_RVcoll.08-H018_Hipparchia_semele_NYMPHALIDAE
EZROM1065-09_RVcoll.08-H037_Hipparchia_semele_NYMPHALIDAE
EZRMN389-09_RVcoll.08-H032_Hipparchia_semele_NYMPHALIDAE
EZRMN420-09_RVcoll.09-V621_Hipparchia_semele_NYMPHALIDAE
EZROM214-08_RV-07-E377_Hipparchia_semele_NYMPHALIDAE
EZROM319-08_RV-07-E681_Hipparchia_fagi_NYMPHALIDAE
EZRMN316-08_RVcoll.08-M465_Hipparchia_syriaca_NYMPHALIDAE
EZROM210-08_RV-06-V696_Hipparchia_fagi_NYMPHALIDAE
EZRMN421-09_RVcoll.09-V622_Hipparchia_fagi_NYMPHALIDAE
EZROM775-08_RVcoll.07-E431_Arethusana_arethusa_NYMPHALIDAE
EZROM024-08_RV-07-E414_Arethusana_arethusa_NYMPHALIDAE
EZROM023-08_RV-06-V702_Arethusana_arethusa_NYMPHALIDAE
EZROM598-08_RV-07-E432_Arethusana_arethusa_NYMPHALIDAE
EZROM067-08_RV-07-E381_Brintesia_circe_NYMPHALIDAE
EZROM610-08_RV-06-V717_Brintesia_circe_NYMPHALIDAE
EZROM819-08_RVcoll.08-M469_Brintesia_circe_NYMPHALIDAE
EZROM822-08_RVcoll.08-M541_Brintesia_circe_NYMPHALIDAE
EZRMN343-08_RVcoll.08-H039_Brintesia_circe_NYMPHALIDAE
EZROM089-08_RV-06-V673_Chazara_briseis_NYMPHALIDAE
EZROM090-08_RV-06-V675_Chazara_briseis_NYMPHALIDAE
EZROM429-08_RV-06-M982_Minois_dryas_NYMPHALIDAE
EZROM603-08_RV-07-F536_Hipparchia_statilinus_NYMPHALIDAE
EZROM341-08_RV-08-A019_Hyponephele_lycaon_NYMPHALIDAE
EZROM1023-08_RVcoll.08-M564_Hyponephele_lycaon_NYMPHALIDAE
EZROM1043-09_RVcoll.07-D062_Lasiommata_maera_NYMPHALIDAE
EZROM231-08_RV-07-D153_Lasiommata_maera_NYMPHALIDAE
EZROM1056-09_RVcoll.07-E341_Lasiommata_maera_NYMPHALIDAE
EZROM1048-09_RVcoll.07-D183_Lasiommata_maera_NYMPHALIDAE
EZROM230-08_RV-07-D092_Lasiommata_maera_NYMPHALIDAE
EZROM229-08_RV-06-M844_Lasiommata_maera_NYMPHALIDAE
EZROM228-08_RV-06-K626_Lasiommata_maera_NYMPHALIDAE
EZROM233-08_RV-06-K633_Lasiommata_megera_NYMPHALIDAE
EZROM633-08_RV-07-D159_Lasiommata_megera_NYMPHALIDAE
EZROM225-08_RV-07-C959_Kirinia_roxelana_NYMPHALIDAE
EZROM455-08_RV-07-C975_Pararge_aegeria_NYMPHALIDAE
EZROM451-08_RV-06-K523_Pararge_aegeria_NYMPHALIDAE
EZRMN156-08_RVcoll.08-M445_Pararge_aegeria_NYMPHALIDAE
EZROM248-08_RV-06-M908_Lopinga_achine_NYMPHALIDAE
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
41


40
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
EZROM102-08_RV-06-M843_Coenonympha_arcania_NYMPHALIDAE
EZROM865-08_RVcoll.08-M457_Coenonympha_arcania_NYMPHALIDAE
EZROM103-08_RV-06-M941_Coenonympha_arcania_NYMPHALIDAE
EZROM872-08_RVcoll.08-M245_Coenonympha_leander_NYMPHALIDAE
EZROM876-08_RVcoll.08-M255_Coenonympha_leander_NYMPHALIDAE
EZROM880-08_RVcoll.08-M619_Coenonympha_rhodopensis_NYMPHALIDAE
EZRMN359-08_RVcoll.08-M726_Coenonympha_tullia_NYMPHALIDAE
EZROM882-08_RVcoll.08-M621_Coenonympha_rhodopensis_NYMPHALIDAE
EZROM109-08_RV-06-K637_Coenonympha_pamphilus_NYMPHALIDAE
EZROM574-08_RV-07-C311_Coenonympha_pamphilus_NYMPHALIDAE
EZROM112-08_RV-07-D208_Coenonympha_pamphilus_NYMPHALIDAE
EZROM111-08_RV-07-D118_Coenonympha_pamphilus_NYMPHALIDAE
EZROM107-08_RV-07-D191_Coenonympha_glycerion_NYMPHALIDAE
EZROM868-08_RVcoll.08-M221_Coenonympha_glycerion_NYMPHALIDAE
EZROM867-08_RVcoll.08-M217_Coenonympha_glycerion_NYMPHALIDAE
EZROM616-08_RV-07-D906_Coenonympha_glycerion_NYMPHALIDAE
EZROM108-08_RV-07-C955_Coenonympha_glycerion_NYMPHALIDAE
EZROM404-08_RV-07-D295_Melitaea_athalia_NYMPHALIDAE
EZRMN100-08_RVcoll.08-M346_Melitaea_athalia_NYMPHALIDAE
EZROM406-08_RV-07-E394_Melitaea_athalia_NYMPHALIDAE
EZROM694-08_RV-07-C314_Melitaea_athalia_NYMPHALIDAE
EZRMN373-08_RVcoll.08-M104_Melitaea_athalia_NYMPHALIDAE
EZROM693-08_RV-07-C312_Melitaea_athalia_NYMPHALIDAE
EZRMN372-08_RVcoll.08-M103_Melitaea_athalia_NYMPHALIDAE
EZROM405-08_RV-07-D976_Melitaea_athalia_NYMPHALIDAE
EZROM402-08_RV-06-K602_Melitaea_athalia_NYMPHALIDAE
EZROM588-08_RV-07-C313_Melitaea_britomartis_NYMPHALIDAE
EZRMN370-08_RVcoll.08-M100_Melitaea_britomartis_NYMPHALIDAE
EZROM407-08_RV-06-K676_Melitaea_aurelia_NYMPHALIDAE
EZROM410-08_RV-07-D421_Melitaea_aurelia_NYMPHALIDAE
EZRMN333-08_RVcoll.08-H005_Melitaea_aurelia_NYMPHALIDAE
EZROM409-08_RV-06-M961_Melitaea_aurelia_NYMPHALIDAE
EZRMN334-08_RVcoll.08-H006_Melitaea_aurelia_NYMPHALIDAE
EZROM411-08_RV-07-C980_Melitaea_aurelia_NYMPHALIDAE
EZROM408-08_RV-06-M931_Melitaea_aurelia_NYMPHALIDAE
EZROM414-08_RV-06-M863_Melitaea_diamina_NYMPHALIDAE
EZRMN111-08_RVcoll.06-M880_Melitaea_diamina_NYMPHALIDAE
EZROM426-08_RV-06-M925_Melitaea_trivia_NYMPHALIDAE
EZRMN123-08_RVcoll.07-E301_Melitaea_trivia_NYMPHALIDAE
EZRMN124-08_RVcoll.08-M407_Melitaea_trivia_NYMPHALIDAE
EZROM428-08_RV-07-E353_Melitaea_trivia_NYMPHALIDAE
EZROM427-08_RV-07-D218_Melitaea_trivia_NYMPHALIDAE
EZROM642-08_RV-07-D481_Melitaea_trivia_NYMPHALIDAE
EZROM425-08_RV-07-E422_Melitaea_phoebe_NYMPHALIDAE
EZRMN117-08_RVcoll.08-M371_Melitaea_phoebe_NYMPHALIDAE
EZROM419-08_RV-07-D004_Melitaea_phoebe_NYMPHALIDAE
EZROM416-08_RV-06-M962_Melitaea_didyma_NYMPHALIDAE
EZRMN113-08_RVcoll.08-M265_Melitaea_didyma_NYMPHALIDAE
EZRMN108-08_RVcoll.08-M308_Melitaea_cinxia_NYMPHALIDAE
EZROM163-08_RV-07-C981_Melitaea_cinxia_NYMPHALIDAE
EZROM412-08_RV-07-D164_Melitaea_cinxia_NYMPHALIDAE
EZRMN106-08_RVcoll.07-D040_Melitaea_cinxia_NYMPHALIDAE
EZRMN109-08_RVcoll.08-M359_Melitaea_cinxia_NYMPHALIDAE
EZROM413-08_RV-07-D473_Melitaea_cinxia_NYMPHALIDAE
EZROM182-08_RV-07-D348_Euphydryas_aurinia_NYMPHALIDAE
EZROM983-08_RVcoll.07-D248_Euphydryas_aurinia_NYMPHALIDAE
EZROM982-08_RVcoll.07-C116_Euphydryas_aurinia_NYMPHALIDAE
EZROM183-08_RV-07-C115_Euphydryas_aurinia_NYMPHALIDAE
EZROM987-08_RVcoll.06-K678_Euphydryas_maturna_NYMPHALIDAE
EZROM184-08_RV-06-K677_Euphydryas_maturna_NYMPHALIDAE
EZROM992-08_RVcoll.08-M363_Euphydryas_maturna_NYMPHALIDAE
EZROM990-08_RVcoll.08-M249_Euphydryas_maturna_NYMPHALIDAE
EZRMN031-08_RVcoll.08-M505_Limenitis_camilla_NYMPHALIDAE
EZROM244-08_RV-06-M853_Limenitis_camilla_NYMPHALIDAE
EZROM247-08_RV-07-C965_Limenitis_populi_NYMPHALIDAE
EZRMN034-08_RVcoll.08-M554_Limenitis_reducta_NYMPHALIDAE
EZRMN025-08_RVcoll.08-M262_Libythea_celtis_NYMPHALIDAE
EZRMN024-08_RVcoll.08-M259_Libythea_celtis_NYMPHALIDAE
EZROM442-08_RV-07-C961_Nymphalis_xanthomelas_NYMPHALIDAE
EZRMN423-09_RVcoll.09-V660_Nymphalis_antiopa_NYMPHALIDAE
EZROM352-08_RV-08-A030_Nymphalis_polychloros_NYMPHALIDAE
EZRMN419-09_RVcoll.07-D628_Nymphalis_l-album_NYMPHALIDAE
EZROM218-08_RV-06-K659_Aglais_io_NYMPHALIDAE
EZROM217-08_RV-06-K655_Aglais_io_NYMPHALIDAE
EZROM1025-08_RVcoll.08-M493_Aglais_io_NYMPHALIDAE
EZRMN205-08_RVcoll.08-M441_Polygonia_c-album_NYMPHALIDAE
EZROM754-08_RVcoll.08-M491_Aglais_urticae_NYMPHALIDAE
EZROM003-08_RV-07-C204_Aglais_urticae_NYMPHALIDAE
EZROM001-08_RV-06-M855_Aglais_urticae_NYMPHALIDAE
EZROM561-08_RV-07-D077_Vanessa_atalanta_NYMPHALIDAE
EZRMN297-08_RVcoll.08-M395_Vanessa_atalanta_NYMPHALIDAE
EZROM560-08_RV-06-M897_Vanessa_atalanta_NYMPHALIDAE
EZROM596-08_RV-07-C360_Vanessa_atalanta_NYMPHALIDAE
EZRMN300-08_RVcoll.07-D934_Vanessa_cardui_NYMPHALIDAE
EZRMN301-08_RVcoll.07-E484_Vanessa_cardui_NYMPHALIDAE
EZROM019-08_RV-06-K528_Araschnia_levana_NYMPHALIDAE
EZROM011-08_RV-06-M869_Apatura_iris_NYMPHALIDAE
EZROM762-08_RVcoll.08-M553_Apatura_iris_NYMPHALIDAE
EZROM763-08_RVcoll.08-M591_Apatura_iris_NYMPHALIDAE
EZROM010-08_RV-06-M898_Apatura_ilia_NYMPHALIDAE
EZRMN371-08_RVcoll.08-M101_Apatura_ilia_NYMPHALIDAE
EZRMN376-08_RVcoll.08-M779_Apatura_metis_NYMPHALIDAE


EZROM025-08_RV-06-M882_Argynnis_adippe_NYMPHALIDAE
EZROM572-08_RV-07-C347_Argynnis_adippe_NYMPHALIDAE
EZROM026-08_RV-06-M950_Argynnis_adippe_NYMPHALIDAE
EZROM776-08_RVcoll.07-C913_Argynnis_adippe_NYMPHALIDAE
EZROM034-08_RV-06-M980_Argynnis_niobe_NYMPHALIDAE
EZRMN337-08_RVcoll.08-H017_Argynnis_niobe_NYMPHALIDAE
EZROM781-08_RVcoll.07-D049_Argynnis_niobe_NYMPHALIDAE
EZRMN340-08_RVcoll.08-H027_Argynnis_paphia_NYMPHALIDAE
EZROM028-08_RV-06-M881_Argynnis_aglaja_NYMPHALIDAE
EZROM600-08_RV-07-D517_Argynnis_aglaja_NYMPHALIDAE
EZROM031-08_RV-07-D935_Argynnis_aglaja_NYMPHALIDAE
EZROM037-08_RV-07-D071_Argynnis_pandora_NYMPHALIDAE
EZROM785-08_RVcoll.08-M404_Argynnis_pandora_NYMPHALIDAE
EZROM784-08_RVcoll.08-M273_Argynnis_pandora_NYMPHALIDAE
EZROM661-08_RV-07-D073_Argynnis_pandora_NYMPHALIDAE
EZROM036-08_RV-07-D069_Argynnis_pandora_NYMPHALIDAE
EZROM032-08_RV-06-M971_Argynnis_laodice_NYMPHALIDAE
EZROM063-08_RV-06-M865_Brenthis_ino_NYMPHALIDAE
EZROM064-08_RV-06-M906_Brenthis_ino_NYMPHALIDAE
EZROM807-08_RVcoll.08-M478_Brenthis_ino_NYMPHALIDAE
EZROM815-08_RVcoll.07-D309_Brenthis_ino_NYMPHALIDAE
EZROM808-08_RVcoll.08-M452_Brenthis_daphne_NYMPHALIDAE
EZROM607-08_RV-06-M910_Brenthis_daphne_NYMPHALIDAE
EZROM059-08_RV-07-D978_Brenthis_daphne_NYMPHALIDAE
EZROM062-08_RV-07-C970_Brenthis_hecate_NYMPHALIDAE
EZROM813-08_RVcoll.08-M516_Brenthis_hecate_NYMPHALIDAE
EZROM061-08_RV-06-M850_Brenthis_hecate_NYMPHALIDAE
EZROM811-08_RVcoll.08-M433_Brenthis_hecate_NYMPHALIDAE
EZROM609-08_RV-07-D543_Brenthis_hecate_NYMPHALIDAE
EZROM632-08_RV-07-D953_Issoria_lathonia_NYMPHALIDAE
EZROM223-08_RV-07-D165_Issoria_lathonia_NYMPHALIDAE
EZROM615-08_RV-07-D313_Boloria_selene_NYMPHALIDAE
EZROM573-08_RV-07-C336_Boloria_selene_NYMPHALIDAE
EZROM097-08_RV-06-K635_Boloria_selene_NYMPHALIDAE
EZROM099-08_RV-07-C421_Boloria_selene_NYMPHALIDAE
EZROM861-08_RVcoll.08-M344_Boloria_selene_NYMPHALIDAE
EZROM301-08_RV-07-E663_Boloria_euphrosyne_NYMPHALIDAE
EZROM614-08_RV-06-N006_Boloria_euphrosyne_NYMPHALIDAE
EZROM859-08_RVcoll.08-M645_Boloria_euphrosyne_NYMPHALIDAE
EZRMN428-09_RVcoll.09-V681_Boloria_titania_NYMPHALIDAE
EZROM091-08_RV-06-K536_Boloria_dia_NYMPHALIDAE
EZROM794-08_RVcoll.08-M523_Boloria_aquilonaris_NYMPHALIDAE
EZROM054-08_RV-07-E575_Boloria_pales_NYMPHALIDAE
EZRMN135-08_RVcoll.08-M507_Neptis_rivularis_NYMPHALIDAE
EZROM437-08_RV-07-D485_Neptis_rivularis_NYMPHALIDAE
EZROM436-08_RV-06-N007_Neptis_rivularis_NYMPHALIDAE
EZROM682-08_RV-06-N017_Neptis_rivularis_NYMPHALIDAE
EZROM438-08_RV-06-K648_Neptis_sappho_NYMPHALIDAE
EZROM644-08_RV-06-M984_Neptis_sappho_NYMPHALIDAE
EZROM440-08_RV-07-E522_Neptis_sappho_NYMPHALIDAE
EZRMN143-08_RVcoll.08-M506_Neptis_sappho_NYMPHALIDAE
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
0.03


Pieridae
EZROM1044-09_RVcoll.07-D065_Pieris_napi_PIERIDAE
EZROM465-08_RV-07-C989_Pieris_napi_PIERIDAE
EZROM1071-09_RVcoll.08-L363_Pieris_napi_PIERIDAE
EZROM466-08_RV-06-K568_Pieris_napi_PIERIDAE
EZROM471-08_RV-07-D101_Pieris_napi_PIERIDAE
EZROM470-08_RV-07-D027_Pieris_napi_PIERIDAE
EZROM472-08_RV-07-E530_Pieris_napi_PIERIDAE
EZRMN171-08_RVcoll.08-M609_Pieris_bryoniae_PIERIDAE
EZROM683-08_RV-06-M978_Pieris_bryoniae_PIERIDAE
EZRMN170-08_RVcoll.08-M496_Pieris_bryoniae_PIERIDAE
EZROM461-08_RV-06-M972_Pieris_bryoniae_PIERIDAE
EZRMN180-08_RVcoll.08-M243_Pieris_napi_PIERIDAE
EZROM476-08_RV-07-D209_Pieris_rapae_PIERIDAE
EZROM474-08_RV-06-K650_Pieris_rapae_PIERIDAE
EZROM477-08_RV-07-C973_Pieris_rapae_PIERIDAE
EZROM591-08_RV-07-C356_Pieris_rapae_PIERIDAE
EZROM473-08_RV-06-K542_Pieris_rapae_PIERIDAE
EZRMN183-08_RVcoll.08-M412_Pieris_rapae_PIERIDAE
EZRMN173-08_RVcoll.08-M248_Pieris_mannii_PIERIDAE
EZROM647-08_RV-07-D357_Pieris_brassicae_PIERIDAE
EZROM459-08_RV-06-M876_Pieris_brassicae_PIERIDAE
EZROM460-08_RV-07-C969_Pieris_brassicae_PIERIDAE
EZROM514-08_RV-07-C942_Pontia_edusa_PIERIDAE
EZROM510-08_RV-06-M919_Pontia_edusa_PIERIDAE
EZROM594-08_RV-07-C378_Pontia_edusa_PIERIDAE
EZROM583-08_RV-07-C368_Gonepteryx_rhamni_PIERIDAE
EZROM204-08_RV-07-D300_Gonepteryx_rhamni_PIERIDAE
EZROM006-08_RV-07-D096_Anthocharis_cardamines_PIERIDAE
EZROM008-08_RV-07-D924_Anthocharis_cardamines_PIERIDAE
EZROM005-08_RV-06-K531_Anthocharis_cardamines_PIERIDAE
EZROM757-08_RVcoll.08-M213_Anthocharis_cardamines_PIERIDAE
EZROM756-08_RVcoll.07-D315_Anthocharis_cardamines_PIERIDAE
EZROM977-08_RVcoll.08-M415_Euchloe_ausonia_PIERIDAE
EZROM770-08_RVcoll.08-M385_Aporia_crataegi_PIERIDAE
EZROM018-08_RV-07-D461_Aporia_crataegi_PIERIDAE
EZROM889-08_RVcoll.07-C950_Colias_crocea_PIERIDAE
EZRMN395-09_RVcoll.08-H036_Colias_crocea_PIERIDAE
EZROM122-08_RV-07-D199_Colias_crocea_PIERIDAE
EZRMN396-09_RVcoll.07-F556_Colias_erate_PIERIDAE
EZROM897-08_RVcoll.07-C136_Colias_myrmidone_PIERIDAE
EZROM131-08_RV-06-V650_Colias_myrmidone_PIERIDAE
EZROM119-08_RV-06-V681_Colias_chrysotheme_PIERIDAE
EZROM116-08_RV-07-D232_Colias_alfacariensis_PIERIDAE
EZROM620-08_RV-07-D363_Colias_alfacariensis_PIERIDAE
EZRMN324-08_RVcoll.08-H004_Colias_alfacariensis_PIERIDAE
EZRMN329-08_RVcoll.08-H035_Colias_alfacariensis_PIERIDAE
EZRMN353-08_RVcoll.08-M720_Colias_hyale_PIERIDAE
EZROM703-08_RV-07-E553_Leptidea_reali_PIERIDAE
EZRMN014-08_RVcoll.08-M322_Leptidea_reali_PIERIDAE
EZROM239-08_RV-06-K554_Leptidea_sinapis_PIERIDAE
EZROM238-08_RV-06-K556_Leptidea_sinapis_PIERIDAE
EZROM237-08_RV-07-C176_Leptidea_morsei_PIERIDAE
1
2
3
4
5
6
7
8
9
10
11
12
13
14
0.02

Chapter IV

Talavera, G., Espadaler, X., Vila, R. Discovered just before extinction? The first endemic ant from the Balearic Islands endangered by climate change. *In prep*

Discovered just before extinction? The first endemic ant from the Balearic Islands endangered by climate change

Gerard Talavera[a,b], Xavier Espadaler[c] and Roger Vila[a]

[a] Institut de Biologia Evolutiva (CSIC-UPF), Passeig Marítim de la Barceloneta, 37, 08003 Barcelona, Spain

[b] Departament de Genètica i Microbiologia, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain

[c] Departament de Biologia Animal, de Biologia Vegetal i d'Ecologia and CREAF, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain

ABSTRACT

It is frequently stated that many species will become extinct before being scientifically described. Climatic models predict global warming to increase in the near future, which will likely suppose a significant threat to biodiversity and accentuate extinctions, although there is limited empirical evidence about current extinction processes directly caused by climatic change. Organisms on islands are especially vulnerable because of limited possibilities for spatial shifts following climatic conditions. This effect is even more severe in islands-within-islands, like populations occurring in ecologically unique island mountaintops. In this work we document the discovery of *Lasius balearicus* sp. nov., the first endemic ant for the Balearic Islands, based on morphological, molecular and ecological evidence. We show that it only inhabits the summits of Mallorca Island, with an extremely restricted and altitudinally constrained distribution. Climate-based distribution modelling, coupled with the extremely low intraspecific genetic diversity, total range and nest densities documented, suggest low probability for short-term survival, thus becoming a model to study real-time climatic-based extinction. Indeed, most climatic scenarios predict that it will become extinct or critically endangered by 2050 or 2080, and only the most optimistic SRES scenario, based on lower energy requirements and emissions, gives some hope for the survival of this species after a strong bottleneck. This case shows that a determined change in human activities would likely have a positive impact for biodiversity and save certain species from extinction.

INTRODUCTION

It is frequently stated that many species will become extinct before being scientifically described. While a substantial number of probable species extinctions have been reported in the last century (IUCN, 2012; Turvey, 2009), documenting cases of current extinction processes and their causes are a highly important contribution to biodiversity estimations, conservation and to increase society consciousness on the impacts of human activities on other organisms. The discovery of new species displaying restricted distributions, together with their conservation status evaluation and future perspectives, are important to estimate and compare biodiversity extinction and discovery rates.

Climate change involves several processes with consequences on seasonality, temperature, rainfall, catastrophic events or CO_2 concentrations, among others, which promote that species develop responses and adaptations (Foden *et al*, 2008). According to Bellard *et al* (2012), those responses may be driven along three distinct by non-exclusive axis: spatial, temporal and self. Spatially, species track new appropriate conditions through dispersion or latitudinal and altitudinal range shifts. Temporally, species can experience changes on the timing of life cycle events (i.e. phenology), and lastly, self adaptations are selected through new requirements in their own local habitat, which would include physiological alterations or behavioural strategies. In fact, recent studies are documenting poleward shifts in species distributions, and other changes in population structure and abundance attributable to climate change (e.g. Parmesan and Yohe, 2003; Root *et al*, 2003; IPCC 2007; Devictor *et al*, 2012). While possibilities to find suitable responses for species in front of climatic alterations may seem to be numerous, failing to adapt along one or several of these three axes shall result in extinction. Relatively few cases of species extinction have been reported because of climatic changes during the Quaternary period (Willis *et al*, 2004; Botkin *et al*, 2007), but future scenarios are gloomy. Indeed, some studies suggest that climate change could surpass habitat destruction as the greatest global threat to biodiversity over the next few decades (Leadley *et al*, 2010; Bellard *et al*, 2012), and projections for future species extinction rates indicate that they will likely surpass background rates estimated from the cenozoic fossil record by more than two orders of magnitude (Mace *et al*, 2005; Pereira *et al*, 2011).

The 20th century witnessed a significant increase in Earth's average surface temperature and greater frequencies and intensities of extreme weather events (IPCC, 2001). Despite of efforts to decrease fossil fuel emissions, the prediction is that these trends will continue throughout the 21st century (IPCC, 2001). Thus, a substantial percentage of the species on Earth might be experiencing the need to cope with this sudden climatic change process, which acts in synergy with other human activities that disturb current environmental conditions (habitat alteration and fragmentation, introduction of invasive species, etc.) (Botkin *et al*, 2007). These rapid changes may often surpass the ability of the species to respond to them (Lavergne *et al*, 2010; Salamin *et al*, 2010), accumulating a climatic debt that poses a significant threat to global biodiversity (Devictor *et al*, 2012).

The described effects of global warming on biodiversity are not expected to be constant across space. For example, it has been argued that organisms on islands, both in a literal and ecological sense, are especially vulnerable (Cronk, 1997; Losos and Ricklefs, 2010). In these cases, except for species with exceptional dispersal abilities, there is practically no possibility of a spatial shift to follow climatic conditions, either latitudinal (in sea or oceanic islands) or altitudinal (for example in mountain tops). This effect is even most severe in islands-within-islands displaying extremely isolated ecological and spatial conditions, like populations occurring in island mountaintops. Thus, these types of habitats and their fauna and flora should be considered at high risk and studies documenting their biodiversity and estimating future scenarios are urgently needed. Sadly enough, island mountaintops represent ideal places where to search for species at the brink of extinction and yet unknown to science.

It is difficult to model extinction processes because of the many factors involved (Botkin *et al*, 2007), and there is limited empirical evidence about current extinctions processes directly caused by climatic changes. In this work we document the discovery of the first endemic ant in the Balearic Islands, and show that it only inhabits the summits of Mallorca Island, with an extremely restricted and altitudinally constrained distribution. The finding of a new non-cryptic endemic ant species in Europe is of remarkable interest because being localized in a relatively well-studied area of the World where ant endemism is not common. Climate-based distribution modelling shows that short-term survival of this species is unlikely, thus becoming a model to study real-time climatic-based extinction. This case illustrates how incomplete our knowledge of biodiversity is, and supports the statement that a number of species will become extinct before we can even study them.

MATERIALS AND METHODS

Sampling

An area covering the whole Serra de Tramuntana range (N. Mallorca, Balearic Islands, Spain) was surveyed between 2008 and 2012. An entire day was devoted to prospect each of the main peaks (10 hours of field work approximately) in order to record ant diversity at different altitudes. Surveys were also extended to other potential areas of Mallorca, nearby islands (Menorca, Eivissa) and nearby mainland (Serra d'Aitana, Alacant, East Iberian Peninsula). Most samples were preserved in 100% ethanol for molecular analysis, but a number of specimens were prepared dry for morphological examination. Identification codes and collection localities for the samples used for molecular analysis are listed in the Supplementary Table S1 and the specimens used for ecological niche modelling are listed in the Supplementary Table S2. Voucher specimens are deposited in the R.V. collection at Institut de Biologia Evolutiva (CSIC-UPF) and in X.E. collection at Universitat Autònoma de Barcelona, both in Barcelona (Spain). Syntypes are deposited in the Museum of Comparative Zoology (Cambridge, MA, US), British Natural History Museum (London, UK), Senckenberg Museum für Naturkunde (Görlitz, Germany) and paratypes at the Museu Balear de Ciències Naturals (Sóller, Mallorca).

Biometric analyses

Measures were taken with a Nikon SMZ-U stereomicroscope with variable magnification (from 7.5x to 75x) and two cold light fiber tubes. A white plastic light diffuser was used when measuring PDCL. A total of 51 individuals of *Lasius balearicus* sp. nov. corresponding to 8 nests from 5 localities were studied. Ten biometric characters were measured (Supplementary Table S3): Head width (HW), head length (HL), scape length (SL), head shape (HL/HW), relative scape length (SL/HL), pubescence distance on clypeus (PDCL), number of standing hairs on scapus (nHS), number of standing hairs on hind tibia (nHHT) and number masticatory dents (MaDe) as defined in Seifert (1992). Some of these characters were also measured for queens (HWQ, HLQ, HLQ/HWQ, SLQ, SLQ/HLQ, SLQ/HWQ, MLQ, MHQ, MHQ/MLQ). Mean values for these variables were extracted from literature (Seifert, 1992; Schlick-Steiner *et al*, 2003) for the *Lasius* s. str. taxa *niger*, *platythorax*, *japonicus*, *grandis*, *cinereus*, *emarginatus, hayashi, productus, sakagamii, alienus, psammophilus, brunneus, austriacus, neglectus, turcicus, lasioides* and from Seifert (1988) for the outgroup *L. mixtus*. Taxa lacking values for some of the variables were measured based on available specimens from different nests covering the range of distribution (Supplementary Table S3).

A t-test (Sokal and Rohlf, 1995) for the measured variables was applied to pairwise comparisons between *L. balearicus* and the most closely related species (*grandis*, *cinereus*, *niger*, *japonics*, *platythorax*, *emarginatus*). Since multiple comparisons were performed between morphological variables. A caution to control for false discovery rates were applied following the procedure in Waite and Campbell (2006) and using a p<0.001 (Supplementary Table S4).

Molecular data for phylogenetic inference

A total of 15 individuals (three specimens for each of five populations) of *L. balearicus* sp. nov. covering the whole known distribution range were used to obtain molecular data, as well as nine individuals from different populations of *L. grandis* and *L. cinereus* (Supplementary Table S1). Genomic DNA from entire ant bodies was extracted using the DNeasy™ Tissue Kit (Qiagen Inc., Valencia, CA) and following the manufacturer's protocols. To amplify a fragment by polymerase chain reaction of the mitochondrial gene *cytochrome oxidase subunit I* (COI), we used primers Lasius-R and Lasius-L as described in Maruyama *et al* (2008). A modified primer named Lasius-2L (5'- TAYCCTCCATTAGCYTCTAA -3') was designed to improve the amplification for *L. balearicus* sp. nov. The mitochondrial region corresponding to *16S rRNA* was also amplified using the primers "16Sar-L" and "16Sar-L2" (Maruyama *et al*, 2008). Sequences that were matching the same *COI* and *16S* regions from the subgenus *Lasius* s. str. and the outgroup *Lasius mixtus* were retrieved from GenBank, corresponding to previous studies by Maruyama *et al* (2008), Steiner *et al* (2006), and Kremer *et al* (2009) (Supplementary Table S1). PCR products were purified and sequenced by Macrogen Inc. and sequences obtained were deposited in GenBank (accession numbers Supplementary Table S1).

Morphological data for phylogenetic inference

A morphological set of characters were used to complement molecular data for phylogenetic inference. A total of 67 characters (see Supplementary Table S5) were defined using various sources: 1) characters that were variable within the genus *Lasius* s. str. in the morphological matrix of Maruyama *et al* (2008) were selected and adapted, and a new character (sides of pronotum for workers)was incorporated; 2) the biometric measurements obtained (Supplementary Table S3) were converted to discrete characters.

Phylogenetic inference and dating

COI and 16S sequences were edited using GENEIOUS PRO v. 4.8.3 (Biomatters Ltd., 2009) and aligned using Muscle (Edgar, 2004). Gblocks v.091 with relaxed parameters (Castresana, 2000; Talavera and Castresana, 2007) was applied to the 16S alignment. These analyses resulted in a final alignment of 864 bp for *COI* and 498 bp for *16S* for a total of 60 and 39 sequences respectively (Supplementary Table S1). BEAST 1.6.0 was used to construct the phylogeny and estimate node ages based on the molecular characters. Since no reference calibration points were available for this genus, we used a substitution rate 1.5% uncorrected pairwise distance per million years, inferred from the entire mitochondrial genome of various arthropod taxa (Queck *et al*, 2004). A constant size coalescent approach and an uncorrelated relaxed clock were used as priors (Drummond *et al*, 2006). Parameters were estimated using two independent runs of 50 million generations each and a burn-in of 5 million generations was applied to get the final tree. A phylogenetic tree combining molecular and morphological data was also reconstructed using Mr.Bayes 3.1.2 (Huelsenbeck and Ronquist, 2001) running two independent chains of 10 million generations each (with a pre-run burn-in of 100,000 generations). Three partitions were set corresponding to each molecular marker and morphological data. In both reconstructions, jModeltest v.0.1 (Posada, 2008) was used to assign GTR + I + G as a model of nucleotide substitution for both molecular markers according to Akaike information criterion (AIC), 6 gamma rate categories were assigned and convergence was checked with the program Tracer v1.5.

Ecological Niche Modelling

Species distribution models were used to estimate distribution limits for *L. balearicus* through time according to environmental climatic data. The layers for 19 climatic variables available in WorldClim (http://www.worldclim.org/ described by Hijmans *et al*, 2005) were used. As WorldClim variables generally show a high collinearity that can lead to statistical bias and model over-prediction (i.e. modelling an overly small distribution by placing too many restrictions on where a species can occur), a subselection of variables was employed. Values for all 19 parameters in a Mallorca island raster dataset were extracted from a 30 arc-seconds (approximately 1 km^2 grid cells) grid using DIVA-GIS software v.7.1.7 (Hijmans *et al*, 2005). Points from the whole raster were used to analyze the level of correlation between pairs of variables in JMP ver. 7.0.2 (SAS Institute 2008). When two variables shared a Pearson correlation coefficient of 0.8 or higher (Rissler and Apodaca, 2007), we selected the biologically most meaningful variable for ants, either the most general one, or the one involving the season when ants are active. Thus, eight out of 19 variables were selected: Bio1 (annual mean temperature), Bio2 (mean diurnal range), Bio3 (Isothermality), Bio4 (Temperature Seasonality), Bio7 (temperature annual range), Bio12 (annual precipitation), Bio15 (precipitation seasonality) and Bio17 (precipitation of driest quarter).

To predict the potential distribution models we employed MAXENT v.3.3.2 (Phillips *et al*, 2006), which uses a machine-learning algorithm to identify the areas in which the environmental conditions are suitable for the species considered. We considered 21 presence records (reduced to 10 cells) for *L. balearicus* and 36 (reduced to 34 cells) for its sister species *L. grandis* (Supplementary Table S2). The default parameter settings were used (maximum number of background points: 10,000, regularization multiplier: 1, auto features, maximum iterations: 500, convergence threshold: 0.00001) as suggested by Phillips *et al* (2006). Distribution models were generated through 100 replicates, using as a subsampling approach with 50 % of the points for training and 50% for testing the model. The logistic output was selected due to the easier interpretation of the results (interpreted as probability of presence of the species) compared to raw and cumulative output formats, and the results were presented with a linear scale (Phillips, 2008; Phillips and Dudík, 2008).

To test the accuracy of the models, the area under the receiver operating characteristic curve (AUC) and the threshold-dependent binomial omission tests calculated by Maxent (Phillips *et al*, 2006; Moffett *et al*, 2007; Pawar *et al*, 2007), were inspected. Values of AUC higher than 0.7 (Pearce and Ferrier, 2000; Elith, 2002; Newbold *et al,* 2009) or 0.85 (Newbold, 2009) are considered acceptable and binomial omission tests are significant at $P < 0.05$.

After calibrating the model for their current distributions based on present climate, we modelled the distribution onto Last Glacial Maximum (LGM) WorldClim data and onto IPCC 4 (Intergovernmental Panel on Climate Change; New *et al*, 1999) future conditions data, using the same eight variables considered in the present models. For LGM, two general atmospheric circulation models were used: CCSM (Community Climate System Model) and MIROC (Model for Interdisciplinary Research on Climate) models from the Paleoclimate Modelling Intercomparison Project Phase II (PMIP2). To predict future potential distributions we used Statistical Downscaling (Delta Method) layers based on CGCM2 and CGCM3 (Coupled Global Climate Models) models created by CCCMA (Canadian Centre for Climate Modelling and Analysis) and under three varied emission SRES scenarios for two different dates corresponding to the years 2050 and 2080: A1b (Maximum energy requirements or emissions) (CGCM31), A2a (High energy requirements or emissions) (CGCM2) and B2a (Lower energy requirements or emissions) (CGCM2) (see IPCC Third Assessment report, 2001 for scenarios).

RESULTS

Habitat

Specimens attributable to *Lasius balearicus* sp. nov. were only collected during prospections in the seven major summits of the Serra de Tramuntana range, between 800 and 1400 meters above sea level. No similar specimens were observed below this altitude in Mallorca or in any other region studied. Habitat was similar in the localities where *L. balearicus* was found: calcareous surface, dry and shrubby vegetation, always above treeline (Supplementary Figure S1). Despite extensive searches, only few nests restricted to small areas were found per population.

Biometric and morphological analyses

In the majority of numeric characters and derived indexes, *L. balearicus* is statistically different from its closest species (Supplementary Table S3-S4). Aside from being smaller than the rest of *Lasius* s. str., except for *L. cinereus,* its yellowish brown colouration allows for an immediate separation. Under the microscope its extreme hairiness shows out immediately as a remarkable trait. Absolute size and hairiness differentiate *L. grandis* and *L. balearicus*. The most biometrically similar species is *L. cinereus*, from which *L. balearicus* differs by its shorter scape and more developed pilosity in the scape, back of head and extensor profile of the hind tibia.

Phylogeny and dating

The two inferred phylogenies generally agree with previous published hypotheses (Janda *et al,* 2004; Seifert *et al,* 2004; Maruyama *et al,* 2008; Cremer *et al,* 2008), displaying two main deeply diverged clades that segregate hairy and non-hairy taxa. *Lasius balearicus* specimens were recovered as a supported clade (*pp* = 1) within the group formed by *L. cinereus*, *L. grandis, L. japonicus, L. platythorax, L. niger and L. emarginatus* in the two trees (Figure 1). The *L. balearicus* clade was well diverged from the rest of species in the group, and the estimated age of this species was 1.45 Myr [95% HPD: 0.85-2.23 Myr]. Remarkably, the intra-specific divergence for *L. balearicus* was almost inexistent (99.9% Pairwise Identity), with only two parsimony informative sites in 1351 mitochondrial bases length, and displaying four different haplotypes for COI and only one for 16S.

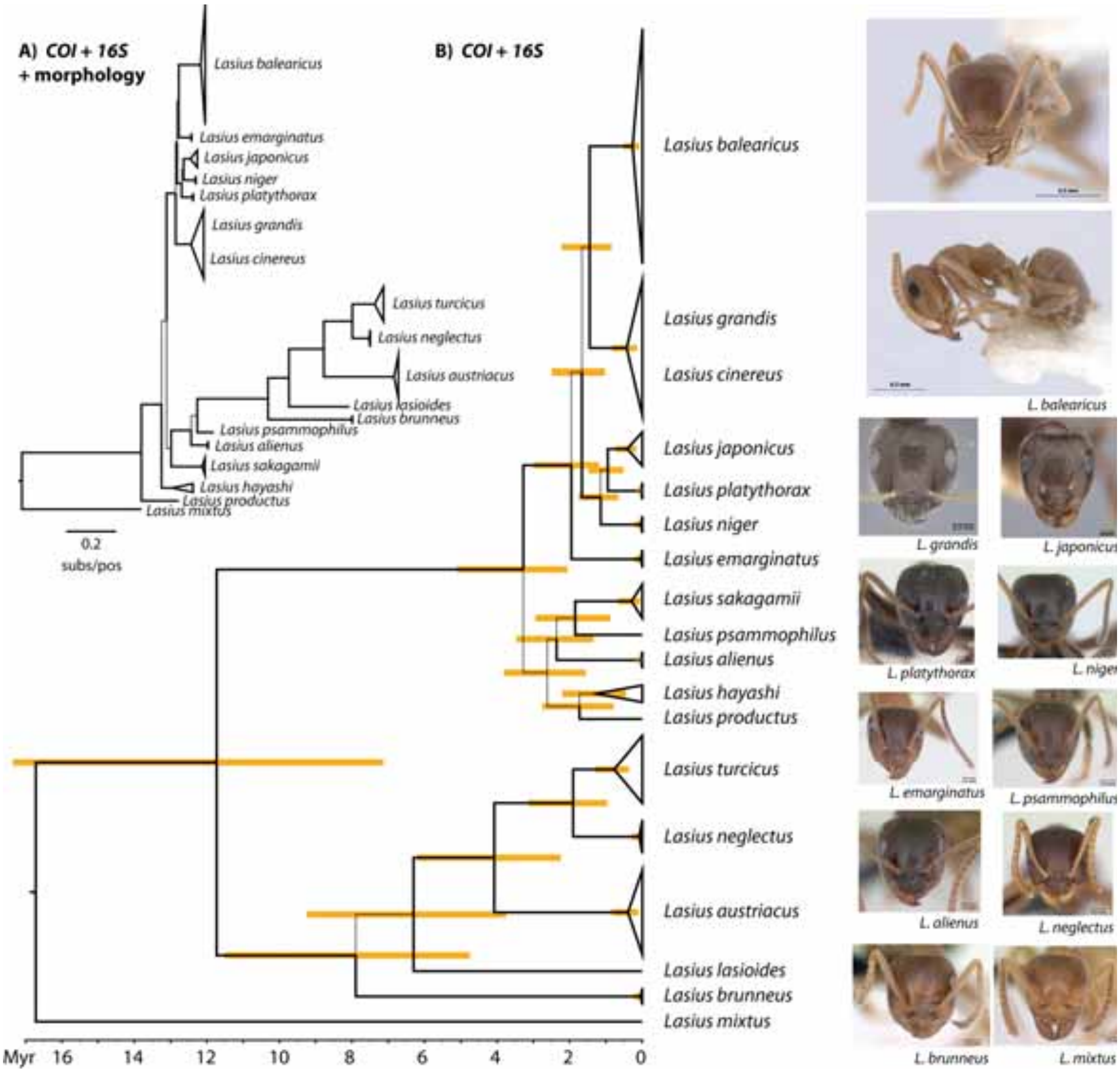


Figure 1. **Phylogenetic inference for the subgenus *Lasius* s. str. A.** Combined molecular + morphology phylogram. **B.** Bayesian chronogram based on *COI* + *16S*. A relaxed clock rate of 1.5% uncorrected pairwise distance per million years was used as prior for dating nodes. Bars indicate the 95% HPD for age estimations according to the axis representing time in millions years before present. Thick lines indicate supported relationships (posterior probabilities ≥ 0.95). Pictures were extracted from www.antweb.org (for credits see Acknowledgements).

Ecological Niche Modelling

Distribution models for *L. balearicus* produced high mean AUC scores (0.990, SD 0.004) and significance for all binomial omission tests, indicating a good performance of the models. The predicted current distribution for *L. balearicus* was limited to the mountainous area of the island, with highest probabilities at the highest altitudes of Serra de Tramuntana (Figure 2).

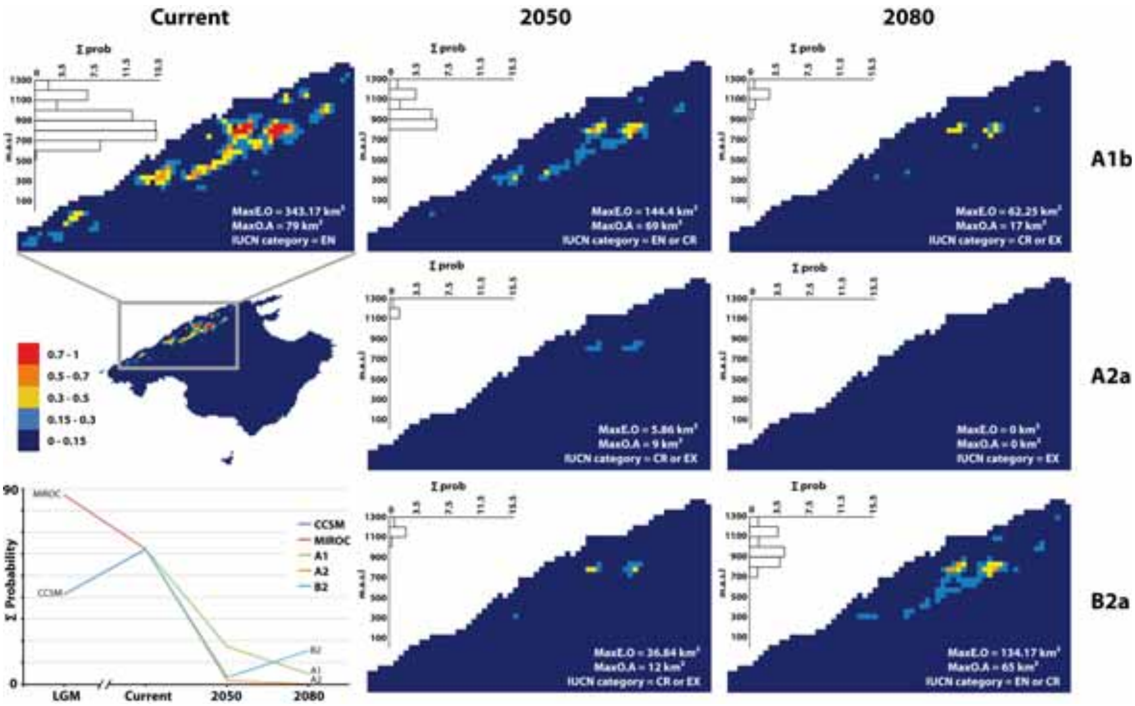


Figure 2. Ecological niche modelling for *Lasius balearicus* sp. nov. Current prediction and future projections for the years 2050 and 2080 are represented for Mallorca Island. Probability ranges are represented in a colour scale. Bar diagrams show the sum of presence probabilities according to altitude. The graph on the down left corner displays the evolution of total presence probabilities through time. Maximum estimates for the Extent of Occurrence (E.O), Area of Occupancy (O.A) and proposed IUCN category are shown according to IUCN criteria. Estimates for the last glacial maximum (LGM) are shown in the Supplementary Figure S2.

A heuristic estimate of relative contributions of the environmental variables to the Maxent model is shown in the Supplementary Table S6. Annual mean temperature was the main contributor to the model (89.1 %) for *L. balearicus*, therefore temperature seems to be a crucial factor affecting its distribution. Jackknife tests of variable importance indicated that annual mean temperature had the highest gain when used in isolation, suggesting that this variable contained the most useful information, and precipitation seasonality decreased the gain the most when omitted, suggesting that it contained the most information not present in the other variables. A comparison of modern to LGM distributions (under both the CCSM and MIROC climatic models) (Supplementary Figure S2), suggested that *L. balearicus* has displayed a generally stable distribution in the interglacial period (Figure 2). On the contrary, future projections show a dramatic reduction of suitable climatic areas under the different scenarios for *L. balearicus*, to the point that it will likely become extinct by 2050 or 2080, especially according to scenario A2a. Modelling for *Lasius grandis* showed completely different tendencies for predicted distribution (Supplementary Figure S3), suggesting an increase of suitable habitat with time. In this case the performance of the model was not satisfactory (AUC score 0.656, SD 0.056; binomial omission tests mostly not significant), probably due to the wide array of niches that this generalist species inhabits. Precipitation seasonality was the variable that contributed the most (46.4%) to the model (Supplementary Table S6).

DISCUSSION

Species status of the discovered taxon

Lasius balearicus sp. nov. presents a unique morphology and colour, as well as a highly specialized habitat that distinguishes it from all other known *Lasius* s.str. species. Molecular phylogeny confirms that it is a genetically isolated taxon that split from the sister lineage about 1.45 Myr ago (Figure 1, Supplementary Table S7). The sister lineage comprises several widespread taxa that suggests that the common ancestor occurred in mainland. The last land bridge between Mallorca and mainland existed no more recently than about 5 Myr ago (Meulenkamp and Sissingh, 2003) and therefore a vicariant speciation event seems not to be the case. The most likely biogeographical scenario is rafting or dispersal by the ancestor from mainland to Mallorca. In fact, other *Lasius* species widespread in mainland are also present in the Balearics (*L. grandis* and *L. lasioides*), proving that dispersal across the Mediterranean is not an uncommon event in this group. Ant endemism in Mediterranean islands is surprisingly low: only a few endemic species are known from Corsica (four species; Casevitz-Weulersse, 1996; 2010) and Sardinia (3 species; Baroni Urbani, 1971; Rigato, 1999), none of them within the genus *Lasius*. Thus, *L. balearicus* sp. nov. represents the first endemic ant described in the Balearic islands, as well as the first endemic *Lasius* species in the Mediterranean islands.

Current habitat and distribution

Lasius balearicus sp. nov. is restricted to the treeless summits of Serra de Tramuntana, always above 800 metres of altitude. Nests seem to be associated to calcareous, exposed and rocky areas, with sparse and shrubby vegetation, frequently composed by the endemic plants *Hypericum balearicum* and *Genista valdes-bermejoi*. Such habitat specialization reduces even more the potential areas to isolated localities within the island. Despite performing intensive surveys in most climatically suitable areas along the mountain range, the species was not recorded in forested extensions, as for example in the Mola de Planícia peak (942 m). This mountain is completely forested, including the top. However, based on climatic modelling, it displays a higher presence probability (44.7%) than its neighbour peak Puig Galatzó (1027 m) (28.7%), where the forest does not reach the top and *L. balearicus* is present. Based only on the climatic model prediction with probabilities > 0.15 (minimum probability with actual presence recorded), we estimate a maximum area of occupancy of 179 km^2. To obtain a more accurate estimation, we calculated the area with both suitable land cover and climatic conditions as shown in Supplementary Figure S4, resulting in 109 km^2 for a maximum occupancy. Even so, the suitable area for the species is probably even lower since most of these pixels include areas with altitudes below 800 m. Based on our observations, the species is restricted to summits and ridges, and it could only be detected in a total area of 8 km^2. In summary, because not all peaks could be explored, the real occupancy area likely lies between 8 and 109 km^2.

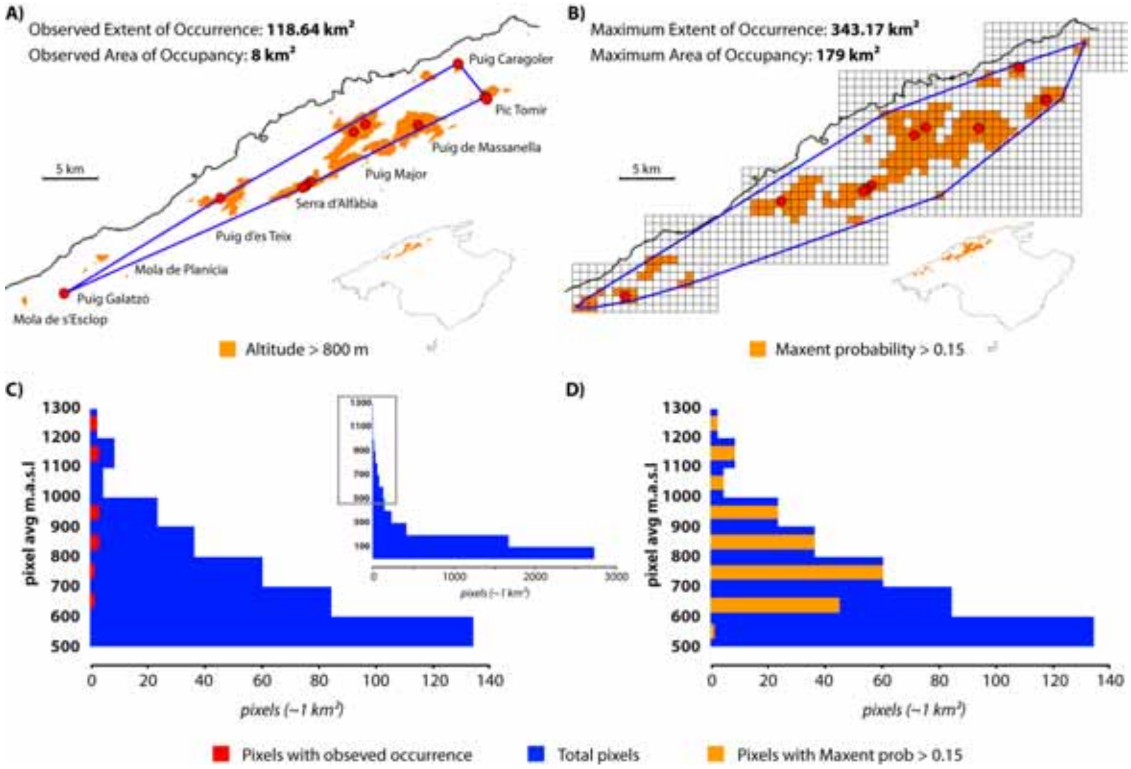


Figure 3. Distribution of *Lasius balearicus* sp. nov. Presence locality points (in red) are displayed on maps for Serra de Tramuntana highlighting in orange the areas above 800 m of altitude (a), and areas with estimated presence probability > 0.15 (minimum probability with actual presence recorded) (b). The altitudinal distributions o areas with observed (c, in red) and estimated presence (d, in orange) are shown on a background distribution of altitudes for the entire island (blue). Extent of occurrence and areas of occupancy based on both observed and estimated presence are indicated according to the IUCN criteria.

Effects of climate change

Global warming is not necessarily detrimental for all organisms and in some instances it could promote species success. Within Mallorca Island, this seems to be the case for *L. grandis*, for which suitable niche areas will apparently increase with climate change (Supplementary Figure S3). However, fragility on island endemic organisms is well documented, especially because of having small population sizes, and a high percentage of documented extinctions occurred on islands (Turvey, 2009; Fordham and Brook, 2010; Losos and Ricklefs, 2010). The reasons are that, small populations cannot deal with big and rapid ecological changes such as deforestation, natural catastrophes, habitat reduction or global warming. *Lasius balearicus* seems to be a good example. Our data shows that, despite a relatively stable distribution maintained since the LGM (Figure 2, Supplementary Figure S2), the species will collapse in the following decades. Certainly, modelling based on future climatic conditions predicts a rapid loss of niche area for this species under all the scenarios studied. The use of three different SRES emission scenarios to assess future climate change allowed capturing to some extent the uncertainty in future climate due to human decisions and, indeed, they produce somewhat different results in agreement to the levels of global surface warming (IPCC, 2001) displayed in the three scenarios. The worst case is shown by the scenario A2a, where the suitable area would be reduced to a minimum by 2050 and to zero in 2080, and thus the extinction of the species would be most likely. This scenario represents a differentiated world that consolidates into a series of economic regions, somehow reproducing the present situation and tendencies. Scenario A1b is based on converging successful economic development that would minimize richness differences among countries. The result is a progressive narrowing of *L. balearicus* potential habitat, and in 2080 the species could only survive with some probability in the two main peaks of the island (*Puig Major* and *Puig de Massanella*). The emission scenario B2a is particularly interesting because, after a drastic niche reduction that will result in nearly extinction by 2050, it reveals a substantial recovery. The B2a SRES scenario is based on environmental, social and economical sustainability, and assumes lower energy requirements and greenhouse gas emissions. Even in this case, it is uncertain whether *L. balearicus* will become extinct before climatic conditions reverse, as well as the capacity of the species to recolonize peaks if they eventually become suitable once again.

Several factors could exacerbate the projected impacts. For example, it has been demonstrated that forested areas can shift towards higher altitudes due to climate change (Walther *et al*, 2005). Given that *L. balearicus* seems unable to inhabit forested areas, this phenomenon would suppose another crucial source of pressure. Also, competition with the closest species in the island, *L. grandis*, should be considered. *Lasius grandis* is a generalist species with completely different ecological requirements and can be found from at sea level to high altitude (Supplementary Figure S3, Supplementary Table S2). Actually, nests were found at short distance from those of *L. balearicus* in a few cases, although the general rule is to find them at lower altitudes. The fact that future climatic conditions seem to favour *L. grandis* (Supplementary Figure S3) serve as a control to show different responses depending on the species, and a future expansion of *L. grandis* opens the door to species interactions effects with unknown repercussions for the future distribution of *L. balearicus*.

The potential of *L. balearicus* to counteract climate change is low for the three axes proposed by Bellard *et al* (2012). An altitudinal shift is observed in our projections (Figure 2), but the species already inhabits the highest peaks and the result is generally the disappearance at the lowest distribution limits. Since there are no more mountains in the Balearic Islands, a spatial shift would imply a virtually impossible dispersal to suitable habitat in Iberian mainland or in other mountainous islands like Corsica or Sardinia. Adaptation to new climatic conditions could be driven through changes in phenology or by intrinsic evolution, but the population sizes and genetic variability upon which natural selection can act are very low in *L. balearicus*. Moreover, *Lasius* ant nests may last for a long time, which indicates long generation times. Thus, all evidence suggests that it is highly improbable that *L. balearicus* could adapt to new habitats in such a short time.

Conservation status and management

Lasius balearicus sp. nov. presents a highly restricted, isolated and fragmented distribution, a very low intraspecific variability coupled with low population densities, and negative trends are suggested by niche modelling, to the point that the species may become extinct by 2050 or 2080 depending on the climatic scenario. According to the IUCN Red List categories and criteria (IUCN 2001), we propose the inclusion of the new taxon to the list under the category Endangered (EN B1ab(i,ii,iv); B2ab(i,ii,iv); C2a(i)), based on the estimation of a suitable area of occupancy between 8 km^2 (observed) and 109 km^2 (maximum based on climate conditions and vegetation cover data, see Supplementary Figure S4), an extent of occurrence between 118.64 km^2 (observed) and 343.17 km^2 (maximum) (Figure 3), and a population size definitely smaller than 2500 mature individuals, with no peak containing more than 250 nests. The species could even be categorized as Critically Endangered (CR B2ab(i,ii,iv); C2a(i)) since the observed area of occupancy is less than 10 km^2 and the population size could easily be fewer than 250 mature individuals in total, with no population containing more than 50 nests. However, we opt for a conservative decision since 8 km^2 for the area of occupancy is probably an underestimation, and the exact number of nests is difficult to be assessed. Estimations for the future suggest that *L. balearicus* could disappear soon after being discovered. Indeed, our projections show that by 2050 the species will likely be critically endangered or extinct and by 2080 it will be most likely extinct, except in the supposition of a moderation in greenhouse gases emissions. This tendency contrasts with a probable rapid distributional expansion of *L. grandis* that will apparently widely colonize the mountain summits, with unknown consequences for *L. balearicus* due to interspecific effects.

Management for the conservation of this species would require in the first place monitoring the known populations and inferring its approximate number of nests, as well as exploring the peaks that have not yet been inspected. Apart from a deeper survey, the habitat needs special protection. Indeed, since the future of *L. balearicus* is uncertain due to global warming, it is most important to avoid any other sources of pressure that could synergistically affect this species. The Serra de Tramuntana is a UNESCO world heritage site and the majority of the range belongs to a local category with medium level of natural protection (Paratge Natural). Although the summit areas are not object of big alterations because of human pressure, they are of extreme fragility for its little extension. The effects of the high density of introduced goats, frequent even in the highest peaks, may be important, especially in dry years. An eventual fire may also represent a serious risk, given their recurrent occurrence in the Mediterranean region during summer. Several endemic plants have found refuge in some of the summits of even just in one (Sáez and Rosselló, 2000; 2001; López *et al*, 2012), which should encourage policymakers to consider protection initiatives at that smaller scale. It is important to highlight that the most optimistic climatic scenario, assuming a tendency to environmental, social and economical sustainability, gives some hope that this endemic ant can survive.

CONCLUSION

All the data gathered (morphological, molecular and ecological) demonstrates the existence of a new endemic species of ant in Mallorca Island high elevations (formal description in Appendix 1). The extremely low intraspecific genetic diversity, total range and nest densities documented, coupled with dramatic future predictions without the apparent possibility of dispersal to suitable habitat due to geographic and altitudinal isolation, suggest a high probability of short-term extinction for this species. As a consequence, we strongly recommend including *Lasius balearicus* to the IUCN Red List of Endangered Species under the category “endangered”. Although prospects for its survival are gloomy, the most optimistic climatic scenario suggests the potential recovery of the species after a strong bottleneck in 2050. Even if reversal of climate tendencies may eventually arrive too late for *L. balearicus*, this case shows that a determined change in human activities would likely have a positive impact and avoid species’ extinctions.

ACKNOWLEDGEMENTS

This paper is dedicated to Cedric Alex Collingwood, who in 1982 collected the first specimens of the new *Lasius* here described. We want to thank Cristina Tur and Anna Traveset for providing complementary records, Joan Garcia-Porta for helpful discussion on Ecological Niche Modelling and Jesús Gómez-Zurita for assistance in taking pictures of the specimens. Credits for www.antweb.org photographs are: Andrea Walker (*L. grandis*), Shannon Hartman (*L. japonicus*) and April Nobile (*L. platythorax*, *L. niger*, *L. emarginatus*, *L. psammophilus*, *L. alienus*, *L. neglectus*, *L. brunneus* and *L. mixtus*). Support for this research was provided by the Spanish MICINN (project CGL2010-21226/BOS to G.T. and R.V., and predoctoral fellowship BES-2008-002054 to G.T.)

REFERENCES

Avise, J.C. **2004**. Molecular Markers, Natural History and Evolution. Sinauer Associates, Sunderland, MA.

Baroni Urbani, C. **1971**. Catalogo delle specie di Formicidae d'Italia. *Memories della Societa Entomologica Italiana* 50:1-287.

Boomsma, J.J., Mabelis, A.A., Verbeek, M.G.M., Los, E. **1987**. Insular biogeography and distribution ecology of ants on the Frisian islands. *Journal of Biogeography* 14:21-37.

Walther, G-R., Beißner, S., Pott, R. 2005. Climate change and high mountain vegetation shifts. *In* Broll, G., Keplin, B [Eds]., *Mountain Ecosystems: Studies in Treeline Ecology* pp 77-95. Springer, Berlin, Heidelberg.

Butchart, S.H., Walpole, M., Collen, B., van Strien, A., Scharlemann, J.P., Almond, R.E., Baillie, J.E., Bomhard, B., Brown, C., Bruno, J., Carpenter, K.E., Carr, G.M., Chanson, J., Chenery, A.M., Csirke, J., Davidson, N.C., Dentener, F., Foster, M., Galli, A., Galloway, J.N., Genovesi, P., Gregory, R.D., Hockings, M., Kapos, V., Lamarque, J.F., Leverington, F., Loh, J., McGeoch, M.A., McRae, L., Minasyan, A., Hernández-Morcillo, M., Oldfield, T.E., Pauly, D., Quader, S., Revenga, C., Sauer, J.R., Skolnik, B., Spear, D., Stanwell-Smith, D., Stuart, S.N., Symes, A., Tierney, M., Tyrrell, T.D., Vié, J.C., Watson, R. **2010**. Global biodiversity: indicators of recent declines. *Science* 328:1164.

Casevitz-Weulersse, J. **1996**. Biogeographical aspects of the ant fauna of Corsica (Hymenoptera: Formicidae). *Pan-Pacific Entomologist* 72:193-201.

Casevitz-Weulersse, J. **2010**. A propos des fourmis de la Corse: *Aphaenogaster corsica* n. sp. (Hymenoptera: Formicidae: Myrmicinae). *Le Bulletin d'Arthropoda* 43:4-8.

Cremer, S., Ugelvig, L.V., Drijfhout, F.P., Schlick-Steiner, B.C., Steiner, F.M., Seifert, B., Hughes, D.P., Schulz, A., Petersen, K.S., Konrad, H., Stauffer, C., Kiran, K., Espadaler, X., d'Ettorre, P., Aktaç, N., Eilenberg, J., Jones, G.R., Nash, D.R., Pedersen, J.S., Boomsma, J.J. **2008**. The evolution of invasiveness in garden ants. *PLoS ONE* 3(12):e3838.

Cronk, Q.C.B. **1997**. Islands: stability, diversity, conservation. *Biodiversity and Conservation* 6: 477-493.

Edgar, R.C. **2004**. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32(5):1792-1797.

Feldhaar, H., Foitzik, S., Heinze, J. **2008**. Life-long commitment to the wrong partner: hybridization in ants. *Philosophical Transactions of the Royal Society B: Biological Sciences* 363: 2891–2899.

Folgarait, P.J. **1998**. Ant biodiversity and its relationship to ecosystem functioning: a review. *Biodiversity and Conservation* 7:1221-1244.

Fordham, D.A., Brook, B.W. **2010**. Why tropical island endemics are acutely susceptible to global change. *Biodiversity and Conservation* 19(2):329-342.

Hasegawa, E. **1998**. Phylogeny and host-parasite relationships in social parasitism in *Lasius* ants. *Entomological Science* 1:133-135.

Huelsenbeck, J.P., Ronquist, F. **2001**. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754–755.

IPCC Thrid Asessment report. **2001**. Climate Change 2001: The scientific basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York.

IUCN. **2001**. *IUCN Red List Categories and Criteria: Version 3.1.* IUCN Species Survival Commission. IUCN, Gland, Switzerland and Cambridge, UK.

IUCN. **2012**. IUCN Red List of Threatened Species. Version 2012.2. <www.iucnredlist.org>. Downloaded on 18 October 2012.

Janda, M., Folkova, D., Zrzavy, J. **2004**. Phylogeny of *Lasius* ants based on mitochondrial DNA and morphology, and the evolution of social parasitism in the Lasiini (Hymenoptera: Formicidae). *Molecular Phylogenetics and Evolution* 33:595-614.

Losos, J.B., Ricklefs, R.E [Eds]. **2010**. The Theory of Island Biogeography Revisited. Princeton University Press, Princeton, NJ.

López, J.M., Martinell, C., Massó, S., Blanché, C., Sáez, L. **2012**. The 'paradigm of extremes': Extremely low genetic diversity in a extremely narrow endemic species, Coristospermum huteri (Umbelliferae). *Plant Systematics and Evolution* In press.

Mace, G.M., Masundire, H., Baillie, J.E.M. **2005**. Biodiversity *in* Ecosystems and Human-Well Being: Current State and Trends, B. Scholes, R. Hassan, Eds. Island Press, Washington, DC, US.

Maruyama, M., Steiner, F.M., Stauffer, C., Stauffer, C., Akino, T., Crozier, R.H., Schlick-Steiner, B.C. **2008**. A DNA and morphology based phylogenetic framework of the ant genus *Lasius* with hypotheses for the evolution of social parasitism and fungiculture. *BMC Evolutionary Biology* 8:237–251.

Meulenkamp, J.E., Sissingh, W. **2003**. Tertiary palaeogeography and tectonostratigraphic evolution of the Northern and Southern Peri-Tethys platforms and the intermediate domains of the African-Eurasian convergent plate boundary zone. *Palaeogeography, Palaeoclimatology, Palaeoecology* 196:209-228.

New, M., Hulme, M., Jones, P. **1999**. Representing twentieth-century space-time climate variability. Part 1. Development of a 1961–90 mean monthly terrestrial climatology. *Journal of Climate* 12, 829–856. Data available at http://ipcc-ddc.cru.uea.ac.uk/cru data/examine/cru climate.html.

Pearson, B. **1983**. Hybridization between the ant species *Lasius niger* and *Lasius alienus* - the genetic evidence. *Insectes Sociaux* 30:402–411.

Peluelas, J., Boada, M. **2003**. A global change-induced biome shift in the Montseny mountains (NE Spain). *Global change Biology* 9(2):131-140.

Phillips, S.J., Anderson, R.P., Schapired, R.E. **2006**. Maximum entropy modeling of species geographic distributions. *Ecological Modelling* 190: 231-259.

Rigato, F. **1999**. *Myrmecina melonii* n. sp., a new ant from Sardinia, with a review of the West Palaearctic Myrmecina (Hymenoptera Formicidae). *Bollettino della Societa Entomologica Italiana* 131: 83-92.

Sáez, L., Rosselló, J.A. **2000**. A new species of Agrostis (Gramineae) belonging to the *A. alpina* complex. *Botanical Journal of the Linnean Society* 133(3):359-370.

Sáez, L., Rosselló, J.A. **2001**. Llibre Vermell de la flora amenaçada de les Illes Balears. Govern de les Illes Balears. Palma de Mallorca.

Schlick-Steiner, B.C., Steiner, F.M., Schödl, S., Seifert, B. **2003**. *Lasius austriacus* sp.n., a Central European ant related to the invasive species *Lasius neglectus*. *Sociobiology* 41:725-736.

Seifert, B. **1988**. A revision of the European species of the ant subgenus *Chthonolasius* (Insecta, Hymenoptera, Formicidae). *Entomologische Abhandlungen Staatlichen Museum für Tierkunde Dresden* 51:143-180.

Seifert B. **1992**. A taxonomic revision of the Palaearctic members of the ant subgenus *Lasius* s. str. (Hymenoptera: Formicidae). *Abh. Ber. Naturkundemus. Görlitz* 66(5):1–67.

Seifert, B. **1999**. Interspecific hybridisations in natural populations of ants by example of a regional fauna (Hymenoptera, Formicidae). *Insectes Sociaux* 46: 45–52.

Seifert, B. **2009**. Cryptic species in ants (Hymenoptera: Formicidae) revisited: we need a change in the alpha-taxonomic approach. *Myrmecological News* 12:149–166.

Steiner, F.M., Schlick-Steiner, B.C., Schödl, S., Espadaler, X., Seifert, B., Christian, E., Stauffer, C. **2004**. Phylogeny and bionomics of *Lasius austriacus* (Hymenoptera, Formicidae). *Insectes Sociaux* 51:24-29.

Sokal, R.R., Rohlf, J.F. **1995**. Biometry. Third edition. Freeman, NY, US.

Turvey, S.T [Ed]. **2009**. Holocene extinctions. Oxford University Press, Oxford, UK.

Umphrey, G.J. **2006**. Sperm parasitism in ants: selection for interspecific mating and hybridization. *Ecology* 87:2148–2159.

Van der Have, T.M., Pedersen, J.S., Boomsma, J.J. **2011**. Mating, hybridisation and introgression in Lasius ants (Hymenoptera: Formicidae). *Myrmecological News* 15:109-115.

Waite, T.A., Campbell, L.G. **2006**. Controlling the false discovery rate and increasing statistical power in ecological studies. *Ecoscience* 13(4):439-442.

Willis, K.J., Bennett, K.D., Walker, D. **2004**. The evolutionary legacy of the ice ages. *Philosophical Transactions of the Royal Society B* 359:155–303.

Wilson, E.O. **1955**. A monographic revision of the ant genus *Lasius*. *Bulletin of the Museum of Comparative Zoology* 113:3-205.

Wysocka, A., Krzysztofiak, L., Krzysztofiak, A., Zolnierkiewicz, O., Ojdowska, E., Sell, J. 2011. Low genetic diversity in Polish populations of sibling ant species: *Lasius niger* (L.) and *Lasius platythorax* Seifert (Hymenoptera, Formicidae). *Insectes Sociaux* 58:191-195.

Appendix 1

Formal description of *Lasius balearicus* Talavera, Espadaler & Vila, sp. nov.

WORKER. *Measurements*: mean; st. dev. (range): HL 0.88; 0.05 (0.73-0.97), HW 0.80; 0.05 (0.65-0.89), SL 0.85; 0.05 (0.67-0.93), HL/HW 1.09; 0.02 (1.04-1.12), SL/HL 0.96; 0.02 (0.91-1.03), PDCL 26.43; 4.87 (19-42), nHS 35.82; 5.38 (26-52), nHHT 29.78; 4.61 (17-40), nBH 22.92; 3.28 (13-30), Mandibular dents 8.55; 0.57 (8-10). N = 52 for all variables but PDCL (n=20) and Mandibular Dents (n=48).

Description of worker

A small *Lasius* s.str. Head longer than wide; marked hairiness characterizes *L. balearicus.* Specifically, the number of standing hairs on dorsal profile of the scape is the highest among known species of Pale arctic *Lasius*. In frontal view the head profile is entirely hairy, up to the mandible base. Clypeal pubescence rather diluted, similar to *L. platythorax*. Tibial hairs number is also very high, only surpassed by the Asiatic *Lasius hirsutus*, from which it differs by the much denser gaster tergite pubescence. Clypeal carina variable within a same nest: from being clearly expressed in the anterior two thirds to totally absent. Scape usually shorter than head length.

Surface characters: frontal head with very visible punctures. Microreticulum is more developed at the posterior third of the head surface, and most developed in the disk of pronotum. Space between punctures and microreticulum is shining.

Colour is also diagnostic. Entire body yellowish brown, with legs clearer. Taken in isolation they would remind recently eclosed *Lasius grandis* or *Lasius cinereus*. All examined workers have the same, consistent, colouration. When dry, the gaster is somewhat darker that the head and alitrunk.

Type material

Syntype workers. 15 workers. SPAIN: Mallorca, Coll d'es Prat, 1194 m; 13 October 2008; R. Vila & G. Talavera leg. Three workers deposited at each of the following institutions: MCZ Cambridge; BMNH London; SMN Görlitz. Three paratype workers from Camí de s'Arxiduc deposited at the Museu Balear de Ciències Naturals (Sóller, Mallorca). Other paratype material in authors collection.

Paratype workers. 15 workers, SPAIN: Mallorca, Valldemossa, Es Teix, Camí de s'Arxiduc, 919 m, 12 October 2008, R. Vila & G. Talavera leg.; 15 workers, SPAIN: Mallorca, pic Tomir, 1048 m, 14 October 2008, R. Vila & G. Talavera leg.; 3 workers, SPAIN: Mallorca, Puig d'en Galileu, 1210 m, R. Vila leg.; 2 workers, SPAIN: Mallorca, Coma de n'Arbona, 1200 m, 2009, A. Traveset leg.; 1 worker, SPAIN: Mallorca, Es Teix, October 1982, C.A. Collingwood leg.; 1 worker, SPAIN: Mallorca, Ses Clotades, 1330 m, 15 June 2011, C. Tur leg. In author's collection.

Etymology. The name is derived from the Balearic Islands.

SUPPLEMENTARY DOCUMENTATION

Discovered just before extinction? The first endemic ant from the Balearic Islands endangered by climate change

Gerard Talavera[a,b], Xavier Espadaler[c] and Roger Vila[a]

[a] Institut de Biologia Evolutiva (CSIC-UPF), Passeig Marítim de la Barceloneta, 37, 08003 Barcelona, Spain

[b] Departament de Genètica i Microbiologia, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain

[c] Departament de Biologia Animal, de Biologia Vegetal i d'Ecologia and CREAF, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain

Supplementary Table S1. Species, specimen codes, GenBank codes, localities and sources of sequences ([1] = this study; [2] = Maruyama *et al* 2008; [3] = Steiner *et al* 2004; [4] = Cremer *et al* 2008).

Species	Sample Code	COI	16S	Locality	Source
L. balearicus	RVcoll.08J418	X	X	Valldemossa, Es Teix. Mallorca. **Spain** *(920 m)*	[1]
L. balearicus	RVcoll.08J419	X	X	Valldemossa, Es Teix. Mallorca. **Spain** *(920 m)*	[1]
L. balearicus	RVcoll.08J420	X	X	Valldemossa, Es Teix. Mallorca. **Spain** *(920 m)*	[1]
L. balearicus	RVcoll.08J421	X	X	Coll des Prat, Escorca. Mallorca. **Spain** *(1194 m)*	[1]
L. balearicus	RVcoll.08J422	X	X	Coll des Prat, Escorca. Mallorca. **Spain** *(1194 m)*	[1]
L. balearicus	RVcoll.08J423	X	X	Coll des Prat, Escorca. Mallorca. **Spain** *(1194 m)*	[1]
L. balearicus	RVcoll.08J424	X	X	Pic Tomir. Mallorca. **Spain** *(1035 m)*	[1]
L. balearicus	RVcoll.08J425	X	X	Pic Tomir. Mallorca. **Spain** *(1041 m)*	[1]
L. balearicus	RVcoll.08J426	X	X	Pic Tomir. Mallorca. **Spain** *(1048 m)*	[1]
L. balearicus	RVcoll.12L413	X	X	Serra d'Alfàbia. Mallorca. **Spain** *(1066 m)*	[1]
L. balearicus	RVcoll.12L414	X	X	Serra d'Alfàbia. Mallorca. **Spain** *(1056 m)*	[1]
L. balearicus	RVcoll.12L419	X	X	Serra d'Alfàbia. Mallorca. **Spain** *(1025 m)*	[1]
L. balearicus	RVcoll.12L421	X	X	Puig Galatzó. Mallorca. **Spain** *(1006 m)*	[1]
L. balearicus	RVcoll.12L424	X	X	Puig Galatzó. Mallorca. **Spain** *(1012 m)*	[1]
L. balearicus	RVcoll.12L425	X	X	Puig Galatzó. Mallorca. **Spain** *(998 m)*	[1]
L. grandis	RVcoll.08J433	X	X	Riera de Fuirosos. Catalonia. **Spain**	[1]
L. grandis	RVcoll.08J435	X	X	Campus UAB. Catalonia. **Spain**	[1]
L. grandis	RVcoll.08J437	X	X	Campus UAB. Catalonia. **Spain**	[1]
L. grandis	RVcoll.08J438	X	X	Valldemossa, Son Moragues. Mallorca. **Spain**	[1]
L. cinereus	RVcoll.08J427	X	X	Riera de Fuirosos. Catalonia. **Spain**	[1]
L. cinereus	RVcoll.08J428	X	X	Riera de Fuirosos. Catalonia. **Spain**	[1]
L. cinereus	RVcoll.08J429	X	X	Riera de Fuirosos. Catalonia. **Spain**	[1]
L. cinereus	RVcoll.08J430	X	X	Pont de Muntanyana. Catalonia. **Spain**	[1]
L. cinereus	RVcoll.08J431	X	X	Coll d'Ares. Catalonia. **Spain**	[1]
L. japonicus	MMANT19	AB371015	AB371061	Kagawa-ken, Takamatsu-shi. **Japan**	[2]
L. japonicus	MMANT55	AB371014	AB371060	Chiba-ken, Kimitsu-shi. **Japan**	[2]
L. japonicus	MMANT76	AB371016	AB371062	Hokkaido,□Sapporo-shi. **Japan**	[2]
L. niger	MMANT26	AB371019	AB371065	Vienna. **Austria**	[2]
L. niger		AY225866		Vienna. **Austria**	[3]
L. platythorax	MMANT28	AB371020	AB371066	Moosbrunn. **Austria**	[2]
L. platythorax		AY225867		Moosbrunn. **Austria**	[3]
L. emarginatus	MMANT41	AB371011	AB371057	Vienna. **Austria**	[2]
L. emarginatus		AY225868		Vienna. **Austria**	[3]
L. hayashi	MMANT46	AB371013	AB371059	Gifu-ken, Kamitakara-mura. **Japan**	[2]
L. hayashi	MMANT54	AB371012	AB371058	Chiba-ken, Kimitsu-shi. **Japan**	[2]
L. productus	MMANT18	AB371021	AB371067	Kagawa-ken, Takamatsu-shi. **Japan**	[2]
L. sakagamii	MMANT29	AB371022	AB371068	Gifu-ken, Gifu-shi. **Japan**	[2]
L. sakagamii	MMANT56	AB371023	AB371069	-to, Edogawa-ku. **Japan**	[2]
L. sakagamii		AY225864		Gifu. **Japan**	[3]
L. alienus	MMANT21	AB371008	AB371054	Braunsberg. **Austria**	[2]
L. alienus		AY225865		Braunsberg. **Austria**	[3]
L. psammophilus		AY225863		Gfo□. **Austria**	[3]
L. brunneus		AY225877		Hof. **Austria**	[3]
L. brunneus	MMANT25	AB371010	AB371056	Rassing. **Austria**	[2]
L. austriacus	MMANT27	AB371009	AB371055	Feldberg. **Austria**	[2]
L. austriacus		AY225870		Feldberg. **Austria**	[3]
L. austriacus		AY225869		Braunsberg. **Austria**	[3]
L. austriacus		AY225873		Retz. **Austria**	[3]
L. austriacus		AY225871		Feldberg. **Austria**	[3]
L. austriacus		AY225872		Feldberg. **Austria**	[3]
L. neglectus	MMANT20	AB371018	AB371064	Budapest. **Hungary**	[2]
L. neglectus		AY225875		Budapest. **Hungary**	[3]
L. neglectus		AY225876		Debrecen. **Hungary**	[3]
L. turcicus		DQ975435		Maltepe. **Turkey**	[4]

L. turcicus	DQ975428		Mueezzinler. **Turkey**	[4]
L. turcicus	DQ975426		Bilecik. **Turkey**	[4]
L. turcicus	DQ975417		Kuelcueler. **Turkey**	[4]
L. turcicus	DQ975401		Pazaryeri. **Turkey**	[4]
L. lasioides	AY225874		Sant Cugat del Vallès. Catalonia. **Spain**	[3]
L. mixtus	AB370988	AB371034	Göpfritz. **Austria**	[3]

Supplementary Table S2. Presence records coordinates used for ecological niche modelling. Sources ([1] = Current study, [2] = www.hormigas.org, [3] = Comín del Río Thesis*)

Locality	Taxa	Longitude	Latitude	Source
Valldemossa (Puig des Teix)	*Lasius balearicus*	39.732428	2.648064	[1]
Coll des Prat (Puig Massanella)	*Lasius balearicus*	39.808294	2.851256	[1]
Coll des Prat (Puig Massanella)	*Lasius balearicus*	39.808347	2.852008	[1]
Coll des Prat (Puig Massanella)	*Lasius balearicus*	39.808483	2.852017	[1]
Pic Tomir	*Lasius balearicus*	39.837189	2.922072	[1]
Pic Tomir	*Lasius balearicus*	39.836853	2.922008	[1]
Pic Tomir	*Lasius balearicus*	39.836767	2.92215	[1]
Pic Tomir	*Lasius balearicus*	39.836967	2.922389	[1]
Pic Tomir	*Lasius balearicus*	39.838011	2.921797	[1]
Pic Tomir	*Lasius balearicus*	39.837967	2.920722	[1]
Coma n'Arbonna (Puig Major)	*Lasius balearicus*	39.801389	2.785833	[1]
Ses Clotades (Puig Major)	*Lasius balearicus*	39.809444	2.797222	[1]
Serra d'Alfàbia	*Lasius balearicus*	39.743472	2.732889	[1]
Serra d'Alfàbia	*Lasius balearicus*	39.74425	2.734944	[1]
Serra d'Alfàbia	*Lasius balearicus*	39.744889	2.735806	[1]
Serra d'Alfàbia	*Lasius balearicus*	39.748917	2.741056	[1]
Puig Galatzó	*Lasius balearicus*	39.633917	2.486889	[1]
Puig Galatzó	*Lasius balearicus*	39.633722	2.4865	[1]
Puig Galatzó	*Lasius balearicus*	39.634222	2.486611	[1]
Puig Caragoler	*Lasius balearicus*	39.87199	2.89341	[1]
Puig Caragoler	*Lasius balearicus*	39.87146	2.89360	[1]
Maria de la Salut	*Lasius grandis*	39.655472	3.077389	[1]
Serra d'Alfàbia	*Lasius grandis*	39.746	2.738306	[1]
Serra d'Alfàbia	*Lasius grandis*	39.740222	2.726667	[1]
Puig Galatzó	*Lasius grandis*	39.642972	2.468611	[1]
Escorca (Es Guix)	*Lasius grandis*	39.814978	2.890689	[1]
Serra de Torrellas, km 36.5	*Lasius grandis*	39.788889	2.779444	[1]
Valldemossa, Son Moragues	*Lasius grandis*	39.731964	2.651947	[1]
Cala Mondragó	*Lasius grandis*	39.3525	3.186944	[2]
Coll de sa Bastida	*Lasius grandis*	39.699444	2.539167	[2]
Cova de Sa Gleda (Manacor)	*Lasius grandis*	39.500278	3.276944	[2]
Es Portixol (Palma)	*Lasius grandis*	39.561667	2.665278	[2]
Selva	*Lasius grandis*	39.766667	2.9	[2]
Ses Aufanes (Campanet)	*Lasius grandis*	39.766667	2.966667	[2]
Son Bunyola	*Lasius grandis*	39.699444	2.539167	[2]
Selva	*Lasius grandis*	39.769722	2.900278	[2]
S'Albufera	*Lasius grandis*	39.779006	3.133164	[3]

Sa Pobla	*Lasius grandis*	39.77145	3.014386	[3]
Salobrar	*Lasius grandis*	39.325875	2.991667	[3]
Esporlas-La Granja	*Lasius grandis*	39.668889	2.56	[3]
Deià (Son Gallard)	*Lasius grandis*	39.740278	2.623333	[3]
Son Sardina	*Lasius grandis*	39.619167	2.655	[3]
Algaida	*Lasius grandis*	39.556014	2.903508	[3]
Inca (Plaça Cerdós)	*Lasius grandis*	39.723339	2.912911	[3]
Coll de Sóller (Sóller)	*Lasius grandis*	39.733333	2.69	[3]
Embassament de Cúber	*Lasius grandis*	39.785833	2.796944	[3]
Valldemossa	*Lasius grandis*	39.7225	2.606389	[3]
S'Albufera	*Lasius grandis*	39.809444	3.108611	[3]
S'Albufera	*Lasius grandis*	39.794167	3.086389	[3]
S'Albufera	*Lasius grandis*	39.774167	3.134722	[3]
Sa Pobla	*Lasius grandis*	39.750556	3.015556	[3]
Sa Pobla	*Lasius grandis*	39.774167	2.989722	[3]
Cala Pi	*Lasius grandis*	39.374117	2.864167	[3]
Colònia Sant Jordi	*Lasius grandis*	39.326944	2.992	[3]
Palma	*Lasius grandis*	39.555142	2.592853	[3]

*Comín del Río, P. **1988**. Estudio de los formícidos de Baleares: Contribución al estudio taxonómico, geográfico y biológico. **Doctoral Thesis**, Universidad de las Islas Baleares.

Supplementary Table S3. Measures for 11 variables for the *Lasius* s.str. species included in our study plus *L. mixtus*. Measures obtained by us are highlighted in red. The rest of the values were extracted from literature (Seifert, 1988; 1992; Schlick-Steiner *et al*, 2003). The number of specimens measured for each species is shown at the top of the columns or specifically the characters if different.

	balearicus	*niger*	*platythorax*	*japonicus*	*grandis*	*cinereus*	*emarginatus*	*hayashi*	*productus*	*sakagamii*	*alienus*	*psammophilus*	*brunneus*	*austriacus*	*neglectus*	*turcicus*	*lasioides*	*mixtus*
	(n=51)	*(n=5)*	*(n=5)*	*(n=5)*	*(n=5)*	*(n=22)*	*(n=5)*		*(n=5)*						*(n=17)*			*(n=10)*
HW	811.30	926.72	952.80	923.21	973.26	831.37	929.41	980.87	1011.52	841.03	781.75	785.57	915.27	666.31	663.55	813.52	793.36	990.10
HL	886.58	981.40	985.20	982.30	1037.50	917.00	1001.90	1010.30	1114.70	914.20	848.20	843.70	947.30	747.60	793.60	896.50	848.10	1056.30
CS	848.94	954.06	969.00	952.76	1005.38	874.18	965.65	995.59	1063.11	877.62	814.98	814.63	931.28	706.96	728.57	855.01	820.73	1023.20
SL	857.56	920.55	937.91	953.81	1021.94	898.66	1003.90	951.70	1207.22	914.20	781.19	796.45	807.10	654.15	743.92	846.03	804.00	913.71
SL/CS	1.01	0.96	0.97	1.00	1.02	1.03	1.04	0.96	1.14	1.04	0.96	0.98	0.87	0.93	1.02	0.99	0.98	0.89
HL/HW	1.09	1.06	1.03	1.06	1.07	1.10	1.08	1.03	1.10	1.09	1.09	1.07	1.04	1.12	1.20	1.10	1.07	1.07
SL/HL	0.97	0.94	0.95	0.97	0.99	0.98	1.00	0.94	1.08	1.00	0.92	0.94	0.85	0.88	0.94	0.94	0.95	0.84
PDCL	26.44 (n=19)	13.00	25.90	18.70	19.40	20.50	27.30	20.80	18.10	14.60	17.30	22.50	27.60	33.17	31.24	32.15	31.40	14.70
nHS	35.82	16.00	21.00	15.70	22.80	22.20	12.80	21.30	10.60	29.90	0.10	0.60	0.00	0.05	0.32	0.06	0.00	5.60
nHHT	29.78	17.70	23.30	16.90	25.70	21.70	20.60	22.40	9.10	27.80	0.90	2.30	0.20	0.16	0.29	0.53	0.00	7.26
MaDe	8.55	8.22	8.25	8.14	8.48	8.00	8.67	8.00	8.45	8.60	8.11	8.21	7.00	7.33	7.35	7.35	7.02	8.30
nBH	22.92	15.80	14.00	15.50 (n=4)	18.80	14.00	13.40	16.40	12.10	23.60	4.70	8.50	2.30	4.83	10.02	8.55	3.20	14.60
nHCl	12.98	5.00	6.40	4.75	9.60	6.40	7.00											
HWQ	1599 (n=2)	1617.30	1568.90	1609.00	1674.80	1508.00	1592.20	1693.00	1825.00	1591.50	1542.00	1612.90	1541.40	1399.00 (n=5)	1340.00	1380.80	1396.30	1495.00
HLQ	1499 (n=2)	1430.10	1388.20	1445.00	1529.80	1400.00	1454.60	1518.00	1671.00	1444.00	1383.20	1417.20	1384.70	1253.00 (n=5)	1200.00	1230.00	1234.90	1365.00
HLQ/HWQ	0.94	0.88	0.88	0.90	0.91	0.94	0.91	0.90	0.92	0.91	0.90	0.88	0.90	0.90	0.89	0.89	0.89	0.91
SLQ	1299 (n=2)	1680.69	1622.11	1696.61	1773.06	1198.00	1632.36	1878.71	1528.00	1682.00	1648.04	1650.59	1754.56	1082.00 (n=5)	1030.00	1428.24	1392.22	1141.00
SLQ/HLQ	0.87	0.85	0.86	0.85	0.86	0.87	0.89	0.81	0.90	0.86	0.84	0.86	0.79	0.86	0.86	0.86	0.89	0.84
SLQ/HWQ	0.81	0.75	0.76	0.76	0.79	0.81	0.81	0.72	0.84	0.78	0.75	0.75	0.71	0.77	0.77	0.77	0.79	0.76
MLQ	2906 (n=1)	3011.00	2781.00	2740.00	3078.00	2763.00	2924.00	2796.00	3233.00	3043.00	2914.00	3041.00	2647.80	2859.00 (n=5)	2473.00	2726.00	2470.90	2213.00
MHQ	1506 (n=1)	1800.58	1479.49	1471.38	1557.47	1362.00	1403.52	1328.10	1600.00	1710.17	1693.03	1766.82	1218.52	1659.00 (n=5)	1273.00	1433.00	1141.56	1279.00
MHQ/MLQ	0.52	0.60	0.53	0.54	0.51	0.49	0.48	0.48	0.49	0.56	0.58	0.58	0.46	0.58	0.51	0.53	0.46	0.58

Supplementary Table S4. t-Student tests, false discovery rate and Bonferroni corrections for ten morphological character pair comparisons between *L. balearicus* and closest *Lasius s. str.* taxa (*niger*, *platythorax*, *japonicus*, *grandis*, *cinereus* and *emarginatus*). Pair comparisons in red indicates p < 0.001 for FDR.

Character	Pair comparison	d.f.	t	P	order	FDR	Bonferroni
HW	bal-nig	51	-15.5958818	0	1	0.000757576	0.000909091
HL	bal-nig	51	-13.45527349	0	2	0.001515152	0.000909091
SL	bal-nig	51	-8.648149874	0	3	0.002272727	0.000909091
HL/HW	bal-nig	51	9.873328064	0	4	0.003030303	0.000909091
SL/HL	bal-nig	51	8.238030783	0	5	0.003787879	0.000909091
PDCL	bal-nig	19	11.68386695	0	6	0.004545455	0.000909091
nHS	bal-nig	51	26.04355663	0	7	0.00530303	0.000909091
nHHT	bal-nig	51	18.51193198	0	8	0.006060606	0.000909091
nBH	bal-nig	51	15.31086278	0	9	0.006818182	0.000909091
nHCl	bal-nig	51	22.21822274	0	10	0.007575758	0.000909091
HW	bal-pla	51	-19.1199749	0	11	0.008333333	0.000909091
HL	bal-pla	51	-13.9944816	0	12	0.009090909	0.000909091
SL	bal-pla	51	-11.03119456	0	13	0.009848485	0.000909091
HL/HW	bal-pla	51	16.99880377	0	14	0.010606061	0.000909091
nHS	bal-pla	51	19.47470703	0	15	0.011363636	0.000909091
nHHT	bal-pla	51	9.933305053	0	16	0.012121212	0.000909091
nBH	bal-pla	51	19.18073394	0	17	0.012878788	0.000909091
nHCl	bal-pla	51	18.32048047	0	18	0.013636364	0.000909091
HW	bal-jap	51	-15.12174253	0	19	0.014393939	0.000909091
HL	bal-jap	51	-13.58298067	0	20	0.015151515	0.000909091
SL	bal-jap	51	-13.21457214	0	21	0.015909091	0.000909091
HL/HW	bal-jap	51	8.448232922	0	22	0.016666667	0.000909091
nHS	bal-jap	51	26.4376876	0	23	0.017424242	0.000909091
nHHT	bal-jap	51	19.73745011	0	24	0.018181818	0.000909091
nBH	bal-jap	51	15.95584131	0	25	0.018939394	0.000909091
nHCl	bal-jap	51	22.91424815	0	26	0.01969697	0.000909091
HW	bal-gra	51	-21.88450487	0	27	0.020454545	0.000909091
HL	bal-gra	51	-21.41568799	0	28	0.021212121	0.000909091
SL	bal-gra	51	-22.56763673	0	29	0.021969697	0.000909091
HL/HW	bal-gra	51	7.878194865	0	30	0.022727273	0.000909091
nHS	bal-gra	51	17.10992118	0	31	0.023484848	0.000909091
nHHT	bal-gra	51	6.256750658	0	32	0.024242424	0.000909091
nBH	bal-gra	51	8.861077525	0	33	0.025	0.000909091
nHCl	bal-gra	51	9.411355285	0	34	0.025757576	0.000909091
SL	bal-cin	51	-5.642338228	0	35	0.026515152	0.000909091
nHS	bal-cin	51	17.89818313	0	36	0.027272727	0.000909091
nHHT	bal-cin	51	12.38434132	0	37	0.028030303	0.000909091
nBH	bal-cin	51	19.18073394	0	38	0.028787879	0.000909091
nHCl	bal-cin	51	18.32048047	0	39	0.029545455	0.000909091
DeMa	bal-cin	47	6.508849562	0	40	0.03030303	0.000909091
HW	bal-ema	51	-15.95840511	0	41	0.031060606	0.000909091
HL	bal-ema	51	-16.36415936	0	42	0.031818182	0.000909091
SL	bal-ema	51	-20.09171245	0	43	0.032575758	0.000909091
SL/HL	bal-ema	51	-9.714202537	0	44	0.033333333	0.000909091
nHS	bal-ema	51	30.24762037	0	45	0.034090909	0.000909091
nHHT	bal-ema	51	14.06942875	0	46	0.034848485	0.000909091
nBH	bal-ema	51	20.47069099	0	47	0.035606061	0.000909091
nHCl	bal-ema	51	16.6500195	0	48	0.036363636	0.000909091

PDCL	bal-jap	19	6.72799567	0.000002	49	0.037121212	0.000909091
PDCL	bal-gra	19	6.119379898	0.000007	50	0.037878788	0.000909091
SL/HL	bal-gra	51	-4.945640561	0.00001	51	0.038636364	0.000909091
DeMa	bal-jap	47	4.861609942	0.00002	52	0.039393939	0.000909091
HL/HW	bal-ema	51	4.457966525	0.00005	53	0.040151515	0.000909091
PDCL	bal-cin	19	5.162983686	0.00006	54	0.040909091	0.000909091
SL/HL	bal-pla	51	4.310979745	0.00008	55	0.041666667	0.000909091
HL	bal-cin	51	-4.317114944	0.00008	56	0.042424242	0.000909091
DeMa	bal-nig	47	3.920330159	0.0003	57	0.043181818	0.000909091
DeMa	bal-pla	47	3.567350241	0.0009	58	0.043939394	0.000909091
SL/HL	bal-cin	51	-3.543122333	0.0009	59	0.04469697	0.000909091
HW	bal-cin	51	-2.711658179	0.0092	60	0.045454545	0.000909091
HL/HW	bal-cin	51	-2.667509183	0.011	61	0.046212121	0.000909091
DeMa	bal-ema	47	-1.374368619	0.18	62	0.046969697	0.000909091
SL/HL	bal-jap	51	-1.018589522	0.32	63	0.047727273	0.000909091
DeMa	bal-gra	47	0.861170865	0.4	64	0.048484848	0.000909091
PDCL	bal-ema	19	-0.749283808	0.47	65	0.049242424	0.000909091
PDCL	bal-pla	19	0.467947735	0.65	66	0.05	0.000909091

Supplementary Table S5. Morphological set of characters used in phylogenetic inference. Blue: criteria to convert biometric measurements to discrete data. Orange: numbers correspond to selected characters from Maruyama *et al* (2008) (character 103 was newly incorporated). Below, resulting character states for the studied species.

Discrete values for biometric mesaures	**HW**	<750(0), >750 and <900 (1), >900 (2)
	HL	<800(0), >800 and <950 (1), <950(2)
	CS	<750(0), >750 and <900 (1), >900 (2)
	SL	<700(0), >700 and <860 (1), >860 and < 1050 (2), >1050 (3)
	SL/CS	<0.95(0), >0.95 and <1.1 (1), >1.1(2)
	HL/HW	<1.05(0), >1.05 and <1.15 (1), >1.15(2)
	SL/HL	<0.90(0), >0.90 and <1.05 (1), >1.05(2)
	PDCL	<15(0), >15 and <25 (1), >25 and < 30 (2), >30 (3)
	nHS	<1 (0), >1 and <20 (1), >20 and <25 (2), >25 (3)
	nHHT	<5(0), >5 and <15 (1), >15(2)
	MaDe	<7.5(0), >7.5(1)
	nBH	<5(0), >5 and <20 (1), >20(2)
	HWQ	<1400(0), >1400 and <1800 (1), >1800(2)
	HLQ	<1300(0), >1300 and <1600 (1), >1600(2)
	HL/HW	<0.89 (0), >0.89 and <0.90 (1), >0.90 and <0.93 (2), >0.93 (3)
	SLQ	<1200 (0), >1200 and <1600 (1), >1600 and <1800 (2), >1800 (3)
	SL/HL	<0.8 (0), >0.8 and <0.83 (1), >0.83 and <0.85 (2), >0.85 and <0.88 (3) >0.88 (4)
	SL/HW	<0.75 (0), >0.75 and <0.8 (1), >0.8 and <0.83 (2), >0.83 (3)
	MLQ	<2400 (0), >2400 and <2600 (1), >2600 and <3200 (2), >3200 (3)
	MHQ	<1300(0), >1300 and <1600 (1), >1600(2)
	MHQ/MLQ	<0.49 (0), >0.49 and <0.55 (1), >0.55 (2)

Maruyama *et al* (2008) (adapted)

1 QM, Body colour: black to brown pigmentation (0), without pigmentation, yellow to reddish yellow (1).

2 W, Body colour: black to brown pigmentation (0), without pigmentation, yellow to reddish yellow (1).

3 WQ, Mouthparts: mandible with last basal tooth not clearly separated from masticatory boarder (0), basal boarder in right angle with masticatory boarder (1).

4 W, Mouthparts: mandible without offset tooth absent (0), with offset tooth as minute projection (1), with offset tooth (2).

7 M, Mouthparts: mandible with preapical cleft (0), without preapical cleft (1).

8 M, Mouthparts: mandible with masticatory margin emarginate (0), convex (1).

11 W, Mouthparts: maxillary palpus reaches to neck cavity (0), does not reach to neck cavity (1).

12 QM, Mouthparts: maxillary palpus long, reaching level of posterior margin of eye (0), reach to level of fronterior margin of eye (1).

13 WQ, Mouthparts: maxillary palpus with 5th and 6th segments almost as long as 4th (0), 5th and 6th segments conspicuously reduced compared to 4th (1).

14 W, Mouthparts: maxillary palpus with 4th segment as long as 3rd (0), 4th conspicuously reduced to 3rd (1).

15 W, Mouthparts: maxillary palpus with 6th segment longer to only slightly shorter than 5th (0), conspicuously shorter than 5th (1).

18 W, Antenna: scape with pubescence erect, rough surface (0), decumbed to suberect surface, with hairs (1), appressed, smooth surface (2).

19 W, Antenna: scape with setae many (2), few (1), absent (0).

20 Q, Antenna: setae on scape many (2), few (1), absent (0).

21 Q, Antenna: scape long, more than 0.7 as long as head width (0), short, less than 0.7 as long as head width (1).

24 W, Genae with setae absent (0), few (1), many (2).

25 Q, Genae setae absent (0), present (1).

28 W, Head: setae simplified (0), more or less flattened and serrate laterally (1).

30 Q, Head: occipital margin almost strait or slightly emarginate (0), conspicuously emarginate (1).

32 WQ, Head: mandibular gland not developed, less than 0.3 as long as head length (0), well developed, more than 0.5 as long as head length (1).

33 QM, Mesosoma broader than head (0), narrower than head (1).

34 W, Mesosoma: setae simplified (0), more or less flattened and serrate laterally (1).

35 Q, Mesosoma: setae simplified (0), more or less flattened and serrate (1).

36 M, Mesosoma: pronotum in lateral view slightly narrowed anteriorly (0), conspicuously narrowed anteriorly part less than 0.5 as high as highest part (1).

38 Q, Mesosoma: mesonotum 0.8-1.2 as high as pronotum in height (0), less than 0.5 as high as pronotum (1).

41 Q, Mesosoma: suture between katepisternum and propodeum clear (0), not clear, poorly differentiated (1).

49 QM, Mesosoma: max. dimension of metapleural gland opening less wide than max. diameter of outer margin of propodeal spiracle (0), more wide (1).

54 M, Mesosoma: metapleural gland with sulcus its anterior (0), without sulcus (1).

57 QM, Wings: hyaline uniformly in basal 1/3 (0), brownish in basal part (1).

61 W, Petiole: in frontal view sides parallel (0), convex (1), diverging dorsad (2).

62 W, Petiole: shape of dorsal crest straight (0), emarginated (1), curved (2).

64 Q, Hind tibia: setae absent (0), few (1), many (2).

65 W, Hind tibia: setae absent (0), few (1), many (2).

66 W, Hind tibia: length of setae less than 30 µm (0), 30 to 60 µm (1), over 60 µm (2).

67 Q, Gaster more than 1.5 as broad as mesosoma (0), less than 1.3 as broad as mesosoma (1).

75 Q, Overall body size not over 5 mm (0), about 5 mm and more (1).

86 Colony founding independent (0), parasitic on Lasius s.str. (1)

87 Activity of workers epigaeic (0), hypogaeic (1).

88 Colony monogynous (0), oligogynous or polygynous (1).

90 M, Number of teeth on masticatory border 1-4 (0), over 6 (1), nil (2).

91 WQ, Maxillary palp: length of 4th segment less than 0.12 head width (0), more than 0.14 head width (1).

92 W, Occipital head margin straight (0), straight to feebly convex (1), strongly convex (2).

99 Q, Scape length/head length ratio below 0.65 (0), between 0.66 and 0.81 (1), between 0.82 and 0.91 (2), over 0.91 (3).

101 Q, Scape pubescence character: fully appressed, smooth surface (0), moderately pubescent, decumbent (1), subdecumbent, rough surface (2).

102 Q, Whole surface of head without setae (0), covered by setae (1).

103 W. Side of pronotum smooth, shining (0), dull, mat (1)

	Maruyama *et al* (2008) (adapted)																						
	1	**2**	**3**	**4**	**7**	**8**	**11**	**12**	**13**	**14**	**15**	**18**	**19**	**20**	**21**	**24**	**25**	**28**	**30**	**32**	**33**	**34**	**35**
L. balearicus	**1**	**1**	**0**	**0**	**1**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**2**	**?**	**?**	**2**	**?**	**1**	**?**	**0**	**0**	**1**	**?**
L. grandis	0	0	0	0	1	1	0	0	0	0	0	0	2	2	1	2	1	1	0	0	0	1	0
L. cinereus	0	0	0	0	1	1	0	0	0	0	0	0	2	2	1	2	1	1	0	0	0	1	0
L. brunneus	0	0	0	0	1	1	0	0	0	0	0	2	0	2	1	0	1	1	0	0	0	1	0
L. neglectus	0	0	0	0	1	1	0	0	0	0	0	1	0	1	1	0	0	1	0	0	0	1	0
L. sakagamii	0	0	0	0	1	1	0	0	0	0	0	0	2	0	1	2	0	1	0	0	0	1	0
L. emarginatus	0	0	0	0	1	1	0	0	0	0	0	0	2	0	1	2	0	1	0	0	0	1	0
L. platythorax	0	0	0	0	1	1	0	0	0	0	0	0	2	0	1	2	0	1	0	0	0	1	0
L. niger	0	0	0	0	1	1	0	0	0	0	0	0	2	0	1	2	0	1	0	0	0	1	0
L. japonicus	0	0	0	0	1	1	0	0	0	0	0	0	2	0	1	2	0	1	0	0	0	1	0
L. alienus	0	0	0	0	1	1	0	0	0	0	0	1	0	1	1	0	1	1	0	0	0	1	0
L. hayashi	0	0	0	0	1	1	0	0	0	0	0	1	2	0	1	1	0	1	0	0	0	1	0
L. productus	0	0	0	0	1	1	0	0	0	0	0	1	2	2	1	1	0	1	0	0	0	1	0
L. austriacus	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1	(01)	0	1	0	0	0	1	0

	Maruyama *et al* (2008) (adapted)																						
	36	**38**	**41**	**49**	**54**	**57**	**61**	**62**	**64**	**65**	**66**	**67**	**75**	**86**	**87**	**88**	**90**	**91**	**92**	**99**	**101**	**102**	**103**
L. balearicus	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**(01)**	**?**	**2**	**2**	**0**	**1**	**0**	**0**	**?**	**2**	**1**	**0**	**?**	**?**	**?**	**0**
L. grandis	1	0	0	1	0	0	1	1	2	2	2	0	1	0	0	0	2	1	0	2	2	1	0
L. cinereus	1	0	0	1	0	0	0	1	2	2	2	0	1	0	0	0	2	1	0	2	2	1	1
L. brunneus	1	0	0	1	0	0	0	1	1	0	4	0	1	0	0	(01)	2	1	0	1	0	0	0
L. neglectus	1	0	0	1	0	0	1	(01)	1	1	2	0	1	0	0	1	2	1	0	2	2	1	0
L. sakagamii	1	0	0	1	0	0	0	(01)	2	2	2	0	1	0	0	1	2	1	0	2	2	1	0
L. emarginatus	1	0	0	1	0	0	1	(012)	2	2	2	0	1	0	0	0	2	1	0	2	2	1	0
L. platythorax	1	0	0	1	0	0	1	(01)	2	2	2	0	1	0	0	0	2	1	0	2	2	1	0
L. niger	1	0	0	1	0	0	1	1	2	2	2	0	1	0	0	0	2	1	0	2	2	1	0
L. japonicus	1	0	0	1	0	0	1	1	2	2	2	0	1	0	0	(01)	2	1	0	2	2	1	0
L. alienus	1	0	0	1	0	0	1	(01)	1	1	0	0	1	0	0	0	2	1	0	2	1	0	0
L. hayashi	1	0	0	1	0	0	1	(02)	2	2	2	0	1	0	0	0	2	1	0	2	2	1	0
L. productus	1	0	0	1	0	0	1	(01)	1	1	2	0	1	0	0	0	2	1	0	2	2	1	0
L. austriacus	1	0	0	1	0	0	1	1	1	0	0	0	1	0	0	0	2	1	0	2	2	0	0

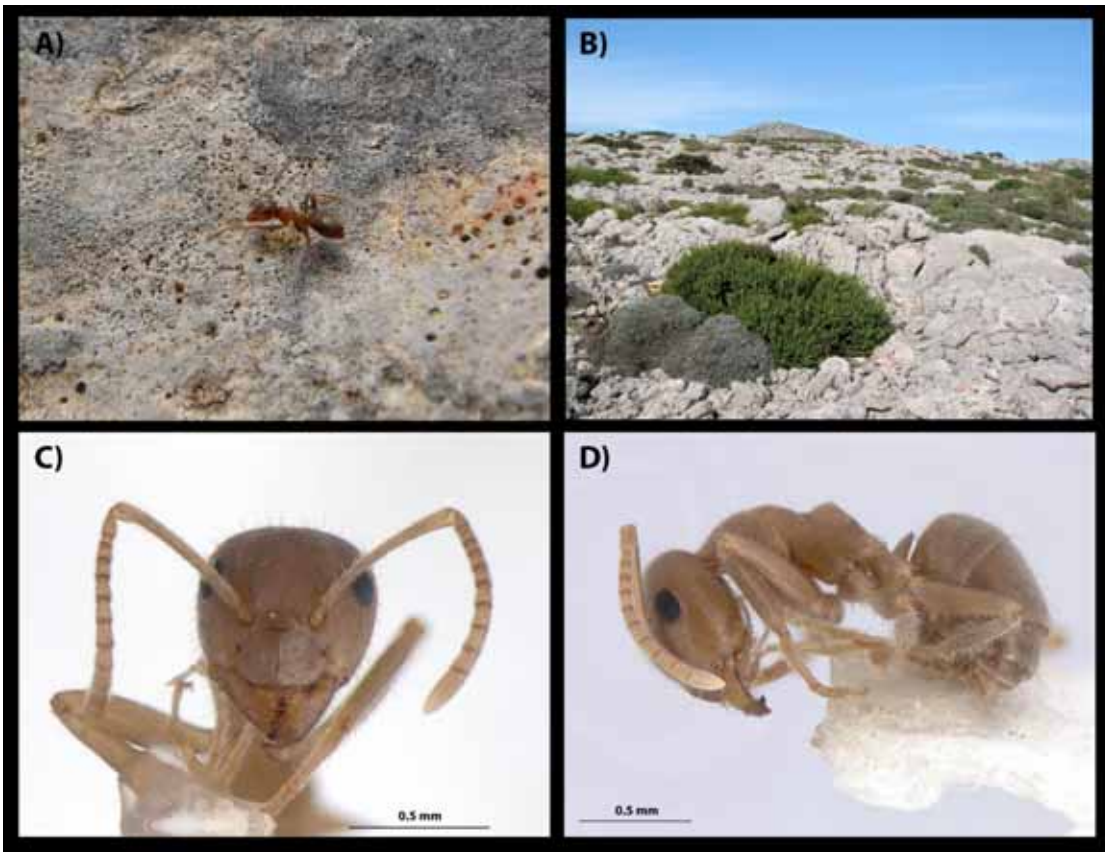
	Discrete values for biometric mesaures																			
	HW	HL	CS	SL	SL/CS	HL/HW	SL/HL	PDCL	nHS	nHHT	MaDe	nBH	HWQ	HLQ	HL/HW	SLQ	SL/HL	SL/HW	MLQ	MHQ	MHQ/MLQ
L. balearicus	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**2**	**3**	**2**	**1**	**2**	**1**	**1**	**3**	**1**	**3**	**2**	**2**	**1**	**1**
L. grandis	2	2	2	2	1	1	1	1	2	2	1	1	1	1	2	2	3	1	2	1	1
L. cinereus	1	1	1	2	1	1	1	1	2	2	1	1	1	1	3	0	3	2	2	1	1
L. brunneus	2	1	2	1	0	0	0	2	0	0	0	0	1	1	1	2	0	0	2	0	0
L. neglectus	0	0	0	1	1	2	1	3	0	0	0	1	0	0	1	0	3	1	1	0	1
L. sakagamii	1	1	1	2	1	1	1	0	3	2	1	2	1	1	2	2	3	1	2	2	2
L. emarginatus	2	2	2	2	1	1	1	2	1	2	1	1	1	1	2	2	4	2	2	1	0
L. platythorax	2	2	2	2	1	0	1	2	2	2	1	1	1	1	0	2	3	1	2	1	1
L. niger	2	2	2	2	1	1	1	0	1	2	1	1	1	1	0	2	3	1	2	2	2
L. japonicus	2	2	2	2	1	1	1	1	1	2	1	1	1	1	1	2	3	1	2	1	1
L. alienus	1	1	1	1	1	1	1	1	0	0	1	0	1	1	1	2	2	1	2	2	2
L. hayashi	2	2	2	2	1	0	1	1	2	2	1	1	1	1	1	3	1	0	2	1	0
L. productus	2	2	2	3	2	1	2	1	1	1	1	1	2	2	2	1	4	3	3	1	1
L. austriacus	0	0	0	0	0	1	0	3	0	0	0	0	0	0	1	0	3	1	2	2	2

Supplementary Table S6. Heuristic estimates of relative contributions of the environmental variables to the MAXENT distribution models for *Lasius balearicus* and *Lasius grandis*.

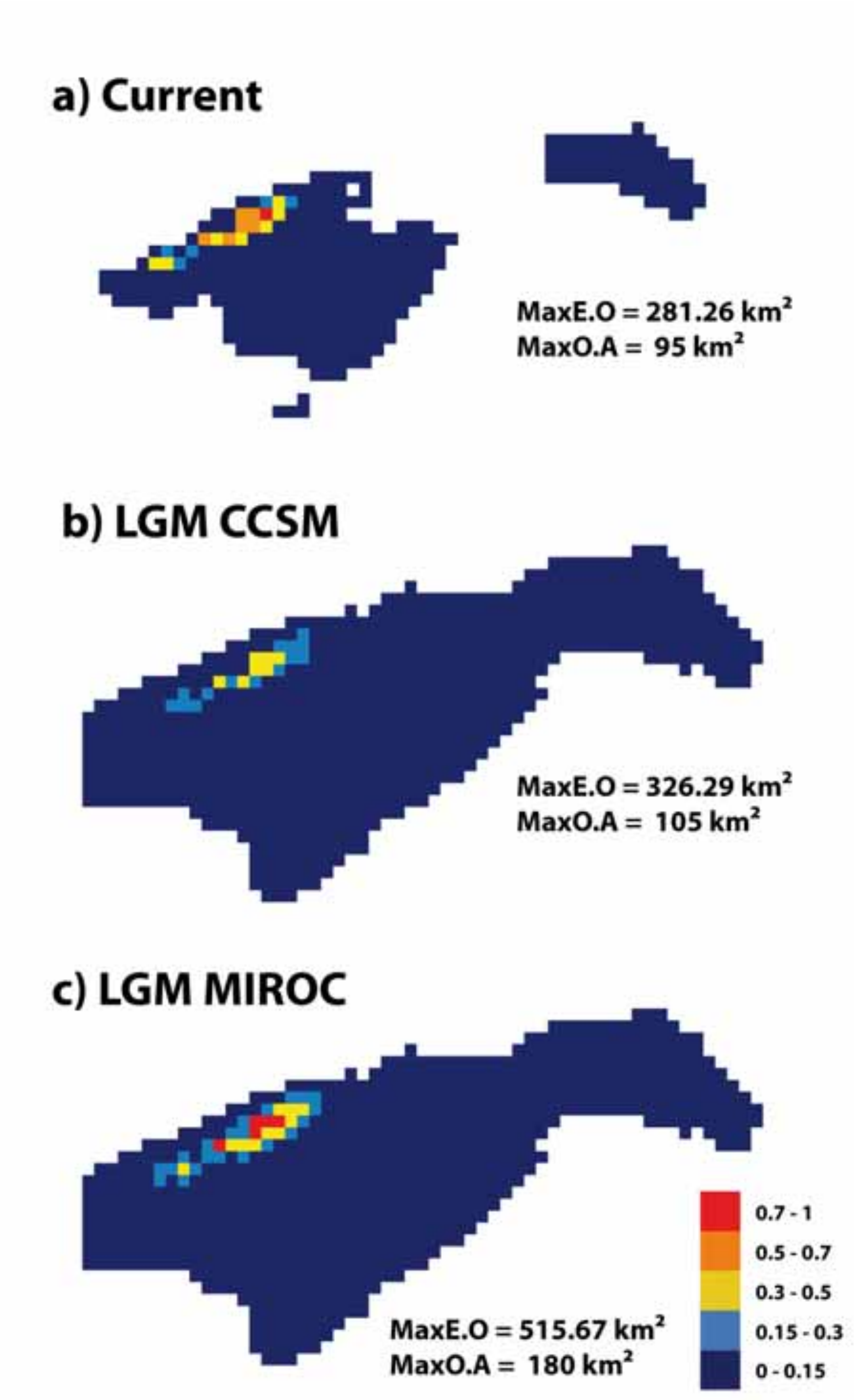
L. balearicus	**Variable**	**Per cent contribution (%)**
	Annual mean temperature (Bio1)	89.1
	Mean diurnal range (Bio2)	1
	Isothermality (Bio3)	3
	Temperature seasonality (Bio4)	0.1
	Temperature annual range (Bio7)	0.5
	Annual precipitation (Bio12)	0.5
	Precipitation seasonality (Bio15)	4.8
	Precipitation of driest quarter	1.1
L. grandis		
	Annual mean temperature (Bio1)	21.8
	Mean diurnal range (Bio2)	2.3
	Isothermality (Bio3)	2.1
	Temperature seasonality (Bio4)	6.9
	Temperature annual range (Bio7)	0.6
	Annual precipitation (Bio12)	1.8
	Precipitation seasonality (Bio15)	46.4
	Precipitation of driest quarter	18.2

Supplementary Table S7. Intra-specific genetic distances, evolutionary ages and monophyly supports estimated for the *Lasius* species studied.

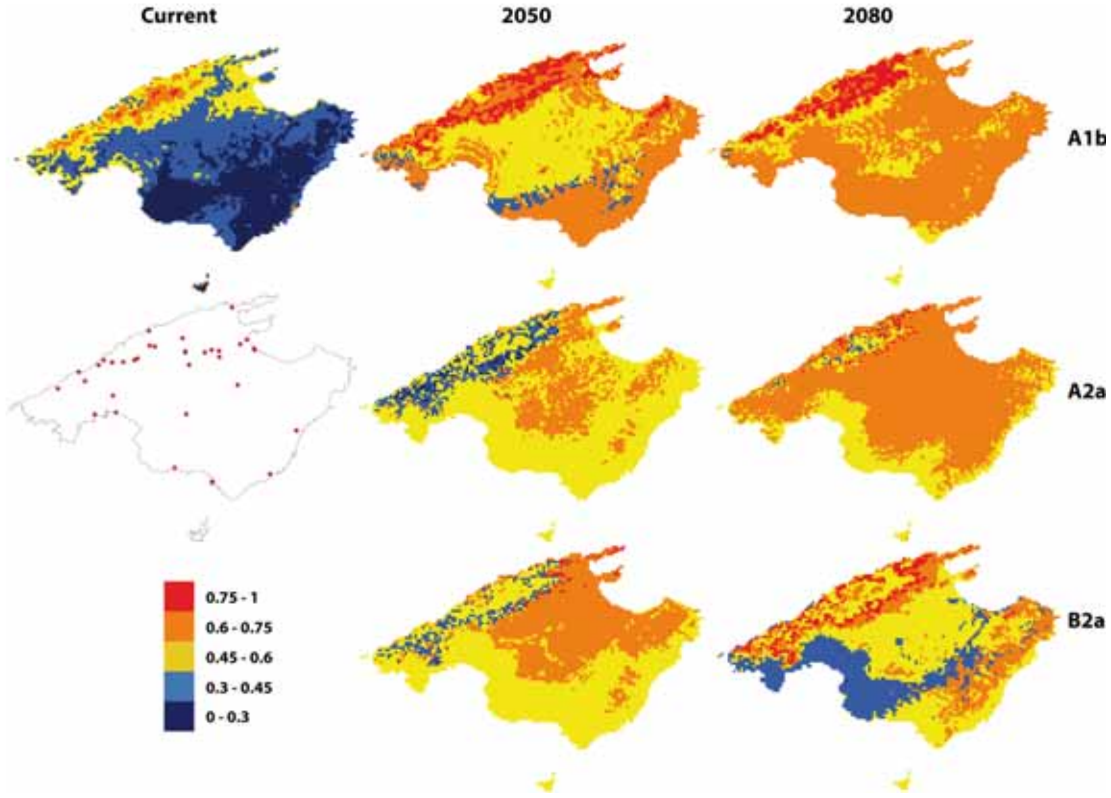
Taxa	**n**	**Monophyly support**	**Intra-specific genetic distance**	**Myr [95% HPD]**
L. balearicus	15	100	0.001	1.45 [0.85-2.23]
L. cinereus	3	paraphyletic	0.004	n.a.
L. grandis	4	paraphyletic	0.001	n.a.
L. japonicus	3	100	0.007	0.95 [0.52-1.49]
L. niger	2	100	0.000	1.15 [0.66-1.76]
L. platythorax	2	100	0.000	0.95 [0.52-1.49]
L. emarginatus	2	100	0.000	1.96 [1.2-2.99]
L. hayashi	2	63	0.007	1.74 [0.45-2.21]
L. sakagamii	3	100	0.004	1.85 [0.88-2.95]
L. alienus	2	100	0.000	2.36 [1.35-3.47]
L. psammophilus	1	monotypic	n.a.	1.85 [0.88-2.95]
L. brunneus	2	100	0.000	7.89 [4.75-11.52]
L. austriacus	6	100	0.003	4.08 [2.25-6.21]
L. neglectus	3	100	0.000	1.90 [0.97-3.14]
L. turcicus	5	100	0.011	1.90 [0.97-3.14]
L. lasioides	1	monotypic	n.a.	6.30 [3.76-9.24]
L. productus	1	monotypic	n.a.	1.74 [0.79-2.76]



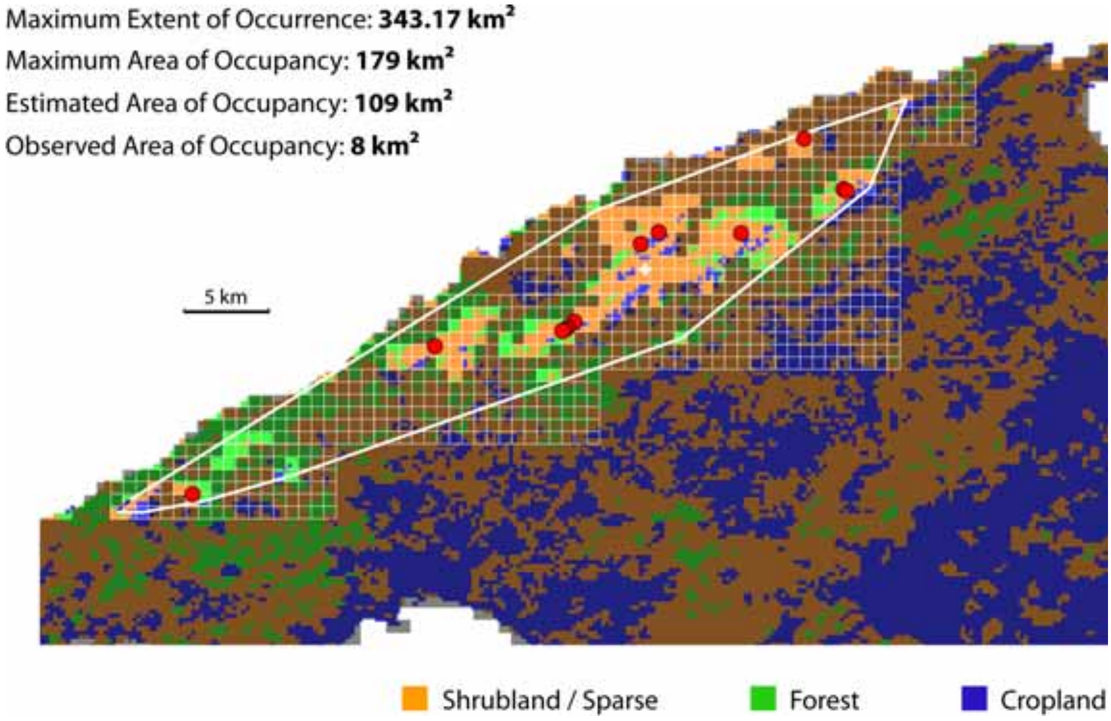
Supplementary Figure S1. Pictures. A) *Lasius balearicus* sp. nov. worker in nature. B) Typical habitat for the species. C) Frontal view and D) lateral view of a worker specimen.



Supplementary Figure S2. LGM projection. Model projection for *L. balearicus* for the present (a) and the Last Glacial Maximum (~21,000 years BP) according to the general atmospheric circulation models CCSM (b) and MIROC (c). Probability ranges are represented in a colour scale. Maximum estimates for the Extent of Occurrence (E.O) and Area of Occupancy (O.A) are shown according to IUCN criteria.



Supplementary Figure S3. Predicted suitable ecological areas for *Lasius grandis* and projections to future conditions in the years 2050 and 2080 under three different SRES scenarios. Probability ranges are represented in a colour scale. Sampling localities are represented with red dots in a separate map.


Maximum Extent of Occurrence: **343.17 km²**
Maximum Area of Occupancy: **179 km²**
Estimated Area of Occupancy: **109 km²**
Observed Area of Occupancy: **8 km²**
5 km
🟧 Shrubland / Sparse
🟩 Forest
🟦 Cropland

Supplementary Figure S4. Area of occupancy. Estimation of the area of occupancy (109 Km^2) for *Lasius balearicus* based on ecological niche modelling probabilities and land cover data (extracted from GLOBCOVER version 2.2). Based on our field observations, only areas that displayed an estimated probability > 0.15 (bright cells) and Shrubland / Sparse vegetation were considered as suitable.

Chapter V

Vila, R., Lukhtanov, V., Talavera, G., Gil-T, F., Pierce, N.E. **2010**. How common are dot-like distributions? Taxonomic oversplitting in western European *Agrodiaetus* (Lepidoptera: Lycaenidae) revealed by chromosomal and molecular markers. *Biological Journal of the Linnean Society* **101**:130-154.

THE LINNEAN SOCIETY of London BIOLOGICAL Journal of the Linnean Society

Biological Journal of the Linnean Society, 2010, **101**, 130–154. With 7 figures

How common are dot-like distributions? Taxonomical oversplitting in western European *Agrodiaetus* (Lepidoptera: Lycaenidae) revealed by chromosomal and molecular markers

ROGER VILA[1,2,3]†, VLADIMIR A. LUKHTANOV[4,5]*†, GERARD TALAVERA[1,3], FELIPE GIL-T.[6] and NAOMI E. PIERCE[7]

[1]*Institute of Evolutionary Biology (UPF-CSIC), Passeig Marítim de la Barceloneta 37-49, 08003 Barcelona, Spain*
[2]*Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Spain*
[3]*Departament de Genètica i Microbiologia, Universitat Autònoma de Barcelona, Edifici C, 08193 Bellaterra, Spain*
[4]*Department of Karyosystematics, Zoological Institute of Russian Academy of Science, Universitetskaya nab. 1, 199034 St. Petersburg, Russia*
[5]*Department of Entomology, St. Petersburg State University, Universitetskaya nab. 7/9, 199034 St. Petersburg, Russia*
[6]*Apartado postal 3042, E-18080 Granada, Spain*
[7]*Department of Organismic and Evolutionary Biology, Harvard University, 26 Oxford Street, Cambridge, Massachusetts 02138, USA*

Received 25 January 2010; revised 27 March 2010; accepted for publication 27 March 2010

Approximately 50 taxa of butterflies in Western Europe have been described as new species or elevated to the level of species during the last 40 years. Many, especially those belonging to the genus *Agrodiaetus*, have unusually localized, 'dot-like' distributional ranges. In the present study, we use a combination of chromosomal and molecular markers to re-evaluate the species status of *Agrodiaetus* distributed west of the 17th meridian. The results obtained do not support the current designations of *Agrodiaetus galloi*, *Agrodiaetus exuberans*, and *Agrodiaetus agenjoi* as endemic species with highly restricted distribution ranges, but indicate that these taxa are more likely to be local populations of a widely distributed species, *Agrodiaetus ripartii*. *Agrodiaetus violetae* is shown to be a polytypic species consisting of at least two subspecies, including *Agrodiaetus violetae subbaeticus* **comb. nov.** and *Agrodiaetus violetae violetae*. *Agrodiaetus violetae* is genetically (but not chromosomally) distinct from *Agrodiaetus fabressei* and has a wider distribution in southern Spain than previously believed. *Agrodiaetus humedasae* from northern Italy is supported as a highly localized species that is distinct from its nearest relatives. We propose a revision of the species lists for *Agrodiaetus* taking these new data into account. The results reported in the present study are relevant to animal conservation efforts in Europe because of their implications for IUCN Red List priorities. © 2010 The Linnean Society of London, *Biological Journal of the Linnean Society*, 2010, **101**, 130–154.

ADDITIONAL KEYWORDS: systematics – species – conservation – distribution – range – reinforcement – phylogeny.

*Corresponding author. E-mail: lukhtanov@mail.ru
†These authors contributed equally to the research.

INTRODUCTION

Comparison of the first comprehensive work on European butterflies (Higgins & Riley, 1970) with more recent publications (de Prins & Iversen, 1996; Tolman, 1997; Kudrna, 2002; Lafranchis, 2004; Dennis *et al.*, 2008) shows that approximately 50 butterfly taxa have been described as new species or elevated to species rank during the last 40 years. Many of these newly-recognized species have extremely local 'dot-like' distributions that are restricted to particular mountain valleys in Spain, Italy, the Balkan Peninsula and Crimea, or to small Mediterranean islands (Kudrna, 2002). Usually, these dot-like distributed taxa are geographically isolated populations whose morphological and ecological differences from their closest relatives have rarely been assessed. In theory, species with such restricted ranges may represent either relicts of species that had much broader distributions in the past, or young species that originated recently and have not yet expanded their ranges. However, before considering these possibilities, a more thorough consideration must be made of whether these nominal taxa are indeed valid species rather than isolated populations of other known species with broader distributions.

Species in the butterfly genera and subgenera *Agrodiaetus*, *Hipparchia*, *Plebejus*, *Lysandra*, and *Polyommatus* make up a large proportion of those with dot-like distributions. These groups are among the most species-rich genera of European butterflies, and a number include taxa in the process of speciation. The genus *Agrodiaetus* (considered by some to be a subgenus of the large genus *Polyommatus*) is especially interesting in this respect. *Agrodiaetus* comprises a taxonomically diverse group of blue butterflies (Forster, 1956–1961; Eckweiler & Häuser, 1997; Wiemers, 2003; Kandul *et al.*, 2004; Wiemers, Keller & Wolf, 2009). The monophyly of the genus is strongly supported by molecular data (Kandul *et al.*, 2002, 2004; Wiemers, 2003; Wiemers *et al.*, 2009). Adults of *Agrodiaetus* have a wingspan of only 2–4 cm, and the sexes are often dimorphic, with females typically brown and males blue on the upper surface of their wings. This blue coloration is plesiomorphic, and is found in many species in closely-related genera of the *Polyommatus* section (Kandul *et al.*, 2004). Phylogenetic evidence suggests that reinforcement of pre-zygotic reproductive isolation is likely to have given rise to different male wing coloration in this group: males can have brown, white, silver, violet, and even orange wings, and quite a few of those with light wing coloration also reflect ultraviolet light (Lukhtanov *et al.*, 2005). Given that the number of species of *Agrodiaetus* (at least 120) is much greater than the variety of colours displayed by males, and other diagnostic morphological characters are scarce, the genus may also include cryptic species.

The most remarkable characteristic of the genus *Agrodiaetus* is its unusual diversity of chromosomal complements, or karyotypes. Species of *Agrodiaetus* exhibit among the highest range in chromosome number in the animal kingdom. The karyotype is generally stable within species, although differences between closely-related species are often high. Haploid chromosome numbers in *Agrodiaetus* range from $n = 10$ in *Agrodiaetus caeruleus* to $n = 134$ in *Agrodiaetus shahrami* (Lukhtanov & Dantchenko, 2002a; Lukhtanov *et al.*, 2005).

Modern lists of European *Agrodiaetus* include 13–22 species, depending on the taxonomic interpretation of species or subspecies status for a number of taxa (De Prins & Iversen, 1996; Dennis, 1997; Kudrna, 2002; Dennis *et al.*, 2008). Some of these taxa have quite broad distributions. However, eleven species of European *Agrodiaetus* (i.e. approximately one-half the current species list) have been described to have dot-like distributions and to be restricted to particular mountains or valleys in Spain, Italy, the Balkan Peninsula, and Crimea. These are: (1) *Agrodiaetus violetae* (southern Spain: Sierra de la Almijara); (2) *Agrodiaetus fulgens* (north-eastern Spain: Catalonia); (3) *Agrodiaetus agenjoi* (north-eastern Spain: Catalonia); (4) *Agrodiaetus exuberans* (north-western Italy: Susa); (5) *Agrodiaetus humedasae* (north-western Italy: Cogne Valley); (6) *Agrodiaetus galloi* (southern Italy: Calabria); (7) *Agrodiaetus nephohiptamenos* (southern Bulgaria and northern Greece: Pirin, Orvilos, Pangeon and Phalakron Mountains); (8) *Agrodiaetus eleniae* (northern Greece: Mount Phalakron); (9) *Agrodiaetus orphicus* (southern Bulgaria and northern Greece: Mount Rhodope); (10) *Agrodiaetus budashkini* (Ukraine: Crimea); and (11) *Agrodiaetus pljushtchi* (Ukraine: Crimea). We analyzed three of these nominal species (*A. budashkini*, *A. pljushtchi*, and *A. fulgens*) in previous studies (Kandul *et al.*, 2004; Lukhtanov, Vila & Kandul, 2006; Lukhtanov & Budashkin, 2007). The present study addresses the status of *A. violetae*, *A. agenjoi*, *A. exuberans*, *A. humedasae*, *A. galloi* and related taxa from south-west Europe, and includes a general analysis of the problem of dot-like species ranges in *Agrodiaetus*.

All these target taxa have brown wing coloration in both males and females, and are difficult to distinguish using traditional morphological characters. The first step to characterize such species typically involves molecular methods. However, the use of standard molecular markers such as short fragments of the mitochondrial gene COI and the noncoding nuclear sequence, internal transcribed spacer 2 (ITS2), is sometimes insufficient to distinguish between evolutionarily young sister species, either because they may be weakly differentiated with respect to these markers (Wiemers, 2003; Kandul *et al.*, 2004; Wiemers & Fiedler, 2007; Lukhtanov *et al.*, 2009) or because they are too polymorphic (Lukhtanov & Shapoval, 2008; Lukhtanov, Shapoval & Dantchenko, 2008). An absence of lineage sorting among species can be frequently a problem for the use of molecular markers in rapidly evolving taxa of *Agrodiaetus*: the time to coalescence for alleles within lineages may be greater than the time subsequent to speciation (Kandul *et al.*, 2004).

Chromosomal characters in many groups may evolve more quickly, and because they are often present as fixed differences, can sometimes provide better markers for recently evolved taxa (King, 1993; Dobigny *et al.*, 2005). The study of the karyotype provides good diagnostic characters for most *Agrodiaetus* species and, as such, has become an important requirement for describing and delimiting new taxa (de Lesse, 1960a; Lukhtanov & Dantchenko, 2002b; Lukhtanov *et al.*, 2003, 2006). As with molecular data, cytological data have their own limitations; they may be incapable of resolving groups of species characterized by extreme chromosomal conservatism. However, molecular and chromosomal approaches are complementary, and applying a combination of these approaches can provide powerful taxonomic insights, especially when considered with morphological and ecological data (Lukhtanov *et al.*, 2006; Descimon & Mallet, 2009).

Dot-like distributed species present practical as well as theoretical difficulties. Increasing the number of such species substantially increases the potential conservation load for European butterflies (Dennis, 1997). Endemic species, those with small or restricted ranges, are in greater danger of becoming extinct through systematic or stochastic changes in the environment than are widely distributed species (Gaston, 1994). Thus, even if restricted range is not the only factor taken into account, it is not surprising that several local European *Agrodiaetus* taxa are listed among species of conservation concern (Van Swaay *et al.*, 2010).

SPECIES AND SUBSPECIES CONCEPTS

SPECIES

In the present study, we adopt a classification based on the biological species concept (BSC) (Poulton, 1904; Mayr, 1963; Häuser, 1987). Under the BSC, actual or potentially reproductively isolated entities are classified as species. Isolation may not necessarily be complete, but it should be strong enough to prevent taxa from merging when they occur in sympatry (Mayr, 1963; Coyne & Orr, 2004). In practice, the existence of isolation can be tested most effectively via the genotypic cluster approach (Mallet, 2001, 2006; Mallet & Willmott, 2003), in which data on morphological, genetic, ecological, and behavioural characters in a local area are used as evidence of distinctness in sympatry. Genotypic clusters in sympatry can be seen in phenotypic data as a bimodal distribution of traits, and in genetic data as a deficit of heterozygotes or as the presence of linkage disequilibrium among genes. Species recognition through linkage disequilibrium analysis of unlinked genetic markers has already been used in *Agrodiaetus* (Lukhtanov & Shapoval, 2008).

However, when taxa are allopatric, the direct application of the BSC may be more difficult. We suggest that allopatric taxa be considered species if they are clearly distinct with respect to characters that contribute to pre- or post-zygotic reproductive isolation. In the case of *Agrodiaetus*, a strong difference in the colour of the upper side of the male wing (e.g. blue versus brown) most likely contributes to pre-zygotic isolation (Lukhtanov *et al.*, 2005).

Chromosome differences can also be considered indirect evidence for reproductive isolation between taxa in allopatry. It is well known that chromosome rearrangements can cause sterility (King, 1993), and even relatively small differences in chromosome structure can result in post-zygotic isolation (Ferree & Barbash, 2009). However, this is not always true and, in some cases, heterozygosity for chromosome rearrangements does not result in sterility (Nagaraju & Jolly, 1986). Indeed, there is no well-established general rule to determine how many or what types of chromosome rearrangements can be tolerated before resulting in infertile offspring.

The chromosome number of *Agrodiaetus* is generally stable within populations of this genus and, in only a few cases, a limited amount of variability in intra-population haploid chromosome number has been observed. The range of this variation has never exceeded four chromosomes, which we infer is the likely upper threshold of chromosome number differences compatible with offspring fertility in this group (Lukhtanov & Dantchenko, 2002b; Lukhtanov, Wiemers & Meusemann, 2003). Thus, empirical observations of *Agrodiaetus* suggest that a fixed difference of five or more chromosomes in haploid number sets (which is equal to ten or more chromosomes in diploid number sets) provides a useful criterion to use in designating allopatric chromosome races as nonconspecific until direct evidence for the presence/absence of reproductive isolation can be obtained.

Molecular data alone, even in the case of a relatively high level of genetic differentiation between the taxa under comparison, are not sufficient to define biological species because the divergence of standard genetic markers between distinct sympatric species can be low or absent, and intraspecific variation can be relatively high (Lukhtanov *et al.*, 2009). However, genetic divergence comparisons may be useful in highlighting potentially interesting monophyletic lineages that deserve further study, and in identifying morphologically similar species that are not closely related. For example, in the present study, the brown-coloured *Agrodiaetus fabressei* is not sister to the morphologically similar *A. violetae*, but to two blue-coloured species, *Agrodiaetus dolus* and *A. fulgens*. Both mitochondrial and nuclear markers support this result, and thus we consider *A. fabressei* and *A. violetae* not to be conspecific.

SUBSPECIES

Diagnosable allopatric entities (populations or groups of populations) with fixed difference(s) in morphological and/or chromosomal characters should be classified as subspecies if they do not correspond to the species criteria specified above. In general, we agree with Descimon & Mallet (2009), that 'there is justification for reviving the rather neglected (and misused) rank of subspecies, with the trend among lepidopterists to consider only more strongly distinct forms (in morphology, ecology, or genetics) as subspecies, and to lump dubious geographic forms as synonyms . . . [This provides] . . . a useful compromise between descriptions of geographic variation, the needs of modern butterfly taxonomy, and Darwin's pragmatic use of the term species in evolutionary studies.'

MATERIAL AND METHODS

TAXON SAMPLING

In the present study, we focus only on those taxa found in Europe west of the 17th meridian. In this region, almost all *Agrodiaetus* taxa and populations are concentrated on the Iberian Peninsula in France and Italy. Except for *Agrodiaetus damon*, they belong to two groups of species: the *Agrodiaetus admetus* group, and the *A. dolus* group, which are sister clades in all published phylogenetic reconstructions (Wiemers, 2003; Kandul *et al.*, 2004, 2007; Lukhtanov *et al.*, 2005). These two groups also include taxa from the Balkan Peninsula, eastern Europe, and western Asia that are not considered in detail in the present study. However, to estimate relationships among western European taxa, we include in our analysis all eastern European and non-European species except the Anatolian–Iranian species, *Agrodiaetus demavendi*, where specimens with unambiguous species determination and precise chromosome number count were not available (Tables 1, 2).

When collecting in the field, we used a protocol that allowed us to obtain molecular and chromosomal information from the same individual specimens (Bulatova *et al.*, 2009). Additionally, we tried to obtain samples from the type localities of each studied taxa in order to connect the chromosomal and molecular data with correct species names. In particular, *A. violetae*, *Agrodiaetus fabressei subbaeticus*, *A. exuberans*, *Agrodiaetus ripartii susae*, *A. humedasae* and *A. galloi* were collected from their type-localities. Specimens RV-03-H463 and RVcoll. 07-F038 of *A. agenjoi* were collected approximately 6.5 km and 125 km, respectively, from the taxon type locality, 'Barcelona, Taradell' [Barcelona province, Catalonia, north-east Spain] (Forster, 1965). Specimens RE-07-G266 and RE-07-G273 of *Agrodiaetus ripartii rippertii* were collected approximately 100 km north-west from the taxon type locality, 'aux environs de Digne' [Alpes de Haute Provence, France]) (Boisduval, 1832).

We also inspected the morphology and taxon identification of samples whose sequences we downloaded from GenBank. In doing so, we found that samples MW01105 and MAT-99-Q878 from Catalonia, previously identified as *A. ripartii* (Wiemers, 2003; Kandul *et al.*, 2004), have no white streak on the underside of the hind wing. Although this character can be labile, if we take it into account in conjunction with the collecting locality, we consider that these specimens actually belong to the nominal species, *A. agenjoi*.

KARYOTYPING

Only fresh adult males were used for karyotyping. Adults were collected in the field, and after they were killed by a sharp pinch to the thorax, testes were immediately excised and placed into 0.5-mL vials with freshly prepared Carnoy fixative (ethanol and glacial acetic acid, 3 : 1). Bodies were preserved in 2-mL plastic vials with 100% ethanol for DNA analysis, and wings were stored in glassine envelopes.

Gonads were stored in fixative for 2–6 months at 4 °C and then stained with 2% acetic orcein for 30 days at 20 °C. Cytogenetic analysis was conducted using a two-phase method of chromosome analysis (Lukhtanov & Dantchenko, 2002a; Lukhtanov *et al.*, 2006). Chromosome preparations are stored in the Department of Entomology of St Petersburg State University, Russia. Butterfly bodies in ethanol, and wings in glassine envelopes are stored in the Lepidoptera DNA and Tissues Collection at the Museum

Table 1. List of the *Agrodiaetus* samples used in the present study

(Traditionally) accepted name and combination	Proposed name and combination	Sample code	Locality
Agrodiaetus admetus	*Agrodiaetus admetus*	AD-00-P016	Armenia, Aiodzor Mts, Gnishyk
Agrodiaetus admetus	*Agrodiaetus admetus*	JC 01014	Greece, Peloponnisos, Mt Taiyetos, 1200–1300 m
Agrodiaetus admetus	*Agrodiaetus admetus*	MW98084	Turkey, Antalya, Cukurelma N Elmali 1300 m
Agrodiaetus admetus anatoliensis	*Agrodiaetus admetus anatoliensis*	VL-01-L101	Turkey, Gümüshane, Torul
Agrodiaetus admetus malievi	*Agrodiaetus admetus malievi*	VL-03-F903	Azerbaijan, Talysh, Zuvand
Agrodiaetus agenjoi	*Agrodiaetus ripartii ripartii*	MAT-99-Q878	Spain, Lleida, Tremp, Rúbies
Agrodiaetus agenjoi	*Agrodiaetus ripartii ripartii*	MW01105	Spain, Tarragona, Santa Coloma de Queralt, 700 m
Agrodiaetus agenjoi	*Agrodiaetus ripartii ripartii*	RV-03-H463	Spain, Barcelona, El Brull, 830 m
Agrodiaetus agenjoi	*Agrodiaetus ripartii ripartii*	RVcoll.07-F038	Spain, Tarragona, Serra de Prades, Barranc de Vinarroig, 920 m
Agrodiaetus ainsae	*Agrodiaetus fulgens ainsae*	MAT-99-Q894	Spain, Lleida, Tremp, Rúbies
Agrodiaetus ainsae	*Agrodiaetus fulgens ainsae*	MW01001	Spain, Álava, Ilarduya, W Eguino, 550 m
Agrodiaetus ainsae	*Agrodiaetus fulgens ainsae*	MW01053	Spain, Huesca, Embalse de la Peña, Sta. María, 500 m
Agrodiaetus ainsae	*Agrodiaetus fulgens ainsae*	MW01078	Spain, Huesca, Embalse de la Peña, Triste, 600 m
Agrodiaetus alcestis	*Agrodiaetus alcestis*	MW98212	Turkey, Adana, Saimbeyli, 1500 m
Agrodiaetus alcestis	*Agrodiaetus alcestis*	MW98315	Turkey, Karaman, Ermenek, Yellibeli Geçidi, 1800 m
Agrodiaetus alcestis karacetinae	*Agrodiaetus alcestis karacetinae*	MW00229	Iran, Zanjan, Qazayd Dagh, 25 km O. Zanjan, 2300 m
Agrodiaetus alcestis karacetinae	*Agrodiaetus alcestis karacetinae*	MW00231	Iran, Zanjan, Qazayd Dagh, 25 km O. Zanjan, 2300 m
Agrodiaetus alcestis karacetinae	*Agrodiaetus alcestis karacetinae*	MW99380	Turkey, Hakkari, 22 km NW Yüksekova, 1800 m
Agrodiaetus alcestis karacetinae	*Agrodiaetus alcestis karacetinae*	VL-03-F669	Iran, Markazi, Khiru
Agrodiaetus aroaniensis	*Agrodiaetus aroaniensis*	JC00040	Greece, Peloponnisos, Mt Helmos, 1350 m
Agrodiaetus damocles krymaeus	*Agrodiaetus damocles krymaeus*	NK-00-P103	Ukraine, Crimea, Kurortnoe
Agrodiaetus damon	*Agrodiaetus damon*	MAT-99-Q841	Spain, Girona, Pyrenees Mts, Urús
Agrodiaetus dantchenkoi	*Agrodiaetus dantchenkoi*	MW99274	Turkey, Van, Gürpinar, Kurubas Geçidi, 2200 m
Agrodiaetus dantchenkoi	*Agrodiaetus dantchenkoi*	MW99276	Turkey, Van, Gürpinar, Kurubas Geçidi, 2200 m
Agrodiaetus dantchenkoi	*Agrodiaetus dantchenkoi*	MW99319	Turkey, Van, 25–32 km N Çatak, 2000–2200 m
Agrodiaetus dantchenkoi	*Agrodiaetus dantchenkoi*	MW99320	Turkey, Van, 25–32 km N Çatak, 2000–2200 m
Agrodiaetus dantchenkoi	*Agrodiaetus dantchenkoi*	VL-01-L342	Turkey, Van, Çatak
Agrodiaetus dolus virgilia	*Agrodiaetus dolus virgilia*	RE-07-G106	Italy, Rocca Pia, 1215 m
Agrodiaetus dolus vittatus	*Agrodiaetus dolus vittatus*	MAT-99-Q923	France, Languedoc Reg, Mende
Agrodiaetus eriwanensis	*Agrodiaetus eriwanensis*	AD-00-P303	Armenia, Aiodzor Mts, Gnishyk
Agrodiaetus erschoffii	*Agrodiaetus erschoffii*	AD-02-L274	Iran, Gorgan, Shahkuh
Agrodiaetus exuberans	*Agrodiaetus ripartii ripartii*	RE-07-G229	Italy, Susa Valley, Urbiano, Mompantero, 720 m
Agrodiaetus fabressei fabressei	*Agrodiaetus fabressei fabressei*	JM00001	Spain, Cuenca, Tragacete, Mogorrita
Agrodiaetus fabressei fabressei	*Agrodiaetus fabressei fabressei*	MAT-99-Q972	Spain, Cuenca, Una, 970 m
Agrodiaetus fabressei fabressei	*Agrodiaetus fabressei fabressei*	MAT-99-Q984	Spain, Albarracín, Puerto de la Losilla
Agrodiaetus fabressei fabressei	*Agrodiaetus fabressei fabressei*	MW01039	Spain, Soria, Sierra de Cabrejas, Abejar, 1100 m
Agrodiaetus fabressei fabressei	*Agrodiaetus fabressei fabressei*	RV-03-H596	Spain, Castelló, Coll d'Ares, 1148 m
Agrodiaetus fabressei subbaeticus	*Agrodiaetus violetae subbaeticus*	RV-03-H554	Spain, Granada, Sierra de la Sagra, 1775 m
Agrodiaetus fabressei subbaeticus	*Agrodiaetus violetae subbaeticus*	RV-03-H555	Spain, Granada, Sierra de la Sagra, 1775 m
Agrodiaetus fabressei subbaeticus	*Agrodiaetus violetae subbaeticus*	RV-03-H556	Spain, Granada, Sierra de la Sagra, 1702 m
Agrodiaetus fabressei subbaeticus	*Agrodiaetus violetae subbaeticus*	RV-03-H557	Spain, Granada, Sierra de la Sagra, 1702 m
Agrodiaetus fabressei subbaeticus	*Agrodiaetus violetae subbaeticus*	RV-03-H558	Spain, Granada, Sierra de la Sagra, 1702 m
Agrodiaetus fabressei subbaeticus	*Agrodiaetus violetae subbaeticus*	RV-03-H560	Spain, Granada, Sierra de la Sagra, 1702 m
Agrodiaetus fulgens	*Agrodiaetus fulgens fulgens*	MAT-99-Q910	Spain, Tarragona, Santa Coloma de Queralt
Agrodiaetus fulgens	*Agrodiaetus fulgens fulgens*	MW01107	Spain, Tarragona, Santa Coloma de Queralt, 700 m
Agrodiaetus galloi	*Agrodiaetus ripartii ripartii*	RE-07-G436	Italy, Calabria, Serra del Prete, Mont Pollino, 1650 m
Agrodiaetus galloi	*Agrodiaetus ripartii ripartii*	RE-07-G437	Italy, Calabria, Serra del Prete, Mont Pollino, 1650 m
Agrodiaetus galloi	*Agrodiaetus ripartii ripartii*	RE-07-G441	Italy, Calabria, Serra del Prete, Mont Pollino, 1650 m
Agrodiaetus galloi	*Agrodiaetus ripartii ripartii*	RE-07-G445	Italy, Calabria, Serra del Prete, Mont Pollino, 1650 m
Agrodiaetus galloi	*Agrodiaetus ripartii ripartii*	RE-07-G447	Italy, Calabria, Serra del Prete, Mont Pollino, 1650 m
Agrodiaetus humedasae	*Agrodiaetus humedasae*	MW99591	Italy, Aosta, Val di Cogne, Pondel, 900 m
Agrodiaetus humedasae	*Agrodiaetus humedasae*	MW99605	Italy, Aosta, Val di Cogne, Pondel, 900 m
Agrodiaetus humedasae	*Agrodiaetus humedasae*	RE-07-G191	Italy, Aosta, Val di Cogne, Ozien-Visyes, 1000 m
Agrodiaetus humedasae	*Agrodiaetus humedasae*	RE-07-G192	Italy, Cogne Valley, Ozien-Visyes, 1000 m
Agrodiaetus humedasae	*Agrodiaetus humedasae*	RE-07-G193	Italy, Cogne Valley, Ozien-Visyes, 1000 m
Agrodiaetus humedasae	*Agrodiaetus humedasae*	RE-07-G194	Italy, Cogne Valley, Ozien-Visyes, 1000 m
Agrodiaetus humedasae	*Agrodiaetus humedasae*	RE-07-G203	Italy, Aosta, Val di Cogne, Ozien-Visyes, 1000 m
Agrodiaetus interjectus	*Agrodiaetus interjectus*	MW99164	Turkey, Erzurum, 5 km NE. Çiftlik, 1900 m
Agrodiaetus khorasanensis	*Agrodiaetus khorasanensis*	VL-03-F526	Iran, Khorasan, Kopetdagh Mts
Agrodiaetus khorasanensis	*Agrodiaetus khorasanensis*	WE02431	Iran, Khorasan, 5 km SW Firizi, 1700–1900 m
Agrodiaetus menalcas	*Agrodiaetus menalcas*	MW98020	Turkey, Fethiye, Gülübeli Geçidi, W. Elmali, 1500 m
Agrodiaetus menalcas	*Agrodiaetus menalcas*	MW98172	Turkey, Sivas, Gökpinar, Gürün, 1700 m
Agrodiaetus menalcas	*Agrodiaetus menalcas*	MW99494	Turkey, Van, Erek Dagi, 2200 m
Agrodiaetus menalcas	*Agrodiaetus menalcas*	VL-01-L122	Turkey, Dilekyolu, Gümüshane
Agrodiaetus ripartii	*Agrodiaetus ripartii ripartii*	AD-00-P033	Russia, Tula Reg, Tatinki

Table 1. *Continued*

(Traditionally) accepted name and combination	Proposed name and combination	Sample code	Locality
Agrodiaetus ripartii pelopi	*Agrodiaetus ripartii ripartii*	JC00043	Greece, Peloponnisos, Mt Helmos, 1350–1500 m
Agrodiaetus ripartii budashkini	*Agrodiaetus ripartii ripartii*	NK-00-P859	Ukraine, Crimea, Karabi yaila
Agrodiaetus ripartii colemani	*Agrodiaetus ripartii colemani*	NK-00-P822	Kazakhstan, West Tian-Shan
Agrodiaetus ripartii paralcestis	*Agrodiaetus ripartii paralcestis*	MW99068	Turkey, Artvin, Kiliçkaya, Yusufeli, 1350 m
Agrodiaetus ripartii paralcestis	*Agrodiaetus ripartii paralcestis*	MW99196	Turkey, Erzincan, 5 km SE Çaglayan, 1500 m
Agrodiaetus ripartii paralcestis	*Agrodiaetus ripartii paralcestis*	MW99263	Turkey, Van, Kurubas Geçidi, Gürpinar, 2200 m
Agrodiaetus ripartii paralcestis	*Agrodiaetus ripartii paralcestis*	MW99264	Turkey, Van, Kurubas Geçidi, Gürpinar, 2200 m
Agrodiaetus ripartii paralcestis	*Agrodiaetus ripartii paralcestis*	AD-00-P337	Armenia, Pambak Mts, Dzhur-dzhur Pass
Agrodiaetus ripartii paralcestis	*Agrodiaetus ripartii paralcestis*	VL-01-L103	Turkey, Gümüshane
Agrodiaetus ripartii paralcestis	*Agrodiaetus ripartii paralcestis*	VL-01-L166	Turkey, Gümüshane, Dilekyolu
Agrodiaetus ripartii ripartii	*Agrodiaetus ripartii ripartii*	MW01014	Spain, Burgos, Ubierna, 20 km N Burgos, 900 m
Agrodiaetus ripartii ripartii	*Agrodiaetus ripartii ripartii*	MW01072	Spain, Huesca, Triste, Embalse de la Pena, 600 m
Agrodiaetus ripartii rippertii	*Agrodiaetus ripartii ripartii*	RE-07-G266	France, Drôme, Col de la Chaudière, 1025 m
Agrodiaetus ripartii rippertii	*Agrodiaetus ripartii ripartii*	RE-07-G273	France, Drôme, Col de la Chaudière, 1025 m
Agrodiaetus ripartii sarkani	*Agrodiaetus ripartii ripartii*	NK-00-P829	Kazakhstan, Dzhungarian, Alatau Mts, Kolbai
Agrodiaetus ripartii sarkani	*Agrodiaetus ripartii ripartii*	NK-00-P848	Kazahkstan, Tarbagatai Mts, Taskeskan
Agrodiaetus ripartii susae	*Agrodiaetus ripartii ripartii*	RE-07-G254	Italy, Torino, Novalesa-Moncenisio, 1155 m
Agrodiaetus ripartii susae	*Agrodiaetus ripartii ripartii*	RE-07-G255	Italy, Torino, Novalesa-Moncenisio, 1155 m
Agrodiaetus rjabovi	*Agrodiaetus rjabovi*	VL-02-X474	Iran, Gilan, Masuleh
Agrodiaetus rjabovi	*Agrodiaetus rjabovi*	VL-03-F816	Azerbaijan, Talysh, Zuvand
Agrodiaetus surakovi	*Agrodiaetus surakovi*	AD-00-P006	Armenia, Aiodzor Mts, Gnishyk
Agrodiaetus urmiaensis	*Agrodiaetus urmiaensis*	VL-04-E365	Iran, Azarbayjan-e-Gharbi
Agrodiaetus valiabadi	*Agrodiaetus valiabadi*	MW00064	Iran, Mazandaran, Pul-e Zanguleh, 15 km NE Kendevan, 2400 m
Agrodiaetus valiabadi	*Agrodiaetus valiabadi*	MW00498	Iran, Mazandaran, 5 km S. Valiabad, 1900 m
Agrodiaetus violetae	*Agrodiaetus violetae violetae*	FGT-05-J629	Spain, Granada, Sierra de la Almijara
Agrodiaetus violetae	*Agrodiaetus violetae violetae*	FGT-05-J630	Spain, Granada, Sierra de la Almijara
Agrodiaetus violetae	*Agrodiaetus violetae violetae*	RVcoll.08-H299	Spain, Andalucía

of Comparative Zoology, Harvard University, and R. Vila's DNA and Tissues Collection at the Universitat Autònoma de Barcelona.

DNA extraction and sequencing

Total genomic DNA was extracted using the DNeasyTM Tissue Kit (Qiagen Inc.) in accordance with the manufacturer's instructions. Published primers were used to amplify mitochondrial cytochrome oxidase subunit *I* (COI), leucine transfer RNA (leu-tRNA), cytochrome oxidase subunit *II* (COII) (Folmer *et al.*, 1994; Simon *et al.*, 1994; Monteiro & Pierce, 2001), and nuclear ITS2 (White *et al.*, 1990). The polymerase chain reaction (PCR) was carried out in 25-mL reactions using a DNA Engine thermal cycler (MJ Research Inc.), and typically contained 0.5 mM of each primer, 0.8 mM dNTPs, 1 ¥ Qiagen PCR buffer with additional $MgCl_2$ to a final concentration of 2 mM and 1.25 units Qiagen Taq DNA polymerase. All reactions were initially denatured at 94 °C for 2 min, and then subjected to 35 cycles of 60 s at 94 °C denaturation, 60 s at 45 °C–56 °C (annealing temperature depended on gene amplified), and 90 s at 72 °C extension. After amplification, double-stranded DNA was purified using QIAquick PCR purification kits (Qiagen).

Primers used for amplification served as sequencing primers. All samples were sequenced in both directions. Cycle sequencing reactions were performed in 12-mL reactions: 1.5 mL of ABI Prism BigDye, version 3.1 (Applied Biosystems Inc.), 1.0 mL of 5 ¥ buffer (buffer: 400 mM Tris at pH 9.0 and 10 mM $MgCl_2$), and 0.33 mL each (10 mM) of primer. The remainder of the mixture was composed of ultra pure water 50–90 ng of template DNA in each reaction. Cycle sequence reaction started with a denaturing step of 94 °C for 2 min, followed by 25 cycles of 10 s at 94 °C, 5 s at annealing temperature, which varied for different gene regions, and 4 min at 60 °C. Sequencing was conducted in a 3100 Genetic Analyzer (Applied Biosystems/Hitachi). Sequences obtained specifically for this study were deposited in GenBank under accession numbers HM210162 to HM210202.

Phylogenetic analysis

For phylogenetic analysis, we used sequences of COI, leu-tRNA, COII and ITS2 original to the present study, as well as sequences obtained from GenBank that had been included in Kandul *et al.* (2004) and Wiemers & Fiedler (2007) (Table 1). We re-edited some of the sequences from previous studies, and a

Table 2. Data used for karyotype and molecular phylogenetic analyses. GenBank codes for sequences obtained specifically for this study, re-edited or with a new fragment sequenced, are highlighted in bold

Taxon (Traditionally) accepted name and combination)	Sample code	Karyotype analysis	Molecular 45-taxa dataset	Molecular 80-taxa dataset	COI genbank code	COII genbank code	ITS2 genebank code
A. admetus	AD-00-P016		X	X	**AY496711** (re-edited)	**AY496711** (re-edited)	
A. admetus	JC 01014	[a]n = 80		X	AY556867		AY556733
A. admetus	MW98084			X	AY556986		
A. admetus anatoliensis	VL-01-L101	[b]n = *ca*80	X	X	AY496710	AY496710	
A. admetus malievi	VL-03-F903	[c]n = 79	X	X	EF104617	EF104617	**HM210176**
A. agenjoi	MAT-99-Q878		X	X	AY496780	AY496780	
A. agenjoi	MW01105			X	AY556962		
A. agenjoi	RV-03-H463		X	X	**EF104603** (re-edited)	**EF104603** (re-edited)	
A. agenjoi	RV-07-F038	[d]n = 90					
A. ainsae	MAT-99-Q894	[k]n = 108-110	X	X	**AY496712** (new part seq)	AY496712	**HM210177**
A. ainsae	MW01001	[k]n = 108-110		X	AY556941		AY556601
A. ainsae	MW01053	[k]n = 108-110		X	AY556954		AY556610
A. ainsae	MW01078	[k]n = 108-110		X	AY556958		
A. alcestis	MW98315	[e]n = 20		X	AY557024		AY556653
A. alcestis	MW98212	[e]n = 21		X	AY557008		AY556641
A. alcestis karacetinae	MW00229	[e]n = *ca*19		X	AY556906		
A. alcestis karacetinae	MW00231	[e]n = *ca*19		X	AY556907		AY556574
A. alcestis karacetinae	MW99380	[e]n = 19		X	AY557090		
A. alcestis karacetinae	VL-03-F669	[b]n = 19	X	X	AY954018	AY954018	
A. aroaniensis	JC00040	[f]n = 48		X	AY556856		AY556725
A. damocles krymaeus	NK-00-P103	[g]n = 26	X	X	**AY496727** (re-edited)	**AY496727** (re-edited)	**HM210178**
A. damon	MAT-99-Q841	[h]n = 45	X	X	**AY496732** (new part seq)	AY496732	**HM210179**
A. dantchenkoi	MW99274	[i]n = 42		X	AY557072		AY556678
A. dantchenkoi	MW99276	[e]n = *ca*40-43		X	AY557073		AY556679
A. dantchenkoi	MW99319	[i]n = 42		X	AY557081		AY556685
A. dantchenkoi	MW99320	[e]n = *ca*40-41		X	AY557082		
A. dantchenkoi	VL-01-L342	[i]n = 42	X	X	**AY496737** (re-edited)	**AY496737** (re-edited)	
A. dolus virgilia	RE-07-G106	[k]n = 122	X	X	**HM210162**	**HM210162**	**HM210180**
A. dolus vittatus	MAT-99-Q923	[k]n = 124-125	X	X	**AY496740** (new part seq)	**AY496740** (re-edited)	**HM210181**
A. eriwanensis	AD-00-P303	[j]n = 32	X	X	**AY496742** (re-edited)	**AY496742** (re-edited)	
A. erschoffii	AD-02-L274	[b]n = 13	X	X	**AY496743** (new part seq)	AY496743	**HM210182**
A. exuberans	RE-07-G229	[d]2n = *ca*180	X	X	**HM210172**	**HM210172**	**HM210183**
A. fabressei fabressei	JM00001	[a]n = 90		X	AY556869		AY556734
A. fabressei fabressei	MAT-99-Q972	[a]n = 90	X	X	**HM210165**	**HM210165**	**HM210184**
A. fabressei fabressei	MAT-99-Q984	[a]n = 90	X	X	**AY496744** (new part seq)	**AY496744** (re-edited)	**HM210185**
A. fabressei fabressei	MW01039	[a]n = 90		X	AY556952		AY556608
A. fabressei fabressei	RV-03-H596	[a]n = 90	X	X	**EF104605** (re-edited)	**EF104605** (re-edited)	**HM210186**
A. fabressei subbaeticus	RV-03-H554	[d]n = 90					
A. fabressei subbaeticus	RV-03-H555	[d]n = 90	X	X	**HM210166**	**HM210166**	**HM210187**
A. fabressei subbaeticus	RV-03-H556	[d]n = 90					
A. fabressei subbaeticus	RV-03-H557	[d]n = 90					
A. fabressei subbaeticus	RV-03-H558	[d]n = 90	X	X	**EF104604** (re-edited)	**EF104604** (re-edited)	**HM210188**
A. fabressei subbaeticus	RV-03-H560	[d]n = 90					
A. fulgens	MAT-99-Q910	[k]n = 109	X	X	**AY496746** (new part seq)	**AY496746** (re-edited)	**HM210189**
A. fulgens	MW01107	[k]n = 109		X	AY556963		AY556615
A. galloi	RE-07-G436	[d]n = 90	X	X	**HM210167**	**HM210167**	**HM210190**
A. galloi	RE-07-G437	[d]n = 90	X	X	**HM210168**	**HM210168**	**HM210191**
A. galloi	RE-07-G441	[d]n = 90					
A. galloi	RE-07-G445	[d]n = 90					
A. galloi	RE-07-G447	[d]n = 90					
A. humedasae	MW99591			X	AY557127		AY556710 .

Table 2. *Continued*

Taxon (Traditionally) accepted name and combination)	Sample code	Karyotype analysis	Molecular 45-taxa dataset	Molecular 80-taxa dataset	COI genbank code	COII genbank code	ITS2 genebank code
A. humedasae	MW99605			X	AY557128		AY556711
A. humedasae	RE-07-G191	[d]n = 39	X	X	**HM210169**	**HM210169**	**HM210192**
A. humedasae	RE-07-G192	[d]n = 39					
A. humedasae	RE-07-G193	[d]n = 39					
A. humedasae	RE-07-G194	[d]n = 39					
A. humedasae	RE-07-G203		X	X	**HM210170**	**HM210170**	**HM210193**
A. interjectus	MW99164	[e]n = 31		X	AY557059		AY556671
A. khorasanensis	VL-03-F526	[b]n = 84	X	X	AY954013	AY954013	
A. khorasanensis	WE02431			X	AY557138		AY556737
A. menalcas	MW98020			X	AY556982		
A. menalcas	MW98172			X	AY557001		AY556635
A. menalcas	MW99494			X	AY557111		
A. menalcas	VL-01-L122	[b]n = 85	X	X	AY496763	AY496763	**HM210194**
A. ripartii	AD-00-P033		X	X	**AY496787** (re-edited)	**AY496787** (re-edited)	
A. ripartii	JC00043			X	AY556858		AY556727
A. ripartii budashkini	NK-00-P859	[l]n = 90	X	X	**AY496779** (re-edited)	**AY496779** (re-edited)	**HM210195**
A. ripartii colemani	NK-00-P822	[m]n = 90	X	X	**AY496781** (re-edited)	**AY496781** (re-edited)	
A. ripartii paralcestis	MW99068	[e]n = *ca*90		X	AY557042		
A. ripartii paralcestis	MW99196			X	AY557064		AY556673
A. ripartii paralcestis	MW99263			X	AY557070		
A. ripartii paralcestis	MW99264			X	AY557071		
A. ripartii paralcestis	AD-00-P337		X	X	**AY496782** (re-edited)	**AY496782** (re-edited)	
A. ripartii paralcestis	VL-01-L103	[b]n = *ca*90	X	X	AY496783	AY496783	
A. ripartii paralcestis	VL-01-L166	[c]n = 90	X	X	AY496784	AY496784	
A. ripartii ripartii	MW01014	[e]n = *ca*90		X	AY556944		AY556603
A. ripartii ripartii	MW01072			X	AY556957		
A. ripartii rippertii	RE-07-G266	[d]n = 90	X	X	**HM210171**	**HM210171**	**HM210196**
A. ripartii rippertii	RE-07-G273	[d]n = 90					
A. ripartii sarkani	NK-00-P829	[m]n = 90	X	X	AY496785	AY496785	
A. ripartii sarkani	NK-00-P848	[m]n = 90	X	X	AY496786	AY496786	
A. ripartii susae	RE-07-G254		X	X	**HM210163**	**HM210163**	**HM210197**
A. ripartii susae	RE-07-G255		X	X	**HM210164**	**HM210164**	**HM210198**
A. rjabovi	VL-02-X474	[b]n = 43	X	X	AY954006	AY954006	
A. rjabovi	VL-03-F816	[b]n = 49	X	X	AY954019	AY954019	
A. surakovi	AD-00-P006	[j]n = 50	X	X	**AY496792** (re-edited)	**AY496792** (re-edited)	**HM210199**
A. urmiaensis	VL-04-E365	[c]n = 19	x	x	**EF104631** (re-edited)	EF104631	
A. valiabadi	MW00064			x	AY556882		AY556557
A. valiabadi	MW00498	[e]n = 23		x	AY556934		AY556594
A. violetae	FGT-05-J629		x	x	**HM210173**	**HM210173**	**HM210200**
A. violetae	FGT-05-J630	[d]n = *ca*90	x	x	**HM210174**	**HM210174**	**HM210201**
A. violetae	RVcoll.08.H299		x	x	**HM210175**	**HM210175**	**HM210202**

[a] The karyotype information for the population studied (but not for this individual) was taken from de Lesse (1960a).
[b] The karyotype of this sample was studied in Lukhtanov *et al.* (2005).
[c] The karyotype of this sample was studied by Lukhtanov (unpublished).
[d] The karyotype of this sample was studied in the present work.
[e] The karyotype of this sample was studied in Wiemers (2003).
[f] The karyotype information for the population studied (but not for the same individual) was taken from Coutsis *et al.* (1999).
[g] The karyotype information for the population studied (but not for the same individual) was taken from Kandul and Lukhtanov (1997).
[h] The karyotype information for the population studied (but not for the same individual) was taken from de Lesse (1960b).
[i] The karyotype of this sample was studied in Lukhtanov *et al.* (2003).
[j] The karyotype information for the population studied (but not for the same individual) was taken from Lukhtanov and Dantchenko (2002b).
[k] The karyotype information for the population studied (but not for the same individual) was taken from Lukhtanov *et al.* (2006).
[l] The karyotype information for the population studied (but not for the same individual) was taken from Kandul *et al.* (2004).
[m] The karyotype of this sample was studied in Lukhtanov and Dantchenko (2002a).

few changes to these were introduced. In two cases, an additional terminal fragment was sequenced using the same specimen. Revised sequences have been updated in GenBank. The final dataset includes 80 specimens representing 37 taxa, including four outgroups. We also analyzed a subset of these taxa: the 45-specimen dataset includes only those samples with little or no missing data.

Sequences were unambiguously aligned using SEQUENCHER, version 3.1 (Genecodes Corporation). For each dataset and gene, regions where more than 50% of the sequences contained missing data were removed using the software GBLOCKS, version 0.91 (Castresana, 2000). The incongruence length difference (ILD) test (Farris *et al.*, 1994) was performed to study the homogeneity between our mitochondrial and nuclear datasets. The test was performed with PAUP* using heuristic searches with tree bisection–reconnection (TBR) branch swapping and 100 random taxon addition replicates, saving no more than ten equally parsimonious trees per replicate. Only parsimony informative sites were included. No significant conflict ($P = 0.98$) was detected by the ILD test between the mitochondrial (COI + tRNALeu + COII) and nuclear (ITS2) data. Thus, we combined mitochondrial and nuclear sequences to improve phylogenetic signal. This resulted in concatenated alignments with a total of 2812 bp for the 45-specimen dataset (mean = 2452 bp, SD = 430.7), and 2691 bp for the 80-specimen dataset (mean = 1843 bp, SD = 788.2).

Phylogenetic relationships were inferred using maximum likelihood (ML), Bayesian Inference (BI) and maximum parsimony (MP). MODELTEST, version 3.6 (Posada & Crandall, 1998) was used to determine substitution models for model-based phylogenetic inferences according to hierarchical likelihood ratio tests (Huelsenbeck & Crandall, 1997).

Maximum likelihood

For ML trees, we used PHYML, version 2.4.4 (Guindon & Gascuel, 2003) with the nucleotide substitution model HKY (Hasegawa, Kishino & Yano, 1985). This software also estimated the Gamma distribution parameter, proportion of invariable sites and nucleotide frequencies. Branch support was assessed using 100 bootstrap replicates.

Bayesian inference

Bayesian analyses were conducted using MRBAYES, version 3.1.2 (Huelsenbeck & Ronquist, 2001). Datasets were partitioned by gene, and by codon position for COI and COII. Substitution models used for each partition were chosen according to MODELTEST (F81 for the second position of COI, GTR for the third position of COI, and HKY for the rest of partitions). Two runs of 1 000 000 generations with four chains (one cold and three heated) were performed. Chains were sampled every 100 generations, and burn-in was determined based on inspection of log likelihood over time plots using TRACER, version 1.4 (available from http://beast.bio.ed.ac.uk/Tracer).

Maximum parsimony

MP analyses were conducted using PAUP, version 4.0b10 (Swofford, 2000). Heuristic searches were performed with TBR branch swapping and 10 000 random taxon addition replicates, saving no more than ten equally parsimonious trees per replicate. To estimate branch support on the recovered topology, nonparametric bootstrap values (Felsenstein, 1985) were assessed with PAUP, version 4.0b10. One hundred bootstrap pseudoreplicates were obtained under a heuristic search with TBR branch swapping with 1000 random taxon addition replicates for the 45 taxon set, saving no more than ten equally parsimonious trees per replicate. Given the long computational time required for the 80-specimen set, 100 random taxon addition replicates were used in this case.

DATING PHYLOGENETIC EVENTS

BEAST, version 1.4.8 (Drummond & Rambaut, 2007) was used to estimate node ages. The analysis was carried out using the 45-taxa COI and COII dataset, with the same conditions described above for Bayesian phylogeny reconstruction. Monophyly constriction was enforced for several nodes according to the topology in Figure 1. Because no external calibration points, either in the form of a fossil or biogeographic event, are available for *Agrodiaetus*, we used a similar approach to that of Kandul *et al.* (2004). We selected two strongly supported nodes: one within the *dolus* species group and one within the *admetus* species group. Both are of an age close to 0.5 Myr, which we consider adequate to minimize the effects of saturation. Mean uncorrected pairwise distances within the two clades were calculated using MEGA4 (Tamura *et al.*, 2007). Dates for the two calibration points were the arithmetic means of the ages obtained applying a molecular clock with two published substitution rates: 1.5% uncorrected pairwise distance per million years estimated using a variety of invertebrates (Quek *et al.*, 2004) for COI, and a faster rate of 2.3% uncorrected pairwise distance per million years for the entire mitochondrial genome of various arthropod taxa (Brower, 1994). A normal prior distribution was used and the standard deviation was tuned so that the 95% central posterior density included the ages obtained with both rates. The dataset was analyzed under the HKY model applying a strict molecular clock along the branches. Base frequencies were estimated and the site heterogeneity

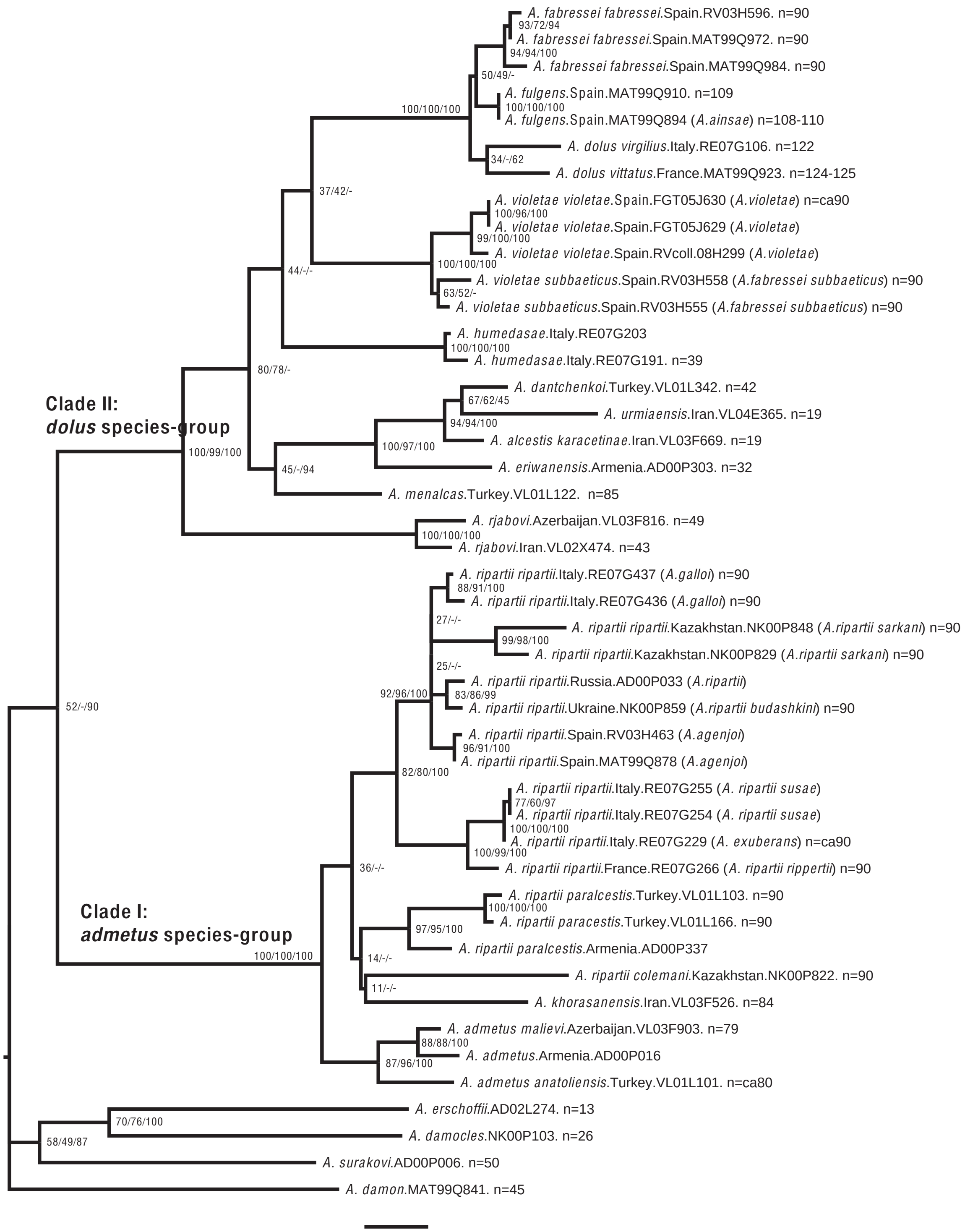

A. fabressei fabressei.Spain.RV03H596. n=90
93/72/94
A. fabressei fabressei.Spain.MAT99Q972. n=90
94/94/100
A. fabressei fabressei.Spain.MAT99Q984. n=90
50/49/-
A. fulgens.Spain.MAT99Q910. n=109
100/100/100
100/100/100
A. fulgens.Spain.MAT99Q894 (A.ainsae) n=108-110
A. dolus virgilius.Italy.RE07G106. n=122
34/-/62
A. dolus vittatus.France.MAT99Q923. n=124-125
37/42/-
A. violetae violetae.Spain.FGT05J630 (A.violetae) n=ca90
100/96/100
A. violetae violetae.Spain.FGT05J629 (A.violetae)
99/100/100
A. violetae violetae.Spain.RVcoll.08H299 (A.violetae)
100/100/100
44/-/-
A. violetae subbaeticus.Spain.RV03H558 (A.fabressei subbaeticus) n=90
63/52/-
A. violetae subbaeticus.Spain.RV03H555 (A.fabressei subbaeticus) n=90
A. humedasae.Italy.RE07G203
100/100/100
A. humedasae.Italy.RE07G191. n=39
80/78/-
A. dantchenkoi.Turkey.VL01L342. n=42
Clade II:
dolus species-group
67/62/45
A. urmiaensis.Iran.VL04E365. n=19
94/94/100
A. alcestis karacetinae.Iran.VL03F669. n=19
100/99/100
100/97/100
A. eriwanensis.Armenia.AD00P303. n=32
45/-/94
A. menalcas.Turkey.VL01L122. n=85
A. rjabovi.Azerbaijan.VL03F816. n=49
100/100/100
A. rjabovi.Iran.VL02X474. n=43
A. ripartii ripartii.Italy.RE07G437 (A.galloi) n=90
88/91/100
A. ripartii ripartii.Italy.RE07G436 (A.galloi) n=90
27/-/-
A. ripartii ripartii.Kazakhstan.NK00P848 (A.ripartii sarkani) n=90
99/98/100
A. ripartii ripartii.Kazakhstan.NK00P829 (A.ripartii sarkani) n=90
25/-/-
A. ripartii ripartii.Russia.AD00P033 (A.ripartii)
92/96/100
83/86/99
52/-/90
A. ripartii ripartii.Ukraine.NK00P859 (A.ripartii budashkini) n=90
A. ripartii ripartii.Spain.RV03H463 (A.agenjoi)
96/91/100
A. ripartii ripartii.Spain.MAT99Q878 (A.agenjoi)
82/80/100
A. ripartii ripartii.Italy.RE07G255 (A. ripartii susae)
77/60/97
A. ripartii ripartii.Italy.RE07G254 (A. ripartii susae)
100/100/100
A. ripartii ripartii.Italy.RE07G229 (A. exuberans) n=ca90
100/99/100
36/-/-
A. ripartii ripartii.France.RE07G266 (A. ripartii rippertii) n=90
A. ripartii paralcestis.Turkey.VL01L103. n=90
Clade I:
100/100/100
admetus species-group
A. ripartii paracestis.Turkey.VL01L166. n=90
97/95/100
A. ripartii paralcestis.Armenia.AD00P337
100/100/100
14/-/-
A. ripartii colemani.Kazakhstan.NK00P822. n=90
11/-/-
A. khorasanensis.Iran.VL03F526. n=84
A. admetus malievi.Azerbaijan.VL03F903. n=79
88/88/100
A. admetus.Armenia.AD00P016
87/96/100
A. admetus anatoliensis.Turkey.VL01L101. n=ca80
A. erschoffii.AD02L274. n=13
70/76/100
A. damocles.NK00P103. n=26
58/49/87
A. surakovi.AD00P006. n=50
A. damon.MAT99Q841. n=45
0.004

Figure 1. Maximum likelihood tree of *Agrodiaetus* based on the combined analysis of the mitochondrial cytochrome oxidase subunit *I* (COI), leucine transfer RNA (leu-tRNA), cytochrome oxidase subunit *II* (COII) and nuclear internal transcribed spacer 2 (ITS2) (2812 bp) from 45 samples of *Agrodiaetus* according to the Hasegawa, Kishino & Yano model (log likelihood score = –8727.72). Traditional names are indicated in parentheses when new names or combinations are proposed. Haploid chromosome numbers (*n*) are indicated after specimen codes. Numbers at nodes indicate maximum likelihood bootstrap/maximum parsimony bootstrap/Bayesian posterior probability, with nonmatching clades using different analyses indicated by '–'. The scale bar represents 0.004 substitutions/position.

model gamma with four categories was used. Parameters were estimated using two independent runs of 10 million generations each (with a pre-run burn-in of 100 000 generations) to ensure convergence, and checked with the software TRACER, version 1.4. Summary trees were generated using TREEANNOTATOR, version 1.4.8 (available from http://beast.bio.ed.ac.uk).

RESULTS

KARYOTYPES

Karyotype of A. violetae

The taxon *A. violetae* is extremely rare. We were able to obtain a limited number of individuals, of which only one sample had metaphase plates suitable for determination of karyotype characteristics. In this preparation, the chromosome number was determined to be *n* = *ca*90 (Table 3). Two chromosomes were especially large (Fig. 2A) in the second metaphase of meiosis (MII) complement, and one chromosome was medium-sized. The two largest chromosomes were nearly of equal size, and the medium-sized chromosome was 1.8–2.0 times smaller than these.

Karyotype of A. fabressei subbaeticus

The haploid chromosome number of *A. fabressei subbaeticus* was found to be *n* = 90 (Fig. 2B, C, Table 3), thus confirming our previous results (Lukhtanov *et al.*, 2006). Three bivalents were especially large (Fig. 2B) in the first metaphase of meiosis (MI) complement. Bivalent 1 was only slightly larger than bivalent 2, and the latter was 1.4–1.8 times larger than bivalent 3. In the MII complement, the two largest chromosomes were nearly of equal size, and chromosome 3 was 1.8–2.0 times smaller than the two biggest chromosomes (Fig. 2C).

Karyotype of A. humedasae

The haploid chromosome number was determined to be *n* = 39 (Table 3). Bivalents in MI and chromosomes in MII were fairly differentiated with respect to their size; however, it is difficult to divide them objectively into size groups because the sizes of the 39 bivalents decrease more or less linearly (Fig. 2D, E, F).

KARYOTYPES OF *A. AGENJOI*, *A. RIPARTII* RIPPERTII, *A. GALLOI*, AND *A. EXUBERANS*

The haploid chromosome number was determined to be *n* = 90 in *agenjoi*, *rippertii*, and *galloi*. In MI, two bivalents were especially large and were situated in the centre of the metaphase plates. Bivalent 1 was 1.4–1.6 times larger than bivalent 2. The sizes of the remaining 88 bivalents decreased more or less linearly (Fig. 2G, H, I, J, K, L). Few meiotic metaphase

Table 3. Number of bivalents and mitotic chromosomes observed in the taxa and specimens studied

Taxon	Specimen code number	Country	Haploid (*n*) or diploid (2*n*) chromosome number	Number of cells with accurately determined bivalent/ chromosome number	Number of large (L) and medium (M) bivalents/chromosomes in haploid complement
violetae	FGT-05-J630	Spain	*n* = *ca*90	–	2L + 1M
subbaeticus	RV-03-H554	Spain	*n* = *ca*90	–	2L + 1M
subbaeticus	RV-03-H555	Spain	*n* = 90	5MI	2L + 1M
subbaeticus	RV-03-H556	Spain	*n* = *ca*90	–	2L + 1M
subbaeticus	RV-03-H557	Spain	*n* = 90	2MII	2L + 1M
subbaeticus	RV-03-H558	Spain	*n* = 90	4MI	2L + 1M
subbaeticus	RV-03-H560	Spain	*n* = 90	2MI, 2MII	2L + 1M
humedasae	RE-7-G191	Italy	*n* = 39	12MI	–
humedasae	RE-7-G192	Italy	*n* = 39	8MI	–
humedasae	RE-7-G193	Italy	*n* = 39	4MII	–
humedasae	RE-7-G194	Italy	*n* = 39	7MI	–
agenjoi	RV-07-F038	Spain	*n* = 90	5MI, 3MII	1L + 1M
rippertii	RE-7-G266	France	*n* = 90	2MI, 2MII	1L + 1M
rippertii	RE-7-G273	France	*n* = 90	3MI	1L + 1M
exuberans	RE-7-G229	Italy	2*n* = *ca*180	–	1L + 1M
galloi	RE-7-G436	Italy	*n* = 90	7MI, 3MII	1L + 1M
galloi	RE-7-G437	Italy	*n* = 90	6MI, 3MII	1L + 1M
galloi	RE-7-G441	Italy	*n* = 90	4MI	1L + 1M
galloi	RE-7-G445	Italy	*n* = 90	4MI	1L + 1M
galloi	RE-7-G447	Italy	*n* = *ca*90	–	1L + 1M

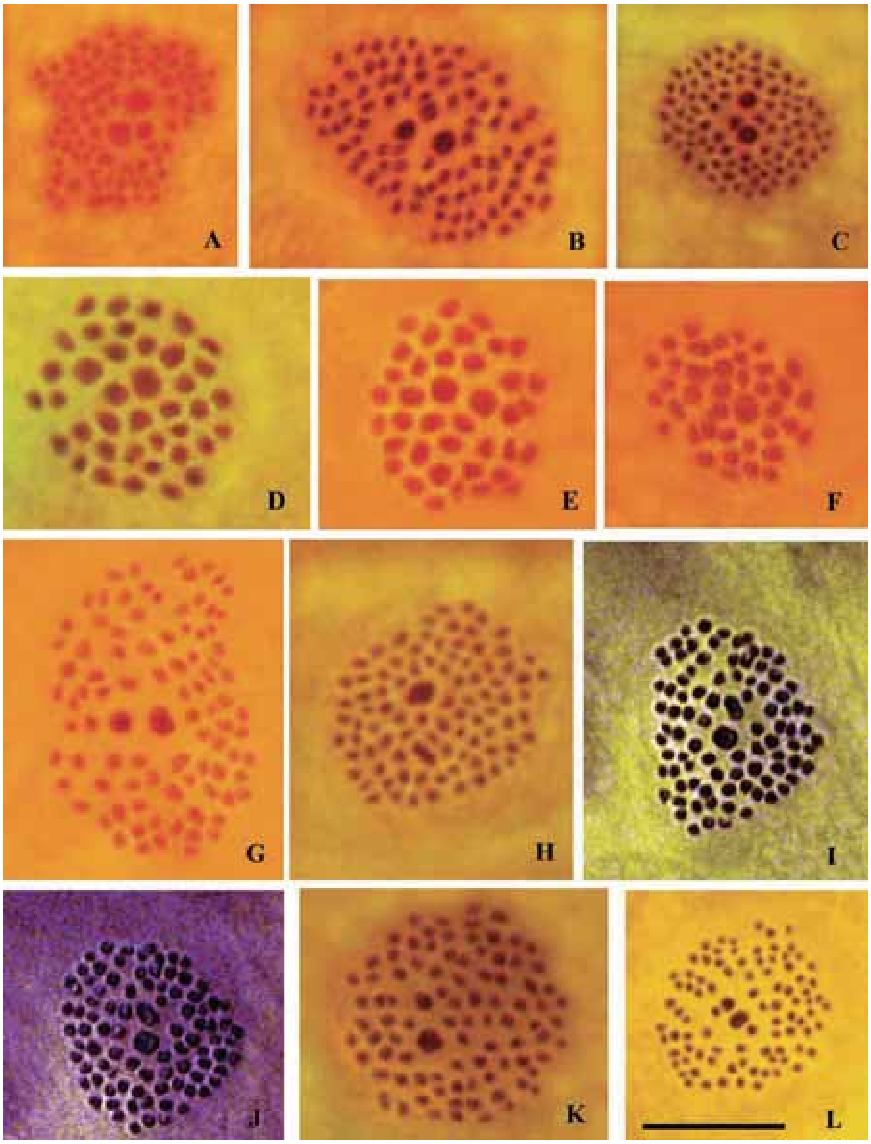


Figure 2. *Agrodiaetus* karyotypes. Scale bar corresponds to 10 mm in all figures. A, *Agrodiaetus violetae violetae* (sample FGT-05-J630). Pole view of a second metaphase of meiosis (MII) plate ($n = ca90$). Two large and one medium-sized chromosome in the centre of the plate can be seen. B, squash preparation of *Agrodiaetus violetae subbaeticus* **comb. nov.** (sample RV-03-H555). First metaphase of meiosis (MI) plate ($n = 90$). Three bivalents are larger than the rest (two large and one medium) in the centre of the metaphase plate. C, *Agrodiaetus violetae subbaeticus* **comb. nov.** (sample RV-03-H560). Pole view of an intact (unsquashed) MII plate ($n = 90$). All the chromosomes are situated in a plane with the largest elements in the centre of the circular metaphase plate clearly separated from each other by gaps. Three chromosomes are larger than the rest (two large + one medium). D, E, F, *Agrodiaetus humedasae*. Pole view of intact (unsquashed) MI plates ($n = 39$). Bivalents are fairly differentiated with respect to their size; however, it is difficult to divide them objectively into size groups because the sizes of the 39 bivalents decrease more or less linearly. D, sample RE-07-G191; E, sample RE-07-G192; F, sample RE-07-G194. G, squash preparation of *Agrodiaetus ripartii agenjoi* (sample RVcoll.07-F038). MI plate ($n = 90$). Two bigger bivalents (one large and one medium) are in the centre of the metaphase plate. H, *Agrodiaetus ripartii rippertii* (sample RE-07-G273). MI plate ($n = 90$). Pole view of a slightly squashed MI plate. Two larger bivalents (one large and one medium) are on the metaphase plate. The original position of the bivalents was altered during preparation, and the medium bivalent is no longer situated in the centre, as it was initially. I, J, K, L, *Agrodiaetus ripartii galloi*. MI plates ($n = 90$). Two bivalents are bigger than the rest (one large and one medium) in the centre of the metaphase plates. I, J, slightly squashed plates of sample RE-07-G436; K, a squashed plate of sample RE-07-G437. L, squash preparation of *Agrodiaetus ripartii galloi* (sample RE-07-G436). MII plate ($n = 90$). Two chromosomes are bigger than the rest (one large and one medium) in the centre of the metaphase plate.

plates were found in *exuberans*, and they were not acceptable for chromosome counts. However, they each displayed one large and one medium bivalent in MI, exactly as it was found in *A. ripartii*. The diploid chromosome number of *exuberans*, however, could be established to be *n* = *ca*180 (with two larger and two medium-sized chromosomes), which would correspond to a haploid number of *n* = *ca*90 with one larger and one medium-sized bivalent (Table 3).

Phylogeny

Analyses for both the 45-specimen dataset and the 80-specimen dataset recover the *admetus* (clade I) and the *dolus* (clade II) species groups as strongly supported (Figs 1, 3). This concords with results of other studies (Kandul *et al.*, 2002, 2004, 2007; Wiemers, 2003). Within each of these two main groups, many clades are well supported, whereas some of the relationships are not fully resolved. If we compare analyses from the 45-specimen dataset and the 80-specimen dataset, we find that the addition of short COI sequences and ITS2 from Wiemers (2003) adds information by expanding the sampling, but generally produces a lowering of node support. This may be explained by the low overlap of these short COI sequences with many of the longer ones, as well as the low variability of the ITS2 marker between closely-related taxa. Indeed, a tree generated exclusively from ITS2 data (not shown), recovers only the deepest nodes defining the *dolus* and the *admetus* species groups, except for *Agrodiaetus valiabadi*, whose placement is unresolved. Within the *dolus* group, ITS2 supports the *dolus–fulgens–fabressei* clade, the close relationship between the taxa *violetae* and *subbaeticus*, as well as the sister relationship between *A. humedasae* and *Agrodiaetus aroanensis*. Thus, the utility of ITS2 is limited, although, because it is a nuclear marker, it independently confirms the main groups obtained using the mitochondrial data.

Dating analysis (Fig. 4) estimated an age of 3.21 Myr (2.25–4.29; error interval covering 95% highest posterior density) for the genus *Agrodiaetus*, similar to the dates obtained in previous studies (Mensi *et al.*, 1994; Kandul *et al.*, 2004). The estimated age for the split between the sister *dolus* and *ripartii* lineages is 2.73 Myr (range 1.89–3.58 Myr). Finer relationships recovered within each species group and their ages are described in detail in the Discussion, together with their taxonomical implications.

DISCUSSION

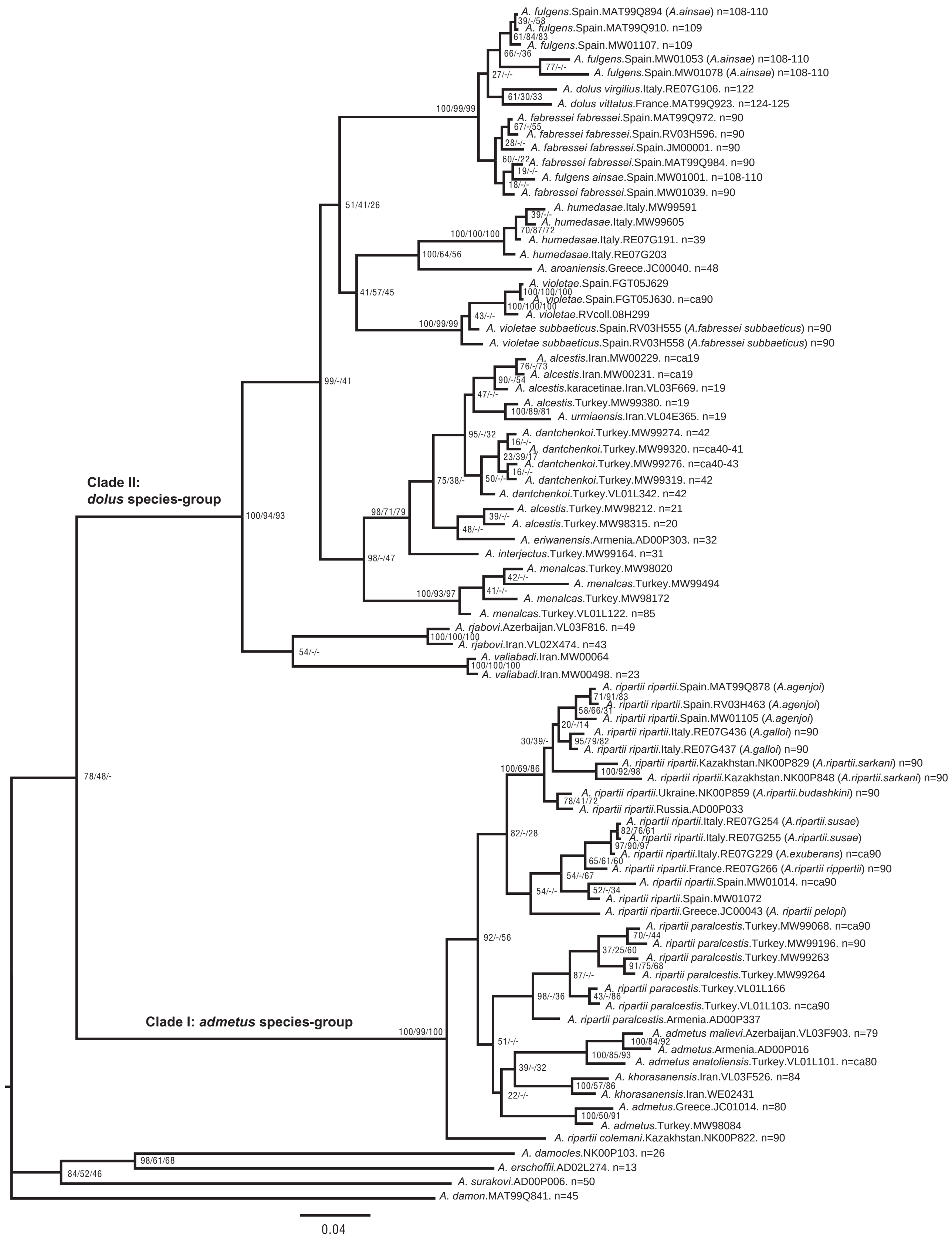
Taxonomical oversplitting in Western European Agrodiaetus

The European *Agrodiaetus* taxa distributed west of the 17th meridian belong to three different phylogenetic lineages (Kandul *et al.*, 2002, 2004, 2007; Wiemers, 2003; our data). One highly differentiated lineage is sister to all other *Agrodiaetus* and consists of a single species, *A. damon*, which has a broad distribution range from Spain to Mongolia (Fig. 5A). This species has no close relatives, and its standing as a good species has never been disputed. All other western European taxa constitute two lineages: the *A. ripartii* lineage, which is part of clade I, and the *A. dolus* lineage, which is part of clade II (Figs 1, 3). The *A. ripartii* lineage includes the taxa *agenjoi*, *exuberans*, *galloi*, *pelopi*, *ripartii*, *rippertii*, and *susae*. The *A. dolus* lineage includes the taxa *ainsae*, *aroaniensis*, *dolus*, *fabressei*, *fulgens*, *humedasae*, *subbaeticus*, *violetae*, *virgilia*, and *vittatus*. The present study supports all previous conclusions about the general taxonomic structure of the *A. admetus* (clade I) and the *A. dolus* (clade II) species groups. At the same time, it sheds light on the taxonomic status and phylogenetic relationships of several western European species whose positions were under debate.

Agrodiaetus ripartii lineage (Fig. 5B)

Agrodiaetus agenjoi: This taxon was described by Forster (1965) from Barcelona (Catalonia, Spain) as a subspecies of the Balkanian–Anatolian species *A. admetus*. Subsequently, de Lesse (1968) and Munguira, Martín & Pérez-Valiente (1995) demonstrated the karyotype similarity of the taxon *agenjoi* and *A. ripartii* (both taxa have *n* = 90, including one large and one medium-sized chromosome pair) and suggested that *A. agenjoi* should be considered a subspecies of *A. ripartii*. Despite these chromosomal studies, and without any explicit justification, *agenjoi* is often treated in the literature as a distinct species with a

Figure 3. Bayesian tree based on the combined analysis of data from mitochondrial cytochrome oxidase subunit *I* (COI), leucine transfer RNA (leu-tRNA), cytochrome oxidase subunit *II* (COII) and nuclear internal transcribed spacer 2 (ITS2) (2691 bp), partitioned by marker and gene codon position, from 80 samples of *Agrodiaetus* (log likelihood score = −7942.31). Traditional names are indicated in parentheses when new names or combinations are proposed. Haploid chromosome numbers (*n*) are indicated after the specimen code numbers. Numbers at nodes indicate Bayesian posterior probability/maximum likelihood bootstrap/maximum parsimony bootstrap, with nonmatching clades among different analysis indicated by '–'. The scale bar represents 0.04 substitutions/position.


A. fulgens.Spain.MAT99Q894 (A.ainsae) n=108-110
39/-/58
A. fulgens.Spain.MAT99Q910. n=109
61/84/83
A. fulgens.Spain.MW01107. n=109
66/-/36
A. fulgens.Spain.MW01053 (A.ainsae) n=108-110
77/-/-
A. fulgens.Spain.MW01078 (A.ainsae) n=108-110
27/-/-
A. dolus virgilius.Italy.RE07G106. n=122
61/30/33
A. dolus vittatus.France.MAT99Q923. n=124-125
100/99/99
A. fabressei fabressei.Spain.MAT99Q972. n=90
67/-/55
A. fabressei fabressei.Spain.RV03H596. n=90
28/-/-
A. fabressei fabressei.Spain.JM00001. n=90
60/-/22
A. fabressei fabressei.Spain.MAT99Q984. n=90
19/-/-
A. fulgens ainsae.Spain.MW01001. n=108-110
18/-/-
A. fabressei fabressei.Spain.MW01039. n=90
51/41/26
A. humedasae.Italy.MW99591
39/-/-
A. humedasae.Italy.MW99605
70/87/72
100/100/100
A. humedasae.Italy.RE07G191. n=39
100/64/56
A. humedasae.Italy.RE07G203
A. aroaniensis.Greece.JC00040. n=48
A. violetae.Spain.FGT05J629
100/100/100
A. violetae.Spain.FGT05J630. n=ca90
41/57/45
100/100/100
A. violetae.RVcoll.08H299
43/-/-
100/99/99
A. violetae subbaeticus.Spain.RV03H555 (A.fabressei subbaeticus) n=90
A. violetae subbaeticus.Spain.RV03H558 (A.fabressei subbaeticus) n=90
A. alcestis.Iran.MW00229. n=ca19
76/-/73
A. alcestis.Iran.MW00231. n=ca19
90/-/54
99/-/41
A. alcestis.karacetinae.Iran.VL03F669. n=19
47/-/-
A. alcestis.Turkey.MW99380. n=19
100/89/81
A. urmiaensis.Iran.VL04E365. n=19
A. dantchenkoi.Turkey.MW99274. n=42
95/-/32
16/-/-
A. dantchenkoi.Turkey.MW99320. n=ca40-41
23/39/17
A. dantchenkoi.Turkey.MW99276. n=ca40-43
16/-/-
Clade II:
dolus species-group
75/38/-
50/-/-
A. dantchenkoi.Turkey.MW99319. n=42
A. dantchenkoi.Turkey.VL01L342. n=42
A. alcestis.Turkey.MW98212. n=21
98/71/79
39/-/-
100/94/93
A. alcestis.Turkey.MW98315. n=20
48/-/-
A. eriwanensis.Armenia.AD00P303. n=32
A. interjectus.Turkey.MW99164. n=31
98/-/47
A. menalcas.Turkey.MW98020
42/-/-
A. menalcas.Turkey.MW99494
41/-/-
100/93/97
A. menalcas.Turkey.MW98172
A. menalcas.Turkey.VL01L122. n=85
A. rjabovi.Azerbaijan.VL03F816. n=49
100/100/100
A. rjabovi.Iran.VL02X474. n=43
54/-/-
A. valiabadi.Iran.MW00064
100/100/100
A. valiabadi.Iran.MW00498. n=23
A. ripartii ripartii.Spain.MAT99Q878 (A.agenjoi)
71/91/83
A. ripartii ripartii.Spain.RV03H463 (A.agenjoi)
58/66/31
A. ripartii ripartii.Spain.MW01105 (A.agenjoi)
20/-/14
A. ripartii ripartii.Italy.RE07G436 (A.galloi) n=90
30/39/-
95/79/82
A. ripartii ripartii.Italy.RE07G437 (A.galloi) n=90
A. ripartii ripartii.Kazakhstan.NK00P829 (A.ripartii.sarkani) n=90
100/69/86
100/92/98
78/48/-
A. ripartii ripartii.Kazakhstan.NK00P848 (A.ripartii.sarkani) n=90
A. ripartii ripartii.Ukraine.NK00P859 (A.ripartii.budashkini) n=90
78/41/72
A. ripartii ripartii.Russia.AD00P033
A. ripartii ripartii.Italy.RE07G254 (A.ripartii.susae)
82/76/61
82/-/28
A. ripartii ripartii.Italy.RE07G255 (A.ripartii.susae)
97/90/97
A. ripartii ripartii.Italy.RE07G229 (A.exuberans) n=ca90
65/61/60
A. ripartii ripartii.France.RE07G266 (A.ripartii rippertii) n=90
54/-/67
A. ripartii ripartii.Spain.MW01014. n=ca90
52/-/34
54/-/-
A. ripartii ripartii.Spain.MW01072
A. ripartii ripartii.Greece.JC00043 (A. ripartii pelopi)
A. ripartii paralcestis.Turkey.MW99068. n=ca90
70/-/44
92/-/56
A. ripartii paralcestis.Turkey.MW99196. n=90
37/25/60
A. ripartii paralcestis.Turkey.MW99263
91/75/68
87/-/-
A. ripartii paralcestis.Turkey.MW99264
A. ripartii paracestis.Turkey.VL01L166
98/-/36
43/-/86
A. ripartii paralcestis.Turkey.VL01L103. n=ca90
Clade I: admetus species-group
A. ripartii paralcestis.Armenia.AD00P337
100/99/100
A. admetus malievi.Azerbaijan.VL03F903. n=79
100/84/92
51/-/-
A. admetus.Armenia.AD00P016
100/85/93
A. admetus anatoliensis.Turkey.VL01L101. n=ca80
39/-/32
A. khorasanensis.Iran.VL03F526. n=84
100/57/86
22/-/-
A. khorasanensis.Iran.WE02431
A. admetus.Greece.JC01014. n=80
100/50/91
A. admetus.Turkey.MW98084
A. ripartii colemani.Kazakhstan.NK00P822. n=90
A. damocles.NK00P103. n=26
98/61/68
A. erschoffii.AD02L274. n=13
84/52/46
A. surakovi.AD00P006. n=50
A. damon.MAT99Q841. n=45
0.04

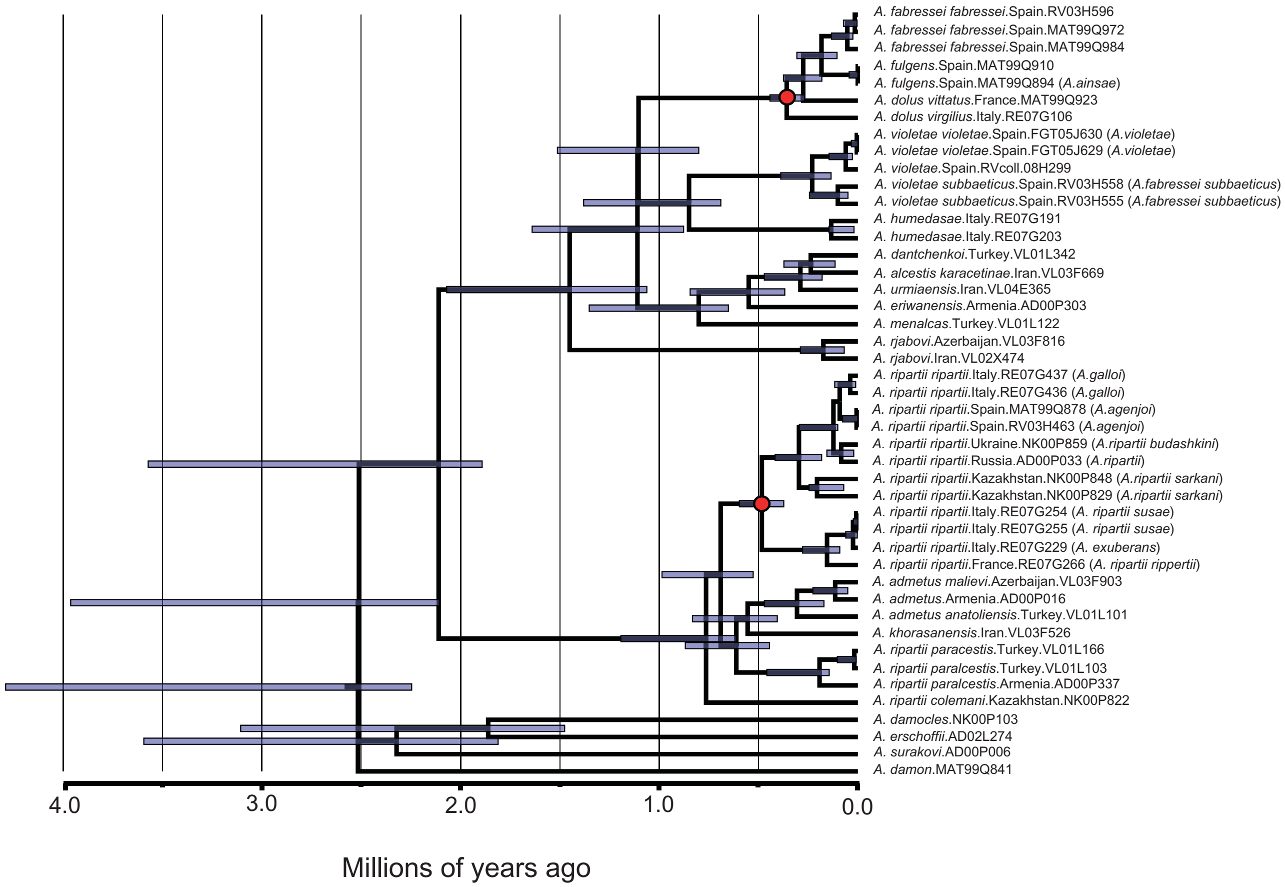

A. fabressei fabressei.Spain.RV03H596
A. fabressei fabressei.Spain.MAT99Q972
A. fabressei fabressei.Spain.MAT99Q984
A. fulgens.Spain.MAT99Q910
A. fulgens.Spain.MAT99Q894 (A.ainsae)
A. dolus vittatus.France.MAT99Q923
A. dolus virgilius.Italy.RE07G106
A. violetae violetae.Spain.FGT05J630 (A.violetae)
A. violetae violetae.Spain.FGT05J629 (A.violetae)
A. violetae.Spain.RVcoll.08H299
A. violetae subbaeticus.Spain.RV03H558 (A.fabressei subbaeticus)
A. violetae subbaeticus.Spain.RV03H555 (A.fabressei subbaeticus)
A. humedasae.Italy.RE07G191
A. humedasae.Italy.RE07G203
A. dantchenkoi.Turkey.VL01L342
A. alcestis karacetinae.Iran.VL03F669
A. urmiaensis.Iran.VL04E365
A. eriwanensis.Armenia.AD00P303
A. menalcas.Turkey.VL01L122
A. rjabovi.Azerbaijan.VL03F816
A. rjabovi.Iran.VL02X474
A. ripartii ripartii.Italy.RE07G437 (A.galloi)
A. ripartii ripartii.Italy.RE07G436 (A.galloi)
A. ripartii ripartii.Spain.MAT99Q878 (A.agenjoi)
A. ripartii ripartii.Spain.RV03H463 (A.agenjoi)
A. ripartii ripartii.Ukraine.NK00P859 (A.ripartii budashkini)
A. ripartii ripartii.Russia.AD00P033 (A.ripartii)
A. ripartii ripartii.Kazakhstan.NK00P848 (A.ripartii sarkani)
A. ripartii ripartii.Kazakhstan.NK00P829 (A.ripartii sarkani)
A. ripartii ripartii.Italy.RE07G254 (A. ripartii susae)
A. ripartii ripartii.Italy.RE07G255 (A. ripartii susae)
A. ripartii ripartii.Italy.RE07G229 (A. exuberans)
A. ripartii ripartii.France.RE07G266 (A. ripartii rippertii)
A. admetus malievi.Azerbaijan.VL03F903
A. admetus.Armenia.AD00P016
A. admetus anatoliensis.Turkey.VL01L101
A. khorasanensis.Iran.VL03F526
A. ripartii paracestis.Turkey.VL01L166
A. ripartii paralcestis.Turkey.VL01L103
A. ripartii paralcestis.Armenia.AD00P337
A. ripartii colemani.Kazakhstan.NK00P822
A. damocles.NK00P103
A. erschoffii.AD02L274
A. surakovi.AD00P006
A. damon.MAT99Q841
4.0
3.0
2.0
1.0
0.0
Millions of years ago

Figure 4. Bayesian ultrametric tree for the 45-taxa dataset obtained with BEAST 1.4.8, based on cytochrome oxidase subunit *I* (COI) and cytochrome oxidase subunit *II* (COII) sequences under the Hasegawa, Kishino & Yano model of DNA substitution. The tree was calibrated at the two nodes indicated (red circles) based on two different published divergence rates for mitochondrial DNA in Arthropoda (1.5% and 2.3% pairwise sequence divergence per million years). For each calibration point, a normal prior distribution was centred on the resulting mean age (and SD) was tuned so that the 95% central posterior density included the ages obtained with both rates. Bars in nodes represent the 95% highest posterior density for age estimations, according to the axis representing time in millions years before present. Traditional names are indicated in parentheses when new names or combinations are proposed.

distributional range restricted to Catalonia in northeast Spain (Kolev & De Prins, 1995; Dennis, 1997; Tolman, 1997; Mazzei *et al.*, 2009) or as a subspecies of *A. fabressei* (Manley & Allcard, 1970) (but see also Munguira *et al.*, (1995) and Eckweiler & Häuser (1997), who considered this taxon a subspecies of *A. ripartii*).

Our molecular phylogeny recovers *A. agenjoi* as an internal clade within one of the *A. ripartii* clades. The monophyly of the *agenjoi* clade has good support in the 45-specimen set, but lower support in the 80-specimen set. Its genetic divergence with respect to *A. ripartii* samples from Russia and Ukraine, as well as with the taxon *A. galloi*, is minimal (0.28–0.56%) and includes only three fixed nucleotide substitutions in 1858 bp of COI-tRNALeu-COII. This difference is extremely small and could even be less when additional individuals and intermediate populations are studied. Our chromosomal data confirm that the karyotype of *A. agenjoi* is indistinguishable from that of *A. ripartii*, and do not support the species status of *A. agenjoi*. Moreover, morphological differences between *A. ripartii* and *A. agenjoi* are subtle and inconstant. The character that is usually used to distinguish between them) the presence of a white stripe on the underside of the hind wing of *A. ripartii*, and its absence in *A. agenjoi*; Tolman, 1997) can be variable in *Agrodiaetus* at the species, population, and individual levels, and its taxonomic significance is also low (Eckweiler & Häuser, 1997; Lukhtanov & Budashkin, 2007). Moreover, although generally absent in *agenjoi*, this streak is present in a low percentage of the Catalonian specimens. Because

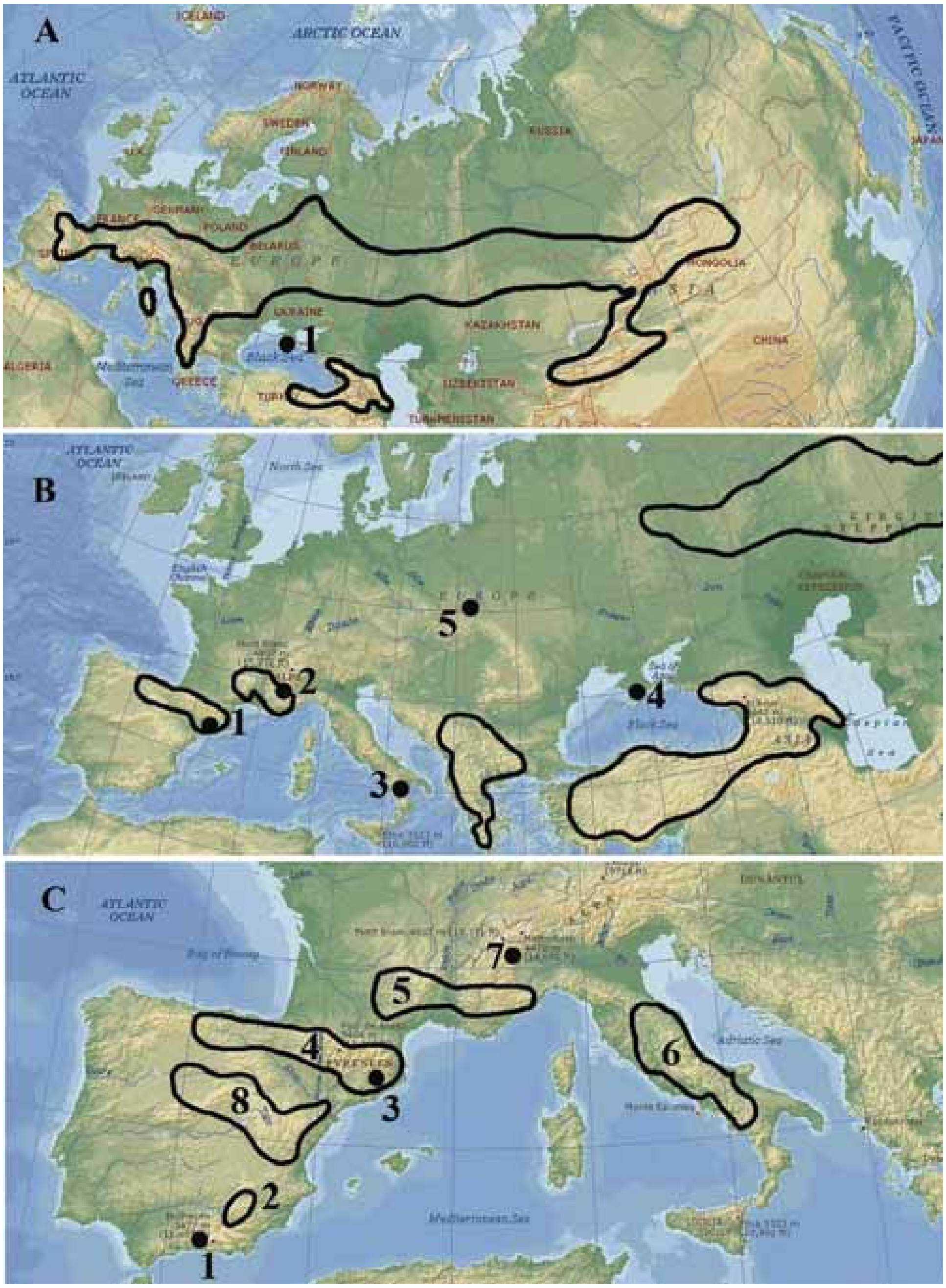


Figure 5. Distribution ranges of western European *Agrodiaetus*, according to data original to the present study, Hesselbarth, Oorchot & Wagener (1995), Kudrna (2002) and García-Barros *et al.* (2004). A, distribution ranges *of Agrodiaetus damon* (closed loops) and *Agrodiaetus pljushtchi* (1). B, distribution ranges of taxa belonging to the *Agrodiaetus ripartii* lineage: 1 – *Agrodiaetus agenjoi* (here assigned to *Agrodiaetus ripartii ripartii*); 2 – *Agrodiaetus exuberans* (here assigned to *Agrodiaetus ripartii ripartii*); 3 – *Agrodiaetus galloi* (here assigned to *A. ripartii ripartii*); 4 – *Agrodiaetus budashkini* (here assigned to *A. ripartii ripartii*); 5 – a geographically isolated population of *A. ripartii* in Poland (Przybylowicz, 2000). Distribution range of the main populations of *A. ripartii* indicated by closed loops. C, distribution ranges of taxa belonging to the *Agrodiaetus dolus* lineage: 1 – *Agrodiaetus violetae violetae*; 2 – *Agrodiaetus violetae subbaeticus* **comb. nov.**; 3 – presumed distribution range of *Agrodiaetus fulgens* before the chromosomal study by Lukhtanov *et al.* (2006); 4 – revised distribution range of *A. fulgens*; 5 – *Agrodiaetus dolus dolus* and *Agrodiaetus dolus vittatus*; 6 – *Agrodiaetus dolus virgilia*; 7 – *Agrodiaetus humedasae*; 8 – *Agrodiaetus fabressei.*

the taxa *ripartii* and *agenjoi* were both described from northern Spain, we consider the name *agenjoi* to be a synonym of *A. ripartii*.

Agrodiaetus galloi: This taxon was described as a distinct species (Baletto & Toso, 1979) from southern Italy on the basis of an extreme difference in karyotype; its chromosome number was established to be $2n = 132$ ($n = 66$), including one pair of large and one pair of medium-sized chromosomes (Troiano & Giribaldi, 1979), whereas *A. ripartii*, geographically and phenotypically the most closely related taxon, has $n = 90$ (de Lesse, 1960b). *Agrodiaetus galloi* has invariably been considered a good species in all studies on European butterflies, with the exception of Eckweiler & Häuser (1997), who questioned the species status of this taxon.

The present study confirms the presence of one large and one medium-sized bivalent in *A. galloi*, but we were unable to confirm the previously reported chromosome number. Without exception, all studied cells and individuals possessed a chromosome number of $n = 90$ (Fig. 2I, J, K, L). We consider that the discrepancy between the earlier chromosome count and ours arises because true MI or MII metaphase cells were not observed in the study by Troiano & Giribaldi (1979). According to their figure 7, which was originally interpreted to be a picture of anaphase I, they in fact observed atypical meiotic divisions. Such atypical divisions occur regularly during male meiosis in all species of Lepidoptera, to the point where during the imaginal stage, they are much more frequent than normal meiotic divisions (Lorkovic, 1990). Generally, atypical divisions display the diploid set; however, the great majority of atypical spermatocytes are not suitable for chromosome counts as a result of multiple nonspecific chromosome conglutinations (Lorkovic, 1990) that can lead to a strong underestimation of the real chromosome number. Thus, we consider that the true number of chromosomes in the samples studied by Troiano & Giribaldi (1979) was also $n = 90$, and that the karyotype of *A. galloi* is in fact indistinguishable from that found in *A. ripartii*.

In our phylogenetic reconstruction, the taxon *galloi* forms a well-supported cluster with *A. ripartii* samples Ukraine, Russia, and Kazakhstan, as well as with the taxon *agenjoi* (Figs 1, 3). Moreover, the genetic divergence of the two studied individuals of *galloi* with respect to the most closely-related *ripartii* and *agenjoi* samples is extremely small and could be even less when additional individuals are studied. The specimens studied have ITS2 sequences identical to those of several *A. ripartii*, excluding the possibility that *galloi* is a diverged taxon that has undergone mitochondrial introgression from *A. ripartii*. Thus, the morphological, chromosomal, and genetic data do not support the treatment of *A. galloi* as a separate species. It should be synonymized with *A. ripartii* or, at most, considered a weakly differentiated local subspecies of *A. ripartii*.

The taxa rippertii, exuberans *and* susae*:* The taxon *rippertii* was described from southern France ('aux environs de Digne') as a separate species by Boisduval (1832). In the original description, however, Boisduval made no reference to Freyer (1830), who established from Spain a morphologically very similar taxon (*ripartii*) 2 years earlier. Therefore, the taxon *rippertii* has been considered a synonym or subspecies of *A. ripartii* in recent literature (Eckweiler & Häuser, 1997).

The taxon *exuberans* was described from 'Oulx' (northern Italy) as a 'race' (i.e. subspecies) of *A. admetus* by Verity (1926). It is similar morphologically to *A. ripartii* and was regarded later as a subspecies or even a synonym of *A. ripartii* (Eckweiler & Häuser, 1997). However, without explicit justification, it has been raised to species rank in most recent studies (Kudrna, 2002; Bertaccini, 2003; Dennis *et al.*, 2008).

The taxon *susae* was described from northern Italy as separate subspecies of *A. ripartii* (Bertaccini, 2003). In accordance with the original description, the taxon *susae* is sympatric with *A. exuberans*, and these two taxa are different in small details of genitalic structure and wing spots. We collected both *exuberans* and *susae* in their exact type locality, and comparison of these individuals showed that the morphological differences between *exuberans* and *susae* are sufficiently subtle so that it is not always possible to distinguish between them in practice (R. Vila & V. A. Lukhtanov, unpubl. observ.). Molecular analysis demonstrated that the taxa *exuberans* and *susae* are almost identical, and genetically similar to *A. ripartii rippertii* from France. These three taxa constitute a well-supported monophyletic clade within the bigger *A. ripartii* clade in the 45-specimen dataset (Fig. 1), although the support of this clade is relatively low in the 80-specimen dataset (Fig. 3). Moreover, nuclear ITS2 sequences of these three taxa are identical, which independently supports the results of the mitochondrial sequences. Chromosomal analysis showed that karyotypes of the taxa *rippertii* and *exuberans* are indistinguishable from those of *A. ripartii* from Europe, Turkey, and Kazakhstan (de Lesse, 1960b; Lukhtanov & Dantchenko, 2002b). We were unable to obtain countable metaphase plates for the taxon *susae* but, taking into account its genetic and morphological similarity to the taxa *rippertii* and *exuberans*, we consider it unlikely to be a separate taxon.

Infraspecific taxonomy of A. ripartii*:* In the molecular phylogeny, *A. ripartii* samples from Asia Minor and Armenia (*Agrodiaetus ripartii paralcestis*) and especially from Central Asia (*Agrodiaetus ripartii colemani*) are genetically distant from European and other Kazakhstani populations (Figs 1, 3). We will not discuss these further here because this is beyond the scope of the present study and our material from these regions is limited. However, all other samples of *A. ripartii* from western Europe, the Balkan Peninsula, European Russia, and the Ukraine (including representatives of the nominal taxa *Agrodiaetus ripartii ripartii*, *A. ripartii rippertii*, *Agrodiaetus ripartii sarkani*, *Agrodiaetus ripartii budashkini*, *A. ripartii susae*, *Agrodiaetus ripartii pelopi*, *A. agenjoi*, *A. exuberans*, and *A. galloi*) form a well-supported clade. The genetically related representatives of this clade display allopatric distributions, are similar in their morphology, and are indistinguishable with respect to karyotype. A more detailed study of *A. ripartii* will be necessary to shed light on relationships between populations and on the total number of subspecies. Given the data available, and until these relationships can be clarified, we provisionally consider all European and North and East Kazakhstani populations to belong to the nominative subspecies *A. ripartii ripartii*. Thus, we recognize three subspecies defined by the main three *A. ripartii* clades: *A. ripartii ripartii*, *A. ripartii paralcestis* and *A. ripartii colemani*.

Agrodiaetus dolus lineage

By contrast to the *A. ripartii* lineage, the *A. dolus* complex is represented in western Europe by a number of distinct taxa that appear to be allopatric in their distribution. All of them are clearly separated from one another by significant chromosomal and/or genetic gaps. Interestingly, two species, *A. dolus and A. fulgens*, are whitish–blue on the upperside of the male wing, and are therefore morphologically different from the rest.

The taxa violetae *and* subbaeticus*:* The taxon *violetae* was described from southern Spain as a new species that is similar to *A. fabresseii*, but differs by the presence of a white stripe on the underside of the hind wing (Gómez-Bustillo, Expósita Hermosa & Martínez Borrego, 1979). The latter character, as already discussed, has low taxonomic significance. The taxon *violetae* is considered in the current literature to be either a valid species (Kudrna, 2002; Gil-T. & Gil-Uceda, 2005; Lafranchis *et al.*, 2007; Gil-T., 2008), a subspecies of *A. fabressei* (Munguira *et al.*, 1995; Eckweiler & Häuser, 1997), a possible subspecies of *A. ripartii* (Tolman, 1997) or a taxon *incertae sedis* (Lukhtanov *et al.*, 2006).

The taxon *subbaeticus* was recently described from southern Spain as a subspecies of *A. fabressei* (Gil-T. & Gil-Uceda, 2005), and its presumed conspecific relationship with *A. fabressei* is supported by chromosomal data (Lukhtanov *et al.*, 2006). In the present study, we analyse for the first time the karyotype of *A. violetae* from the type locality, and show that it is similar to that of the karyotypes of *subbaeticus* and *fabressei*. Thus, from the point of view of karyology, the species status of *A. violetae* is not supported. Our phylogenetic analysis showed that *A. violetae* is unexpectedly quite distant from *Agrodiaetus fabressei fabressei*: these two species are not even sister taxa (Figs 1, 3). On the other hand, the taxa *violetae* and *subbaeticus* form a distinct, highly supported (99–100% bootstrap and BI support) monophyletic clade in all reconstructions. Importantly, the taxa *violetae* and *subbaeticus* have identical ITS2 sequences, and these are quite different from that of *A. fabressei fabressei*. Thus, both nuclear and mitochondrial sequences agree in the close relationship between *violetae* and *subbaeticus*. These results suggest that the taxon *violetae* is a separate species that includes at least two subspecies: *Agrodiaetus violetae violetae* and *Agrodiaetus violetae subbaeticus* **comb. nov.** The subspecific status of *subbaeticus* with respect to *A. violetae* from the type locality is based on morphological differences in the adults (intensity of wing underside spots and female background colour), and in the caterpillars (different colour of the lateral stripes) (Gil-T. & Gil-Uceda, 2005; Gil-T., 2008). These two taxa are allopatric and feed on different subspecies of *Onobrychis argentea* Boiss. (Lafranchis, Gil-T. & Lafranchis, 2007). The present study includes one specimen of a newly discovered, isolated population of *A. violetae*, which is located in a mountain approximately 100 km far from the type locality, and approximately 100 km far from *A. violetae subbaeticus* populations. This population is genetically closer to *A. violetae violetae* and its discovery and status will be described in a future publication (S. Ibáñez & F. Gil-T., unpubl. data). We thus conclude that *A. violetae* is a good local species whose distribution in the south of the Iberian Peninsula is not dot-like, but substantially wider than previously believed.

The taxa dolus *and* virgilia*:* On the basis of karyotype analysis, *A. dolus* consists of two populations with a minor but fixed chromosomal difference between them: the populations from France (*Agrodiaetus dolus dolus* and *Agrodiaetus dolus vittatus*) have $n = 123$–125, with a modal chromosome number of 124, and the populations from central Italy (*Agrodiaetus dolus virgilia*) have $n = 122$ (de Lesse, 1966). Usually, populations are considered to be conspecific. However, sometimes they are treated as separate species, as in *A. dolus* (*n* = 123–125) and *A. virgilia* (*n* = 122) (de Prins & Iversen, 1996; Dennis, 1997). *A. dolus vittatus* and *A. dolus virgilia* are recovered as sister taxa in our phylogenetic analysis. Although their genetic divergence is intermediate and larger than the *fabressei–fulgens* divergence, the fixed difference in one or two chromosome pairs seems at present insufficient to separate *virgilia* from *dolus* at the species level. Given our current knowledge of reproductive isolation between populations of Lepidoptera with variable karyotypes (Lukhtanov & Dantchenko, 2002b), we consider it more likely that these chromosomal forms are still interfertile.

The taxa ainsae *and* fulgens*:* These two taxa were already shown to be conspecific based on the lack of genetic or karyotypic differences, with similar morphology and ecology (Lukhtanov *et al.*, 2006). The taxon *ainsae* was then considered to be a subspecies of *fulgens* because of small morphological differences, including a higher percentage of specimens with a white band on the underside hindwing and slightly paler male uppersides. However, many new populations between the two type localities have been discovered, and it is difficult to draw a line that defines two subspecies. It appears that a cline exists involving intensity and prevalence of the characters mentioned, and that it probably extends to the west to include the taxa *pseudovirgilius* de Lesse, 1962 and *leonensis* Verhulst, 2004 (not studied here). We thus consider *ainsae* to be a junior subjective synonym of *fulgens*.

Agrodiaetus humedasae: This taxon was described from N. Italy (Toso & Balletto, 1976). Its karyotype was found to be *n* = 38 (Troiano, Balletto & Toso, 1979), which is different from that of other representatives of the *A. dolus* and *A. admetus* species groups. Therefore, *A. humedasae* has almost always considered a distinct species. The present study slightly modifies the chromosome number of *A. humedasae* to *n* = 39. In the molecular phylogeny, *A. humedasae* samples form a monophyletic and genetically well-differentiated clade, which is sister to *A. aroaniensis* from Greece (Fig. 3). Interestingly, *A. aroaniensis* also has a relatively low chromosome number (*n* = 48) (Coutsis, Puplesiene & De Prins, 1999). The fact that these two allopatric taxa are chromosomally distinct supports their status as separate species.

PHYLOGEOGRAPHY

A comparison of the distribution ranges of the *A. dolus* and *A. riparii* lineages reveals an interesting pattern (Fig. 5B, C). These two complexes are represented by two groups of geographical isolates with similar population distributions: each lineage has one isolate in the Balkan and Apennine Peninsulas, one isolate in the southern Alps, and from one to four isolates in the Iberian Peninsula. Such a pattern could be considered evidence for similar ecological preferences or parallel histories for these groups. The last assumption may be easily refuted: a comparison of branch lengths on the phylogenetic reconstructions as well as the dating of relevant nodes show that the isolates of these two groups are of different ages and are likely to have originated at different periods of the Pleistocene.

Analysis of distribution and phylogeny in the *A. dolus* lineage shows that the phylogeograpic history of this complex involved a combination of dispersal and vicariance events with a clear general trend of dispersal from the East (Iran), where the group most likely arose, to the West (western Europe) (Fig. 6): The first split, approximately 1.55 Mya (range 1.06–2.07 Mya; error interval covering 95% highest posterior density), was between the Iranian lineage and the rest; the second split, approximately 1.24 Mya (range 0.88–1.64 Mya), was between the Anatolian and European lineages. After this, the European lineage probably spread throughout southern Europe, and approximately 1.15 Mya (range 0.80–1.51 Mya), separated into three clades located in the Balkan Mountains and Alps, southern Spain, and the Iberian–Italian region, respectively. The relatively early separation between the main clades within the *A. dolus* group is in good agreement with their high level of karyotype divergence: the clade had time to develop different chromosome numbers from *n* = 39 in *A. humedasae* to *n* = 125 in *A. dolus*. However, it is interesting to note that the speciation of the taxa *dolus*, *fulgens*, and *fabressei* occurred as recently as 0.36 Mya (range 0.27–0.44 Mya). We specifically discuss the possible origins of these three species below.

Although the *A. ripartii* lineage also has a clear Asian origin, its phylogeographic history seems quite different, especially since it appears to have entered and dispersed in Europe more recently, approximately 0.76 Mya (range 0.53–0.99 Mya). Genetic distance (and correspondingly divergence age) is much lower between *A. ripartii* isolates (Fig. 7). The time of origin of the main *A. ripartii* lineages in Europe and NW Asia can be estimated as approximately 0.48 Mya (range 0.37–0.60 Mya). The alleles of the *COI* gene in the Spanish and Russian- northern Kazakhstani lineages show no lineage sorting, and samples from Spanish populations belong to different haplotype groups (e.g. MW01105 and MW01014; Fig. 3). This absence of lineage sorting can be explained not only by relatively recent origin of lineages, but also by introgression

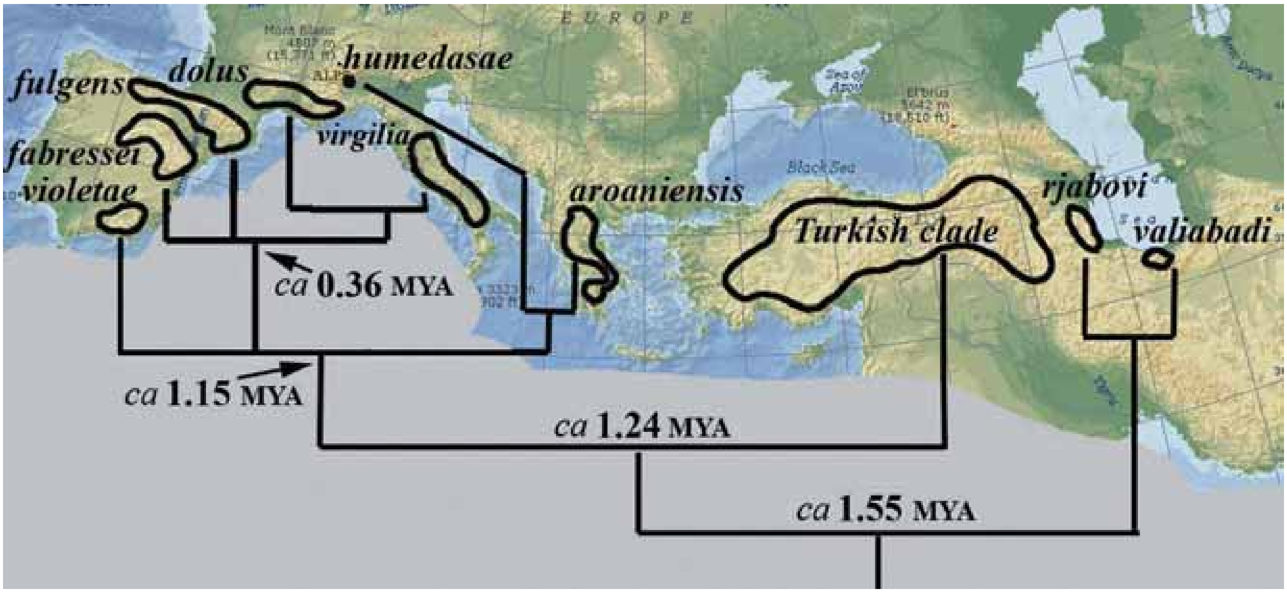


Figure 6. Biogeographical hypothesis describing the first split of the *Agrodiaetus dolus* lineage in the Iranian–Anatolian region, dispersal to Europe and diversification in southern Europe during the Pleistocene.

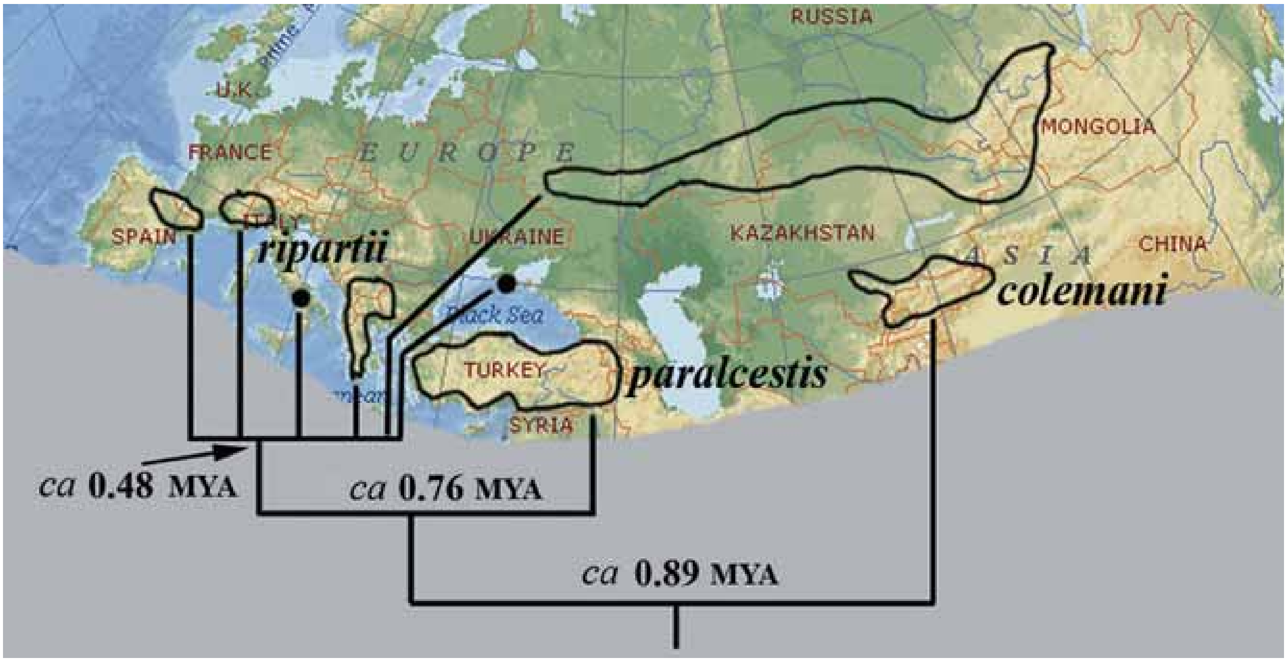


Figure 7. Biogeographical hypothesis describing the first splits of the *Agrodiaetus ripartii* lineage in Asia in the late Pleistocene, and dispersal to Europe and north-west Asia, followed by distribution range fragmentation.

events. Additional indirect evidence supporting the recent divergence hypothesis is the fact that all the clades of the European *A. ripartii* lineage are karyotypically undifferentiated. To conclude, it appears most likely that when *A. ripartii* reached Europe, the Balkan, Apennine and Iberian Peninsulas were already populated by representatives of the *A. dolus* group. Our taxonomic conclusions reflect this difference in biogeographic histories: the older *A. dolus* lineage is represented in western Europe by several species, whereas the younger *A. ripartii* lineage is represented by a group of conspecific populations.

DOT-LIKE DISTRIBUTION RANGES AND CONSERVATION

Our taxonomical revision based on chromosomal and molecular data supports the species status (and consequently the dot-like distribution) of *A. humedasae*. An earlier study likewise supported the dot-like distribution range for *A. pljushtchi* from Crimea (Fig. 5A) (Lukhtanov & Budashkin, 2007). By contrast, we were unable to confirm dot-like distributions for the rest of the studied taxa. The taxa *galloi*, *exuberans*, and *agenjoi* most likely represent local populations of a single species, *A. ripartii*. The same conclusion was earlier obtained (and supported here) for *A. budashkini*, which was described and considered a distinct species from Crimea, but in fact represents an isolated population of *A. ripartii* that is most closely related to the populations in European Russia (Fig. 5B) (Kandul *et al.*, 2004).

Current evidence also supports *A. violetae* as a fairly restricted, good species, but without a dot-like distribution because it consists of at least three groups of populations located in different mountains in the south of the Iberian Peninsula (Fig. 5C). A similar situation was found for *A. fulgens*, which was once considered a species with a very restricted distribution, but later shown to be conspecific with *A. ainsae* (Lukhtanov *et al.*, 2006). Thus, *A. fulgens* must be considered a species with a relatively broad distribution in northern Spain.

In conclusion, of the initial 11 potential cases of dot-like distributed *Agrodiaetus* species in Europe, six are not supported (*A. agenjoi*, *A. budashkini*, *A. exuberans*, *A. fulgens*, *A. galloi*, and *A. violetae*), two are supported (*A. humedasae* and *A. pljushtchi*), and three Balkan taxa remain to be analyzed (*A. nephohiptamenos*, *A. eleniae*, and *A. orphicus*).

Among the studied species, the taxa *A. violetae*, *A. galloi*, and *A. humedasae* are listed as species of conservation concern (Van Swaay *et al.*, 2010) because of their restricted distribution ranges. Two of them (*A. galloi* and *A. humedasae*) are also included in both the IUCN Red List of Threatened Species ((http://www.iucnredlist.org/apps/redlist/details/17939/0; http://www.iucnredlist.org/apps/redlist/details/17941/0) and in the Bern Convention on the Conservation of European Wildlife and Natural Habitats (http://conventions.coe.int/treaty/FR/Treaties/Html/104-2.htm). The results of the present study support the inclusion of *A. humedasae* on these lists. As for *A. galloi*, we show that this taxon is a population of the widely distributed *A. ripartii*, rather than a separate species. This population is geographically strongly isolated and may nevertheless be an important unit for conservation purposes. However, in the light of the data obtained, it is questionable whether it should be prioritized on protection lists above other endangered species.

The classical effect of incorrect taxonomy on conservation efforts is to underestimate the level of biological diversity and, as a consequence, to fail to recognize important conservation units in time (Duagherty *et al.*, 1990; DeSalle & Amato, 2004). By contrast, the present study illustrates a case of overestimation of biological diversity, leading to an inflated number of protected species. This has direct implications for conservation efforts because the protection of invalid species can result in inequitable spending of resources, which are always limited, and divert the attention of biologists and politicians away from species that require more urgent protection.

Species are important practical units in evolution, ecology and conservation, and a complete list of species existing in nature is a fundamental requirement of biodiversity-related studies and their application in all fields of biology. However, every species list contains uncertainties as a result of (1) the evolutionary nature of species, (2) subjectivity in species delimitation, and (3) imperfect taxonomy (Isaac, Mallet & Mace, 2004). The uncertainties of the first type depend on the continuous process of Darwinian evolution giving rise to intermediate forms, or incipient taxa that fail to meet unambiguous criteria for species delimitation (Descimon & Mallet, 2009). The uncertainties of the second type reflect the fact that species have been described and species lists have been created in different taxonomic cultures using different species concepts. These lists are particularly badly affected by extremes of ‘splitter’ or ‘lumper’ approaches (Isaac *et al.*, 2004). The first two types of uncertainties are inherent properties of species lists that can probably never be truly eliminated, although species lists can be made more useful if ambiguities are minimized. The third factor, imperfect taxonomy, should in theory be the easiest to uncover, although it frequently results in self-perpetuating error cascades in biological sciences and conservation efforts (Bortolus, 2008). Cases of imperfect taxonomy are unfortunately not rare, even among popular groups such as butterflies, and we advocate that lists of protected butterflies deserve careful revision with the use of modern techniques and consistently applied criteria for species recognition.

Taxonomic conclusion

We propose the following taxonomic arrangement of European representatives (west of the 17th meridian) of the *A. dolus* and *A. ripartii* lineages (chromosome numbers in parentheses when known):

A. dolus lineage:

A. dolus (Hübner, [1823])

ssp. *dolus* (Hübner, [1823]) (*n* = 123–125)
ssp. *vittatus* (Oberthür, 1892) (*n* = 124–125)
ssp. *virgilia* (Oberthür, 1910) (*n* = 122)
ssp. *gargano* (Wimmers, 1931) (*n* = 122) (not studied in this paper, probably a synonym of *virgilia*)
ssp. *paravirgilia* Verity, 1943 (*n* unknown) (not studied in this paper, probably a synonym of *virgilia*)
A. fulgens (Sagarra, 1925) (*n* = 108–110) (= *ainsae* Forster, 1961)
taxon *pseudovirgilius* de Lesse, 1962 (*n* = 108) (= *magnabrillata* Gómez-Bustillo, 1971) (not studied in the present study, probably a synonym of *fulgens*)
taxon *leonensis* Verhulst, 2004 (*n* unknown) (not included in the present study, probably a synonym of *fulgens*)
A. fabressei (Oberthür, 1910) (*n* = 90)
A. violetae Gómez-Bustillo *et al.*, 1979
ssp. *violetae* Gómez-Bustillo *et al.*, 1979 (*n* = 90)
ssp. *subbaeticus* Gil-T. & Gil-Uceda, 2005 (*n* = 90)
A. humedasae Toso & Balletto, 1976 (*n* = 39)

A. ripartii lineage:
A. ripartii Freyer, 1830
ssp. *ripartii* Freyer, 1830 (= *agenjoi* Forster, 1965; = *budashkini* Kolev & de Prins, 1995; = *exuberans* Verity, 1926; = *montanesa* Gómez-Bustillo, 1971; = *mozuelica* Agenjo, 1973; = *pelopi* Brown, 1976; = *ramonagenjo* Koçak & Kemal, 2001; = *rippertii* Boisduval, 1832; = *sarkani* Lukhtanov & Dantchenko, 2002; = *susae* Bertaccini, 2003) (*n* = 90)

ACKNOWLEDGEMENTS

We thank Alexander Dantchenko, Vlad Dinca, Rod Eastwood, Alan Heath, Nikolai Kandul, Mark Travassos, and Santi Viader for general discussion and/or help in collecting material, and Benjamin Goldman-Huertas for assistance in the laboratory. We are also grateful to Jim Mallet for discussion over the years of problems of species delimitation and species uncertainty. Support for this research was provided by the Russian Foundation for Basic Research (grants RFFI 09-04-01234, 08-04-00295, and 08-04-00787), by grant NSH-3332.2010.4 (Leading Scientific Schools), and by the programmes of the Presidium of Russian Academy of Science 'Gene Pools and Genetic diversity' and 'Origin of biosphere and evolution of geo-biological systems' to V.A.L..; by NSF DEB-0447242 to NEP; and by the Spanish Ministerio de Ciencia e Innovación (CGL2007-60516/BOS) to R.V. and G.T. The collecting trip to Italy, France, and Spain in 2007 was supported by a collecting grant from the Putnam Expeditionary Fund of the Museum of Comparative Zoology, Harvard University.

REFERENCES

Baletto E, Toso GG. 1979. On a new species of *Agrodiaetus* (Lycaenidae) from southern Italy. *Nota Lepidopterologica* 2: 13–22.

Bertaccini E. 2003. Prima segnalazione in piemonte di *Agrodiaetus ripartii* (Freyer, 1831) e descrizione di *A. ripartii susae* ssp. nova (Insecta Lepidoptera Lycaenidae). *Quaderno di Studi e Notizie di Storia Naturale della Romagna* **17** (Suppl.): 127–138.

Boisduval J. 1832 [1832–1834]. Icones historique des Lépidoptères nouveaux ou peu connus. 1. Rhopalocères. Paris. 1–251. Plates 1–47.

Bortolus A. 2008. Error cascades in the biological sciences: the unwanted consequences of using bad taxonomy in ecology. *Ambio* **37**: 114–118.

Brower AVZ. 1994. Rapid morphological radiation and convergence among races of the butterfly *Heliconius erato* inferred from patterns of mitochondrial DNA evolution. *Proceedings of the National Academy of Sciences of the United States of America* **91**: 6491–6495.

Bulatova NS, Searle JB, Nadjafova RS, Pavlova SV, Bystrakova NV. 2009. Field protocols for the genomic era. *Comparative Cytogenetics* **3**: 57–62.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution* **17**: 540–552.

Coutsis JG, Puplesiene J, De Prins W. 1999. The chromosome number and karyotype of *Polyommatus* (*Agrodiaetus*) *ripartii* and *Polyommatus* (*Agrodiaetus*) *aroaniensis* from Greece (Lepidoptera: Lycaenidae). *Phegea* **27**: 81–84.

Coyne JA, Orr AH. 2004. *Speciation.* Sunderland, MA: Sinauer.

De Prins W, Iversen F. 1996. Family Lycaenidae. In: Karsholt O, Razowski J, eds. *The lepidoptera of Europe: a distributional checklist.* Stenstrup: Apollo Books, 205–209.

Dennis RLH. 1997. An inflated conservation load for European butterflies: increases in rarity and endemism accompany increases in species richness. *Journal of Insect Conservation* **1**: 43–62.

Dennis RLH, Dapporto L, Shreeve TG, John E, Coutsis JG, Kudrna O, Saarinen K, Ryrholm N, Williams WR. 2008. Butterflies of European islands: the implications of the geography and ecology of rarity and endemicity for conservation. *Journal of Insect Conservation* **12**: 205–236.

DeSalle R, Amato G. 2004. The expansion of conservation genetics. *Nature Reviews Genetics* **5**: 702–712.

Descimon H, Mallet J. 2009. Bad species. In: Settele J, Konvicka M, Shreeve T, Dennis R, Van Dyck H, eds. *Ecology of butterflies in Europe.* Cambridge: Cambridge University Press.

Dobigny G, Aniskin V, Granjon L, Cornette R, Volobouev V. 2005. Recent radiation in West African *Taterillus* (Rodentia, Gerbillinae): the concerted role of chromosome and climatic changes. *Heredity* **95**: 358–368.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* **7**: 214.

Duagherty CH, Cree A, Hay JM, Thompson MB. 1990. Neglected taxonomy and continuing extinctions of tuatara (*Sphenodon*). *Nature* **347**: 177–179.

Eckweiler W, Häuser C. 1997. An illustrated checklist of *Agrodiaetus* Hübner, a subgenus of *Polyommatus* Latreille, 1804 (Lepidoptera: Lycaenidae). *Nachrichten des Entomologischen Vereins Apollo, Suppl.* **16**: 113–168.

Farris JS, Källersjö M, Kluge AG, Bult C. 1994. Testing significance of congruence. *Cladistics* **10**: 315–319.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**: 783–791.

Ferree PM, Barbash DA. 2009. Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in *Drosophila*. *PLoS Biology* **7**: e1000234. doi:10.1371/journal.pbio.1000234.

Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek RC. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology* **3**: 294–299.

Forster W. 1956–1961. Bausteine zur Kenntnis der Gattung *Agrodiaetus* Scudd. (Lep. Lycaen). *Zeitschrift der Wiener Entomologischen Gesellschaft* **41**: 42–61, 70–89, 118–127; **45**: 105–142; **46**: 8–13, 38–47, 74–94, 110–116.

Forster W. 1965. *Agrodiaetus admetus agenjoi* ssp. nov. *Entomologische Zeitschrift. Frankfurt am Main* **75**: 198–199.

Freyer CF. 1830. *Beiträge zur Geschichte europäischer Schmetterlinge mit Abbildungen nach der Natur*. Nürnberg.

García-Barros E, Munguira ML, Martín Cano JM, Romo Benito HR, Garcia-Pereira P, Maravalhas ES. 2004. *Atlas de las mariposas diurnas de la Península Ibérica e islas Baleares (Lepidoptera: Papilionoidea & Hesperioidea). Monografías de la Sociedad Entomológica Aragonesa*, Vol. 11. Zaragoza, Spain: SEA, UAM & MEC.

Gaston KJ. 1994. *Rarity*. London: Chapman & Hall.

Gil-T. F. 2008. Description of the pre-imaginal stages of *Agrodiaetus violetae* (Gómez-Bustillo, Expósito & Martínez, 1979) and notes about compared ecology and morphology (Lepidoptera: Lycaenidae). *Atalanta* **39**: 343–346, 422–423.

Gil-T. F, Gil-Uceda T. 2005. *Agrodiaetus violetae* (Gómez-Bustillo, Expósito & Martínez, 1979): Morfología comparada y descripción de Agrodiaetus fabressei subbaeticus ssp. nov. del sureste de la Península Ibérica (Lepidoptera, Lycaenidae). *Boletín Sociedad Entomológica Aragonesa* **36**: 357–364.

Gómez-Bustillo MR, Expósita Hermosa A, Martínez Borrego P. 1979. Una nueva especie para la Ciencia: *Agrodiaetus violetae* (Lep. Lycaenidae). *SHILAP Revista de Lepidopterologia* 7: 47–54.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* 52: 696–704.

Hasegawa M, Kishino H, Yano TA. 1985. Dating of the human ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* 22: 160–174.

Häuser CL. 1987. The debate about the biological species concept – a review. *Zeitschrift fur zoologische Systematik und Evolutionsforschung* 25: 241–257.

Hesselbarth G, Oorchot H, Wagener S. 1995. *Die Tagfalter der Türkei unter Berücksichtigung der angrenzenden Länder*. Bocholt: Selbstverlag Siegbert Wagener.

Higgins LG, Riley ND. 1970. *A field guide to the butterflies of Britain and Europe*. London: Collins Publishers.

Huelsenbeck JP, Crandall KA. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. *Annual Review of Ecology and Systematics* 28: 437–466.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**: 754–755.

Isaac NJB, Mallet J, Mace GM. 2004. Taxonomic inflation: its influence on macroecology and conservation. *Trends in Ecology and Evolution* **19**: 464–469.

Kandul NP, Lukhtanov VA. 1997. Karyotype variability and systematics of blue butterflies of the species groups *Polyommatus* (*Agrodiaetus*) *poseidon* and *Polyommatus* (*Agrodiaetus*) *dama* (Lepidoptera, Lycaenidae). *Zoologicheskii Zhurnal* **76**: 63–69.

Kandul NP, Lukhtanov VA, Dantchenko AV, Coleman J, Haig D, Sekercioglu C, Pierce NE. 2002. The evolution of karyotype diversity: a molecular phylogeny of *Agrodiaetus* Hübner, 1822 (Lepidoptera: Lycaenidae) inferred from mtDNA sequences for COI and COII. *4th International Conference on the Biology of Butterflies*, Leeuwenhorst, The Netherlands, 33–34.

Kandul NP, Lukhtanov VA, Dantchenko AV, Coleman JWS, Sekercioglu CH, Haig D, Pierce NE. 2004. Phylogeny of *Agrodiaetus* Hübner, 1822 (Lepidoptera: Lycaenidae) inferred from mtDNA sequences of COI and COII and nuclear sequences of EF1-a: karyotype diversification and species radiation. *Systematic Biology* **53**: 278–298.

Kandul NP, Lukhtanov VA, Pierce NE. 2007. Karyotypic diversity and speciation in *Agrodiaetus* butterflies. *Evolution* **61**: 546–559.

King M. 1993. *Species evolution: the role of chromosomal change*. Cambridge: Cambridge University Press.

Kolev Z, De Prins W. 1995. A new species of the 'brown *Agrodiaetus*' complex from the Crimea (Lepidoptera: Lycaenidae). *Phegea* 23: 119–132.

Kudrna O. 2002. The distribution atlas of European butterflies. *Oedippus* **20**: 1–342.

Lafranchis T. 2004. *Butterflies of Europe*. Paris: Diatheo.

Lafranchis T, Gil-T. F, Lafranchis A. 2007. New data on the ecology of 8 taxa of *Agrodiaetus* Hübner, 1822 from Greece and Spain: hostplants, associated ants and parasitoids (Lepidoptera: Lycaenidae. Hymenoptera. Diptera). *Atalanta* 38: 189–197, 313.

de Lesse H. 1960a. Spéciation et variation chromosomique chez les Lépidoptères Rhopalocères. *Annales des Sciences Naturelles (Series 12)* 2: 1–223.

de Lesse H. 1960b. Les nombres de chromosomes dans la classification du groupe *d'Agrodiaetus ripartii* Freyer (Lepidoptera, Lycaenidae). *Revue française d'Entomologie* 27: 240–264.

de Lesse H. 1962. Cohabitation en Espagne *d'Agrodiaetus ripartii* Freyer et *A. fabressei* Oberthür (Lepidoptera, Lycaenidae). *Revue française d'Entomologie* 28: 50–53.

de Lesse H. 1966. Variation chromosomique chez *Agrodiaetus dolus* Hübner (Lep., Lycaenidae). *Annales de la Société Entomologique de France* 2: 209–214.

de Lesse H. 1968. *Agrodiaetus ripartii* Frey. dans la region de Barcelone (Lycaenidae). *Alexanor* 5: 203–205.

Lorkovic Z. 1990. The butterfly chromosome and their application in systematics and phylogeny. In: Kudrna O, ed. *Butterflies of Europe*, Vol. 2. Wiesbaden: Aula-Verlag, 332–396.

Lukhtanov VA, Budashkin YI. 2007. The origin and taxonomic position of the crimean endemic *Agrodiaetus pljushtchi* (Lepidoptera, Lycaenidae) based on the data on karyology, ecology, and molecular phylogenetics. *Zoologicheskii Zhurnal* 86: 839–845.

Lukhtanov VA, Dantchenko AV. 2002a. Principles of highly ordered metaphase I bivalent arrangement in spermatocytes of *Agrodiaetus* (Lepidoptera). *Chromosome Research* **10**: 5–20.

Lukhtanov VA, Dantchenko AV. 2002b. Descriptions of new taxa of the genus *Agrodiaetus* Hübner, [1822] based on karyotype investigation (Lepidoptera, Lycaenidae). *Atalanta* 33: 81–107, 224–225.

Lukhtanov VA, Kandul NP, Plotkin JB, Dantchenko AV, Haig D, Pierce NE. 2005. Reinforcement of pre-zygotic isolation and karyotype evolution in *Agrodiaetus* butterflies. *Nature* **436**: 385–389.

Lukhtanov VA, Shapoval NA. 2008. Detection of cryptic species in sympatry using population analysis of unlinked genetic markers: a study of the *Agrodiaetus kendevani* species complex (Lepidoptera: Lycaenidae). *Doklady Biological Sciences* 423: 432–436.

Lukhtanov VA, Shapoval NA, Dantchenko AV. 2008. *Agrodiaetus shahkuhensis* sp. n. (Lepidoptera, Lycaenidae), a cryptic species from Iran discovered by using molecular and chromosomal markers. *Comparative Cytogenetics* 2: 99–114.

Lukhtanov VA, Sourakov A, Zakharov EV, Hebert PDN. 2009. DNA barcoding Central Asian butterflies: increasing geographical dimension does not significantly reduce the success of species identification. *Molecular Ecology Resources* **9**: 1302–1310.

Lukhtanov VA, Vila R, Kandul NP. 2006. Rearrangement of the *Agrodiaetus dolus* species group (Lepidoptera, Lycaenidae) using a new cytological approach and molecular data. *Insect Systematics and Evolution* **37**: 325–334.

Lukhtanov VA, Wiemers M, Meusemann K. 2003. Description of a new species of the 'brown' *Agrodiaetus* complex from south-east Turkey (Lycaenidae). *Nota Lepidopterologica* 26: 65–71.

Mallet J. 2001. Species, concepts of. In: Levin S, ed. *Encyclopedia of biodiversity*, Vol. 5. New York, NY: Academic Press, 427–440.

Mallet J. 2006. Species concepts. In: Fox CW, Wolf JB, eds. *Evolutionary genetics: concepts and case studies*. Oxford: Oxford University Press, 367–373.

Mallet J, Willmott K. 2003. Taxonomy: renaissance or Tower of Babel? *Trends in Ecology and Evolution* 18: 57–59.

Manley WBL, Allcard HG. 1970. *A field guide to the butterflies and burnets of Spain*. Hampton: Classey.

Mayr E. 1963. *Animal species and evolution*. Cambridge, MA: Harvard University Press.

Mazzei P, Morel D, Panfili R, Pimpinelli I, Reggianti D. 2009. *Moths and butterflies of Europe and North Africa*. Available at http://www.leps.it/

Mensi P, Lattes A, Cassulo L, Balletto E. 1994. Biochemical taxonomy and evolutionary relationships in *Polyommatus* (subgenus *Agrodiaetus*) (Lepidoptera, Lycaenidae). *Nota Lepidopterologica. Supplement* 5: 105–114.

Monteiro A, Pierce NE. 2001. Phylogeny of *Bicyclus* (Lepidoptera: Nymphalidae) inferred from COI, COII, and EF -1alpha gene sequences. *Molecular Phylogenetics and Evolution* 18: 264–281.

Munguira ML, Martín J, Pérez-Valiente M. 1995. Karyology and distribution as tools in the taxonomy of Iberian *Agrodiaetus* butterflies (Lepidoptera: Lycaenidae). *Nota Lepidopterologica* **17**: 125–140.

Nagaraju J, Jolly MS. 1986. Interspecific hybrids of *Antheraea roylei* and *A. pernyi* – a cytogenetic reassessment. *Theoretical and Applied Genetics* 72: 269–273.

Posada D, Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**: 817–818.

Poulton EB. 1904. What is a species? *Procedings of the Entomological Society of London* **1903**: lxxvii–lxcxvi.

Przybylowicz L. 2000. Polish butterflies of the subgenus *Polyommatus* (*Agrodiaetus*) (Lepidoptera: Lycaenidae). *Polskie Pismo Entomologiczne* **69**: 329–334.

Quek SP, Davies SJ, Itino T, Pierce NE. 2004. Codiversification in an ant-plant mutualism: Stem texture and the evolution of host use in *Crematogaster* (Formicidae: Myrmicinae) inhabitants of *Macaranga* (Euphorbiaceae). *Evolution* 58: 554–570.

Simon C, Frati F, Beckenbach A, Crespi B, Liu H, Flook P. 1994. Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. *Annals of the Entomological Society of America* 87: 651–701.

Swofford DL. 2000. *PAUP*. Phylogenetic analysis using parsimony (*and other methods)*. Sunderland, MA: Sinauer Associates.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* **24**: 1596–1599.

Tolman T. 1997. *Butterflies of Britain & Europe*. London: Harper Collins Publishers.

Toso GG, Balletto E. 1976. Una nuova specie del *Agrodiaetus* Hübn. (Lepidoptera: Lycaenidae). *Annali del Museo Civico di Storia Naturale Giacomo Doria* 81: 124–130.

Troiano G, Balletto E, Toso GG. 1979. The karyotype of *Agrodiaetus humedasae* Toso & Balletto, 1976. *Bollettino della Societa Entomologica Italiana* **111**: 141–143.

Troiano G, Giribaldi MA. 1979. Karyotypic analysis. *Nota lepidopterologica* 2: 22–23.

Van Swaay C, Cuttelod A, Collins S, Maes D, López Munguira M, Šašić M, Settele J, Verovnik R, Verstrael T, Warren M, Wiemers M, Wynhoff I. 2010. *European red list of butterfies*. Luxembourg: Publications Office of the European Union.

Verity R. 1926. Zygaenae, Grypocera and Rhopalocera of the Cottian Alps compared with other races. *Entomologist's Record and Journal of Variation* 38: 101–106, 120–126, 170–176.

White TJ, Bruns S, Lee S, Taylor J. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis MA, Gelfandm DH, Snisky JJ, White TJ, eds. *PCR protocols: a guide to methods and applications*. New York, NY: Academic Press, 315–322.

Wiemers M. 2003. *Chromosome differentiation and the radiation of the butterfly subgenus Agrodiaetus (Lepidoptera: Lycaenidae: Polyommatus) – a molecular phylogenetic approach*. PhD thesis. University Bonn. 143 p. Available at http://hss.ulb.uni-bonn.de:90/2003/0278/0278-1.pdf

Wiemers M, Fiedler K. 2007. Does the DNA barcoding gap exist? A case study in blue butterflies (Lepidoptera: Lycaenidae). *Frontiers in Zoology* **4**: 8.

Wiemers M, Keller A, Wolf M. 2009. *ITS2* secondary structure improves phylogeny estimation in a radiation of blue butterflies of the subgenus *Agrodiaetus* (Lepidoptera: Lycaenidae: *Polyommatus*). *BMC Evolutionary Biology* **9**: 300–327.

Chapter VI

Talavera, G., Lukhtanov, V., Pierce, N.E., Vila, R. In the shadow of phylogenetic uncertainty: the recent diversification of the *Lysandra* butterflies through chromosomal changes. *In prep.*

In the shadow of phylogenetic uncertainty: the recent diversification of the *Lysandra* butterflies through chromosomal changes

Gerard Talavera[a,b], Vladimir A. Lukhtanov[c,d], Naomi E. Pierce[e] and Roger Vila[a,*]

[a]Institut de Biologia Evolutiva (CSIC-UPF), Passeig Marítim de la Barceloneta, 37, 08003 Barcelona, Spain

[b]Departament de Genètica i Microbiologia, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain

[c]Department of Karyosystematics, Zoological Institute of Russian Academy of Science, Universitetskaya nab. 1, 199034 St Petersburg, Russia

[d]Department of Entomology, St Petersburg State University, Universitetskaya nab. 7/9, 199034 St Petersburg, Russia

[e]Department of Organismic and Evolutionary Biology and Museum of Comparative Zoology, Harvard University, 26 Oxford Street, Cambridge, Massachusetts 02138, USA

[*]Corresponding author: roger.vila@csic.es

ABSTRACT

The controversial taxonomy of the butterfly genus *Lysandra* (Lycaenidae, Polyommatinae) defies both molecular and morphological approaches, and could be related to speciation linked to karyotype instability. Here we aim at reconstructing the *Lysandra* species tree based on seven independent genetic markers and using multi-locus coalescent-based methods. While the genus is substantially old, the diversification of the extant lineages was extremely recent (ca. 1.45 Mya) and involved multiple chromosomal rearrangements. We find that the system is generally blurred by uncertainty due to both incomplete lineage sorting and hybridization. We minimize the impact of reticulation in the species tree inference by testing for mitochondrial introgression events. As a result we obtain a partially resolved tree with three main supported clades: *L. punctifera* + *L. bellargus*, the *corydonius* taxa, and *L. coridon* + the Iberian taxa, plus three independent lineages with no apparently close relatives (*L. ossmar*, *L. syriaca* and *L. dezina*). Based on these results and new karyotypic data, we propose a rearrangement recognizing ten species within the genus. Finally, we hypothesize that chromosomal instability may have played a crucial role in the *Lysandra* species radiation. New chromosome rearrangements might be fixed in populations after severe bottlenecks, which at the same time might promote rapid sorting of neutral molecular markers. Thus population bottlenecks might be a prerequisite for chromosomally unstable lineages to speciate.

INTRODUCTION

Main karyotypic features of organisms, such as the number of chromosomes, are usually stable within species (White, 1973; King, 1993). This stability is in good correspondence with the fact that new chromosomal rearrangements usually originate as heterozygotes and are often - although not always - associated with heterozygote disadvantage. Therefore, their spread to fixation within a large population has low probability (King, 1993). Lepidoptera (butterflies and moths) are good example of this stability, and a modal haploid number of chromosomes (n) of n=31 or n=30 is preserved in a great majority of Lepidoptera families (Robinson, 1971; Stekolnikov et al, 2000). The blue butterflies (the family

Lycaenidae) represent an exception because most species display an haploid chromosome number of either 23 or 24 (de Lesse, 1960; Lorkovic, 1990). However, chromosomal instability resulting in a great range of derived chromosome numbers has independently arisen several times in Lepidoptera as a whole (Lorkovic, 1941; Robinson, 1971; Dincă *et al*, 2011; Lukhtanov *et al*, 2011; Brown *et al*, 2012), and at least three times within the Lycaenidae (Kandul *et al*, 2004).

The genus *Lysandra* is one of the lineages within the Polyommatina (Lycaenidae) displaying fast changes in chromosome number. *Lysandra* is exclusively Palaearctic and it has two main centres of biodiversity in the Iberian Peninsula and the Middle East. The precise number of species is unknown due to poor morphological differentiation and *Lysandra* is sometimes cited as an example of difficult taxonomic resolution (De Bast 1985; Mensi *et al*, 1988; Schurian, 1989; Lelièvre, 1992; Wiemers, 2003; Descimon and Mallet, 2009). For example, the specific status of the taxa *caelestissima, gennargenti*, and *nufrellensis*, within the *coridon* group, or the taxa *arzanovi*, *sheikh* and *melamarina*, within the *corydonius* group, are unclear. No systematic revision for the genus has been based on phylogenetic relationships, and current classifications usually rely on chromosome number, number of annual generations and male wing colour (Schurian, 1989). Similarly to the related *Polyommatus* subgenera *Agrodiaetus* and *Plebicula*, it displays an extreme interspecific chromosome number variability, between n=24 and n=93 (de Lesse, 1969; Coutsis *et al*, 2001). As a perceptible difference with the *Agrodiaetus* and *Plebicula*, *Lysandra* karyotypes often display a number close to twice as large as other taxa, a particularity that has lead to the hypothesis of sequential polyploidy events (Lorkovic 1941, 1949; Robinson, 1971). It is also remarkable the particular case of *Lysandra coridon*, that presents intraspecific chromosome number variability in a cline across Europe, with numbers apparently fixed in each population (de Lesse, 1969).

Understanding recent speciation history requires merging phylogenetic and population genetics approaches, taking into account either the persistence of ancestral polymorphisms and possible hybridization traces. Non tree-like evolution is strongly related to the coalescent process, where gene discordance is common among closely related species. Hybridization between *Lysandra* species seems to be common in nature. In fact, potential hybrid specimens have been reported between *L. bellargus* and *L. coridon*, between *L. coridon* and other Iberian taxa, and between *L. corydonius* and *L. ossmar* (Schurian, 1989; Lelièvre, 1992; Hesselbarth *et al*, 1995; Gil-T, 2007; Descimon and Mallet, 2009). Establishing a link between gene genealogy and population or species divergence history requires the incorporation of the coalescence process, as well as the possibility of secondary exchanges after population splits. Distinguishing between these two major causes of conflicting signal across loci is of major importance, but notoriously difficult. Several methods to identify introgression events in a phylogenetic framework have been developed. While most of these methods either do not simultaneously account for the potential existence of incomplete lineage sorting (e.g. Bryant and Moulton, 2004; Jin *et al*, 2006; Gauthier and Lapointe, 2007), or do not distinguish the nature of the discordance (Ané *et al*, 2007), a few incorporate the coalescence of lineages while attempting to assess the possibility of gene introgression (Buckley *et al*, 2006; Joly *et al*, 2009; Kubatko, 2009). Although any genomic regions may be affected by introgression, literature reports that reticulate evolution induced by introgression in animals mostly concern mitochondrial DNA (mtDNA) (e.g. Ferris *et al*, 1983; Ruedi *et al*, 1997; Roca *et al*, 2005; Berthier *et al*, 2006; Melo-Ferreira *et al*, 2012), which results in strong conflicting phylogenetic signals between nuclear and mtDNA markers (e.g Buckley *et al*, 2006; Bossu and Near, 2009; Spinks and Shaffer, 2009).

Here we aim at reconstructing the *Lysandra* species tree based on seven genetic markers and using multi-locus coalescent-based methods. We infer divergence times and demographic history. We observe very low resolution in the selected markers, and generally discordant genealogies. Also, our results show that mitochondrial introgression within *Lysandra* is common and that not counteracting it can lead to wrong phylogenetic and taxonomic conclusions. By accounting for both introgression and incomplete lineage sorting, we obtain a partially resolved tree with three main supported clades and we discuss the role of chromosomal evolution in the *Lysandra* species radiation.

MATERIALS AND METHODS

Taxon sampling

We used 48 representatives of the *Lysandra* species-group covering its entire distribution and including several specimens corresponding to each potential species described. Only in the case of the rare *L. dezina* and *L. syriaca* we were unable to obtain more than a single specimen. The samples are stored in the DNA and Tissues Collection of the Museum of Comparative Zoology (Harvard University, Cambridge, MA, USA) and in the Butterfly Diversity and Evolution Lab (Institut de Biologia Evolutiva, Barcelona, Spain). Three outgroup taxa (*Polyommatus amandus*, *Polyommatus myrrha* and *Neolysandra diana*) were used for phylogenetic analyses, selected according to the general Polyommatina phylogeny of Talavera *et al* (2012). All specimens used in this study are listed in the Supplementary Table S1.

Molecular data

Genomic DNA was extracted from a leg or from a piece of the abdomen of each specimen using DNeasy^{TM} Tissue Kit (Qiagen Inc., Valencia, CA, USA) and following the manufacturer's protocols. Fragments from three mitochondrial genes (here treated as a single marker) –*cytochrome oxidase I (COI)* + *leu-tRNA* + *cytochrome oxidase II (COII)*; and from six nuclear markers – *28S ribosome unit (28S)*, *histone H3 (H3)*, *wingless (Wg)*, *carbamoyl-phosphate synthetase2 / aspartate transcarbamylase / dihydroorotase (CAD)*, *internal transcribed spacer 2 (ITS2)* and *ribosomal protein L5 (Rpl5)* were amplified by polymerase chain reaction and sequenced as described in Vila *et al* (2011). The primers employed are shown in the Supplementary Table S2. The sequences obtained were submitted to GenBank (accession numbers in Supplementary Table S3).

Karyotype analyses

Gonads were stored in Carnoy fixative (ethanol and glacial acetic acid, 3:1) for 2–6 months at 4 °C and then stained with 2% acetic orcein for 30 days at 20 °C. Cytogenetic analysis was conducted as previously described (Lukhtanov *et al*, 2005, 2008; Vershinina and Lukhtanov, 2010). In this study, we have counted the haploid chromosome numbers (n) in metaphase II of male meiosis and the number of bivalents in metaphase I of male meiosis. In total, preparations from 9 specimens and 5 taxa were analysed (Table 2).

Phylogenetic and species tree inference

A molecular matrix was generated for each independent marker by editing and aligning using Geneious 4.8.3 (Biomatters Ltd., 2009). Phylogenetic resolution was evaluated by performing single-gene, mitochondrial genes and nuclear genes phylogenies (Figure 1, Supplementary Figures S1-S6) using the maximum likelihood criterion with the software Phyml 3.0 (Guindon *et al*, 2010) and creating model partitions by marker for the concatenated datasets. jModeltest ver. 0.118 (Posada, 2008) was executed to select the best-fitting DNA substitution models for each marker dataset according to the Akaike information criterion (AIC). As a result, the GTR+I+G model was used for *COI+tRNA-leu+COII*, GTR+G for *ITS2*, GTR for *28S*, HKY+I for *H3* and *Rpl5* and TN+I for *Wg* and *CAD*.

A bayesian coalescent-based multilocus species tree approach was used to infer phylogenetic relationships among species. *BEAST 1.7.2 (Heled and Drummond, 2010) with a strict clock and a linear piecewise demographic model was set for a Markov chain Monte Carlo of 100 million generations sampled every 1000 runs. Two independent runs were performed and convergence was checked using Tracer v1.5. A substitution rate of 1.5% uncorrected pairwise distance per million years (Queck *et al*, 2004), was applied to the mitochondrial partition. Specimens were attributed to 13 different species (*L. albicans*, *L. caelestissima*, *L. hispana*, *L. coridon*, *L. bellargus*, *L. punctifera*, *L. melamarina*, *L. arzanovi*, *L. corydonius*, *L. sheikh*, *L. ossmar*, *L. dezina* and *L. syriaca*), and some combinations of them were also explored due to uncertainty about their taxonomic status within the group. The taxa *gennargenti*, *nufrellensis* and *philippi* were defined as a *L. coridon* specimens according to their position in the ML phylogenies. Population sizes were extracted from *BEAST species trees inference using the Python package Biopy (http://code.google.com/p/biopy/). Since a piecewise linear model was used in *BEAST, posterior population sizes were variable along branches, resulting in values for the beginning and the end of each lineage.

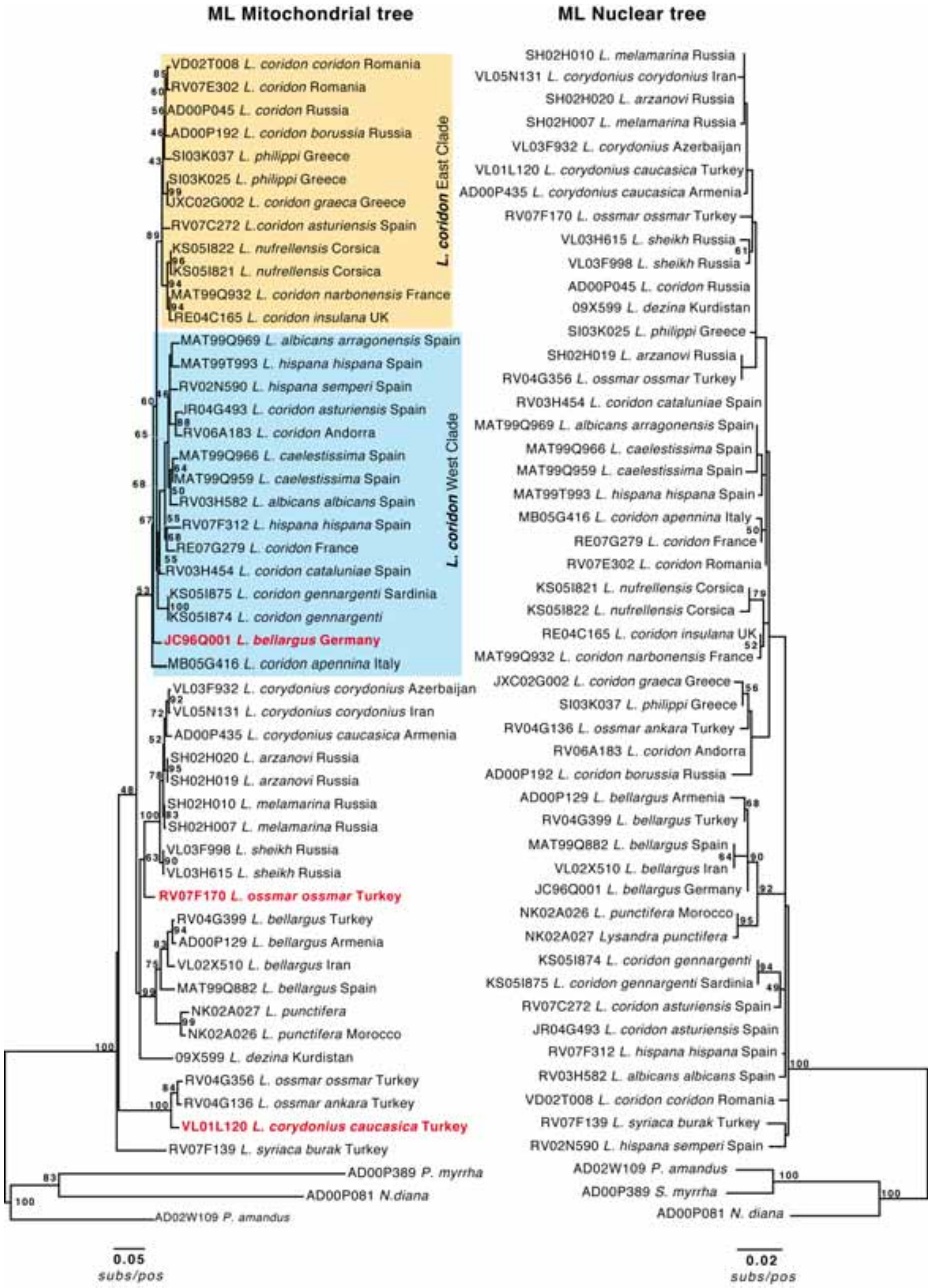

ML Mitochondrial tree
ML Nuclear tree
VD02T008 L. coridon coridon Romania
RV07E302 L. coridon Romania
AD00P045 L. coridon Russia
AD00P192 L. coridon borussia Russia
SI03K037 L. philippi Greece
SI03K025 L. philippi Greece
JXC02G002 L. coridon graeca Greece
RV07C272 L.coridon asturiensis Spain
KS05I822 L. nufrellensis Corsica
KS05I821 L. nufrellensis Corsica
MAT99Q932 L. coridon narbonensis France
RE04C165 L. coridon insulana UK
L. coridon East Clade
MAT99Q969 L. albicans arragonensis Spain
MAT99T993 L. hispana hispana Spain
RV02N590 L. hispana semperi Spain
JR04G493 L. coridon asturiensis Spain
RV06A183 L. coridon Andorra
MAT99Q966 L. caelestissima Spain
MAT99Q959 L. caelestissima Spain
RV03H582 L. albicans albicans Spain
RV07F312 L. hispana hispana Spain
RE07G279 L. coridon France
RV03H454 L. coridon cataluniae Spain
KS05I875 L. coridon gennargenti Sardinia
KS05I874 L. coridon gennargenti
JC96Q001 L. bellargus Germany
MB05G416 L. coridon apennina Italy
L. coridon West Clade
VL03F932 L. corydonius corydonius Azerbaijan
VL05N131 L. corydonius corydonius Iran
AD00P435 L. corydonius caucasica Armenia
SH02H020 L. arzanovi Russia
SH02H019 L. arzanovi Russia
SH02H010 L. melamarina Russia
SH02H007 L. melamarina Russia
VL03F998 L. sheikh Russia
VL03H615 L. sheikh Russia
RV07F170 L. ossmar ossmar Turkey
RV04G399 L. bellargus Turkey
AD00P129 L. bellargus Armenia
VL02X510 L. bellargus Iran
MAT99Q882 L. bellargus Spain
NK02A027 L. punctifera
NK02A026 L. punctifera Morocco
09X599 L. dezina Kurdistan
RV04G356 L. ossmar ossmar Turkey
RV04G136 L. ossmar ankara Turkey
VL01L120 L. corydonius caucasica Turkey
RV07F139 L. syriaca burak Turkey
AD00P389 P. myrrha
AD00P081 N.diana
AD02W109 P. amandus
89
60
56
46
43
99
89
96
94
94
60
46
65
88
58
64
50
67
55
68
55
53
100
92
72
52
95
78
83
100
48
90
63
94
82
71
99
99
100
84
100
83
100
0.05
subs/pos
SH02H010 L. melamarina Russia
VL05N131 L. corydonius corydonius Iran
SH02H020 L. arzanovi Russia
SH02H007 L. melamarina Russia
VL03F932 L. corydonius Azerbaijan
VL01L120 L. corydonius caucasica Turkey
AD00P435 L. corydonius caucasica Armenia
RV07F170 L. ossmar ossmar Turkey
VL03H615 L. sheikh Russia
VL03F998 L. sheikh Russia
AD00P045 L. coridon Russia
09X599 L. dezina Kurdistan
SI03K025 L. philippi Greece
SH02H019 L. arzanovi Russia
RV04G356 L. ossmar ossmar Turkey
RV03H454 L. coridon cataluniae Spain
MAT99Q969 L. albicans arragonensis Spain
MAT99Q966 L. caelestissima Spain
MAT99Q959 L. caelestissima Spain
MAT99T993 L. hispana hispana Spain
MB05G416 L. coridon apennina Italy
RE07G279 L. coridon France
RV07E302 L. coridon Romania
KS05I821 L. nufrellensis Corsica
KS05I822 L. nufrellensis Corsica
RE04C165 L. coridon insulana UK
MAT99Q932 L. coridon narbonensis France
JXC02G002 L. coridon graeca Greece
SI03K037 L. philippi Greece
RV04G136 L. ossmar ankara Turkey
RV06A183 L. coridon Andorra
AD00P192 L. coridon borussia Russia
AD00P129 L. bellargus Armenia
RV04G399 L. bellargus Turkey
MAT99Q882 L. bellargus Spain
VL02X510 L. bellargus Iran
JC96Q001 L. bellargus Germany
NK02A026 L. punctifera Morocco
NK02A027 Lysandra punctifera
KS05I874 L. coridon gennargenti
KS05I875 L. coridon gennargenti Sardinia
RV07C272 L. coridon asturiensis Spain
JR04G493 L. coridon asturiensis Spain
RV07F312 L. hispana hispana Spain
RV03H582 L. albicans albicans Spain
VD02T008 L. coridon coridon Romania
RV07F139 L. syriaca burak Turkey
RV02N590 L. hispana semperi Spain
AD02W109 P. amandus
AD00P389 S. myrrha
AD00P081 N. diana
61
50
79
52
56
68
64
90
92
95
94
49
100
100
100
0.02
subs/pos

Figure 1. Maximum likelihood phylogenetic trees inferred from mitochondrial and nuclear data independently. Mitochondrial sequences from specimens highlighted in red were removed to avoid introgression noise in species tree inference. Boostrap supports higher than 40% are shown at nodes. Scale bars represent substitutions/position.

Testing for hybridization events

Although *BEAST incorporates the uncertainty of the coalescent process in the estimate of the phylogeny, it assumes that no gene flow occurred after the initial split. To quantify the potential impact of horizontal gene flow, causes of discordance were investigated using coalescent approaches to test hybridization as an alternative explanation to incomplete lineage sorting. The method of Joly *et al* (2009) was used as implemented in the software JML (Joly, 2012). The program calculates the minimum distance between sequences of two species and test whether it is smaller than expected under a simulated scenario that includes incomplete lineage sorting but does not account for hybridization. Therefore, JML was used to performing 10000 simulations for each gene tree by using the Seq-Gen code (Rambaut and Grassly, 1997). JML was run independently for each marker and parameters were extracted from the Phyml output for gene genealogies, including nucleotide frequencies, proportion of invariant sites and gamma shape parameter when selected. The relative mutation rate mean of the species tree posterior distribution was used for each locus. A cutoff of 0.05 was applied and *BEAST inference was repeated after removing the sequences of the detected cases of hybridization from the dataset to finally obtain the most accurate species tree possible. To detect residual hybridization, a cutoff of 0.15 was also explored.

RESULTS

The resolution of the phylogenies inferred by standard phylogenetic methods was very poor, suggesting non-treelike evolutionary relationships among *Lysandra* species (Figure 1, Supplementary Figures S1-S6). Parsimony informative sites for each gene are shown in Supplementary Table S4. Maximum likelihood estimates for the individual nuclear genes generally show no tree structure, most likely because of having incompletely sorted alleles. Also, gene phylogenies showed generalized discordances among them, suggesting that the species share sequences at a high degree. Given such discordances among nuclear loci, phylogenetic inference of the species tree based on their concatenation would be prone to errors (Edwards, 2009). Thus we used the multi-species/multi-locus coalescent method implemented in *BEAST (Heled and Drummond, 2010) to estimate species trees from the distribution of single gene trees, co-estimating divergence times and the effective population sizes of tip and ancestral taxa. Initial results reported high levels of uncertainty as shown in Figure 2A. Only two clades recovered strong statistical support, corresponding to the Iberian and the Middle East species-groups. Basal relationships were especially blurred and no root could be recognized.

When testing for hybridization, 11 distances between mitochondrial alleles were smaller than the 5th quantile for the posterior predictive distributions of JML (Table 1) and 16 more if a permissive P-value < 0.15 was considered. No case of introgression involving nuclear genes was detected, excluding those between specimens that lacked specific sequences or contained large missing data on them, which may produce artifactual predictions in JML. The *L. bellargus* specimen JC96Q001 was pointed as the most conflicting because the mitochondrial sequence was unexpectedly similar to that of *L. coridon* and, with a smaller *P*-value, to that of other species. To a lesser extent *L. ossmar* (RV07F170), also presented small genetic distances from apparently distant taxa (*L. coridon* and *L. bellargus*). Although such distances were not statistically significant at the 5% level (i.e. incomplete lineage sorting cannot be significantly rejected), they were recurrent and small enough to cautiously consider this specimen as conflictive. Indeed, this specimen, similarly to the *L. bellargus* specimen JC96Q001, did not cluster with the rest of conspecifics in the mitochondrial tree (Figure 1). As a consequence, both mitochondrial sequences were removed from the final dataset.

Table 1. JML results. List of distances with *P*-values < 0.05 and *p* < 0.15 according to the posterior predictive distributions for mitochondrial sequences are shown. Distances including sequences with large missing data were not considered. Specimens in bold indicate the ones that were considered conflicting and we removed from the original dataset.

Individual 1	Individual 2	Obs. Distance	*P*-value
JML Testing Hybridization (p<0.05)			
JC96Q001 *L. bellargus*	JXC02G002 *L. coridon*	0.00739713	0.0055
JC96Q001 *L. bellargus*	SI03K025 *L. coridon*	0.00739713	0.0055
JC96Q001 *L. bellargus*	KS05I874 *L. coridon*	0.00785945	0.0076
JC96Q001 *L. bellargus*	KS05I875 *L. coridon*	0.00785945	0.0076
JC96Q001 *L. bellargus*	RV07E302 *L. coridon*	0.00832178	0.0122
JC96Q001 *L. bellargus*	KS05I821 *L. coridon*	0.0087841	0.0186
JC96Q001 *L. bellargus*	RV07C272 *L. coridon*	0.00924642	0.0268
JC96Q001 *L. bellargus*	VD02T008 *L. coridon*	0.00924642	0.0268
JC96Q001 *L. bellargus*	AD00P045 *L. coridon*	0.00924642	0.0268
JC96Q001 *L. bellargus*	RE07G279 *L. coridon*	0.00970874	0.0384
JC96Q001 *L. bellargus*	MAT99Q959 *L. caelestissima*	0.0106334	0.0496
JML Testing Hybridization (p<0.15)			
JC96Q001 *L. bellargus*	RE04C165 *L. coridon*	0.0106334	0.073
JC96Q001 *L. bellargus*	MAT99T993 *L. hispana*	0.011558	0.1008
JC96Q001 *L. bellargus*	SH02H019 *L. arzanovi*	0.0166436	0.1158
JC96Q001 *L. bellargus*	SH02H020 *L. arzanovi*	0.0166436	0.1158
JC96Q001 *L. bellargus*	VL03H615 *L. sheikh*	0.0166436	0.117
JC96Q001 *L. bellargus*	MAT99Q969 *L. albicans*	0.0120203	0.1186
RV07F170 *L. ossmar*	RV07E302 *L. coridon*	0.0157189	0.1242
RV07F170 *L. ossmar*	VL02X510 *L. bellargus*	0.0166436	0.1298
RV07F170 *L. ossmar*	RV04G399 *L. bellargus*	0.0166436	0.1298
JC96Q001 *L. bellargus*	SH02H010 *L. melamarina*	0.0171059	0.1337
RV07F170 *L. ossmar*	RV07C272 *L. coridon*	0.0161812	0.1416
RV07F170 *L. ossmar*	JXC02G002 *L. coridon*	0.0161812	0.1416
RV07F170 *L. ossmar*	SI03K025 *L. coridon*	0.0161812	0.1416
RV06A183 *L. coridon*	MAT99Q959 *L. caelestissima*	0.00554785	0.1418
RV07F170 *L. ossmar*	RV04G399 *L. bellargus*	0.0171059	0.1466
JC96Q001 *L. bellargus*	SH02H007 *L. melamarina*	0.0175682	0.1497

In addition to the two specimens suggested by JML, we also excluded the mitochondrial sequence of the *L. corydonius* specimen VL01L120. This sequence is almost identical to those of typical *L. ossmar* specimens, and indeed clusters with two *L. ossmar* specimens with high support in the mitochondrial tree (Figure 1). On the contrary, the nuclear sequences of this specimen are identical to those of other *L. corydonius*. *Lysandra ossmar* and *L. corydonius* are parapatric in East Turkey and they are known to locally hybridize (Hesselbarth *et al*, 1995). We carefully examined wing morphology and we did not detect any traces of hybridization. This result is highly suggestive of genetic introgression, although JML could not significantly discard the possibility of incomplete lineage sorting.

Indeed, hybridization can often be difficult to detect for JML when dealing with recently diverged species (Joly, 2012). Similarly, introgression might have occurred within the *coridon*-clade, but it may be too recent for JML to unambiguously pinpoint it.

Species trees retained high levels of uncertainty after removing potentially introgressed mitochondrial sequences as shown in Figure 2B, but relevant changes were observed mainly involving two phylogenetic clusters. First, the grouping *L. bellargus* + *L. punctifera* considerably increased its posterior probability (from 0.82 to 0.98), and second, *Lysandra ossmar* lost its position as sister taxa of the *corydonius* clade by dramatically decreasing posterior probability from 0.96 to 0.47, and was now recovered in an unresolved phylogenetic position. As a result, three main supported clades were obtained: 1) the *coridon* clade, 2) the *corydonius* clade and 3) *L. punctifera* + *L. bellargus* (Figure 2B).

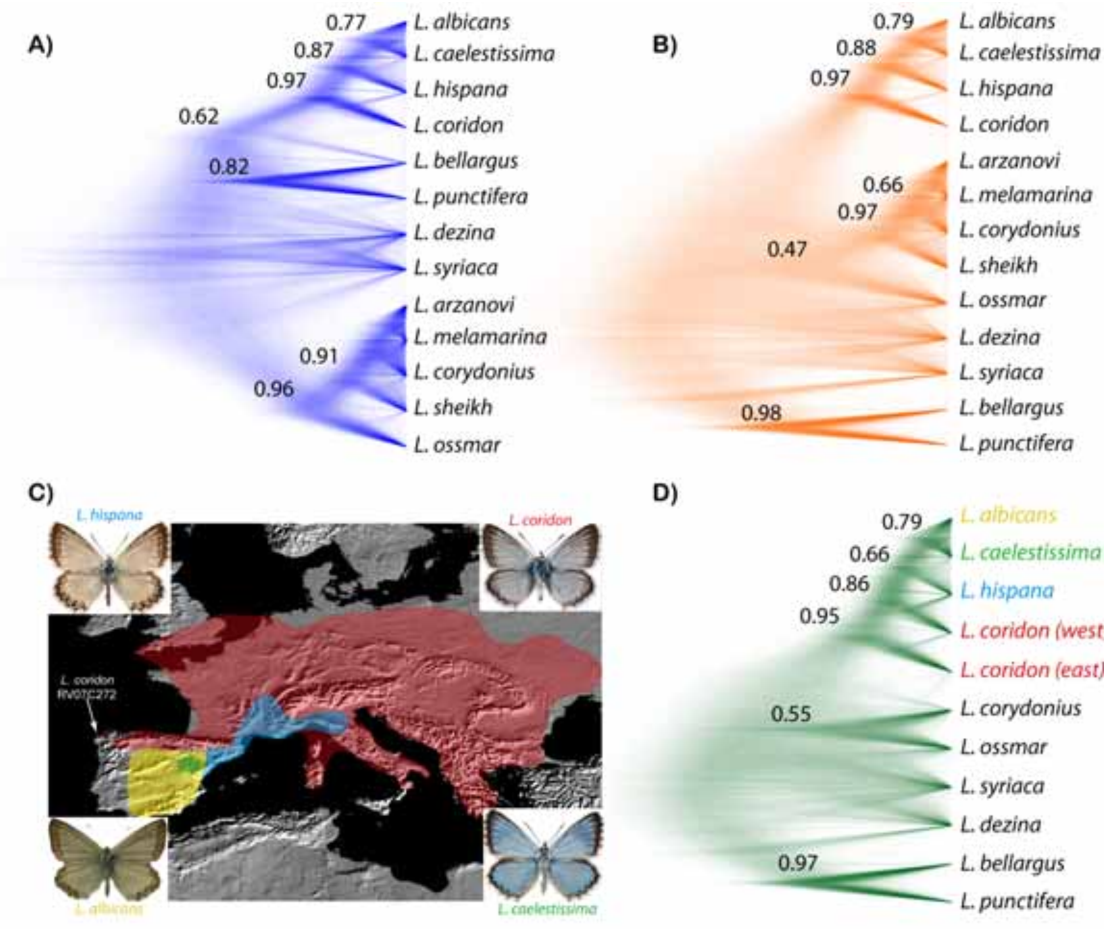


Figure 2. *BEAST species trees for *Lysandra* based on seven independent genetic markers, A) without accounting for hybridization, B) after removing apparently introgressed mitochondrial sequences, C) map of distributions for the Iberian taxa where *L. coridon* specimen RV07C272 is shown in the periphery of the distribution of any other Iberian *Lysandra* taxon and D) species tree after removing apparently introgressed mitochondrial sequences, considering two *L. coridon* groups, and considering the taxa of the *corydonius* group as conspecific. Trees are figured with DensiTree (Bouckaert, 2010), displaying a subsample of the Markov chain Monte Carlo of 10000 trees. Higher levels of certainty are represented by higher densities, and posterior probabilities > 0.50 are shown in the nodes.

Karyotype results

The karyotypes of western European and North African species of *Lysandra*, as well as two Asian taxa (*L. corydonius caucasica* and *L. syriaca syriaca*), were studied by de Lesse (1960). However, karyotypes of the representatives of the genus from South-East Turkey, Azerbaijan and South of Russia have not been studied so far except for one publication (Stradomsky and Shchurov, 2005) (but see our comments on this publication below). In our research we were able to study the karyotypes of the following taxa: *L. syriaca burak*, *L. melamarina*, *L. sheikh* and *L. corydonius corydonius*. We also analysed the karyotype of *L. bellargus* from the easternmost parts of its distribution range (Table 2, Figure 3).

Table 2. **Karyotype results**. Data and results of chromosomally studied material.

Taxon	Specimen Code	Chromosome number (n)	Country	Locality
L. bellargus	VL508	n=45	Iran	Gilan Prov. Masuleh (1900-2100m)
L. bellargus	VL510	n=ca45	Iran	Gilan Prov. Masuleh (1900-2100m)
L. bellargus	F938	n=45	Azerbaijan	Talysh, Zuvand, Mistan (1700-1800m)
L. bellargus	F941	n=45	Azerbaijan	Talysh, Zuvand, Mistan (1700-1800m)
L. corydonius corydonius	F932	n=84	Azerbaijan	Talysh, Zuvand, Mistan (1700-1800m)
L. corydonius melamarina	SH-2002-08	n=84	Russia	Krasnodar Region, Gelendjik, Betta Mts (150m)
L. corydonius sheikh	F998	n=84	Azerbaijan	East Caucasus, Altyagach (1300m)
L. corydonius sheikh	F999	n=84	Azerbaijan	East Caucasus, Altyagach (1300m)
L. syriaca burak	07-F139	n=30	Turkey	Adana, 13 km N. of Saimbeily

For the rare taxon *L. syriaca burak,* we were able to collect and analyse only a single male specimen. In a first meiotic division, 30 chromosome units were observed in most cells (Figure 3) and apparently all these units were bivalents. Therefore, we estimate the haploid chromosome number (n) of *L. syrica burak* as n=30. The chromosomal units vary in size gradually, without the existence of discrete size types. Thus, this taxon differs from *L. syriaca syriaca* (n=24; de Lesse, 1960) by at least 6 fixed chromosomal fissions.

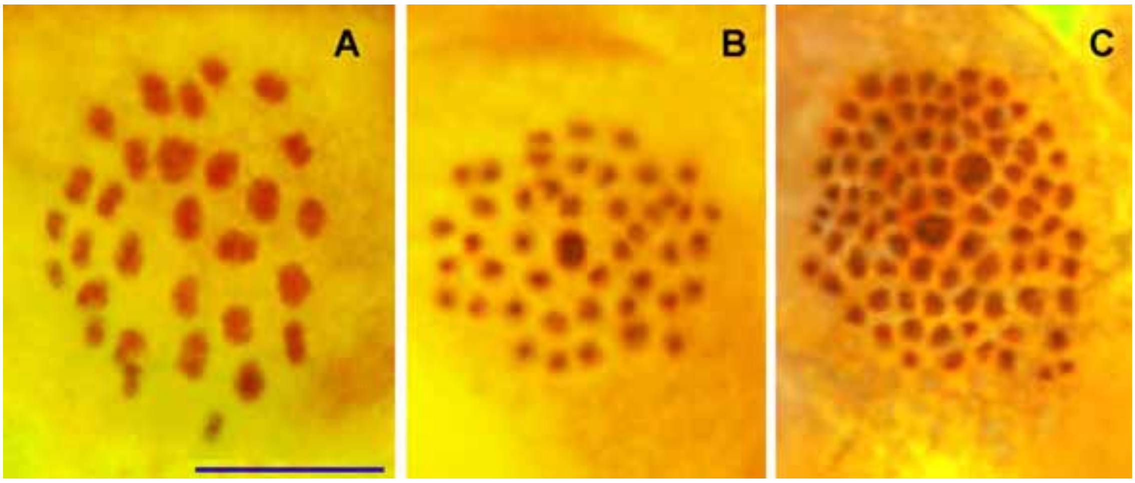


Figure 3. Meiotic karyotypes of *Lysandra* samples. A) Meiosis I metaphase plate of *L. syriaca burak* (specimen 07F139,Turkey, Adana) displays 30 bivalents, a new chromosome number for the genus (n=30); B) Meiosis II metaphase plate of *L. bellargus* (specimen F941, Azerbaijan, Talysh) displays 45 chromosomes (n=45). C) Meiosis I metaphase plate of *L. corydonius sheikh* (specimen F999, Azerbaijan, Altyagach) displays 84 bivalents (n=84). Scale bar is 10 µm.

For the taxa *L. melamarina, L. sheikh and L. corydonius corydonius,* similar karyotypes with n=84 were found (Table 2), including two large bivalents always observed in the centre of MI metaphase plates and numerous small bivalents. This result disagrees with previously assigned counts of n=24-27 for *L. melamarina* (Stradomsky and Shchurov, 2005), which are probably due to chromosome counts in atypical cell divisions that were not suitable for karyotype analysis. For the same reason we are also uncertain about their karyotype estimation of n=19-20 for *L. arzanovi* (Stradomsky and Shchurov, 2005), which we prudently consider unknown until further evidence is obtained.

Thus, the taxa *L. corydonius corydonius* (Azerbaijan), *L. melamarina* (*Russia*) and *L. sheikh* (Russia) (Figure 3C) are chromosomally undistinguishable from *L. corydonius caucasica* (East Turkey) (de Lesse, 1960), all displaying a karyotype with n=84, including two large chromosome units. Although the chromosome number is identical, *L. ossmar* (n=84) from West and Central Turkey differs from the *corydonius*-group in having three large chromosome units (de Lesse, 1960) instead of two.

In *L. bellargus* from Azerbaijan and Iran 45 chromosomes, including one large chromosome in centre of the metaphase plate, were observed (Figure 3B). Thus, these results confirm published data about the karyotype of *L. bellargus* (Lorkovic, 1941; de Lesse, 1960)

DISCUSSION

Phylogenetic relationships

Previous attempts to establish phylogenetic relationships or a taxonomic classification of the taxa within *Lysandra*, either based on morphological data, allozymes, or DNA sequences, all coincided in highlighting the difficulties that this genus entails (De Bast, 1985; Mensi *et al*, 1988; Schurian, 1989; Lelièvre, 1992; Wiemers, 2003). Based on morphology and karyotype we could consider three hypothetic species-groups within the genus: 1) *syriaca* (n=24), 2) *bellargus* (n=45), and 3) *coridon* (n=82-93). The attribution of the taxon *L. punctifera* to one of these groups is unclear under this hypothesis, since it is morphologically very close to *L. bellargus* and yet it displays the same chromosomal number as *L. syriaca* (n=24). The *coridon* superspecies may include (a) the Iberian taxa (*albicans*, *hispana*, *calestissima*), (b) *coridon* sensu stricto and the taxa *gennargenti*, *nufrellensis* and *philippi*, (c) the *corydonius* group (*corydonius*, *sheikh*, *melamarina*, *arzanovi*), (d) *ossmar*, and (e) *dezina*. These taxa might be also grouped because of sharing a very high chromosome number (n > 82, unknown in *L. dezina*). Our molecular results are unexpected in some regards and only partially support this hypothesis.

We recover three well-differentiated clades plus three species with apparently no close relative (*L. syriaca*, *L. dezina* and *L. ossmar*). One of the strongly supported clades is formed by *L. punctifera* and *L. bellargus*, a grouping that corresponds very well to morphology. In fact *L. punctifera* and *L. bellargus* are so similar in their wing patterns that the taxon *punctifera* was described (Oberthür, 1876) and long time considered as a subspecies of *L. bellargus*. Much later, de Lesse (1959) rose *punctifera* to species status based on the discovered difference in chromosome number. These two taxa split ca 0.74 Mya, probably because of dispersal across the West Mediterranean (Gibraltar Strait?), since *L. bellargus* is widespread in the Iberian Peninsula and across Europe into Western Asia, while *L. punctifera* is present in the Southwestern Mediterranean shore (Morocco, N. Algeria and NW. Tunisia). The removal of the conflicting signal created by the mitochondrial sequence of *L. bellargus* JC96Q001 substantially increases the posterior probability of this clade (*pp* from 0.82 to 0.98). This specimen was collected in Germany, where *L. bellargus* flies syntopically and synchronically with *L. coridon*. Indeed, the introgressed mitochondrial haplotype has already been shown to be very widespread in Romania, where nine out of ten sequenced Romanian *L. bellargus* specimens carried the introgression (Dincă *et al*, 2011).

Another supported clade includes all Iberian taxa (*L. albicans*, *L. coelestissima*, and *L. hispana*) + *coridon* sensu stricto group (including the taxa *gennargenti*, *nufrellensis*, and *philippi*) (Figure 2A-2B). Two differentiated clades are recovered in the mitochondrial tree (Figure 1). One, which we will call the eastern clade, is well supported and comprises only *L. coridon*, mainly from central and eastern Europe. In the other clade, the rest of *L. coridon* samples (from Iberian Peninsula and S. France) cluster together with the other three Iberian taxa. The support for this western clade is very low and internal relationships display a high degree of phylogenetic uncertainty. Notably, the four Iberian taxa and *L. coridon* are not recovered as monophyletic and introgressive hybridization is likely at play within this recent clade, despite JML could not significantly differentiate it from the uncertainty created by incomplete lineage sorting. As shown in Figure 2C, the *L. coridon* specimens from the western clade were usually collected in sympatry or proximity to some of the other three Iberian taxa. A macropopulation structure dividing *L. coridon* into an eastern and a western European form has been previously suggested (Schmitt and Seitz, 2001; Schmitt *et al*, 2002; Schmitt and

Zimmermann, 2011). While our results generally agree with the proposed distribution of these two forms, remarkable additions to the known phylogeography of the species are worth being discussed. First of all, a *L. coridon* specimen from Monte Pollino (Calabria, S. Italy) displays a highly diverged haplotype with unresolved position in the tree, and it could represent a relict lineage that has survived in this rather isolated locality. Similarly, the Sardinian taxon *gennargenti*, while apparently related to the eastern clade, is also substantially diverged, which suggests that it has remained isolated for a substantial amount of time. Rather surprisingly, the geographically close taxon *nufrellensis* from Corsica belongs instead a to the eastern clade and its mitochondrial sequence is close to that of one French *L. coridon narbonensis* specimen studied, among others. Thus, *nufrellensis* is apparently the outcome of a more recent colonization of Corsica from mainland, despite having some phenotypic differentiation that could be either the result of a founder effect, drift or adaptation to insular conditions. The other insular population studied is that from the UK, which also belongs to the eastern clade and is almost identical to the French *L. coridon narbonensis* specimen. Thus, we can conclude that *L. coridon* has apparently colonized Great Britain quite recently from mainland. Our results show that the clear-cut division proposed for the two *L. coridon* forms, with a contact zone in north-eastern Germany, along the mountain ranges of the German–Czech border and throughout the eastern Alps (Schmitt *et al*, 2012), is much more complex, at least for mitochondrial markers. Indeed, at least two specimens from the eastern clade were collected in surprisingly western locations: one was *L. coridon narbonensis* MAT99Q932 from Mende (Languedoc region, France). The other specimen is *L. coridon asturiensis* RV07C272 collected in the extreme northwestern Iberian Peninsula. Worth noting, this novel population, which represents the westernmost locality for *L. coridon*, occurs in an isolated small cape situated outside the area of influence of any other Iberian *Lysandra* species.

The two *L. coridon* clades most probably are the outcome of two different glacial refugia (the western clade in the Iberian Peninsula and the eastern clade in the Balkans), and postglacial dispersal created the current distribution (Schmitt *et al*, 2002). However, we propose that isolation during the last glaciation has not generated most of the intraspecific divergence detected in *L. coridon*, but that the four taxa in the Iberian refugium shared mitochondrial sequences in a high degree. The hypothesis of introgression between *L. albicans*, *L. coelestissima*, *L. hispana* and western *L. coridon* is further supported by the species tree treating separately the *L. coridon* eastern and western clades, which recovers them as paraphyletic lineages because the western clade is sister to the other three Iberian taxa (Figure 2C). While introgression within the Iberian Peninsula could not be significantly detected by JML, this hypothesis is more likely than solely incomplete lineage sorting being the cause. Indeed, the presence of the eastern form in the extreme northwest Iberian Peninsula can hardly be explained by recent long-range dispersal. This isolated and ecologically unique population, located outside the area of influence of the other Iberian taxa and of other *L. coridon* populations, is most probably a relict of "pure" Iberian *L. coridon* that was not introgressed, which would mean that the western clade is actually an effect of introgression.

Chromosomal fusion-fission events seem to be extremely common within *L. coridon*, but fixed within populations and displaying a longitudinal chromosome number cline across Europe, as de Lesse (1969) already pointed out. Thus, *L. coridon* seems to be the only *Lysandra* species with widespread instraspecific variability in chromosome number (but our results for *L. syriaca burak* might suggest a similar case). Neighbouring *L. coridon* chromosomal races typically differ in a single chromosomal reorganization and no apparent morphological or ecological differences exist among them, and thus in this case there is no additional evidence pointing to speciation. Nevertheless, a deep populational study incorporating informative nuclear data would be required to fully understand the unusual case of *L. coridon*. Within the *coridon*-group, we consider the somewhat morphologically differentiated taxa *gennargenti* and *nufrellensis* as *L. coridon* subspecies because their genetic divergences fall within that of *L. coridon* sensu stricto, no data is available on their chromosome number and they are allopatric. The taxon *philippi*, which flies in Northern Greece, is genetically identical to *L. coridon graeca* from Central Greece. The taxon *philippi* was described as a separate species due an erroneous chromosome number count: Brown and Coutsis (1978) determined its chromosome number as n=20-26, though no figures were provided.

Coutsis *et al* (2001) restudied the karyotype of *philippi* and found n=88-90 with one large chromosome. As its chromosome number seems to finally be not different from nearby *L. coridon*, there is no evidence suggesting a status of species. The generally parapatric, but sometimes locally sympatric, Iberian taxa are of more complex assessment. Even if they frequently hybridize, they seem to present stable differences in the chromosome number, and we tentatively consider *albicans, hispana and coelestissima* as three recent species.

Lastly, we recover the *corydonius* group as monophyletic. This includes four taxa from the Caucasus (*corydonius*, *arzanovi*, *melamarina* and *sheikh*) that are often considered subspecies of a single species (Vodolazhsky and Stradomsky, 2008). Their divergence is minimal and they are estimated to have diverged ca. 0.25 Mya. They all seem to be allopatric, and differ in certain morphological characteristics and voltinism. However, since the chromosomal data we provide (n=84 for *L. melamarina, L. sheikh and L. corydonius*) cast doubt on previously reported chromosomal differences within this complex, we propose provisionally treating these taxa as conspecific until further evidence is obtained. Indeed, no predominant fixed barriers seem to exist between them, and the fact of being isolated parapatric populations with some degree of phenotypic differentiation encourage to treat them as subspecies sensu Braby *et al* (2012).

The case of *L. ossmar*, a taxon that is usually considered as closely related to the parapatric *L. corydonius* (Schurian, 1989; Hesselbarth *et al*, 1995), is especially interesting. While it was recovered in a morphological and ecological analysis by Schurian (1989, p. 158) as sister to the *L. corydonius* clade, in our dataset this is shown to be an effect produced by the detected cases of mitochondrial introgression between these two taxa. Indeed, when removing the two potentially introgressed sequences, *L. ossmar* is recovered as an independent lineage with no close relative and unresolved position, similarly to the result for the middle-eastern taxa *L. dezina* and *L. syriaca*.

We show that, at least in the case of the genus *Lysandra*, using a species tree approach that includes incomplete lineage sorting is not enough to obtain a phylogeny that faithfully reflects evolution, and that horizontal gene transfer in the form of introgression needs to be addressed to avoid incorrect taxonomic conclusions. Even by concatenating the JML test to the species tree approach, several internal relationships cannot be solved despite using seven independent markers. *Lysandra* represents a major challenge even for the latest phylogenetic methods because it is an extremely recent radiation (ca. 1.45 Mya), with incomplete lineage sorting for most clades, and widespread hybridization. We argue that in a deep coalescent scenario for the entire genus, the three supported clades have experienced fast speciation events that fixed observable molecular plesiomorphic characters. However, the fact of finding certain levels of introgression, in addition to regular difficulties to morphologically distinguish some of the taxa, open the discussion about what are the limits for species delimitation in this group. We suggest using the following criteria to define species in the genus *Lysandra*:

1) Clusters of individuals that can be distinguished by morphological and/or molecular and/or chromosomal characters and can preserve their identity in sympatry or parapatry despite occasional hybridization should be considered species.

2) In the case of allopatric taxa, reciprocal monophyly, especially in combination with distinct differences in karyotype should be considered as evidence for different species (=non-conspecifity) (e.g. *L. punctifera* and *L. bellargus*: reciprocal monophyly and strong discontinuity in chromosome number (n=24 and 45, respectively).

3) Stability in nomenclature (if we no clear evidence to change the generally accepted status of a taxon is found, we keep it (e.g. *L. dezina*).

Whether chromosome number differences suppose a postzygotic barrier in Lepidoptera is a question under debate (Kandul *et al*, 2007; Lukhtanov *et al*, 2011). Mitochondrial introgression requires backcrosses to be realized and this would be impossible if all F1 hybrids were completely infertile However, in animals most cases of hybridization that resulted in viable offspring reported involved taxa with small karyotype number differences (King, 1993) and, generally, a cumulative effect was observed (i.e. fertility decreased proportionally with the level of chromosomal differences) (King, 1993; but see Lyapunova *et al*, 2010). In our study case, we report cases of introgression between taxa with twice the number of chromosomes (Table 1). This scenario suggests two possibilities, either certain hybrids may retain some degree of fertility, or the hybridization occurred before the chromosomal reorganizations were fixed. The first option is most likely given evidence of current natural hybrids (King, 1993) and the small divergence of the introgressed sequences compared to those of the donor species, at least in the case of *L. bellargus* introgressed from *L. coridon* (see Dincă *et al*, 2011).

Karyotype evolution

The different chromosome numbers within *Lysandra* are roughly multiple of the minimum number within the genus, n=24 (*L. syriaca* and *L. punctifera*), which is the usual number within the family *Lycaenidae* (Robinson, 1971; Stekolnikov *et al*, 2000). *L. bellargus* is n=45, and the rest of species present a chromosome number of approximately the quadruple of the basal (n=82-93). This chromosomal series might suggest a case of polyploidy. However, polyploidy, although a frequent mechanism of speciation in plants, is highly unusual among bisexually reproducing animals. No direct measures of DNA content in *Lysandra* have been published, but inspection of chromosome sets indicates that genome size is more or less the same in species with different chromosome numbers. In fact, the higher the chromosome number, the smaller the chromosomes are. Additionally, the figures of meiosis in putative natural hybrids between *L. coridon* and *L. bellargus* present intermediate number of chromosome elements of very different size, which is also more consistent with a fusion-fission hypothesis (de Lesse, 1960, p. 162-164). This leads to the hypothesis that a general fission of chromosomes would be the most probable process accounting for karyotype evolution in *Lysandra* (De Bast, 1985).

Our results do not support the hypothesis of polyploidy. In the first place, the recovered phylogeny does not recover the taxa with n=24 (*L. punctifera* and *L. syriaca*) as closely related. *Lysanda bellargus* (n=45) is recovered as the sister of *L. punctifera* with high support, and thus it cannot represent the intermediate form (2x24 polyploid) leading to the 4x24 polyploids. Moreover, we have found an intermediate state regarding *L. syriaca* (n=30) for the subspecies *burak*, located in

East Turkey (Supplementary Table S1) that, together with the geographic chromosomal cline within the species *L. coridon*, can be only explained as a result of fusion-fission processes.

Chromosomal instability seems to be a rule prevailing within the genus, however the moment when it originated is not clear. According to our results, the most likely explanation points to at least three independent origins of chromosomal instability (Figure 4). One on the split of *L. bellargus* (n=45) from the common ancestor with *L. punctifera* (n=24), another within *L. syriaca* (supposedly at the split between the nominotypical subspecies (n=24) and the subspecies *L. s. burak* (n=30), and at least one more that produced the large numbers displayed by most species.

The occurrence of extensive karyotype diversity among species with little or no genetic and morphological divergence implies the possibility of chromosomal speciation (White, 1973; King, 1993). According to our data, similarly to *Agrodiaetus* (Kandul *et al*, 2004, 2007) and *Plebicula* (data not published), the genetic divergence among *Lysandra* species is small, even when differences in chromosome numbers between the same species are large. Also, although the genus *Lysandra* split from the *Polyommatus* genus at ca. 4.9 Mya, a burst of diversification generating current species diversity did not apparently start until much more recently (1.45 Mya). Since the common ancestor of extant species most likely had 24 chromosomes (as almost all the rest of Lycaenids have), the hypothesis of karyotype diversification driving speciation in *Lysandra* is reinforced.

Chromosomal changes within a lineage may promote the chance of speciation. If new chromosomal rearrangements are underdominant, their fixation (i.e. transition from heterozygous condition to homozygous state) may be promoted by population bottlenecks. At the same time, other consequences from these bottlenecks can be derived, as a rapid lineage sorting of neutral molecular markers. We argue that the *Lysandra* case could likely respond to this hypothesis. First, demographic history inferred from the species tree suggests that the lineages of this genus may have suffered substantial variation in their population sizes (Figure 4, Supplementary Table S5), possibly in the form of bottlenecks. On the other side, we find strong phylogenetic support in three nodes, supposedly because of the fixation of unique characters by rapid lineage sorting, a remarkable situation in a context of generalized incomplete lineage sorting. It is also suggestive the fact that *L. coridon*, a species displaying notorious chromosomal changes that apparently have not produced speciation, exhibits the largest values on population sizes. Overall, this scenario suggests the hypothesis that deep coalescence correlates with karyotype stability and rapid lineage sorting with karyotype instability. Testing this hypothesis would require a much larger population genetic study that could unravel the role of demographic bottlenecks along the history of *Lysandra* diversification.

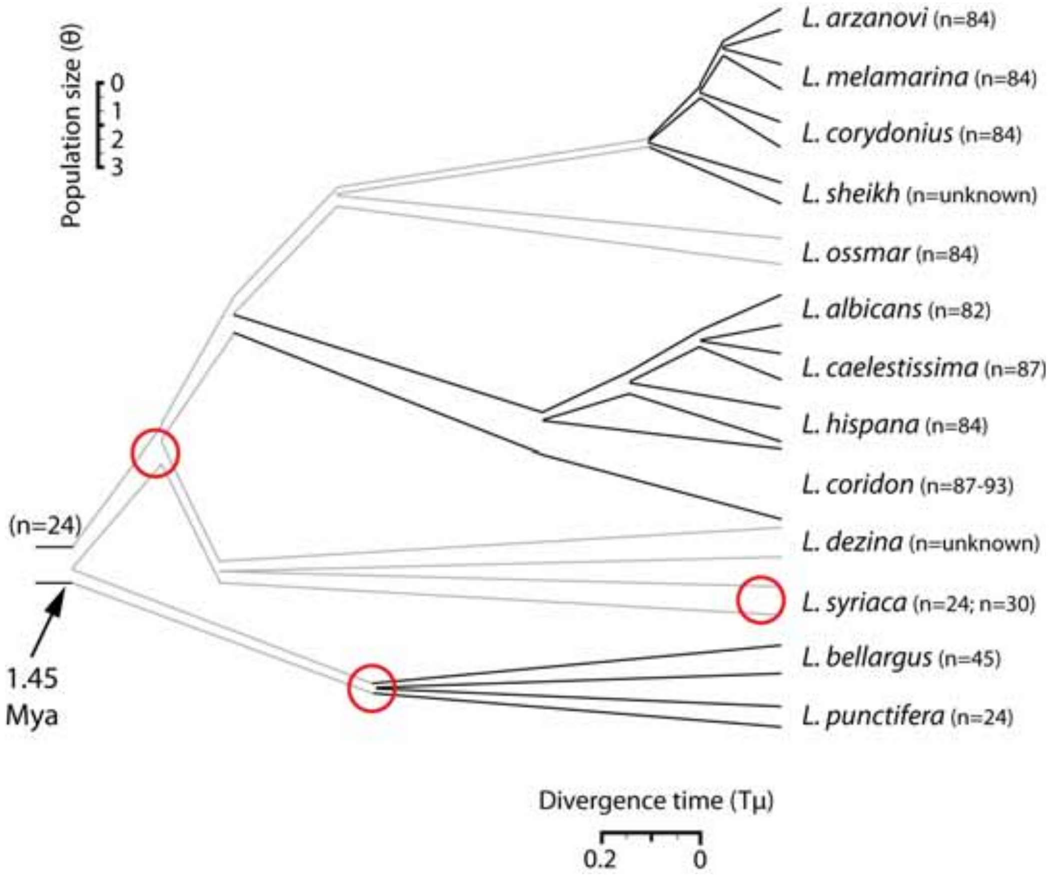


Figure 4. Demographic history drawn on a consensus species tree. Variable population sizes along branches (piecewise linear model in *BEAST) are represented as width (in the Y-axis) according to the Biopy summary from *BEAST inference (after JML application). The X-axis represents divergence time in Mya

according to a 1.5% evolutionary rate for insect mitochondrion. Red circles show the three estimated origins of chromosomal instability. Strong putative population bottlenecks along the phylogeny could promote rapid lineage sorting and fixation of chromosomal changes, to finally lead to speciation.

CONCLUSIONS

The genus *Lysandra* forms a clade displaying recent diversification events that started around 1.4 Mya. We obtain a partially resolved tree with three main supported clades: *L. punctifera* + *L. bellargus*, the *corydonius* taxa and *L. coridon* + the Iberian taxa, plus three taxa without close relatives recovered as lineages with unresolved position. Predominant incomplete lineage sorting seems to blur basal relationships. Our new karyotype findings within the *corydonius* group does not reveal any differences among species, and given the low genetic divergences observed we consider them as subspecific taxa. We show that mtDNA introgression in *Lysandra* is widespread and that not accounting for hybridization in species tree inference can lead to erroneous phylogenetic and taxonomic conclusions. *Lysandra coridon* displays two paraphyletic lineages roughly corresponding to eastern and western Europe, most likely because of massive introgression events between western specimens and the Iberian closely related taxa. However, we argue that the Iberian taxa correspond to true species, although extremely recent, given their apparently stable karyotype differences. Finally we show that chromosomal instability has originated at least three independent times within the group. We also hypothesize that chromosomal instability may have played a crucial role in the *Lysandra* species radiation, where strong population bottlenecks may promote fixing new chromosomal rearrangements at the same time that a rapid lineage sorting of neutral molecular markers occurred, in a generalized scenario of incomplete lineage sorting.

ACKNOWLEDGEMENTS

We want to thank M. Bollino, P. Casula, J. Coleman, J. Coutsis, A.V. Dantchenko, V. Dincă, R. Eastwood, N. Kandul, J. Rubio, V. Shchurov, K. Schurian, and M.A. Travassos for providing specimens. We are especially indebted to Juan Pino, who discovered the genetically interesting *L. coridon* population from Galicia (NW Spain) and offered us samples. We acknowledge Lukas Rieppel for valuable discussion. Support for this research was provided by the Spanish MICINN (project CGL2010-21226 / BOS to G.T. and R.V. and predoctoral fellowship BES-2008- 002054 to G.T.), by the Russian Foundation for Basic Research (grants 12-04-00490, 11-04-00076, 11-04-00734 and 11-04-01119), by grant 16.518.11.7070 (Ministry of Education and Science of the Russian Federation), and by the programs of Russian Academy of Science "Dynamics and conservation of gene pools" and "Origin of biosphere and evolution of geo-biological systems"; grants from the Baker Foundation, the Green Memorial Fund of Harvard University and the Putnam Expeditionary Fund of the Museum of Comparative Zoology to N.E.P. and R.V. and NSF DEB-0447242 to N.E.P.

REFERENCES

Ané C., Larget B., Baum D.A., Smith S.D., Rokas A. **2007**. Bayesian es- timation of concordance among gene trees. *Molecular Biology and Evolution* 24:412–426.

Berthier P., Excoffier L., Ruedi M. **2006**. Recurrent replacement of mtDNA and cryptic hybridization between two sibling bat species Myotis myotis and Myotis blythii. *Proceedings of the Royal Society of London B: Biological Sciences* 273:3101– 3109.

Bossu, C.M., Near, T.J. **2009**. Gene trees reveal repeated instances of mitochondrial DNA introgression in orangethroat darters (Percidae: Etheostoma). *Systematic Biology* 58:114–129.

Bouckaert, R. R. **2010**. DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics* 26:1372–1373.

Braby, M.F., Eastwood, R., Murray, N. **2012.** The subspecies concept in butterflies: has its application in taxonomy and conservation biology outlived its usefulness? *Biological Journal of the Linnean Society* 106(4):699-716.

Bryant D., Moulton V. **2004**. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution* 21:255–265.

Buckley, T.R., Cordeiro, M., Marshall, D.C., Simon, C. **2006**. Differentiating between hypotheses of lineage sorting and introgression in New Zealand alpine cicadas (Maoricicada Dugdale). *Systematic Biology* 55:411–425.

Coutsis, J., de Prins, J., de Prins, W. **2001**. The chromosome number and karyotype of the two morphs of *Polyommatus* (*Lysandra*) *coridon* from Greece (Lepidoptera:Lycaenidae). *Phegea* **29**:63-71.

De Bast, B. **1985**. La notion d'espéce dans le genre *Lysandra* Hemming, 1933 (Lepidoptera Lycaenidae). *Linneana Belgica* **10**:98–110.

Descimon, H., Mallet, J. **2009**. Bad species. *In* Ecology of Butterflies in Europe. Settele, J., Shreeve, T.G., Konvicka, M., Van Dyck, H. [Eds]. Cambridge University Press, Cambridge.

de Lesse H. **1959**. Sur la valeur spécifique de deux sous-espèces d'Agrodiaetus (Lep. Lycaenidae) récemment descrites. *Bulletin mensuel Societé Linnéenne de Lyon* 28:312-315.

de Lesse H. **1960**. Spéciation et variation chromosomique chez les Lépidoptères Rhopalocères. *Annales des Sciences Naturelles* 2, 1–223.

de Lesse, H. **1969**. Les Nombres des Chromosomes dans le Groupe de *Lysandra coridon* (Lep. *Lycaenidae*). *Annales de la Société Entomologique de France* 5:469-532.

Dincă, V., Zakharov, E. V., Hebert, P. D. N., Vila, R. **2011**. Complete DNA barcode reference library for a country's butterfly fauna reveals high performance for temperate Europe. *Proceedings of the Royal Society of London B: Biological Sciences* 278:347-355.

Edwards, S. V. **2009**. Is a new and general theory of molecular systematics emerging? *Evolution* 63:1-19.

Ferris, S.D., Sage, R.D., Huangj, C.M., Nielsen, J.T., Ritte, U., Wilson, A.C. **1983**. Flow of mitochondrial DNA across species boundary. *Proceedings of the National Academy of Sciences USA* 80:2290–2294.

Gauthier, O., Lapointe, F.J. **2007**. Seeing the trees for the network: consensus, information content, and superphylogenies. *Systematic Biology* 56:345–355.

Gil-T., F. **2007**. A natural hybrid of *Polyommatus bellargus* (Rottemburg, 1775) × *P. albicans* (Herrich-Schäffer, 1852) and notes about a probable hybrid of *P. punctifera* (Oberthür, 1876) **×** *P. albicans* (Lepid.: Lycaenidae). *Nachrichten des Entomologischen Vereins Apollo* 28 (1/2):11-13.

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O. **2010**. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies:

Assessing the Performance of PhyML 3.0. *Systematic Biology* 59:307-21.

Heled, J., Drummond, A. **2010**. Bayesian Inference of Species Trees from Multilocus Data. *Molecular Biology and Evolution* 27:570-580.

Hesselbarth, G., Van Oorschot, H., Wagener, S. **1995**. Die Tagfalter der Türkei unter Berücksichtigung der angrenzenden Länder.

Jin, G., Nakhleh, L., Snir, S., Tuller T. **2006**. Maximum likelihood of phylogenetic networks. *Bioinformatics* 22:2604–2611.

Joly, S., McLenachan, P. A., Lockhart, P. J. **2009**. A statistical approach for distinguishing hybridization and incomplete lineage sorting. *American Naturalist* 174:e54–e70.

Joly, S. **2012**. JML: Testing hybridization from species trees. *Molecular Ecology Resources* 12:179-184.

Kandul, N.P., Coleman, J.W.S., Lukhtanov, V.A., Dantchenko, D.A., Sekercioglu C., Haig, D., Pierce, N.E. **2004**. Phylogeny of *Agrodiaetus* Hübner 1822 (Lepidoptera: Lycaenidae) inferred from mtDNA Sequences of *COI* and *COII*, and Nuclear Sequences of *EF1-α*: Karyotype diversification and species radiation. *Systematic Biology* 53:278-298.

King, M. **1993**. Species Evolution: The Role of Chromosomal Change Cambridge: Cambridge University Press.

Kubatko, L.S. **2009**. Identifying hybridization events in the presence of coalescence via model selection. *Systematic Biology* 58:478–488.

Lorković, Z. **1941**. Die Chromosomenzahlen in der Spermatogenese der Tagfalter. *Chromosoma* 2:155-191.

Lorković, Z. **1949**. Chromosomen-Vervielfaschung bei Schmetterlingen und ein neuer Fall fünffacher Zahl. *Revue Suisse de Zoologie* 56: 243-249.

Lorković, Z. **1990**. The butterfly chromosomes and their application in systematics and phylogeny. *In* Butterflies of Europe 2:332-396. Kudrna O [Ed]. Wiesbaden: Aula-Verlag.

Lukhtanov, V. A., Dantchenko, A. V. **2002**. Principles of highly ordered metaphase I bivalent arrangement in spermatocytes of *Agrodiaetus* (Lepidoptera). *Chromosome Research* 10:5–20.

Lukhtanov, V.A., Kandul, N.P., Plotkin, J.B., Dantchenko, A.V., Haig, D., Pierce, N.E. **2005**. Reinforcement of pre-zygotic isolation and karyotype evolution in *Agrodiaetus* butterflies. *Nature* 436:385–389.

Lukhtanov, V.A., Vila, R., Kandul, N.P. **2006**. Rearrangement of the *Agrodiaetus dolus* species group (Lepidoptera, Lycaenidae) using a new cytological approach and molecular data. *Insect Systematics and Evolution* 37:325–334.

Lukhtanov, V.A., Shapoval, N.A., Dantchenko, A.V. **2008**. *Agrodiaetus shahkuhensis* sp. n. (Lepidoptera, Lycaenidae), a cryptic species from Iran discovered by using molecular and chromosomal markers. *Comparative Cytogenetics* 2(2):99-114.

Lukhtanov, V.A., Dinca, V., Talavera, G., Vila, R. **2011**. Unprecedented within-species chromosome number cline in the Wood White butterfly Leptidea sinapis and its significance for karyotype evolution and speciation. *BMC Evolutionary Biology* 11:109.

Lyapunova, E.A., Bakloushinskaya, I.Y., Saidov, A.S., Saidov, K.K. 2010. Dynamics of chromosome variation in mole voles Ellobius tancrei (Mammalia, Rodentia) in Pamiro-Alai in the period from 1982 to 2008. *Russian Journal of Genetics* 46:566-571.

Melo-Ferreira, J., Boursot, P., Carneiro, M., Esteves, P.J., Farelo, L., Alves, P.C. **2012**. Recurrent introgression of mitochondrial DNA among hares (*Lepus* spp.) revealed by species-tree inference and coalescent simulations. *Systematic Biology* 61(3):367-381.

Mensi, P., Lattes, A., Salvidio, S., Balleto, E. **1988**. Taxonomy, evolutionary biology and biogeography of South West European Polyommatus coridon (Lepidoptera: Lycaenidae). *Zoological Journal of the Linnean Society* 93:259-271.

Oberthür, C. 1876. Faunes entomologiques; descriptions d'insects nouveaux ou peu connus. Imprimerie Oberthür. Rennes.

Posada, D. **2008**. jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution* 25:1253–1256.

Quek, S.P., Davies, S.J., Itino, T., Pierce, N.E. **2004**. Codiversification in an ant-plant mutualism: Stem texture and the evolution of host use in Crematogaster (Formicidae: Myrmicinae) inhabitants of Macaranga (Euphorbiaceae). *Evolution* 58:554-570.

Rambaut, A., Grassly, N. C. **1997**. Seq-Gen: an application for the monte carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences* 13:235–238.

Robinson, R. **1971**. Lepidoptera Genetics. Pergamon Press, Oxford.

Roca, A.L., Georgiadis, N., O'Brien, S.J. **2005**. Cytonuclear genomic dissociation in African elephant species. *Nature Genetics* 37:96–100.

Ruedi, M., Smith, M.F., Patton, J.L. **1997**. Phylogenetic evidence of mito- chondrial DNA introgression among pocket gophers in New Mexico (family Geomyidae). *Molecular Ecology* 6:453–462.

Schmitt, T., Seitz, A. **2001**. Allozyme variation in Polyommatus coridon (Lepidoptera: Lycaenidae): identification of ice-age refugia and reconstruction of post-glacial expansion. *Journal of Biogeography* 28:1129–1136.

Schmitt, T., Gießl, A., Seitz, A. **2002**. Postglacial colonisation of western Central Europe by *Polyommatus coridon* (Poda 1761) (Lepidoptera: Lycaenidae): evidence from population genetics. *Heredity* 88:26–34.

Schmitt, T., Zimmermann, M. **2012**. To hybridize or not to hybridize: what separates two genetic lineages of the Chalk-hill Blue *Polyommatus coridon* (Lycaenidae, Lepidoptera) along their secondary contact zone throughout eastern Central Europe? *Journal of Zoological Systematics and Evolutionary Research* 50:106-115.

Schurian, K.G. **1989**. Revision der Lysandra-Gruppe des Genus *Polyommatus* Latr. (Lepidoptera: Lycaenidae). *Neue Entomologische Nachrichten* 24:1-181

Spinks, P.Q., Shaffer, H.B. **2009**. Conflicting mitochondrial and nuclear phylogenies for the widely disjunct ewmys (Testudines: Emydidae) species complex, and what they tell us about biogeography and hybridization. *Systematic Biology* 58:1–20.

Stekolnikov, A.A., Ivanov, V.A., Kuznetzov, V.I., Lukhtanov, V.A. **2000**. Evolution of the chromosome mechanism, wing articulation, male genitalia and phylogeny of Butterflies (Lepidoptera: Hesperioidea, Papilionoidea). *Entomologicheskoe Obozrenie* 79:123–149 [in Russian; English translation in *Entomol. Rev.*].

Stradomsky, B.V., Shchurov, V.I. **2005**. Notes on the status of the Caucasia taxa of the group *Polyommatus* (*Meleageria*) *coridon* (sensu de Lesse) with description of a new species from the high-mountain area of West Caucasia (Lepidoptera: Lycaenidae). *Phegea* 33:69-75.

Talavera, G., Lukhtanov, V.A., Pierce, N.E., Vila, R. **2012**. Establishing criteria for higher-level classification using molecular data: the systematics of *Polyommatus* blue butterflies (Lepidoptera, Lycaenidae). *Cladistics* published online (doi: 10.1111/j.1096-0031.2012.00421.x)

Tshikolovets, V. **2011**. Butterflies of Europe and the Mediterranean area. Pardubice, Czech Republic: Tshikolovets Publications.

Vershinina A.O, Lukhtanov V.A. **2010**. Geographical distribution of the cryptic species Agrodiaetus alcestis alcestis, A. alcestis karacetinae and A. demavendi (Lepidoptera, Lycaenidae) revealed by cytogenetic analysis. *Comparative Cytogenetics* 4(1):1-11.

Vila, R., Lukhtanov, V.A., Talavera, G., Gil, T.F., Pierce, N.E. **2010**. How common are dot-like distribution ranges? Taxonomical oversplitting in Western European Agrodiaetus (Lepidoptera, Lycaenidae) revealed by chromosomal and molecular markers. *Biological Journal of the Linnean Society* 101:130–154.

Vila, R., Bell, C.D., Macniven, R., Goldman-Huertas, B., Ree, R.H., Marshall, C.R., Bálint, Z., Johnson, K., Benyamini, D., Pierce, N.E. **2011**. Phylogeny and palaeoecology of Polyommatus blue butterflies show Beringia was a climate-regulated gateway to the New World. *Proceedings of the Royal Society of London B: Biological Sciences* 278:2737–2744.

Vodolazhsky, D.I., Stradomsky, B.V. **2008**. A study of blues butterflies of the group of *Lysandra corydonius* (Herrich-Schäffer, 1804) (Lepidoptera: Lycaenidae) with the use of mtDNA markers. *Caucasian Entomological Bulletin* 4:353-355.

White, M.J.D. **1973**. Animal Cytology and Evolution Cambridge: Cambridge University Press.

Wiemers, M. 2003. Chromosome differentiation and the radiation of the butterfly subgenus *Agrodiaetus* (Lepidoptera: Lycaenidae: Polyommatus) – a molecular phylogenetic approach. PhD thesis, University of Bonn. Available at: http://hss.ulb.uni-bonn.de/2003/0278/0278.htm

SUPPLEMENTARY INFORMATION

In the shadow of phylogenetic uncertainty: the recent diversification of the *Lysandra* butterflies through chromosomal changes

Gerard Talavera[a,b], Vladimir A. Lukhtanov[c,d], Naomi E. Pierce[e] and Roger Vila[a,*]

[a]Institut de Biologia Evolutiva (CSIC-UPF), Passeig Marítim de la Barceloneta, 37, 08003 Barcelona, Spain

[b]Departament de Genètica i Microbiologia, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain

[c]Department of Karyosystematics, Zoological Institute of Russian Academy of Science, Universitetskaya nab. 1, 199034 St Petersburg, Russia

[d]Department of Entomology, St Petersburg State University, Universitetskaya nab. 7/9, 199034 St Petersburg, Russia

[e]Department of Organismic and Evolutionary Biology and Museum of Comparative Zoology, Harvard University, 26 Oxford Street, Cambridge, Massachusetts 02138, USA

[*]Corresponding author: roger.vila@csic.es

Table S1. Samples used in this study: taxon name, sample accession number at MCZ and sample collection locality.

Genus	Species & ssp.	Sample code	Locality
Lysandra	*albicans albicans*	RV03H582	Puebla de Don Fadrique, 1295 m, Granada, **Spain**
Lysandra	*albicans arragonensis*	MAT99Q969	Una, Cuenca, 970m, **Spain**
Lysandra	*arzanovi*	SH02H019	Aibga-1 Pass. 1850m, Krasnaya Polyana, Aibga Mts., Sotch, Krasnodar Region, **Russia**
Lysandra	*arzanovi*	SH02H020	Aibga-1 Pass. 1850m, Krasnaya Polyana, Aibga Mts., Sotch, Krasnodar Region, **Russia**
Lysandra	*bellargus*	AD00P129	Aragatz Mt., Amberd Valley, 2300m, Transcaucasus, **Armenia**
Lysandra	*bellargus*	JC96Q001	Gambach, Bavaria, **Germany**
Lysandra	*bellargus*	MAT99Q882	Rúbies, Catalonia, **Spain**
Lysandra	*bellargus*	RV04G399	Saimbeyli Valley, 1445m (Adana), **Turkey**
Lysandra	*bellargus*	VL02X510	Masuleh, 1900-2100m, Gilan, **Iran**
Lysandra	*caelestissima*	MAT99Q959	Ciudad Encantada, 1440m, Uña, Cuenca, **Spain**
Lysandra	*caelestissima*	MAT99Q966	Uña, Cuenca, 970m, **Spain**
Lysandra	*coridon apennina*	MB05G416	Mt. Pollino, Calabria, **Italy**
Lysandra	*coridon asturiensis*	JR04G493	Albelda, 900m, La Rioja, **Spain**
Lysandra	*coridon asturiensis*	RV07C272	Cedeira, Capelada, Galicia, **Spain**
Lysandra	*coridon borussia*	AD00P192	Tula region, Tatinki, 120 m., W. **Russia**
Lysandra	*coridon cataluniae*	RV03H454	El Brull, Catalonia, **Spain**
Lysandra	*coridon coridon*	VD02T008	**Romania**
Lysandra	*coridon gennargenti*	KS05I874	Orgosolo, 1250m, vic. Monte Novo S. Giovanni, Sardignia Is.
Lysandra	*coridon gennargenti*	KS05I875	Orgosolo, 1250m, vic. Monte Novo S. Giovanni, Sardignia Is.
Lysandra	*coridon insulana*	RE04C165	Therfield Heath, Royston, **UK**
Lysandra	*coridon narbonensis*	MAT99Q932	Mende, 780m, Languedoc region, **France**
Lysandra	*coridon*	AD00P045	Volgograd region, Kamyshinsky v., 200 m., Low Volga, South **Russia**
Lysandra	*coridon*	RE07G279	NE Bezandun-sur-Bine, 735 m, Drome, **France**
Lysandra	*coridon*	RV06A183	Sorteny, **Andorra**
Lysandra	*coridon*	RV07E302	Baile Herculane, Pecinisca, 220-320m, Caras-Severin, **Romania**
Lysandra	*coridon graeca*	JXC02G002	Mt. Timfristos (=Mt. Veluhi), 1300-1500m, Sterea Ellas, **Greece**
Lysandra	*corydonius caucasica*	VL01L120	Hasköy, 12 km SW Gümüshane, Gümüshane Prov., **Turkey**
Lysandra	*corydonius caucasica*	AD00P435	Aiodzor Mts., Gnishyk 1800 m., Transcaucasus, **Armenia**
Lysandra	*corydonius corydonius*	VL03F932	Talysh Mts, SE **Azerbeijan**
Lysandra	*corydonius corydonius*	VL05N131	**Iran** Azerbaijan-e Sharqi, pass 25 km NW Varzaqan; 2050-2170 m
Lysandra	*dezina*	09X500	**Kurdistan**
Lysandra	*hispana hispana*	MAT99T993	Coll d'Estenalles, 870m, Parc Natural de Sant Llorenç del Munt, **Spain**
Lysandra	*hispana hispana*	RV07F312	El Mont, Albanyà, Alt Empordà, Girona, Spain, 860m
Lysandra	*hispana semperi*	RV02N590	Ares del Maestre, 1150m, Castello, **Spain**

Lysandra	*melamarina*	SH02H007	Gelendjik, Betta Mts., 150m, Krasnodar Region, **Russia**
Lysandra	*melamarina*	SH02H010	Gelendjik, Betta Mts., 150m, Krasnodar Region, **Russia**
Lysandra	*nufrellensis*	KS05I821	**Corsica**, 1300m
Lysandra	*nufrellensis*	KS05I822	**Corsica**, 1300m
Lysandra	*ossmar ankara*	RV04G136	Kargasekmez Geçidi, Kizilcahamam, 1150m (Ankara) **Turkey**
Lysandra	*ossmar ossmar*	RV04G356	3Km NW Urgüp, 1140m (Kapadokya) **Turkey**
Lysandra	*ossmar ossmar*	RV07F170	Yelatan, 15 km S. of Çamardi, Nidge, **Turkey**, 1330m
Lysandra	*philippi*	SI03K025	Mt. Phalakro, 1600 m, District (Nomos) Drama, **Greece**
Lysandra	*philippi*	SI03K037	Mt. Phalakro, 600 m, District (Nomos) Drama, **Greece**
Lysandra	*punctifera*	NK02A026	Ait-b-Yahya, 1900m, Rich, **Morocco**
Lysandra	*punctifera*	NK02A027	Col Taghzoum, 1900m, High Atlas Range, **Morocco**
Lysandra	*sheikh*	VL03F998	Altyagach,1300m, Azerbeijan near the border with Dagestan, **Russia**
Lysandra	*sheikh*	VL03H615	Altyagach,1300m, Azerbeijan near the border with Dagestan, **Russia**
Lysandra	*syriaca burak*	RV07F139	13 km N. of Saimbeily, 1505m (Adana) **Turkey**
Polyommatus	*amandus amurensis*	AD02W109	Primorski Krai, S. Ussuri, Khanka Lake, Poganichnoye, **Russia**
Neolysandra	*diana*	AD00P081	Gegamsky Mts., 1800m, Gegadyr, **Armenia**
Polyommatus	*myrrha cinyraea*	AD00P389	Zangezur Mts., Akhtchi, **Armenia**

Table S2. Primer sequences. mt: mitochondrial, n: nuclear. T = thymine, A = adenine, G = guanine, C = cytosine, K = G+T, W = A+T, M = A+C, Y = C+T, R = A+G, S = G+C, V = G+A+C, I = Inosine, N = A+C+G+T.

Primer location	Primer name	Direction	Sequence (5' to 3')
mt *COI*	LCO1490[1]	forward	GGTCAACAAATCATAAAGATATTGG
mt *COI*	Ron[2,3]	forward	GGATCACCTGATATAGCATTCCC
mt *COI*	Nancy[3]	reverse	CCCGGTAAAATTAAAATATAAACTTC
mt *COI*	Tonya[3]	forward	GAAGTTTATATTTTAATTTTACCGGG
mt *COI*	Hobbes[3]	reverse	AAATGTTGNGGRAAAAATGTTA
mt *COI*	TN2126[4]	forward	TTGAYCCTGCAGGTGGWGGAG
mt *COII*	George[3,5]	forward	ATACCTCGACGTTATTCAGA
mt *COII*	Phyllis[3,5]	reverse	GTAATAGCIGGTAARATAGTTCA
mt *COII*	Strom[3,5]	forward	TAATTTGAACTATYTTACCIGC
mt *COII*	Eva[3,5]	reverse	GAGACCATTACTTGCTTTCAGTCATCT
mt *COII*	JL3146[4]	forward	GAGTTTCACCTTTAATAGAACA
mt *COII*	B-tLys[2]	reverse	GTTTAAGAGACCAGTACTTG
mt *COII*	JL2532[4]	forward	ACAGTAGGAGGATTAACAGGAG
n *CAD*	CAD787F[6]	forward	GGDGTNACNACNGCNTGYTTYGARCC
n *CAD*	CADFa[7]	forward	GDATGGTYGATGAAAATGTTAA
n *CAD*	CADRa[7]	reverse	CTCATRTCGTAATCYGTRCT
n *H3*	H3F[8]	forward	ATGGCTCGTACCAAGCAGACVGC
n *H3*	H3R[8]	reverse	ATATCCTTRGGCATRATRGTGAC
n *ITS-2*	ITS-3[9]	forward	GCATCGATGAAGAACGCAGC
n *ITS-2*	ITS-4[9]	reverse	TCCTCCGCTTATTGATATGC
n *wg*	LepWg1[10]	forward	GARTGYAARTGYCAYGGYATGTCTGG
n *wg*	LepWg2E[7]	reverse	ACNACGAACATGGTCTGCGT
n *wg*	Wg1n[11]	forward	CGGAGATGCGMCAGGARTGC
n *wg*	Wg2n[11]	reverse	CTTTTTCCGTSCGACACAGYTTGC

n *28S*	S3660[12]	forward	GAGAGTTMAASAGTACGTGAAAC
n *28S*	A335[12]	reverse	TCGGARGGAACCAGCTACTA
n Rpl5	F44[13]	forward	TCCGACTTTCAAACAAGGATG
n Rpl5	Lys3R[14]	reverse	ACAGCTCTGGCGCAGCGAAG

[1] Folmer, O., Black, M., Hoeh, W., Lutz, R., & Vrijenhoek, R.C. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol. Marine Biol. Biotech.* 3, 294-299.

[2] Simon, C., Frati, F., Beckebach, A., Crespi, B., Liu, H. & Flook, P. 1994. Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. *Annals of the Entomological Society of America* 87(6), 651-701.

[3] Monteiro, A. & Pierce, N.E. 2001. Phylogeny of *Bicyclus* (Lepidoptera: Nymphalidae) inferred from COI, COII, and EF-1alpha gene sequences. *Molecular Phylogenetics and Evolution* 18, 264-281.

[4] Canfield M.R., Greene E., Moreau C.S., Chen N., & Pierce N.E. 2008. Exploring phenotypic plasticity and biogeography in emerald moths: A phylogeny of the genus *Nemoria* (Lepidoptera: Geometridae). *Molecular Phylogenetics and Evolution* 49(2), 477-87.

[5] Brower, A.V.Z. 1994. Phylogeny of *Heliconius* butterflies inferred from mitochondrial DNA sequences (Lepidoptera: Nymphalidae). *Molecular Phylogenetics and Evolution* 3(2), 159-174.

[6] Moulton, J.K. & Wiegmann, B.M. 2004. Evolution and phylogenetic utility of cad (rudimentary) among Mesozoic-aged eremoneuran Diptera (Insecta). *Molecular Phylogenetics and Evolution* 31, 363-378.

[7] Vila, R., Bell, C.D., Macniven, R., Goldman-Huertas, B., Ree, R.H., Marshall, C.R., Bálint, Z., Johnson, K., Benyamini, D., & Pierce, N.E. 2011. Phylogeny and palaeoecology of *Polyommatus* blue butterflies show Beringia was a climate-regulated gateway to the New World. Proceedings of the Royal Society B 278(1719), 2737-2744.

[8] Colgan, D.J., McLauchlan, A., Wilson, G.D.F., Livingston, S.P., Edgecombe, G.D., Macaranas, J., Cassis G., & Gray, M.R. 1998. Histone H3 and U2 snRNA DNA sequences and arthropod molecular evolution. *Australian Journal of Zoology* 46, 419-437.

[9] White, T.J., Bruns, S., Lee, S., & Taylor, J. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics in *PCR protocols: a guide to methods and applications*, edited by M.A. Innis, Gelfandm D.H., J.J. Snisky, & T. J. White. Academic Press, New York, pp. 315-322.

[10] Brower, A.V.Z. & DeSalle, R. 1998. Patterns of mitochondrial versus nuclear DNA sequence divergence among nymphalid butterflies: the utility of wingless as a source of characters for phylogenetic inference. *Insect Molecular Biology* 7(1), 73-82.

[11] Designed by Ada Kalizewska (Harvard University, Cambridge, MA, USA).

[12] Sequeira, A.S., Normark, B.B., & Farrell, B. 2000. Evolutionary assembly of the conifer fauna: Distinguishing ancient from recent associations in bark beetles. *Proceedings of the Royal Entomological Society (London) B* 267, 2359-2366.

[13] Mallarino R, Bermingham E, Willmott KR, Whinnett A. and CD Jiggins. 2005. Molecular systematics of the butterfly genus Ithomia (Lepidoptera: Ithomiinae): a composite phylogenetic hypothesis based on seven genes. *Molecular Phylogenetics and Evolution* 34, 625-644.

[14] Designed in this study

Table S3. Genbank accession codes. GenBank codes used in this study.

Taxon	Specimen Code	COI + COII	Wg	CAD	ITS2	H3	28S	Rpl5
L. albicans albicans	RV03H582	X (COI)	X	X	X	X	X	X
L. albicans arragonensis	MAT99Q969	X	X	X		X	X	
L. arzanovi	SH02H019	X	X	X		X	X	X
L. arzanovi	SH02H020	X	X	X	X	X	X	X
L. bellargus	AD00P129	X	X	X		X	X	X
L. bellargus	JC96Q001	X	X	X	X	X	X	X
L. bellargus	MAT99Q882	X	X	X	X		X	X
L. bellargus	RV04G399	X	X	X	X	X	X	X
L. bellargus	VL02X510	X	X	X	X	X	X	
L. caelestissima	MAT99Q959	X	X	X	X	X	X	X
L. caelestissima	MAT99Q966	X	X	X	X	X	X	X
L. coridon apennina	MB05G416	X	X	X	X	X	X	X
L. coridon asturiensis	JR04G493	X	X	X	X	X	X	X
L. coridon asturiensis	RV07C272	X	X	X	X	X	X	X
L. coridon borussia	AD00P192	X			X	X	X	X
L. coridon cataluniae	RV03H454	X			X			X
L. coridon coridon	VD02T008	X	X	X	X	X	X	X
L. coridon gennargenti	KS05I874	X	X	X	X	X	X	X
L. coridon gennargenti	KS05I875	X	X	X	X	X	X	
L. coridon insulana	RE04C165	X	X	X	X	X	X	X
L. coridon narbonensis	MAT99Q932	X	X	X	X	X	X	X
L. coridon	AD00P045	X	X	X	X	X	X	
L. coridon	RE07G279	X	X	X	X	X	X	
L. coridon	RV06A183	X	X	X	X	X	X	
L. coridon	RV07E302	X	X	X	X	X	X	
L. coridon graeca	JXC02G002	X	X	X	X	X	X	X
L. corydonius caucasica	VL01L120	X	X	X	X	X	X	X
L. corydonius caucasica	AD00P435	X	X	X	X	X	X	X
L. corydonius corydonius	VL03F932	X	X	X	X	X	X	X
L. corydonius corydonius	VL05N131	X	X	X	X	X	X	X
L. dezina	08X599	X					X	
L. hispana hispana	MAT99T993	X	X	X	X	X	X	X
L. hispana hispana	RV07F312	X	X	X	X	X	X	X
L. hispana semperi	RV02N590	X	X	X	X	X	X	X
L. melamarina	SH02H007	X	X	X	X	X	X	X

L. melamarina	SH02H010	X	X		X	X	X	X
L. nufrellensis	KS05I821	X	X	X	X	X	X	X
L. nufrellensis	KS05I822	X	X	X	X	X	X	X
L. ossmar ankara	RV04G136	X	X	X	X	X	X	
L. ossmar ossmar	RV04G356	X	X	X	X	X	X	
L. ossmar ossmar	RV07F170	X	X	X	X	X	X	
L. philippi	SI03K025	X	X		X	X	X	
L. philippi	SI03K037	X	X	X	X	X	X	X
L. punctifera	NK02A026	X (COI)	X	X	X	X	X	X
L. punctifera	NK02A027	X	X	X	X	X	X	X
L. sheikh	VL03F998	X	X		X	X	X	X
L. sheikh	VL03H615	X	X	X	X	X	X	X
L. syriaca burak	RV07F139	X	X	X	X	X	X	X
P. amandus amurensis	AD02W109	X	X	X	X	X	X	X
N. diana	AD00P081	X	X	X	X	X	X	X
P. myrrha cinyraea	AD00P389	X	X	X	X	X	X	X

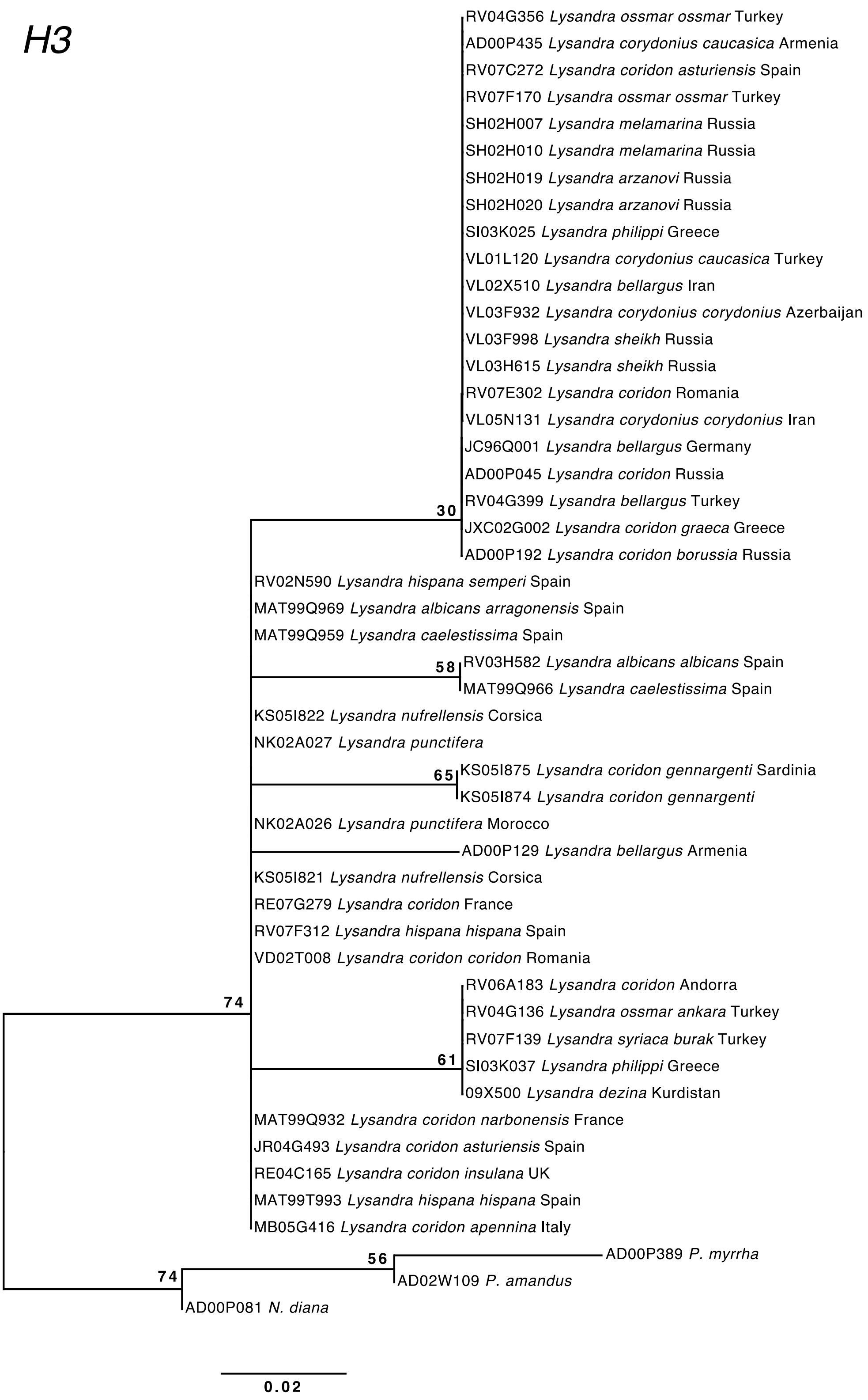
Table S4. Parsimony informative sites and number of positions per each loci for the *Lysandra* species.

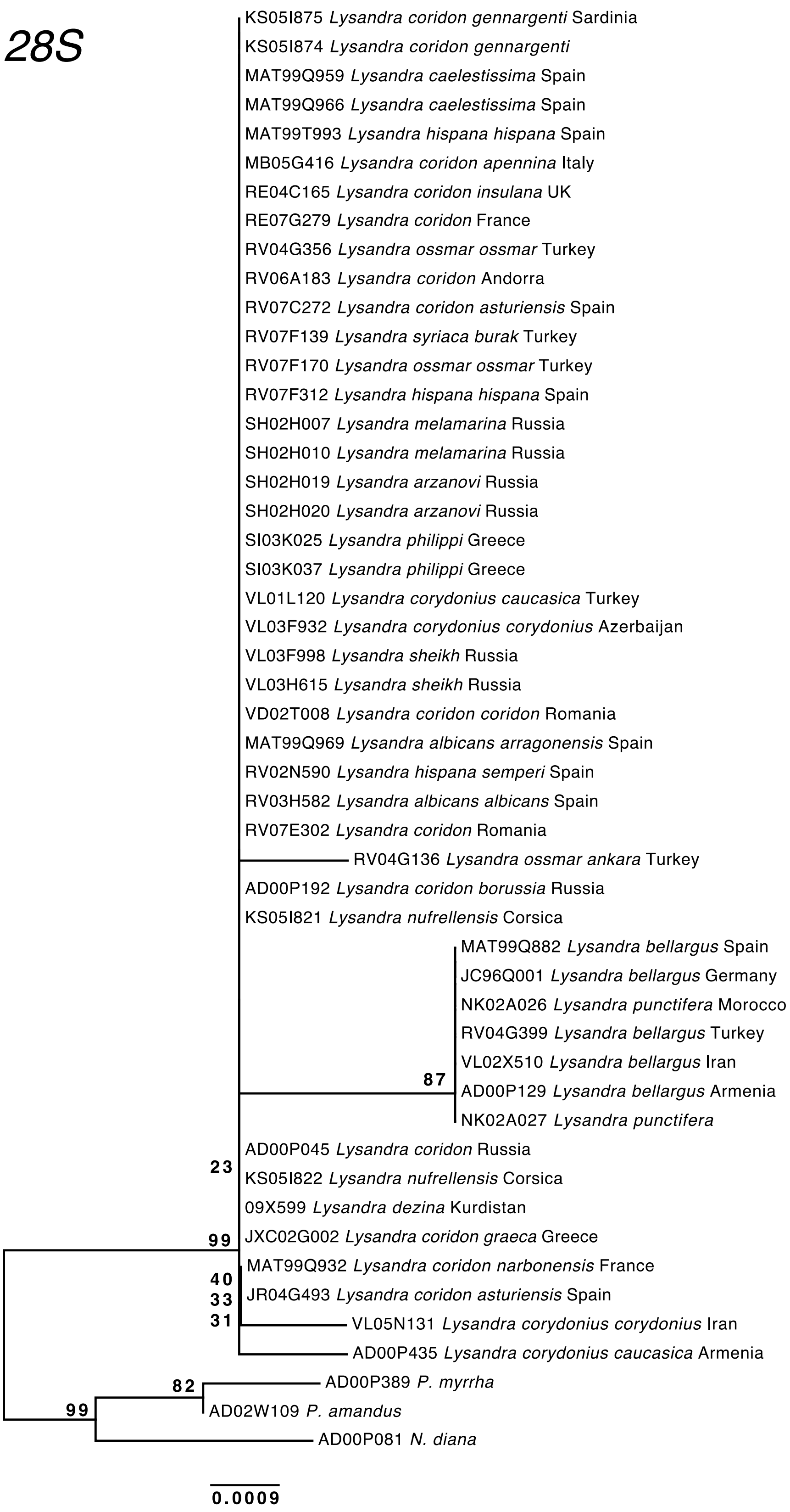
Gene	Parsimony informative sites	Number of positions
CO	153	2164
CAD	2	745
Wg	15	403
ITS2	10	635
28S	3	821
H3	4	329
Rpl5	21	873

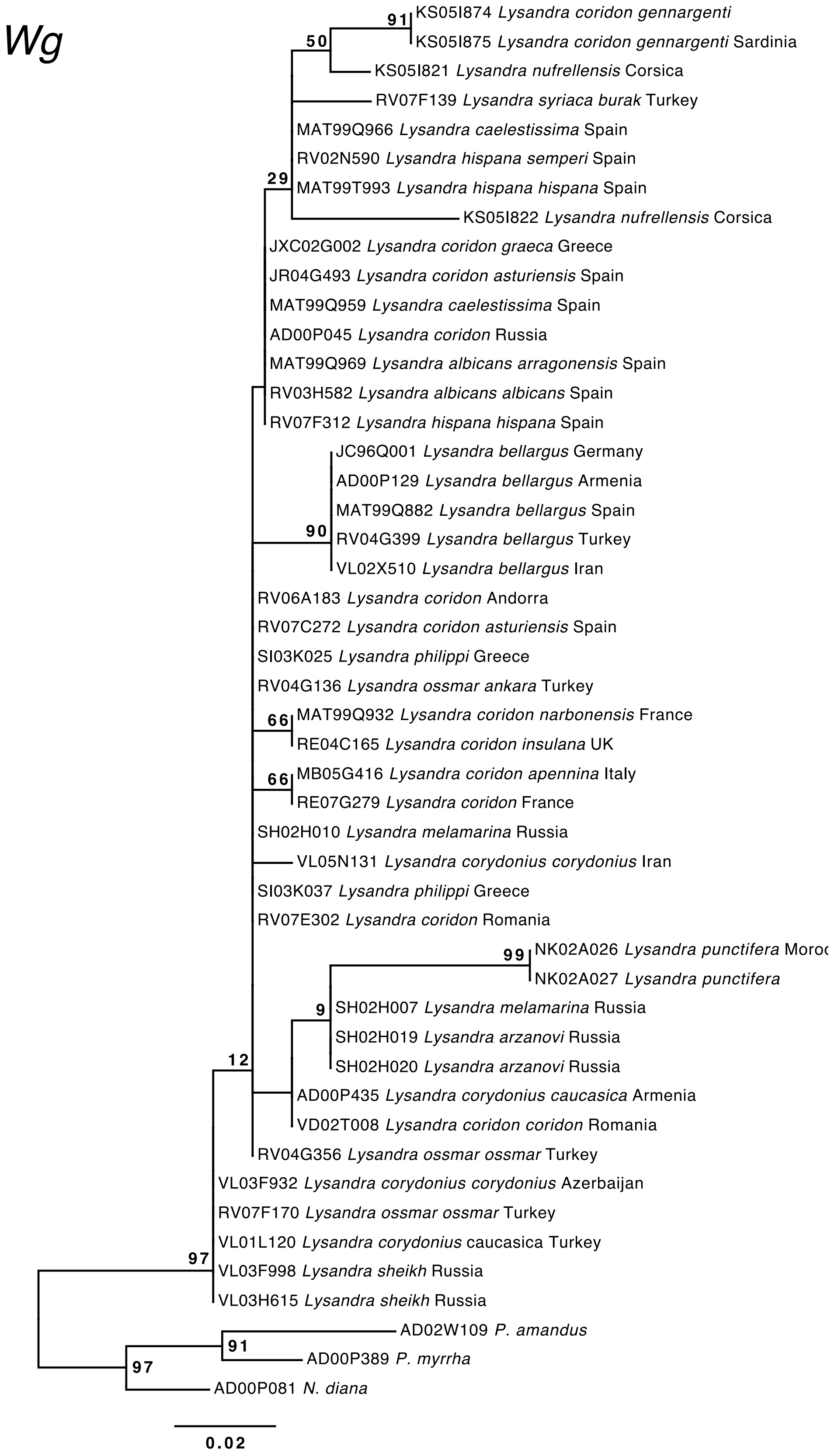
Table S5. Demographic history from *BEAST species tree inference. Values are extracted from Biopy consensus tree summaries. The X axis represents divergence time and Y axis represents relative population sizes width. Demographic values (dmv) for coalescence beginnings and ending points in branches are shown according to piecewise linear model used in *BEAST. Node ages are summarized using TreeAnnotator.

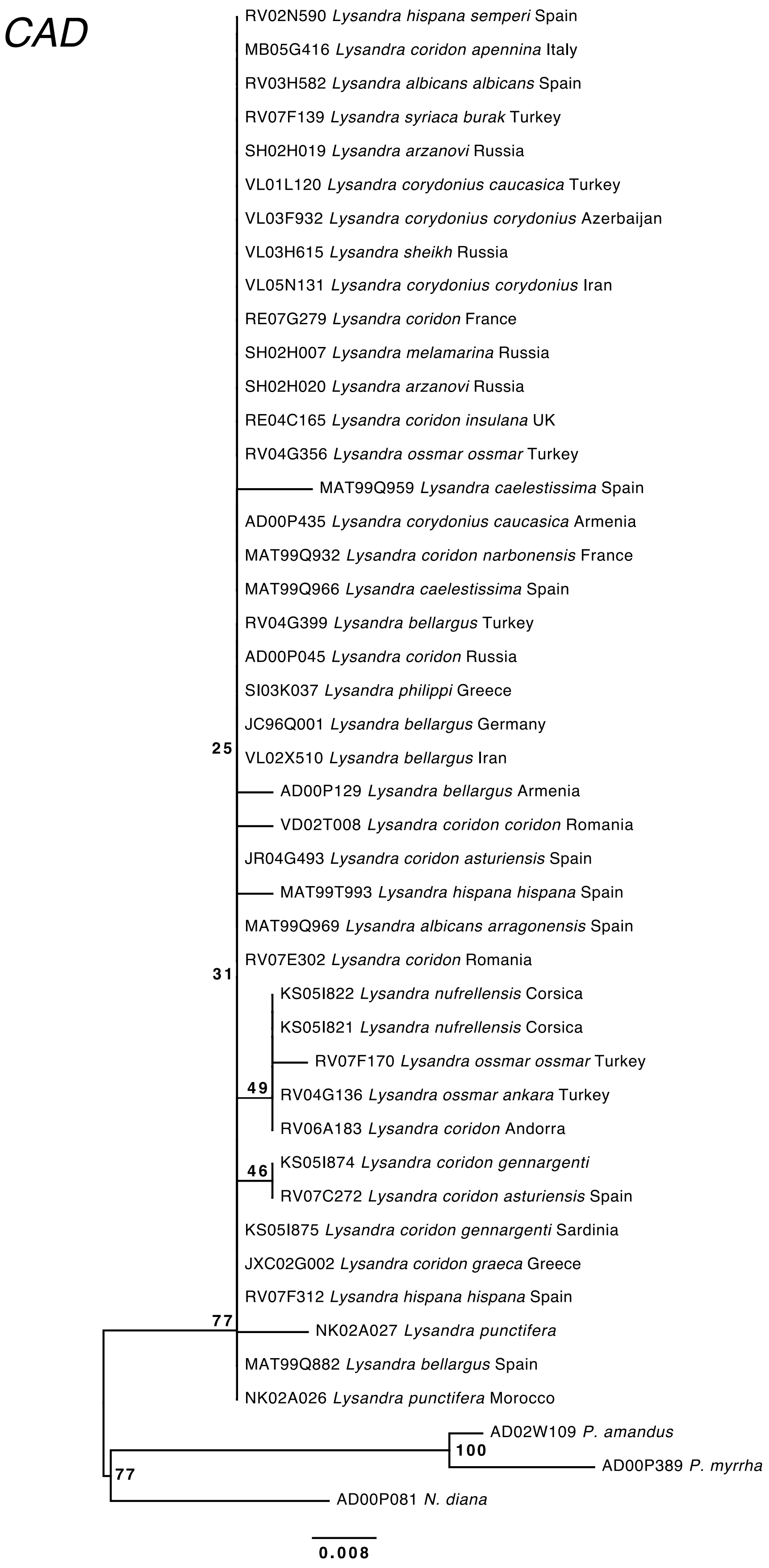
	dmv_b	dmv_e	dmv95_b	dmv95_e	Node age
arz	0.78	0.22	0.19	1.95	0.07
mel	0.90	0.25	0.23	2.12	0.07
arz-mel	0.47	0.26	0.18	1.75	0.15
cory	0.90	0.19	0.22	1.86	0.15
cory-(arz-mel)	0.45	0.13	0.12	1.18	0.25
she	0.76	0.16	0.16	1.80	0.25
she-(cory-(arz-mel))	0.29	0.26	0.14	0.83	0.60
oss	0.93	0.39	0.33	1.84	0.60
oss-(she-(cory-(arz-mel)))	0.65	0.61	0.35	1.67	0.89
alb	1.07	0.33	0.31	2.40	0.12
cael	0.94	0.23	0.24	1.99	0.12
cael-alb	0.57	0.40	0.24	1.88	0.25
his	1.18	0.39	0.38	2.22	0.25
his-(cael-alb)	0.78	0.28	0.25	1.49	0.38
cor	2.43	1.18	1.24	2.82	0.38
cor-(his-(cael-alb))	1.47	0.66	0.60	2.33	0.89
(cor-(his-(cael-alb)))-(oss-(she-(cory-(arz-mel)))	1.27	0.58	0.50	2.38	1.18
dez	1.02	0.35	0.30	2.38	1.18
syr	1.02	0.38	0.32	2.22	1.18
dez-syr	0.73	0.78	0.41	2.59	1.04
(dez-syr)-(cor-(his-(cael-alb)))-(oss-(she-(cory-(arz-mel))))	1.36	0.71	0.52	2.14	1.04
bel	1.00	0.18	0.26	1.26	0.72
punc	0.72	0.18	0.19	1.53	0.72
bel-punc	0.36	0.50	0.23	1.47	1.40
(bel-punc)-rest	1.21	1.21	0.77	2.45	1.40

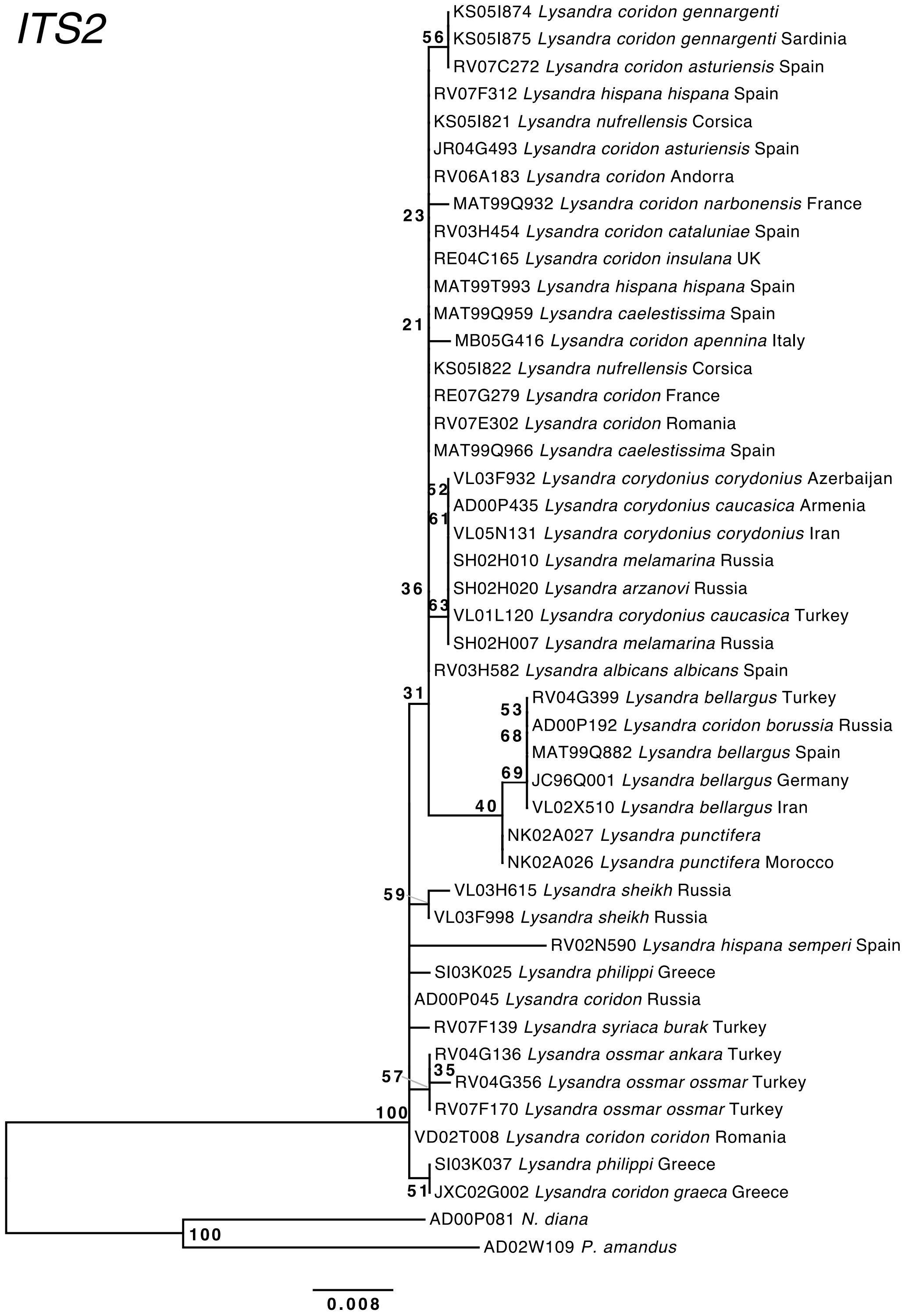
Figures S1-S6. Maximum Likelihood nuclear gene trees (*H3*, *28S*, *Wg*, *CAD*, *ITS2* and *Rpl5*). Highest values for bootstrap support are shown at nodes. Scale bar represents substitutions per position.

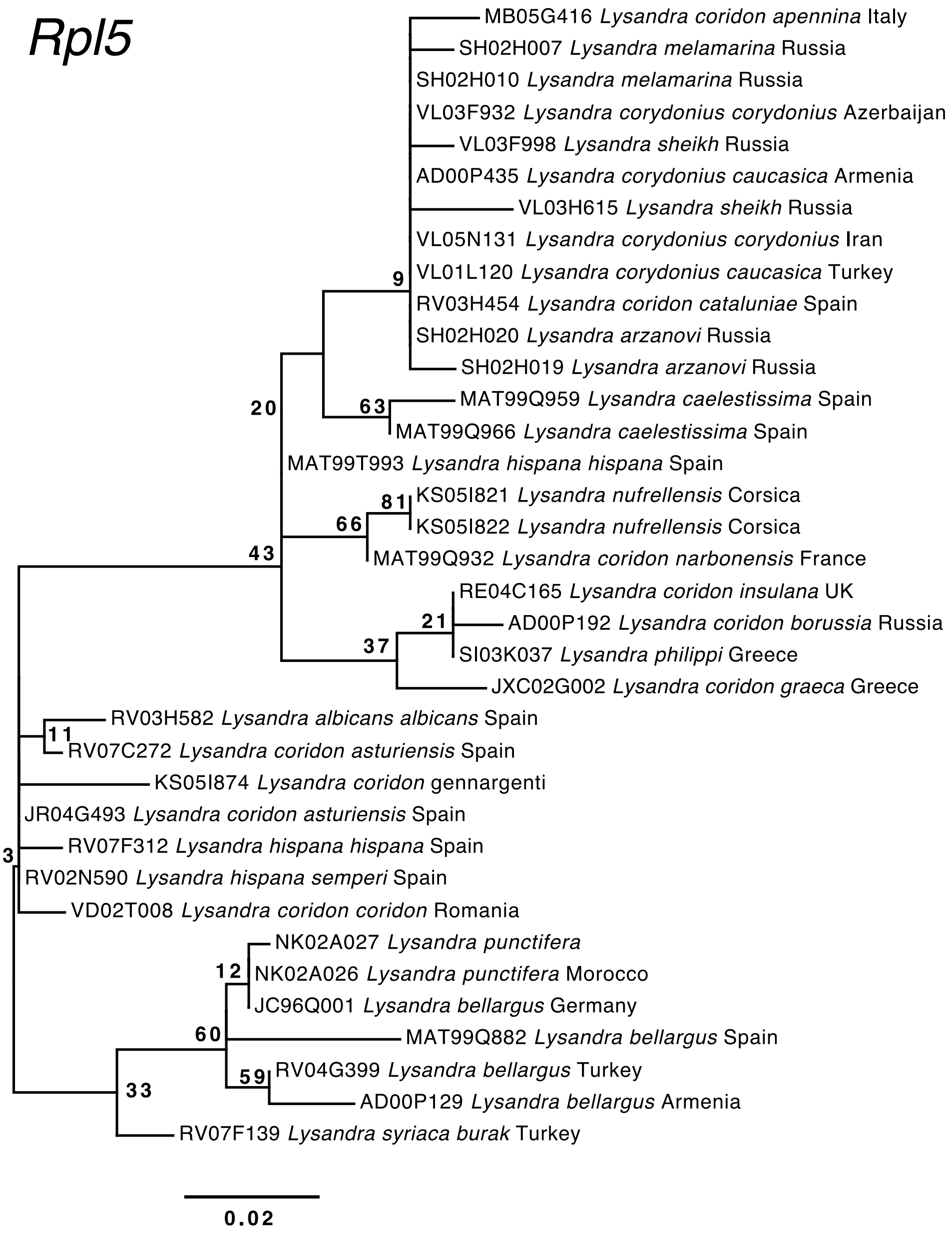

H3
RV04G356 Lysandra ossmar ossmar Turkey
AD00P435 Lysandra corydonius caucasica Armenia
RV07C272 Lysandra coridon asturiensis Spain
RV07F170 Lysandra ossmar ossmar Turkey
SH02H007 Lysandra melamarina Russia
SH02H010 Lysandra melamarina Russia
SH02H019 Lysandra arzanovi Russia
SH02H020 Lysandra arzanovi Russia
SI03K025 Lysandra philippi Greece
VL01L120 Lysandra corydonius caucasica Turkey
VL02X510 Lysandra bellargus Iran
VL03F932 Lysandra corydonius corydonius Azerbaijan
VL03F998 Lysandra sheikh Russia
VL03H615 Lysandra sheikh Russia
RV07E302 Lysandra coridon Romania
VL05N131 Lysandra corydonius corydonius Iran
JC96Q001 Lysandra bellargus Germany
AD00P045 Lysandra coridon Russia
RV04G399 Lysandra bellargus Turkey
30
JXC02G002 Lysandra coridon graeca Greece
AD00P192 Lysandra coridon borussia Russia
RV02N590 Lysandra hispana semperi Spain
MAT99Q969 Lysandra albicans arragonensis Spain
MAT99Q959 Lysandra caelestissima Spain
58
RV03H582 Lysandra albicans albicans Spain
MAT99Q966 Lysandra caelestissima Spain
KS05I822 Lysandra nufrellensis Corsica
NK02A027 Lysandra punctifera
65
KS05I875 Lysandra coridon gennargenti Sardinia
KS05I874 Lysandra coridon gennargenti
NK02A026 Lysandra punctifera Morocco
AD00P129 Lysandra bellargus Armenia
KS05I821 Lysandra nufrellensis Corsica
RE07G279 Lysandra coridon France
RV07F312 Lysandra hispana hispana Spain
VD02T008 Lysandra coridon coridon Romania
RV06A183 Lysandra coridon Andorra
74
RV04G136 Lysandra ossmar ankara Turkey
RV07F139 Lysandra syriaca burak Turkey
61
SI03K037 Lysandra philippi Greece
09X500 Lysandra dezina Kurdistan
MAT99Q932 Lysandra coridon narbonensis France
JR04G493 Lysandra coridon asturiensis Spain
RE04C165 Lysandra coridon insulana UK
MAT99T993 Lysandra hispana hispana Spain
MB05G416 Lysandra coridon apennina Italy
56
AD00P389 P. myrrha
74
AD02W109 P. amandus
AD00P081 N. diana
0.02


28S
KS05I875 *Lysandra coridon gennargenti* Sardinia
KS05I874 *Lysandra coridon gennargenti*
MAT99Q959 *Lysandra caelestissima* Spain
MAT99Q966 *Lysandra caelestissima* Spain
MAT99T993 *Lysandra hispana hispana* Spain
MB05G416 *Lysandra coridon apennina* Italy
RE04C165 *Lysandra coridon insulana* UK
RE07G279 *Lysandra coridon* France
RV04G356 *Lysandra ossmar ossmar* Turkey
RV06A183 *Lysandra coridon* Andorra
RV07C272 *Lysandra coridon asturiensis* Spain
RV07F139 *Lysandra syriaca burak* Turkey
RV07F170 *Lysandra ossmar ossmar* Turkey
RV07F312 *Lysandra hispana hispana* Spain
SH02H007 *Lysandra melamarina* Russia
SH02H010 *Lysandra melamarina* Russia
SH02H019 *Lysandra arzanovi* Russia
SH02H020 *Lysandra arzanovi* Russia
SI03K025 *Lysandra philippi* Greece
SI03K037 *Lysandra philippi* Greece
VL01L120 *Lysandra corydonius caucasica* Turkey
VL03F932 *Lysandra corydonius corydonius* Azerbaijan
VL03F998 *Lysandra sheikh* Russia
VL03H615 *Lysandra sheikh* Russia
VD02T008 *Lysandra coridon coridon* Romania
MAT99Q969 *Lysandra albicans arragonensis* Spain
RV02N590 *Lysandra hispana semperi* Spain
RV03H582 *Lysandra albicans albicans* Spain
RV07E302 *Lysandra coridon* Romania
RV04G136 *Lysandra ossmar ankara* Turkey
AD00P192 *Lysandra coridon borussia* Russia
KS05I821 *Lysandra nufrellensis* Corsica
MAT99Q882 *Lysandra bellargus* Spain
JC96Q001 *Lysandra bellargus* Germany
NK02A026 *Lysandra punctifera* Morocco
RV04G399 *Lysandra bellargus* Turkey
VL02X510 *Lysandra bellargus* Iran
87
AD00P129 *Lysandra bellargus* Armenia
NK02A027 *Lysandra punctifera*
AD00P045 *Lysandra coridon* Russia
23
KS05I822 *Lysandra nufrellensis* Corsica
09X599 *Lysandra dezina* Kurdistan
99
JXC02G002 *Lysandra coridon graeca* Greece
MAT99Q932 *Lysandra coridon narbonensis* France
40
33
JR04G493 *Lysandra coridon asturiensis* Spain
31
VL05N131 *Lysandra corydonius corydonius* Iran
AD00P435 *Lysandra corydonius caucasica* Armenia
82
AD00P389 *P. myrrha*
99
AD02W109 *P. amandus*
AD00P081 *N. diana*
0.0009


Wg
91
KS05I874 Lysandra coridon gennargenti
KS05I875 Lysandra coridon gennargenti Sardinia
50
KS05I821 Lysandra nufrellensis Corsica
RV07F139 Lysandra syriaca burak Turkey
MAT99Q966 Lysandra caelestissima Spain
RV02N590 Lysandra hispana semperi Spain
29
MAT99T993 Lysandra hispana hispana Spain
KS05I822 Lysandra nufrellensis Corsica
JXC02G002 Lysandra coridon graeca Greece
JR04G493 Lysandra coridon asturiensis Spain
MAT99Q959 Lysandra caelestissima Spain
AD00P045 Lysandra coridon Russia
MAT99Q969 Lysandra albicans arragonensis Spain
RV03H582 Lysandra albicans albicans Spain
RV07F312 Lysandra hispana hispana Spain
JC96Q001 Lysandra bellargus Germany
AD00P129 Lysandra bellargus Armenia
MAT99Q882 Lysandra bellargus Spain
90
RV04G399 Lysandra bellargus Turkey
VL02X510 Lysandra bellargus Iran
RV06A183 Lysandra coridon Andorra
RV07C272 Lysandra coridon asturiensis Spain
SI03K025 Lysandra philippi Greece
RV04G136 Lysandra ossmar ankara Turkey
66
MAT99Q932 Lysandra coridon narbonensis France
RE04C165 Lysandra coridon insulana UK
66
MB05G416 Lysandra coridon apennina Italy
RE07G279 Lysandra coridon France
SH02H010 Lysandra melamarina Russia
VL05N131 Lysandra corydonius corydonius Iran
SI03K037 Lysandra philippi Greece
RV07E302 Lysandra coridon Romania
99
NK02A026 Lysandra punctifera Moroc
NK02A027 Lysandra punctifera
9
SH02H007 Lysandra melamarina Russia
SH02H019 Lysandra arzanovi Russia
SH02H020 Lysandra arzanovi Russia
12
AD00P435 Lysandra corydonius caucasica Armenia
VD02T008 Lysandra coridon coridon Romania
RV04G356 Lysandra ossmar ossmar Turkey
VL03F932 Lysandra corydonius corydonius Azerbaijan
RV07F170 Lysandra ossmar ossmar Turkey
VL01L120 Lysandra corydonius caucasica Turkey
97
VL03F998 Lysandra sheikh Russia
VL03H615 Lysandra sheikh Russia
AD02W109 P. amandus
91
AD00P389 P. myrrha
97
AD00P081 N. diana
0.02


CAD
RV02N590 Lysandra hispana semperi Spain
MB05G416 Lysandra coridon apennina Italy
RV03H582 Lysandra albicans albicans Spain
RV07F139 Lysandra syriaca burak Turkey
SH02H019 Lysandra arzanovi Russia
VL01L120 Lysandra corydonius caucasica Turkey
VL03F932 Lysandra corydonius corydonius Azerbaijan
VL03H615 Lysandra sheikh Russia
VL05N131 Lysandra corydonius corydonius Iran
RE07G279 Lysandra coridon France
SH02H007 Lysandra melamarina Russia
SH02H020 Lysandra arzanovi Russia
RE04C165 Lysandra coridon insulana UK
RV04G356 Lysandra ossmar ossmar Turkey
MAT99Q959 Lysandra caelestissima Spain
AD00P435 Lysandra corydonius caucasica Armenia
MAT99Q932 Lysandra coridon narbonensis France
MAT99Q966 Lysandra caelestissima Spain
RV04G399 Lysandra bellargus Turkey
AD00P045 Lysandra coridon Russia
SI03K037 Lysandra philippi Greece
JC96Q001 Lysandra bellargus Germany
25
VL02X510 Lysandra bellargus Iran
AD00P129 Lysandra bellargus Armenia
VD02T008 Lysandra coridon coridon Romania
JR04G493 Lysandra coridon asturiensis Spain
MAT99T993 Lysandra hispana hispana Spain
MAT99Q969 Lysandra albicans arragonensis Spain
RV07E302 Lysandra coridon Romania
31
KS05I822 Lysandra nufrellensis Corsica
KS05I821 Lysandra nufrellensis Corsica
RV07F170 Lysandra ossmar ossmar Turkey
49
RV04G136 Lysandra ossmar ankara Turkey
RV06A183 Lysandra coridon Andorra
KS05I874 Lysandra coridon gennargenti
46
RV07C272 Lysandra coridon asturiensis Spain
KS05I875 Lysandra coridon gennargenti Sardinia
JXC02G002 Lysandra coridon graeca Greece
RV07F312 Lysandra hispana hispana Spain
77
NK02A027 Lysandra punctifera
MAT99Q882 Lysandra bellargus Spain
NK02A026 Lysandra punctifera Morocco
AD02W109 P. amandus
100
AD00P389 P. myrrha
77
AD00P081 N. diana
0.008


ITS2
KS05I874 *Lysandra coridon gennargenti*
56
KS05I875 *Lysandra coridon gennargenti* Sardinia
RV07C272 *Lysandra coridon asturiensis* Spain
RV07F312 *Lysandra hispana hispana* Spain
KS05I821 *Lysandra nufrellensis* Corsica
JR04G493 *Lysandra coridon asturiensis* Spain
RV06A183 *Lysandra coridon* Andorra
MAT99Q932 *Lysandra coridon narbonensis* France
23
RV03H454 *Lysandra coridon cataluniae* Spain
RE04C165 *Lysandra coridon insulana* UK
MAT99T993 *Lysandra hispana hispana* Spain
MAT99Q959 *Lysandra caelestissima* Spain
21
MB05G416 *Lysandra coridon apennina* Italy
KS05I822 *Lysandra nufrellensis* Corsica
RE07G279 *Lysandra coridon* France
RV07E302 *Lysandra coridon* Romania
MAT99Q966 *Lysandra caelestissima* Spain
VL03F932 *Lysandra corydonius corydonius* Azerbaijan
52
AD00P435 *Lysandra corydonius caucasica* Armenia
61
VL05N131 *Lysandra corydonius corydonius* Iran
SH02H010 *Lysandra melamarina* Russia
36
SH02H020 *Lysandra arzanovi* Russia
63
VL01L120 *Lysandra corydonius caucasica* Turkey
SH02H007 *Lysandra melamarina* Russia
RV03H582 *Lysandra albicans albicans* Spain
31
RV04G399 *Lysandra bellargus* Turkey
53
AD00P192 *Lysandra coridon borussia* Russia
68
MAT99Q882 *Lysandra bellargus* Spain
69
JC96Q001 *Lysandra bellargus* Germany
40
VL02X510 *Lysandra bellargus* Iran
NK02A027 *Lysandra punctifera*
NK02A026 *Lysandra punctifera* Morocco
59
VL03H615 *Lysandra sheikh* Russia
VL03F998 *Lysandra sheikh* Russia
RV02N590 *Lysandra hispana semperi* Spain
SI03K025 *Lysandra philippi* Greece
AD00P045 *Lysandra coridon* Russia
RV07F139 *Lysandra syriaca burak* Turkey
RV04G136 *Lysandra ossmar ankara* Turkey
35
57
RV04G356 *Lysandra ossmar ossmar* Turkey
RV07F170 *Lysandra ossmar ossmar* Turkey
100
VD02T008 *Lysandra coridon coridon* Romania
SI03K037 *Lysandra philippi* Greece
51
JXC02G002 *Lysandra coridon graeca* Greece
AD00P081 *N. diana*
100
AD02W109 *P. amandus*
0.008


Rpl5
MB05G416 *Lysandra coridon apennina* Italy
SH02H007 *Lysandra melamarina* Russia
SH02H010 *Lysandra melamarina* Russia
VL03F932 *Lysandra corydonius corydonius* Azerbaijan
VL03F998 *Lysandra sheikh* Russia
AD00P435 *Lysandra corydonius caucasica* Armenia
VL03H615 *Lysandra sheikh* Russia
VL05N131 *Lysandra corydonius corydonius* Iran
9
VL01L120 *Lysandra corydonius caucasica* Turkey
RV03H454 *Lysandra coridon cataluniae* Spain
SH02H020 *Lysandra arzanovi* Russia
SH02H019 *Lysandra arzanovi* Russia
63
MAT99Q959 *Lysandra caelestissima* Spain
20
MAT99Q966 *Lysandra caelestissima* Spain
MAT99T993 *Lysandra hispana hispana* Spain
81
KS05I821 *Lysandra nufrellensis* Corsica
66
KS05I822 *Lysandra nufrellensis* Corsica
43
MAT99Q932 *Lysandra coridon narbonensis* France
RE04C165 *Lysandra coridon insulana* UK
21
AD00P192 *Lysandra coridon borussia* Russia
37
SI03K037 *Lysandra philippi* Greece
JXC02G002 *Lysandra coridon graeca* Greece
RV03H582 *Lysandra albicans albicans* Spain
11
RV07C272 *Lysandra coridon asturiensis* Spain
KS05I874 *Lysandra coridon* gennargenti
JR04G493 *Lysandra coridon asturiensis* Spain
RV07F312 *Lysandra hispana hispana* Spain
33
RV02N590 *Lysandra hispana semperi* Spain
VD02T008 *Lysandra coridon coridon* Romania
NK02A027 *Lysandra punctifera*
12
NK02A026 *Lysandra punctifera* Morocco
JC96Q001 *Lysandra bellargus* Germany
60
MAT99Q882 *Lysandra bellargus* Spain
59
RV04G399 *Lysandra bellargus* Turkey
33
AD00P129 *Lysandra bellargus* Armenia
RV07F139 *Lysandra syriaca burak* Turkey
0.02

Conclusions

CONCLUSIONS

Along the different chapters of this thesis it has been shown the potential of phylogenetic methodologies based on molecular data, combined with the use of data from additional sources, to deal with a broad variety of questions, in this case regarding the insects. The following conclusions are obtained:

- It is demonstrated, for the first time, that with a proper methodology, mitochondrial data can be reconciled with the most generally accepted hypotheses on inter-ordinal relationships of insects based on morphological and nuclear data. Thus, it is confirmed the non-monophyly of Homoptera within Paraenoptera and the Strepsiptera as a sister order to Coleoptera. More controversial, mitogenomes support the Hymenoptera-Mecopterida association and a Psocodea paraphyly.
- The use of amino acid sequences instead of DNA is more appropriate at the inter-ordinal level, and the use of the site-heterogeneous mixture model (CAT) under a Bayesian framework substantially avoids LBA artifacts. Inferring phylogenies above the super-order level of insects constitutes the limit of the phylogenetic signal contained in insect mitochondrial genomes for currently available methods.

- The generic phylogenetic relationships within *Polyommatina* have been resolved for first time after reconstructing a comprehensive multilocus molecular phylogeny for all genera and subgenera described.
- The estimation of molecular ages have helped to rearrange the higher-level taxonomy of *Polyommatina* by establishing a set criteria of criteria based on delimiting a flexible temporal frame that accommodates phylogenetic uncertainty, a distinguishing combination of morphological characters for each genus, and certain levels of current taxonomic designations.

- The GMYC model for species delimitation successfully identified 80% of the species in our dataset, but that percentage can be elevated to 90% if not accounting for intrinsic data limitations such as non-monophyly.

- The GMYC model is remarkably stable under a wide array of circumstances, including high singleton presence (up to 95%), taxon level (as low as between three to five species), and presence of gaps in intraspecific sampling coverage (removal of all intermediate haplotypes).

- *Lasius balearicus* is demonstrated to be a new species of ant from Mallorca island high elevations, given morphological, molecular and ecological evidence. It represents the first endemic ant from the Balearic Islands.
- A extremely low intraspecific genetic diversity, total range and nest densities documented, coupled with dramatic future predictions without the apparent possibility of dispersal to suitable habitat due to geographic and altitudinal isolation, suggest a high probability of short-term extinction for *Lasius balearicus*.

- Out of eight *Agrodiaetus* taxa with dot-like distributions only two represent valid species (*A. humedasae* and *A. pljushtchi*), while four of them turned out to be synonyms of *ripartii* and two of them include other taxa and have a wider distribution. Our results do no support the current designations of *A. galloi*, *A. exuberans* and *A. agenjoi* as endemic species. *A. violetae* is shown to be a polytypic species consisting of at least two subspecies. *A. violetae* is genetically (but not chromosomally) distinct from *A. fabressei* and has a wider distribution in southern Spain than previously believed.
- The overestimation of biological diversity can have negative implications for conservation efforts. *A. galloi* should be excluded from the IUCN Red List of Threatened Species as we show that it represents an isolated population the widely distributed species *A. ripartii*.
- The *Agrodiaetus ripartii* lineage is karyotypically undifferentiated and its origin is more recent than the *A. dolus* lineage (karyotypically variable). Both lineages display similar biogeographical histories at different time ranges, involving a combination of dispersal and vicariance events from

Asia to western Europe that imply that the *ripartii* lineage most likely encountered Europe already populated by *A. dolus* representatives.

- The genus *Lysandra* forms a monophyletic clade with extremely low morphological and molecular variability, and phylogenetic signal is vastly blurred due to mixed incomplete lineage sorting and frequent hybridization. It is shown that not accounting for potential hybridization in species tree inference can lead to precipitate taxonomic conclusions.
- *Lysandra coridon* displays two paraphyletic lineages between western and eastern individuals, most likely because of being seriously influenced by massive introgression events between western specimens and the Iberian taxa. The large estimated population sizes for *L. coridon* may play a crucial role on maintaining the viability of different chromosomal races within the species.
- Chromosomal instability within *Lysandra* originated at least on three occasions. A new chromosomic form (n=30) is found for *L. syriaca*, typically n=24. All the taxa from the caucasian group *corydonius* share the same chromosome number (n=84), for which are considered to conform a single species.

Annexes

- Other publications -

Syst. Biol. 56(4):564–577, 2007
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150701472164

Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments

GERARD TALAVERA AND JOSE CASTRESANA

Department of Physiology and Molecular Biodiversity, Institute of Molecular Biology of Barcelona, CSIC, Jordi Girona 18, 08034 Barcelona, Spain; E-mail: jcvagr@ibmb.csic.es (J.C.)

Abstract.—Alignment quality may have as much impact on phylogenetic reconstruction as the phylogenetic methods used. Not only the alignment algorithm, but also the method used to deal with the most problematic alignment regions, may have a critical effect on the final tree. Although some authors remove such problematic regions, either manually or using automatic methods, in order to improve phylogenetic performance, others prefer to keep such regions to avoid losing any information. Our aim in the present work was to examine whether phylogenetic reconstruction improves after alignment cleaning or not. Using simulated protein alignments with gaps, we tested the relative performance in diverse phylogenetic analyses of the whole alignments versus the alignments with problematic regions removed with our previously developed Gblocks program. We also tested the performance of more or less stringent conditions in the selection of blocks. Alignments constructed with different alignment methods (ClustalW, Mafft, and Probcons) were used to estimate phylogenetic trees by maximum likelihood, neighbor joining, and parsimony. We show that, in most alignment conditions, and for alignments that are not too short, removal of blocks leads to better trees. That is, despite losing some information, there is an increase in the actual phylogenetic signal. Overall, the best trees are obtained by maximum-likelihood reconstruction of alignments cleaned by Gblocks. In general, a relaxed selection of blocks is better for short alignment, whereas a stringent selection is more adequate for longer ones. Finally, we show that cleaned alignments produce better topologies although, paradoxically, with lower bootstrap. This indicates that divergent and problematic alignment regions may lead, when present, to apparently better supported although, in fact, more biased topologies. [Bootstrap support; Gblocks; phylogeny; sequence alignment.]

Methods for the simultaneous generation of multiple alignments and phylogenetic trees are actively being pursued (Fleissner et al., 2005; Lunter et al., 2005; Redelings and Suchard, 2005; Wheeler, 2001), but, at present, common practice of phylogenetic analysis requires, as a first step, the generation of a multiple alignment of the sequences to be analyzed. It has been repeatedly shown that the quality of the alignment may have an enormous impact on the final phylogenetic tree (Kjer, 1995; Morrison and Ellis, 1997; Ogden and Rosenberg, 2006; Smythe et al., 2006; Xia et al., 2003). This is particularly true when sequences compared are very divergent and of different length, which makes necessary the introduction of gaps in the alignments.

Due to the computational requirements of optimal algorithms for multiple sequence alignments, different heuristic strategies have been proposed. The most widely used approach has been the progressive method of alignment (Feng and Doolittle, 1987) that, together with enhancements related to the introduction of gap penalties, was implemented in ClustalW (Thompson et al., 1994). In progressive methods, an initial dendrogram generated from the pairwise comparisons of the sequences is used to recursively build the multiple alignment, using dynamic programming (Needleman and Wunsch, 1970) in the last step. Dynamic programming is an exact algorithm that assures the best possible alignments for given gap penalties but, due to heavy computational requirements, it is only used for pairs of sequences or pairs of clades of the dendrogram and not for the whole multiple alignment. Several other heuristic multiple alignment methods have been recently introduced. They include T-Coffee (Notredame et al., 2000), Mafft (Katoh et al., 2005; Katoh et al., 2002), Muscle (Edgar, 2004), Probcons (Do et al., 2005), and Kalign (Lassmann and Sonnhammer, 2005), among others. All of them are based on the progressive method but include several iterative refinements to construct the final multiple alignment. The latter methods have been shown to outperform purely progressive methods in terms of alignment accuracy and, some of them, even in computational time. However, it has not been shown whether the greater alignment accuracy of more sophisticated methods leads to a significant improvement in phylogenetic reconstruction.

Proteins have some regions that, due to their functional or structural importance, are very well conserved, whereas other regions evolve faster both in terms of nucleotide substitutions and insertions or deletions (Henikoff and Henikoff, 1994; Herrmann et al., 1996; Pesole et al., 1992). That is, evolutionary rate heterogeneity affects to whole regions in addition to single positions. This type of regional rate heterogeneity is very challenging for phylogenetic reconstruction, not only in terms of homoplasy due to saturation (Yang, 1998), but also in terms of errors in homology during alignment.

Dealing with regions of problematic alignment is a matter of active debate in phylogenetics. Although some authors consider that it is best to remove such regions before the tree analysis (Castresana, 2000; Grundy and Naylor, 1999; Löytynoja and Milinkovitch, 2001; Rodrigo et al., 1994; Swofford et al., 1996), others think that there is an important loss of information upon removal of any fragment of the sequences already obtained (Aagesen, 2004; Lee, 2001) and that this practice should only be used as the last resource (Gatesy et al., 1993). A third, intermediate option, is the recoding of such regions using different strategies (Geiger, 2002; Lutzoni et al., 2000; Young and Healy, 2003), which allows the use of at least part of the information. Although these coded characters are most commonly analyzed with parsimony, it is also possible to use them as independent partitions in Bayesian or likelihood frameworks.

In the present work we test, by using simulated protein alignments with gaps, which are the best alignment strategies for optimal phylogenetic reconstruction. Two preliminary considerations are necessary here. First, simulations of sequences may not cover all the complexity of evolution but have the advantage over real sequences that we know the tree from which they have been generated. There are some alignment sets curated from structural information that can be used to test alignment accuracy (Thompson et al., 2005), but the phylogenetic tree is unknown in these sets, thus making problematic their use for proving phylogenetic accuracy. Second, we have been working with simulated sequences that try to reflect the evolutionary patterns of proteins, and thus many of the conclusions extracted from our work cannot be directly extrapolated to other markers such as rRNA, which show very different evolutionary constraints (Gutell et al., 1994; Kjer, 1995; Xia et al., 2003).

In our analysis we used different alignment strategies of the simulated sequences to test if they make any difference in the final phylogenetic tree. We have selected ClustalW as the currently most used progressive alignment method (Thompson et al., 1994) and Mafft (Katoh et al., 2005) and Probcons (Do et al., 2005) as examples of more recently developed methods that have been shown to obtain very high scores in terms of alignment accuracy (Blackshields et al., 2006; Nuin et al., 2006). Simultaneously with the performance of the alignment programs, we tested whether removing blocks of problematic alignment actually leads to more accurate trees. We used for this purpose our previously developed Gblocks program (Castresana, 2000), which selects blocks following a reproducible set of conditions. Briefly, selected blocks must be free from large segments of contiguous nonconserved positions, and flanking positions must be highly conserved to ensure alignment accuracy. Several parameters can be modified to make the selection of blocks more or less stringent. Phylogenetic trees made by maximum likelihood (ML), neighbor joining (NJ), and parsimony of the reconstructed alignments show that, in almost all conditions tested, and at least for alignments that are not too short, the elimination of problematic regions by Gblocks leads to significantly better phylogenetic trees.

Materials and Methods

We simulated protein sequences by means of Rose (Stoye et al., 1998). This program allows the simulation of different substitution rates in different positions with a predetermined spatial pattern. This is a very important feature for testing the behavior of a program like Gblocks, which selects from alignments blocks of contiguous conserved positions with few nonconserved positions inside. This is the reason why a program that simulates among-site rate heterogeneity, but not regional heterogeneity, would not be valid to test the behavior of Gblocks. Thus, an important preliminary step in our simulations was the selection from real proteins of spatial patterns of site rates in order to use these parameters with Rose.

Selection of Evolutionary Rate Patterns

We extracted patterns of rate heterogeneity from real protein alignments using the program TreePuzzle (Strimmer and von Haeseler, 1996) with a model of among-site rate heterogeneity that assumed a Gamma distribution of rates. This distribution was approximated with 16 rate categories, which is the maximum number allowed in TreePuzzle. In particular, we took, from each position, the category and associated relative rate that contributed the most to the likelihood. Positions with rates >1 receive more mutations than the average and positions with rates <1 receive fewer mutations. This list of relative rates (whose average should be 1) were given to Rose to simulate different positions with different rates, creating conserved and divergent regions with lengths and boundaries that approximated those of a real protein. Proteins for extracting rate patterns were NAD2 and NAD4 (subunits 2 and 4 of the mitochondrial NADH dehydrogenase) from several metazoans (Castresana et al., 1998b), and COG0285 from the COG database, which includes mainly bacterial sequences (Tatusov et al., 2003). The three selected profiles produced similar conclusions regarding the best block selection strategy, and we used the NAD2 pattern to perform most of the tests. This pattern contained 361 positions but, after the introduction of further gaps by the simulation algorithm, the final simulated alignments reached approximately 400 positions. In order to simulate alignments of different length, independent simulations obtained with this pattern were concatenated 1, 2, 3, 4, and 8 times to generate final alignments of, approximately, 400, 800, 1200, 1600, and 3200 positions, respectively. The PAM evolutionary model (Dayhoff et al., 1978) was used to simulate the evolution of amino acids.

Selection of Phylogenetic Trees

Simulations with Rose were performed along phylogenetic trees of 16 tips with three different topologies, a purely asymmetric tree (Fig. 1a), an intermediate tree (Fig. 1b), and a symmetric tree (Fig. 1c). These known trees or "real trees" were manually constructed. The average and maximum length from the root to the tips was, for the asymmetric tree, 0.89 and 1.30 substitutions/position, respectively. The other trees had very similar values. The branch lengths of the three trees in Figure 1 were multiplied by factors of 0.5, 1, and 2, respectively, so that we used in total 9 phylogenetic trees. These trees had several short internal branches that made them difficult to resolve; thus, they are trees where the alignment strategy as well as the phylogenetic algorithm used were differentially effective. Simpler trees in terms of longer internodes were easily and equally reproduced by all methods and were not used here. Similarly, trees with a total smaller divergence tended to produce conserved alignments where the alignment method was not an issue and also not used here. Finally, these trees did

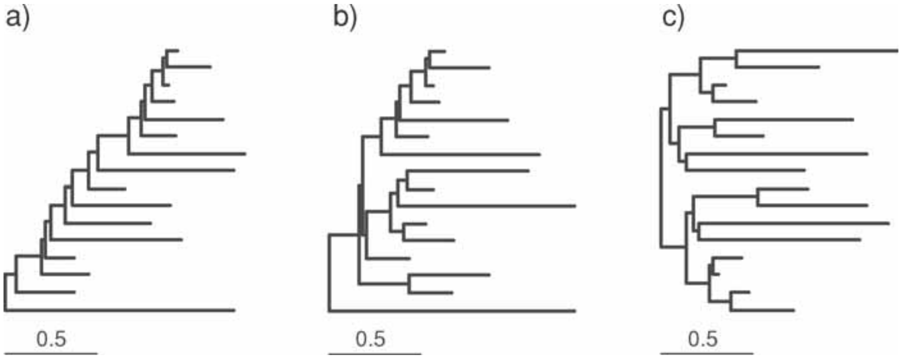


FIGURE 1. Asymmetric (a), intermediate (b), and symmetric (c) trees used in the simulations. The scale bar, in substitutions/position, corresponds to the trees with a divergence ×1.

not contain many closely related sequences, since we wanted to specifically measure differences in reproducing the overall shape of the tree and not differences in recovering the relationships among close sequences.

Gaps Introduced during the Simulations

The Rose program does not have any specific model for the introduction of gaps along the alignment. Rather, gaps are introduced with equal probability in all positions with a relative rate ≥ 1 (Stoye et al., 1998), which is a limitation of this program. To try to overcome this limitation, we used two different gap strategies within Rose. First, we used a single gap threshold for the whole alignment. After several trials, we considered a threshold of 0.0007 as a reasonable one for the divergence levels we analyzed, as deduced from visual inspection of the alignment (that is, eyeing that blocks of divergence and conservation were not so different from the real proteins used to construct the rate profiles). Even so, this threshold tended to produce too many gaps in conserved regions (not shown). In addition, we also generated alignments with two different gap thresholds, 0.001 and 0.0001, which we associated, respectively, to divergent and to conserved regions of the profiles. For doing so, we divided the rate profiles in blocks of homogeneous divergence (that is, each block was either mostly conserved or mostly divergent, which resulted in around 10 to 20 blocks for the different profiles). Then, we did the simulations for each block separately, and with its own gap threshold (high for divergent blocks and low for more conserved blocks). Finally, the different simulated blocks were concatenated. The phylogenetic results were similar with both gap strategies, but we mostly worked with simulations that had the two different gap thresholds, which we considered more realistic. In all cases we chose a vector of indels of the form [0.5, 0.4, 0.3, 0.2, 0.1], which reflects the relative frequency of indels with lengths from 1 to 5 amino acids, respectively.

Realignments of Simulated Sequences

Alignments generated by Rose were cleaned from gaps and new alignments were reconstructed using ClustalW version 1.83 (Thompson et al., 1994), Mafft version 5.531 (Katoh et al., 2002, 2005), and Probcons version 1.1 (Do et al., 2005). Default parameters were used in ClustalW and Probcons. All defaults were also used in Mafft except that a neighbor joining instead of a UPGMA tree was used as guide tree (option –nj). Alignments were cleaned from problematic alignment blocks using Gblocks 0.91 (Castresana, 2000), for which two different parameter sets were used. In one of them, which we call here stringent selection, and which is the default one in Gblocks 0.91, “Minimum Number of Sequences for a Conserved Position” was 9, “Minimum Number of Sequences for a Flank Position” was 13, “Maximum Number of Contiguous Nonconserved Positions” was 8, “Minimum Length of a Block” was 10, and “Allowed Gap Positions” was “None”. In the second set, which we call relaxed selection, we changed “Minimum Number of Sequences for a Flank Position” to 9, “Maximum Number of Contiguous Nonconserved Positions” to 10, “Minimum Length of a Block” to 5, and “Allowed Gap Positions” to “With Half”. The latter option allows the selection of positions with gaps when they are present in less than half of the sequences.

Original simulated alignments and Mafft realignments for 30 example simulations (the first five simulations generated with the symmetric and asymmetric trees) are provided as supplementary information (available online at http://systematicbiology.org).

Phylogenetic Reconstruction

Phylogenetic trees from the complete and the two different Gblocks alignments were estimated by ML, NJ, and parsimony. For ML trees we used the Phyml program version 2.4.4 (Guindon and Gascuel, 2003), with the Jones-Taylor-Thornton model of protein evolution (Jones et al., 1992) and four rate categories in the Gamma distribution. The Gamma distribution parameter and the proportion of invariable sites were estimated by the program. For NJ trees we used Protdist of the Phylip package version 3.63 (Felsenstein, 1989) with the Jones-Taylor-Thornton model to calculate pairwise protein distances, and Neighbor of the same package to calculate the NJ tree. For parsimony we used Protpars of the Phylip package (Felsenstein, 1989) with 50 random initializations to ensure a thorough tree search. If no parsimony tree was obtained, which occurred in less than 1% of the simulations, the corresponding simulation was totally excluded from the analysis. When several equally parsimonious trees were found, only the first one was used. We did not do Bayesian trees because of the enormous computational time required for doing enough number of generations of all simulations performed.

For each alignment length, alignment strategy, and phylogenetic method, 300 simulations were run in a grid of 24 processors. The symmetric difference or Robinson-Foulds (Robinson and Foulds, 1981) topological distance from the calculated tree to the real tree was obtained using Vanilla 1.2 (Drummond and Strimmer, 2001), and the average of all simulations calculated. This program reports half the number of total discordant clades between two trees. For bootstrap analyses, 100 bootstraps were calculated. Due to heavy computational requirements of the bootstrap analyses, the number of simulations was reduced to 150. We checked that a higher number of bootstraps and simulations did not improve the accuracy of the bootstrap results. Bootstrap values were separately calculated for right and wrong partitions of the tree with the help of Bioperl functions (Stajich et al., 2002). Statistical differences among Robinson-Foulds distances in different alignment conditions were detected by the Tukey-Kramer test with an alpha level of 0.05 using the JMP package version 5.1 (SAS Institute, Cary, NC).

Results and Discussion

General Alignment Strategy: Complete versus Gblocks Alignments

The differences in alignments produced by different methods can be appreciated in Figure 2. A fragment of the alignment of simulated sequences (Fig. 2a) was stripped of gaps and realigned by ClustalW (Fig. 2b), Mafft (Fig. 2c), and Probcons (Fig. 2d). As it has been noted before (Higgins et al., 2005), ClustalW tends to produce more compact alignments. That is, ClustalW generates many divergent regions that are almost devoid of gaps, resulting in a relatively simple alignment (Higgins et al., 2005). This can be clearly appreciated in the most problematic region in the center of this alignment (Fig. 2b). Although Mafft also tends to make alignments more compact than the real ones (Fig. 2c), the deviation from the real situation is not as large as with ClustalW, at least with default gap penalties. Probcons produces the least compact alignments of the three programs tested (Fig. 2d). For example, simulations from asymmetric trees with divergence ×1, which had an average original length of 1097 positions, were compacted to an average of 966 positions by Probcons, to 904 positions by Mafft and to 862 positions by ClustalW (Table 1). Similar relative degrees of compression were obtained in other types of simulations.

Gblocks removes problematic regions of a multiple alignment according to a number of rules. First, blocks selected for inclusion must be free from a large number of contiguous nonconserved positions, must be flanked by highly conserved positions, and must have a minimum length, as controlled by the corresponding parameters (see Materials and Methods). In addition, positions with gaps can be removed either always or only when more than half of the sequences contain gaps (Castresana, 2000). The latter parameter has a large influence on the total number of selected positions. We have used Gblocks in simulated realigned sequences with two different conditions. The condition that we call stringent does not allow any gap position. The relaxed condition allows gap positions if they are present in less than half of the sequences, and it is also less restrictive in the other parameters (see Materials and Methods). The effect of the two different parameter sets of Gblocks selection can be appreciated in Figure 2, for ClustalW (Fig. 2b), Mafft (Fig. 2c), and Probcons alignments (Fig. 2d). In both cases, the relaxed parameters (grey blocks) allow the selection of more positions than the stringent parameters (white blocks). Table 1 shows the average number of positions of the complete alignments and the percentage of positions left after treatment with Gblocks with the two different parameter sets. Values in this table are for the asymmetric tree, but similar values were found for other trees.

In order to infer which type of alignment algorithm (ClustalW, Mafft, or Probcons) and which treatment of the resulting alignment (no treatment or Gblocks treatment with stringent or relaxed conditions) was best for phylogenetic analysis, we calculated phylogenetic trees from all these alignments, and measured the topological distance with respect to the real tree. Figure 3 shows, for the simulations with the asymmetric tree, the average topological distances to the real tree from the trees generated with ClustalW alignments, with and without the use of Gblocks. In addition, the distance to the tree obtained from the Gblocks complementary alignment (that is, the alignment resulting after concatenation of all the blocks rejected by Gblocks) is also shown.

TABLE 1. Average number of positions of the complete alignments and the average percentage of positions selected by Gblocks with relaxed and stringent conditions. Simulation of sequences was done following the asymmetric tree and the heterogeneity pattern of the NAD2 protein concatenated two times.

	ClustalW			Mafft			Probcons		
Divergence	Total length	% Gblocks relaxed	% Gblocks stringent	Total length	% Gblocks relaxed	% Gblocks stringent	Total length	% Gblocks relaxed	% Gblocks stringent
×0.5	826.6	79.4	54.3	852.5	74.2	51.6	871.8	70.3	50.9
×1	862.4	64.2	42.0	903.7	59.0	39.8	966.4	51.8	37.6
×2	901.8	46.4	30.2	961.7	42.9	28.4	1117.9	34.7	24.5

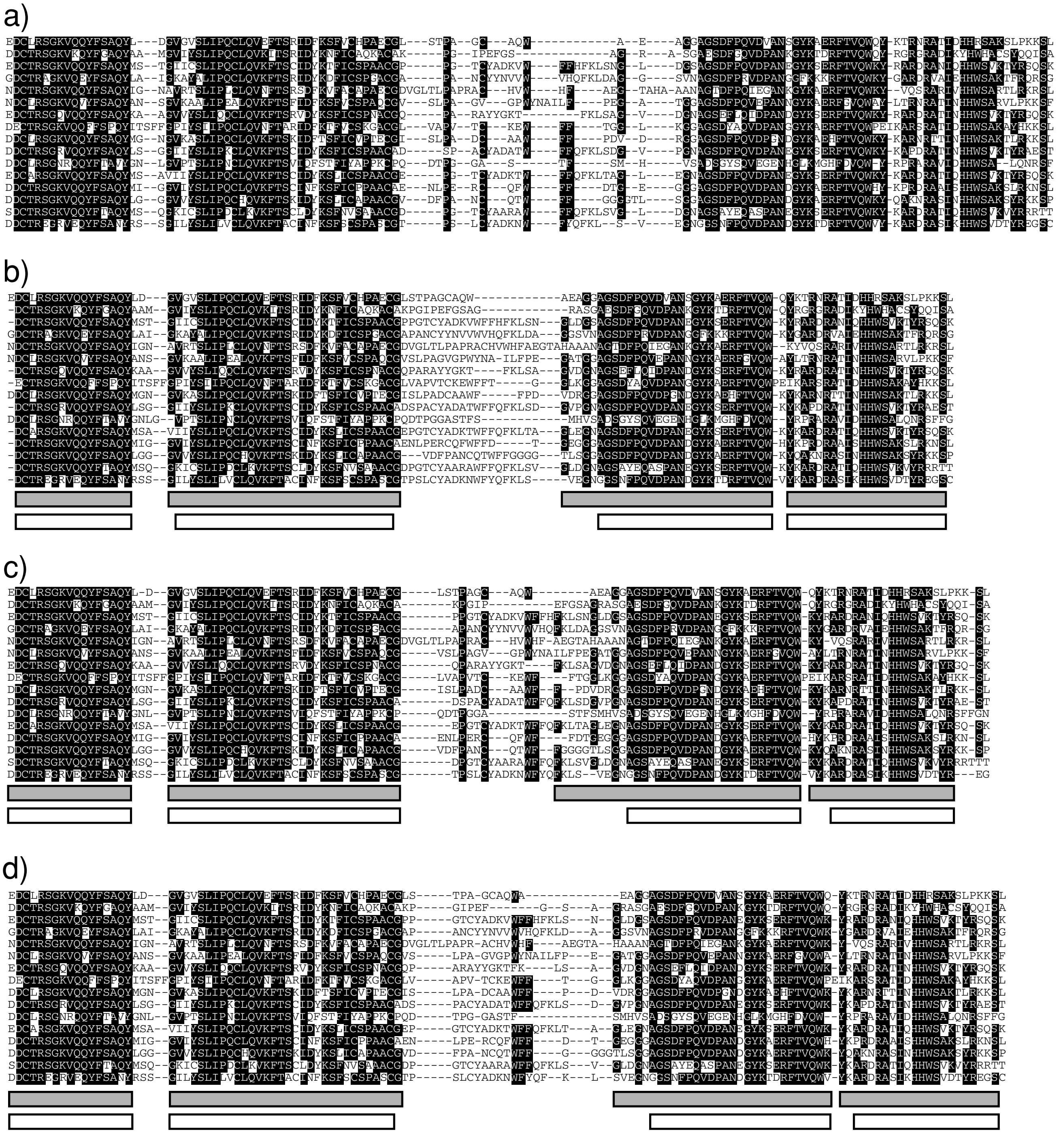


FIGURE 2. Fragment of a simulated alignment (a) and the realignment of the same sequences (after gap removal) by ClustalW (b), Mafft (c), and Probcons (d). The simulation corresponds to an asymmetric tree with divergence ×1. The blocks below each alignment represent the fragments selected by Gblocks with relaxed conditions (grey blocks) and with stringent conditions (white blocks). Positions of the alignments where more than 50% of the sequences are identical are shown with black boxes.

Figure 4 represents for each tree (and for two representative lengths, 800 and 3200 amino acids, as representatives of single-gene and concatenated-gene phylogenies) the best alignment strategies after statistically comparing the average topological distances by means of the Tukey-Kramer test. An overview of these two figures shows that, when the alignments are cleaned by Gblocks with any of the two parameter sets used (dotted lines in Figure 3), the topological distance to the real tree decreases with respect to the complete alignment (solid, red line) in almost all divergences and alignment lengths tested, and with the three tree reconstruction methods used:

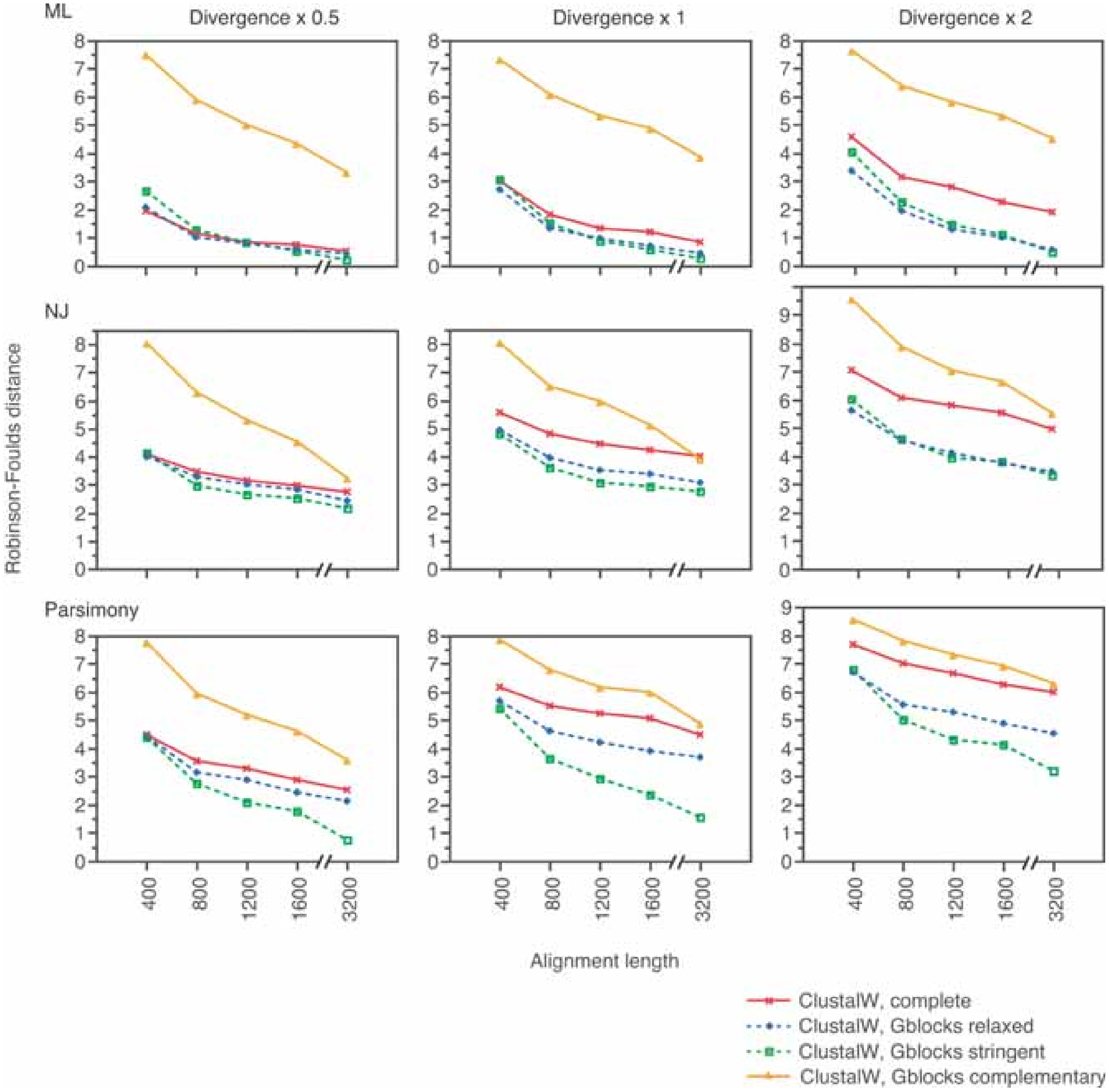
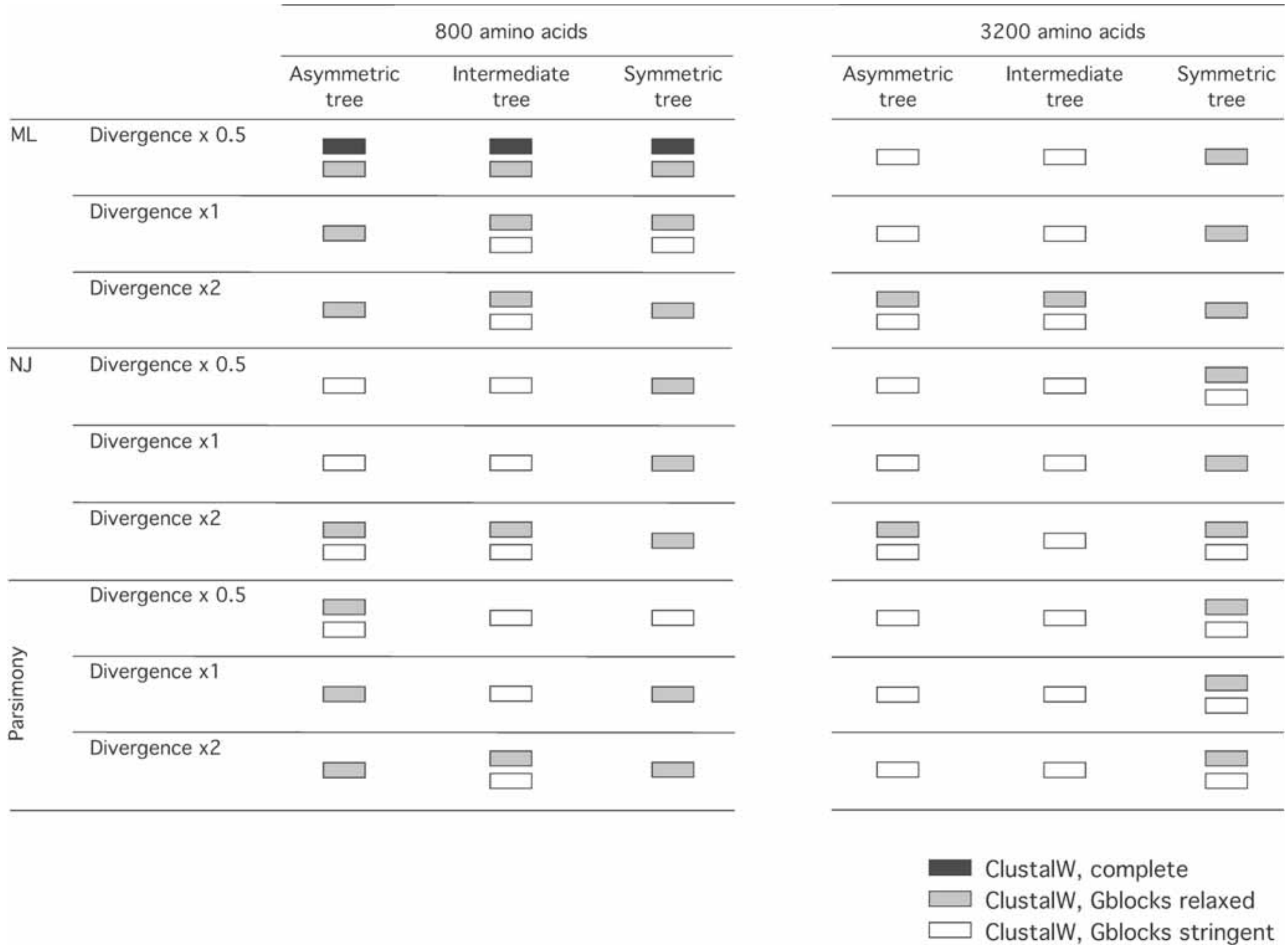


FIGURE 3. Average Robinson-Foulds distances to the real tree from the tree calculated with ClustalW complete alignments (solid, red line with crossed symbols), the same alignments after treatment with Gblocks relaxed (dotted, blue line with diamonds) and stringent (dotted, green line with squared symbols) conditions, and the complementary alignments of the Gblocks relaxed alignment (solid, orange line with triangles). The asymmetric tree with three different divergence levels was used for the simulations with different alignment lengths. Trees were reconstructed by ML, NJ, and parsimony.

ML, NJ, and parsimony. The improvement in topological accuracy upon Gblocks treatment is more noticeable for the highest divergences (×2). This is expected since there are more problematic blocks in these alignments, as shown by the lower percentage of positions selected by Gblocks (Table 1). In addition, the improvement from Gblocks treatment is particularly large for NJ and parsimony. These two methods produce quite poor topologies when using the complete alignments but, upon using Gblocks, particularly with the most stringent conditions (green line, squared symbols), there is a substantial gain in topological accuracy. ML produces the overall best trees (see also below) although, in the lowest divergence (×0.5), there is almost no difference in topological quality between the Gblocks and the complete alignments. In fact, for short genes (400 to 800 amino acids) the complete alignment gives rise to better trees than the Gblocks alignments, although there is no statistical difference between the complete alignment and the Gblocks alignment with relaxed parameters (Fig. 4).



		800 amino acids			3200 amino acids		
		Asymmetric tree	Intermediate tree	Symmetric tree	Asymmetric tree	Intermediate tree	Symmetric tree
ML	Divergence x 0.5	■ ▨	■ ▨	■ ▨	□	□	▨
	Divergence x1	▨	▨ □	▨ □	□	□	▨
	Divergence x2	▨	▨ □	▨	▨ □	▨ □	▨
NJ	Divergence x 0.5	□	□	▨	□	□	▨ □
	Divergence x1	□	□	▨	□	□	▨
	Divergence x2	▨ □	▨ □	▨	▨ □	□	▨ □
Parsimony	Divergence x 0.5	▨ □	□	□	□	□	▨ □
	Divergence x1	▨	□	▨	□	□	▨ □
	Divergence x2	▨	▨ □	▨	□	□	▨ □

■ ClustalW, complete
▨ ClustalW, Gblocks relaxed
□ ClustalW, Gblocks stringent

FIGURE 4. ClustalW alignment strategies that give rise to the statistically best topologies. When two or more strategies do not show statistical differences in Robinson-Foulds distances, all equivalent strategies are represented. The complete alignment is represented by a black block, and the relaxed and stringent Gblocks strategies by grey and white blocks, respectively.

It is thus shown from the example above that the removal of divergent and problematic regions of an alignment is, in principle, beneficial for phylogenetic analyses of relatively divergent sequences. In fact, it is true, as previously argued (Aagesen, 2004; Lee, 2001), that there is some phylogenetic information in the blocks removed by methods like Gblocks. This can be appreciated in Figure 3, which shows the topological distances to the real trees from the trees obtained with the blocks excluded by Gblocks (complementary alignment; solid, orange line). These distances, although very large, become quite reduced for long alignments, indicating that trees obtained from the complementary regions are not random; that is, there is some phylogenetic information in the regions rejected by Gblocks. However, what seems to matter is not the total phylogenetic signal but the signal-to-noise ratio. Despite the relatively simple simulations performed, regions excluded by Gblocks seem to add more noise than signal, thus lowering the quality of the trees from the complete alignments with respect to the Gblocks-cleaned alignments.

Similar conclusions about the beneficial effect of Gblocks can be drawn from Mafft alignments of the same asymmetric trees (Figs. 5 and 6). In this case, Gblocks is not an advantage over the complete alignment in the two most conserved alignments (×0.5 and ×1) when using the ML method although, again, Gblocks relaxed and the complete alignments are not statistically different. The picture for Probcons (Fig. 1 of the online Appendix, available at http://systematicbiology.org) is similar to that for Mafft. Figure 2 of the online Appendix shows a comparison of the three alignment programs with default gap costs, using the trees produced after Gblocks cleaning with relaxed conditions. Under the conditions of these simulations, ClustalW is slightly worse, regarding the trees produced, than the two other programs. The performances of Mafft and Probcons are very similar, and only for NJ and parsimony Probcons alignments work slightly better. Probcons, however, is highly demanding in computational time. Thus, for the rest of the tests we only compared the performances of ClustalW and Mafft.

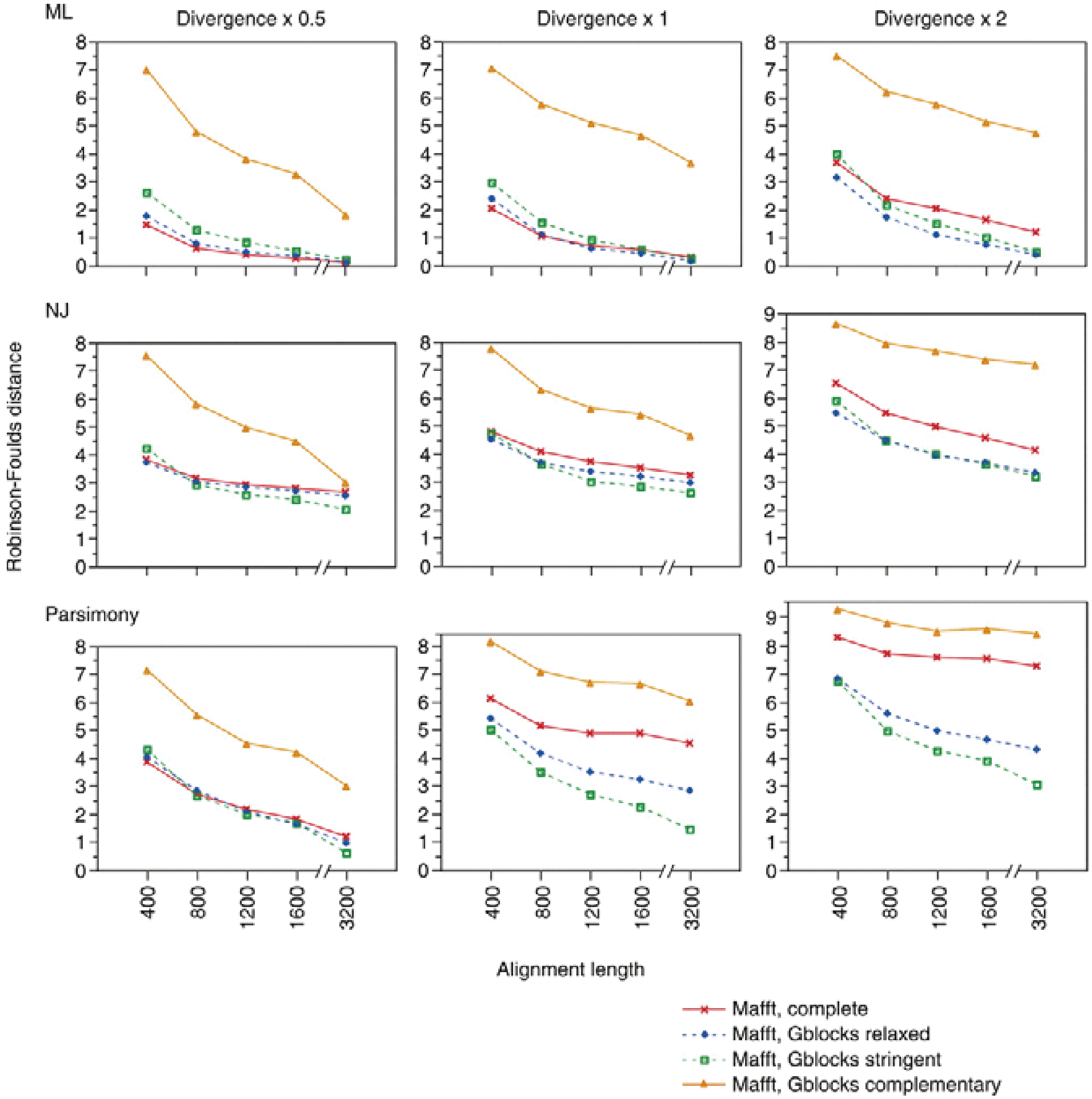


FIGURE 5. Average Robinson-Foulds distances to the real tree from the tree calculated with Mafft complete alignments (solid, red line with crossed symbols), the same alignments after treatment with Gblocks relaxed (dotted, blue line with diamonds) and stringent (dotted, green line with squared symbols) conditions, and the complementary alignments of the Gblocks relaxed alignment (solid, orange line with triangles). The asymmetric tree with three different divergence levels was used for the simulations with different alignment lengths. Trees were reconstructed by ML, NJ, and parsimony.

The results for the symmetric and intermediate trees of both alignment algorithms are shown in the corresponding columns of Figures 4 and 6 for the ClustalW and Mafft methods, respectively (and in Figures 3 to 6 in the online Appendix for all alignment lengths). Two results are noteworthy from these analyses. First, differences in phylogenetic performance between different alignments derived from symmetric trees are quantitatively smaller, in agreement with a previous work (Ogden and Rosenberg, 2006). See, for example, the similarity of the three graphs of ML trees of ClustalW alignments (Fig. 3 in the online Appendix). Second, in these trees there are two conditions where the Gblocks alignments produce ML trees that are statistically worse than the complete alignments: the symmetric and intermediate trees of divergence ×1 with Mafft alignments of 800 amino acids (Fig. 6). These are the only two conditions where we observed this. However, we do not think that this justifies

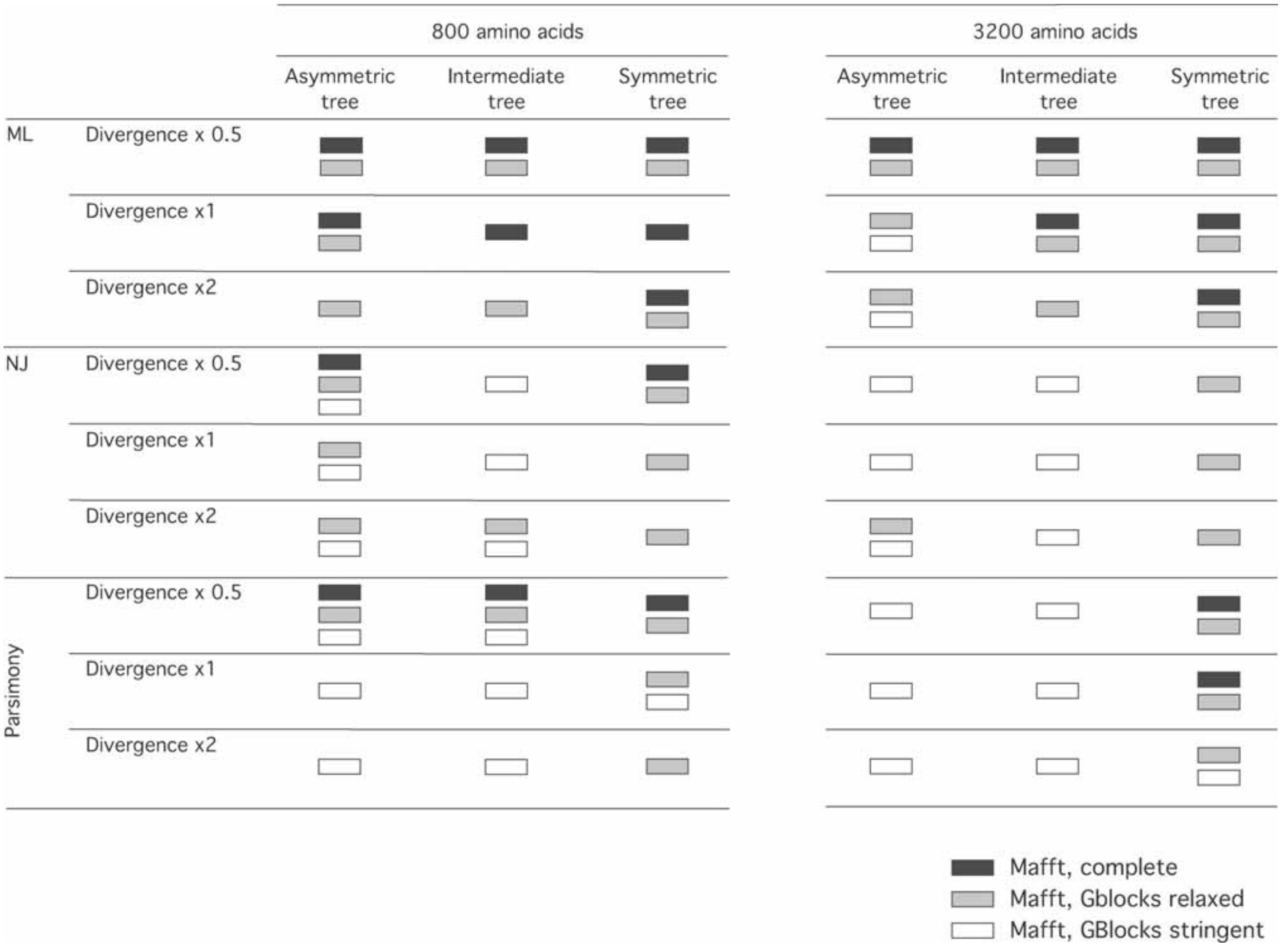


FIGURE 6. Mafft alignment strategies that give rise to the statistically best topologies. When two or more strategies do not show statistical differences in Robinson-Foulds distances, all equivalent strategies are represented. The complete alignment is represented by a black block, and the relaxed and stringent Gblocks strategies by grey and white blocks, respectively.

not using Gblocks in these types of trees, even if we could know the shape of the tree in advance. In real alignments, evolution must be much more complex than what we simulated. For example, we did not simulate biased amino acid compositions (Castresana et al., 1998a) or different models of evolution in different parts of trees (Philippe and Laurent, 1998), all of which will have stronger biasing effects in nonconserved blocks. Because the difference in topological accuracy between the Gblocks and the complete alignments is very small in these two conditions, it is very likely that the addition of any of these effects in the simulations would have made both the Gblocks relaxed and complete alignments of at least equal performance.

All simulations shown so far were performed following a pattern of rate variation of the NAD2 protein. To test the influence of different rate patterns, we used in the simulations profiles derived from two other proteins (NAD4 and COG0285). From the Mafft alignments of these simulations we calculated the corresponding ML trees (Fig. 7 in the online Appendix). Different patterns (and thus different percentages of block selection) gave rise to different performances of the complete and the Gblocks alignments, but the results were similar in relative terms. We also tested the performance of a different gap model, in which gaps were introduced homogeneously along the alignment, instead of using two different gap thresholds in different regions of the alignments (see Materials and Methods). The results were again similar with the simpler gap strategy, as shown for the ML reconstruction of the asymmetric trees (Fig. 8 of the Online appendix).

Phylogenetic Methods Used

The data shown above indicate that ML is the phylogenetic method that best extracts reliable information from problematic alignment regions, since trees derived from complete alignments are relatively good. This contrasts with the trees obtained by NJ and parsimony, which are quite poor from the complete alignments, indicating that they greatly benefited from the use of Gblocks. ML is also the method that produces the overall best trees, in agreement with previous simulation analysis (see references

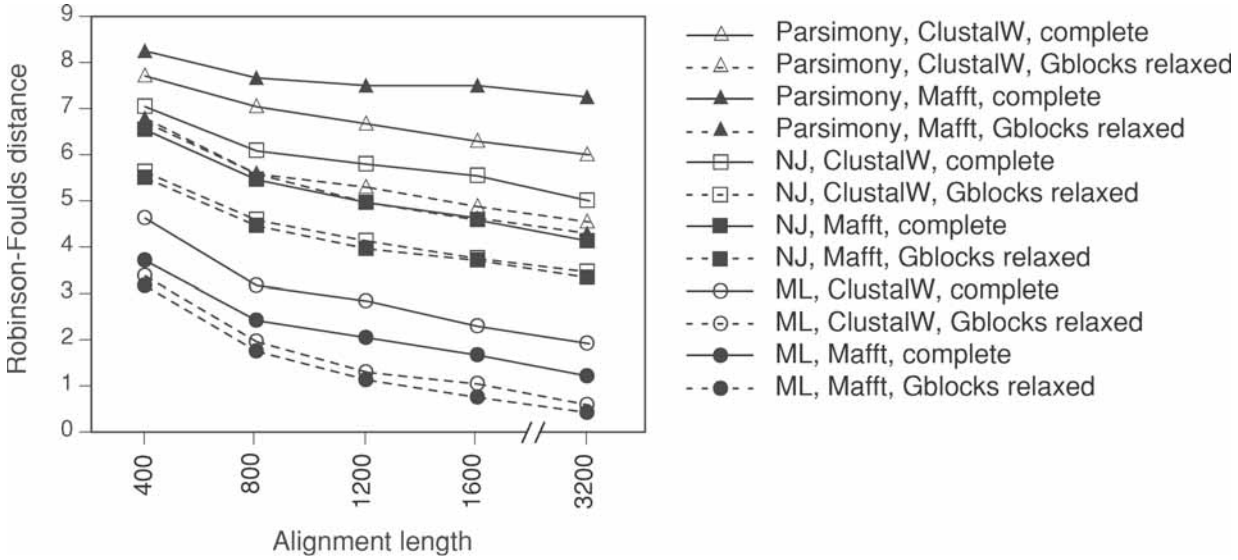


FIGURE 7. Average Robinson-Foulds distances to the real tree from the tree calculated with Mafft complete (solid line, solid symbols) and ClustalW complete alignments (solid line, empty symbols). The tree distances obtained with the same alignments after treatment with Gblocks with relaxed conditions (dotted lines) are also shown. Trees were reconstructed by ML (circles), NJ (squares), and parsimony (triangles). The most divergent asymmetric tree was used for the simulations.

in Felsenstein, 2004). To show this, Figure 7 presents the superimposed graphs for the most divergent asymmetric tree as an example. The better performance of ML in all alignment conditions is clearly appreciated in this graph.

Short versus Long Alignments

Alignment length turned out to be a very important factor to be taken into account when deciding the best alignment cleaning strategy. Figures 3 and 5 show that, in general, for shorter alignments the best Gblocks condition is the relaxed one, whereas for longer alignments the stringent condition tends to work better. This can also be appreciated by comparing the slopes of the graphs corresponding to the complete alignments, and those of the Gblocks alignments with relaxed and stringent conditions. The slope downwards (towards better trees) is less pronounced for the complete alignments and more pronounced for Gblocks with stringent conditions. This means that for single genes (400 to 800 amino acids) the gain in signal-to-noise ratio after elimination of problematic blocks may not compensate the total loss of information. However, for longer alignments, for example, those used in phylogenomic studies where several genes are concatenated (Delsuc et al., 2005; Jeffroy et al., 2006), there is enough total information so that selecting the best pieces with Gblocks using the stringent conditions allows to get closer to the real tree. This basic tendency is observed under all simulation conditions we tested.

Bootstrap Support in Trees Obtained from Gblocks Alignments

Previous performance tests of Gblocks with real data showed that Gblocks alignments obtained less support in ML analysis, because the number of trees not significantly different from the ML tree was smaller in the complete alignment than in the Gblocks alignment (Castresana, 2000). Later, in numerous studies in our group and in other groups, the same effect was observed using bootstrap values of NJ trees, which were lower in the Gblocks alignments. Our simulations reproduced the same behavior again. In NJ trees obtained from 100 bootstrap samples, the average bootstrap support of all partitions was higher for the complete alignments, and lower for Gblocks alignments (Fig. 8). However, the same simulations (see topological distances of NJ trees in Figures 3 and 5) showed that the best trees were obtained with Gblocks conditions and the worse topologies with the complete alignments, thus following the opposite direction, regarding quality, to the bootstrap values, at least for the maximum divergence. A similar trend was found for NJ trees of simulations with symmetric trees (Fig. 9 of the online Appendix) and for bootstrapped ML trees (Fig. 10 of the online Appendix). One may think that the bootstraps of Gblocks trees are lower due to the smaller length of the Gblocks alignments, but it is still very paradoxical that the best topology is associated to a lower bootstrap.

The explanation for this contradictory behavior of Gblocks may be that divergent and problematic alignment regions are biased towards an erroneous topology (Lake, 1991). This could happen if the initial guide tree used in the progressive alignment methods is conducting very strongly the alignment in the divergent and most gappy regions, where alignment programs may easily create similarity at the expense of homology (Higgins et al., 2005). In addition, when alignment software is faced with an ambiguous alignment decision, the algorithmic solution makes consistent but arbitrary decisions

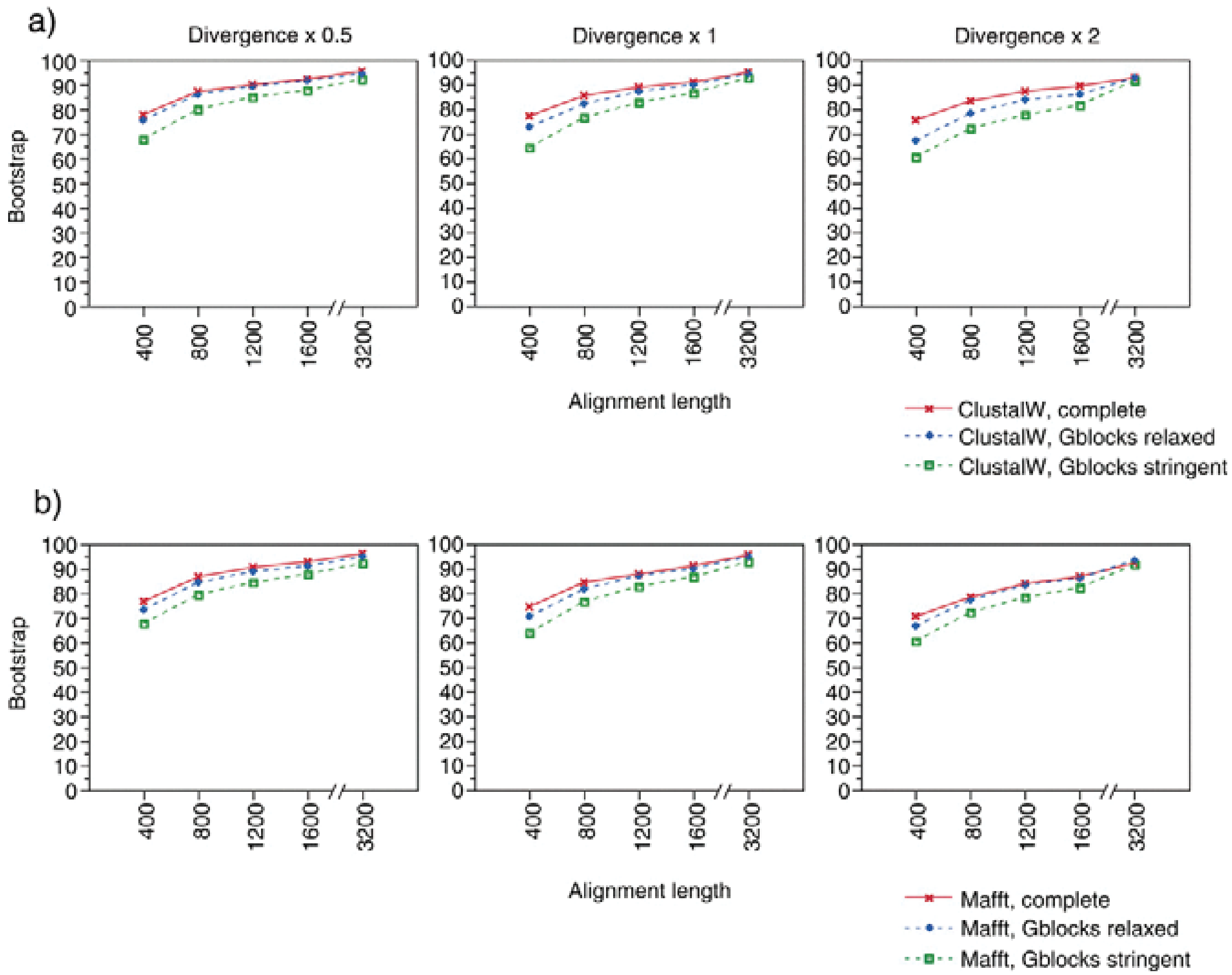


FIGURE 8. Average bootstrap values of NJ trees obtained from ClustalW (a) and Mafft (b) alignments simulated from the asymmetric tree with three different divergence levels. Complete (solid, red line), Gblocks relaxed (dotted, blue line with diamonds), and Gblocks stringent (dotted, green line with squared symbols) alignments are shown.

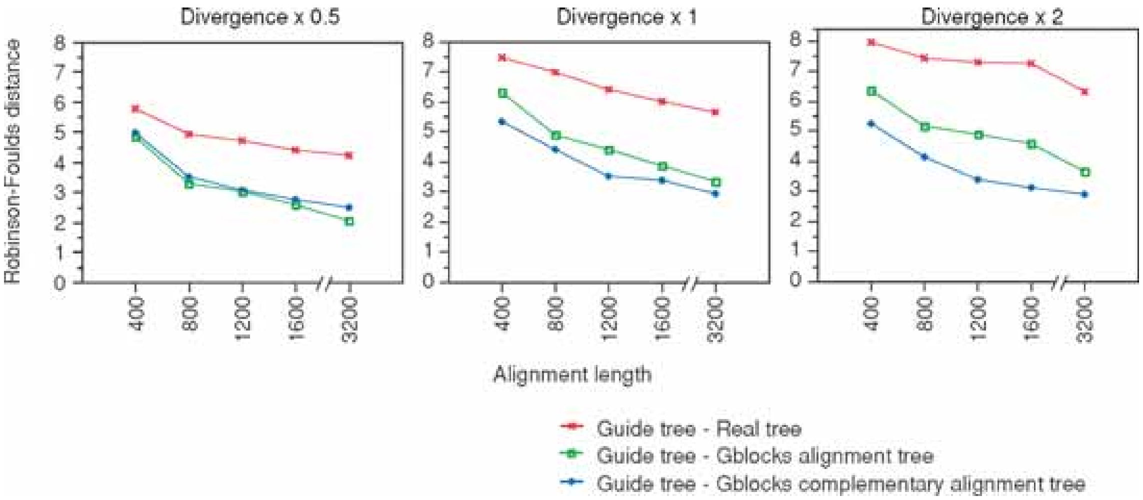


FIGURE 9. Average Robinson-Foulds distances from the ClustalW guide tree to the real tree (red line with crossed symbols), from the guide tree to the NJ tree of the Gblocks alignment with relaxed conditions (green line with squared symbols), and from the guide tree to the NJ tree of the complementary positions of the same Gblocks alignment (blue line with diamonds). The asymmetric tree with three different divergence levels was used for the simulations.

that bias the support indices. That is, this repeated alignment decisions will increase the bootstrap support, and this bias will be stronger in the most divergent regions, where there is more uncertainty. Three results are consistent with this possibility. Firstly, we have observed in our simulations that the initial guide dendrogram used by ClustalW is indeed very different from the real tree, as measured by the Robinson-Foulds distance of both trees (Fig. 9). If all divergent regions tend to easily reproduce this initial dendrogram, we would expect that the guide tree is more similar to the tree obtained from the Gblocks excluded regions than to the Gblocks alignment. Figure 9 shows that this is the case, particularly in the most divergent simulations. Secondly, we see that the effect of increased bootstrap support in the complete alignment with respect to the Gblocks alignments is higher in ClustalW, which highly depends on the initial dendrogram, than in Mafft (Fig. 8). For example, in simulations of 400 amino acids and at ×2 divergence, there is an increase from 60% to 76% bootstrap support in ClustalW when comparing the Gblocks stringent and complete alignments, and only from 60% to 70% in Mafft. In the latter method, the successive iterations of the alignment algorithm may make the final alignment more independent from the initial crude dendrogram, thus explaining that trees generated from these alignments are slightly less biased. And thirdly, when we calculated separately bootstraps of right and wrong partitions for each tree we observe, apart from lower values for wrong partitions, a slightly higher bias in them (Fig. 11 of the online Appendix). The bias is also present in the right partitions, probably because some of the recurrent software decisions in the divergent regions are actually correct. Thus, the bias coming from divergent regions seems to increase the bootstrap of all partitions, although the effect is slightly larger in the wrong ones. All this indicates that bootstrap support cannot be used as a measure of reliability of the tree topology when divergent regions are present in the alignment.

Conclusions

We have shown, under the conditions of these simulations, that the information contained in divergent and ambiguously aligned regions of multiple alignments is, in general, not beneficial for phylogenetic reconstruction. Thus, using Gblocks or a similar method for removing problematic blocks seems to be justified for phylogenetic analysis, particularly for divergent alignments. In this work, we have used simulations of moderately divergent and very heterogeneous proteins, which are typically used in deep phylogenies (i.e., bacterial groups, eukaryotes lineages, metazoan phyla). However, we do not know how removal of blocks would affect more conserved and less heterogeneous alignments. We have also not tested how a finer tuning of parameters of alignment programs and Gblocks may improve the phylogenies. Although we have only used protein alignments, the same conclusions are expected to apply to protein-coding DNA alignments of similar divergence. On the other hand, although we predict that the general conclusion that ambiguously aligned regions in any data set are best excluded when they provide more noise than signal, rRNA alignments as well as alignments from noncoding DNA have very different features from coding alignments, and our simulations were not specifically designed to explore the properties of these kinds of sequences. However, our purpose in this work is not giving strict rules about the best alignment strategy and associated parameters. Rather, our simulations are mainly informative about general tendencies. Thus, in the following we summarize important tendencies observed in our simulations and give some general rules regarding the best alignment strategy that can be applied to real situations of protein alignments.

NJ and parsimony seem to be unable to extract useful phylogenetic information from the problematic alignment regions, because the complete alignments are always much worse than the Gblocks treated alignments, so using Gblocks seems particularly advisable for these methods. Most probably, these two methods are not able to take into account the multiple substitutions that occur in these excessively saturated blocks. On the other hand, ML, less affected by saturation, is able to extract some information from these blocks, since in some conditions the complete alignments are similar or even better than the Gblocks alignments. However, the misidentified homology that may occur in these regions affects all phylogenetic methods, which may explain why using Gblocks is more beneficial at high divergences for all methods.

Regarding the use of stringent or relaxed conditions for Gblocks, two important rules can be extracted from our analysis. First, for ML trees relaxed conditions of Gblocks seem to give rise to better trees, whereas for NJ and parsimony stringent conditions are better. Second, alignment length is a crucial parameter to be taken into account. For short alignments, such as in studies of single short genes, the removal of blocks by Gblocks may leave too few positions, so in these cases it may be better to use very relaxed conditions of Gblocks. In the shortest alignments, which have very little information, use of Gblocks may be even detrimental. At any rate, one should be aware that with this type of short alignments it is only possible to obtain a very approximate topology, possibly quite distant from the real tree. For phylogenomic studies, where there is enough information from the concatenation of several genes (Jeffroy et al., 2006), the use of Gblocks with stringent conditions tends to give rise to the best phylogenetic trees.

Acknowledgments

This work was supported financially by a research grant in bioinformatics from the Fundación BBVA (Spain), and grant number BIO2002-04426-C02-02 from the Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica (I+D+I) of the MEC, cofinanced with FEDER funds. We thank V. Soria-Carrasco for useful technical assistance, and three anonymous reviewers, K. Kjer, and R.D.M. Page for critical comments that helped improve the manuscript.

REFERENCES

Aagesen, L. 2004. The information content of an ambiguously alignable region, a case study of the trnL intron from the Rhamnaceae. Organ. Divers. Evol. 4:35–49.

Blackshields, G., I. M. Wallace, M. Larkin, and D. G. Higgins. 2006. Analysis and comparison of benchmarks for multiple sequence alignment. In Silico Biol. 6:321–339.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17:540–552.

Castresana, J., G. Feldmaier-Fuchs, and S. Pääbo. 1998a. Codon reassignment and amino acid composition in hemichordate mitochondria. Proc. Natl. Acad. Sci. USA 95:3703–3707.

Castresana, J., G. Feldmaier-Fuchs, S. Yokobori, N. Satoh, and S. Pääbo. 1998b. The mitochondrial genome of the hemichordate *Balanoglossus carnosus* and the evolution of deuterostome mitochondria. Genetics 150:1115–1123.

Dayhoff, M. O., R. M. Schwartz, and B. C. Orcutt. 1978. A model of evolutionary change in proteins. Pages 345–352 *in* Atlas of protein sequence structure (M. O. Dayhoff, ed.) National Biomedical Research Foundation, Washington, D.C.

Delsuc, F., H. Brinkmann, and H. Philippe. 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. 6:361–375.

Do, C. B., M. S. Mahabhashyam, M. Brudno, and S. Batzoglou. 2005. ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Res. 15:330–340.

Drummond, A., and K. Strimmer. 2001. PAL: An object-oriented programming library for molecular evolution and phylogenetics. Bioinformatics 17:662–663.

Edgar, R. C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Felsenstein, J. 1989. PHYLIP—Phylogeny inference package (version 3.4). Cladistics 5:164–166.

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates, Sunderland, Massachusetts.

Feng, D. F., and R. F. Doolittle. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J. Mol. Evol. 25:351–360.

Fleissner, R., D. Metzler, and A. von Haeseler. 2005. Simultaneous statistical multiple alignment and phylogeny reconstruction. Syst. Biol. 54:548–561.

Gatesy, J., R. DeSalle, and W. Wheeler. 1993. Alignment-ambiguous nucleotide sites and the exclusion of systematic data. Mol. Phylogenet. Evol. 2:152–157.

Geiger, D. L. 2002. Stretch coding and block coding: Two new strategies to represent questionably aligned DNA sequences. J. Mol. Evol. 54:191–199.

Grundy, W. N., and G. J. Naylor. 1999. Phylogenetic inference from conserved sites alignments. J. Exp. Zool. 285:128–139.

Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52:696–704.

Gutell, R. R., N. Larsen, and C. R. Woese. 1994. Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. Microbiol. Rev. 58:10–26.

Henikoff, S., and J. G. Henikoff. 1994. Protein family classification based on searching a database of blocks. Genomics 19:97–107.

Herrmann, G., A. Schon, R. Brack-Werner, and T. Werner. 1996. CONRAD: A method for identification of variable and conserved regions within proteins by scale-space filtering. Comput. Appl. Biosci. 12:197–203.

Higgins, D. G., G. Blackshields, and I. M. Wallace. 2005. Mind the gaps: Progress in progressive alignment. Proc. Natl. Acad. Sci. USA 102:10411–10412.

Jeffroy, O., H. Brinkmann, F. Delsuc, and H. Philippe. 2006. Phylogenomics: The beginning of incongruence? Trends Genet. 22:225–231.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8:275–282.

Katoh, K., K. Kuma, H. Toh, and T. Miyata. 2005. MAFFT version 5: Improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33:511–518.

Katoh, K., K. Misawa, K. Kuma, and T. Miyata. 2002. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059–3066.

Kjer, K. M. 1995. Use of rRNA secondary structure in phylogenetic studies to identify homologous positions: an example of alignment and data presentation from the frogs. Mol. Phylogenet. Evol. 4:314-330.

Lake, J. A. 1991. The order of sequence alignment can bias the selection of tree topology. Mol. Biol. Evol. 8:378–385.

Lassmann, T., and E. L. Sonnhammer. 2005. Kalign—An accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics 6:298.

Lee, M. S. 2001. Unalignable sequences and molecular evolution. Trends Ecol. Evol. 16:681–685.

Löytynoja, A., and M. C. Milinkovitch. 2001. SOAP, cleaning multiple alignments from unstable blocks. Bioinformatics 17:573–574.

Lunter, G., I. Miklos, A. Drummond, J. L. Jensen, and J. Hein. 2005. Bayesian coestimation of phylogeny and sequence alignment. BMC Bioinformatics 6:83.

Lutzoni, F., P. Wagner, V. Reeb, and S. Zoller. 2000. Integrating ambiguously aligned regions of DNA sequences in phylogenetic analyses without violating positional homology. Syst. Biol. 49:628–651.

Morrison, D. A., and J. T. Ellis. 1997. Effects of nucleotide sequence alignment on phylogeny estimation: A case study of 18S rDNAs of apicomplexa. Mol. Biol. Evol. 14:428–441.

Needleman, S. B., and C. D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48:443–453.

Notredame, C., D. G. Higgins, and J. Heringa. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302:205–217.

Nuin, P. A., Z. Wang, and E. R. Tillier. 2006. The accuracy of several multiple sequence alignment programs for proteins. BMC Bioinformatics 7:471.

Ogden, T. H., and M. S. Rosenberg. 2006. Multiple sequence alignment accuracy and phylogenetic inference. Syst. Biol. 55:314–328.

Pesole, G., M. Attimonelli, G. Preparata, and C. Saccone. 1992. A statistical method for detecting regions with different evolutionary dynamics in multialigned sequences. Mol. Phylogenet. Evol. 1:91–96.

Philippe, H., and J. Laurent. 1998. How good are deep phylogenetic trees? Curr. Opin. Genet. Dev. 8:616–623.

Redelings, B. D., and M. A. Suchard. 2005. Joint Bayesian estimation of alignment and phylogeny. Syst. Biol. 54:401–418.

Robinson, D. F., and L. R. Foulds. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.

Rodrigo, A. G., P. R. Bergquist, and P. L. Bergquist. 1994. Inadequate support for an evolutionary link between the Metazoa and the Fungi. Syst. Biol. 43:578–584.

Smythe, A. B., M. J. Sanderson, and S. A. Nadler. 2006. Nematode small subunit phylogeny correlates with alignment parameters. Syst. Biol. 55:972–992.

Stajich, J. E., et al. 2002. The Bioperl toolkit: Perl modules for the life sciences. Genome Res. 12:1611–1618.

Stoye, J., D. Evers, and F. Meyer. 1998. Rose: Generating sequence families. Bioinformatics 14:157–163.

Strimmer, K., and A. von Haeseler. 1996. Quartet puzzling: A quartet maximum-likelihood method for reconstructing tree topologies. Mol. Biol. Evol. 13:964–969.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. Pages 407–514 *in* Molecular systematics (D. M. Hillis, C. Moritz, and B. K. Mable, eds.). Sinauer Associates, Sunderland, Massachusetts.

Tatusov, R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Thompson, J. D., P. Koehl, R. Ripp, and O. Poch. 2005. BAliBASE 3.0: Latest developments of the multiple sequence alignment benchmark. Proteins 61:127–136.

Wheeler, W. 2001. Homology and the optimization of DNA sequence data. Cladistics 17:S3–S11.

Xia, X., Z. Xie, and K. M. Kjer. 2003. 18S ribosomal RNA and tetrapod phylogeny. Syst. Biol. 52:283–295.

Yang, Z. 1998. On the best evolutionary rate for phylogenetic analysis. Syst. Biol. 47:125–133.

Young, N. D., and J. Healy. 2003. GapCoder automates the use of indel characters in phylogenetic analysis. BMC Bioinformatics 4:6.

First submitted 7 February 2007; reviews returned 6 March 2007; final acceptance 24 March 2007
Associate Editor: Karl Kjer
Editors: Rod Page and Jack Sullivan

BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 21 2007, pages 2954–2956
doi:10.1093/bioinformatics/btm466

Phylogenetics

The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees

Víctor Soria-Carrasco, Gerard Talavera, Javier Igea and Jose Castresana*

Department of Physiology and Molecular Biodiversity, Institute of Molecular Biology of Barcelona, CSIC,
Jordi Girona 18, 08034 Barcelona, Spain

Received on June 26, 2007; revised on August 13, 2007; accepted on September 6, 2007

Advance Access publication September 22, 2007

Associate Editor: Keith Crandall

ABSTRACT

Summary: We introduce a new phylogenetic comparison method that measures overall differences in the relative branch length and topology of two phylogenetic trees. To do this, the algorithm first scales one of the trees to have a global divergence as similar as possible to the other tree. Then, the branch length distance, which takes differences in topology and branch lengths into account, is applied to the two trees. We thus obtain the minimum branch length distance or K tree score. Two trees with very different relative branch lengths get a high K score whereas two trees that follow a similar among-lineage rate variation get a low score, regardless of the overall rates in both trees. There are several applications of the K tree score, two of which are explained here in more detail. First, this score allows the evaluation of the performance of phylogenetic algorithms, not only with respect to their topological accuracy, but also with respect to the reproduction of a given branch length variation. In a second example, we show how the K score allows the selection of orthologous genes by choosing those that better follow the overall shape of a given reference tree.

Availability: http://molevol.ibmb.csic.es/Ktreedist.html

Contact: jcvagr@ibmb.csic.es

1 INTRODUCTION

In phylogenetic reconstruction, the application of different methods or the use of different genes may lead to the estimation of different phylogenetic trees (Castresana, 2007; Hillis *et al*., 2005; Huerta-Cepas *et al*., 2007). In order to analyze if the resulting trees are congruent, it is fundamental to be able to quantify differences between such trees. Normally, only topology is taken into account for such task, for example, by means of the symmetric difference (Robinson and Foulds, 1981). Few methods have been developed that also take branch length information into account (Hall, 2005; Kuhner and Felsenstein, 1994). These methods have been successfully applied to quantify the performance of different phylogenetic methods in simulated alignments, but they have the drawback that they are not directly applicable to trees with different evolutionary rates. Here, we introduce a new phylogenetic comparison measure that takes branch length information into account after scaling the trees so that they have comparable global evolutionary rates.

2 METHOD

The basis of our method to compare two phylogenetic trees, T and T', is the branch length distance (BLD) introduced by Kuhner and Felsenstein (Felsenstein, 2004; Kuhner and Felsenstein, 1994). This distance is sensitive to the similarity in branch lengths of both trees. Consider the set of partitions present in both trees, that is, the whole set of partitions present in T plus the set of partitions present in T' but not in T. Partitions for external branches are also included. For tree T, we can define an array B of branch lengths associated to each partition ($b_1, b_2, \ldots, b_N$). Branches that do not appear in T (corresponding to partitions that are only present in T') are assigned to 0 in such array. We can similarly define the array B' associated to tree T'. The BLD between trees T and T' is the squared root of the sum of $(b_i - b_i')^2$ for all partitions. However, BLD depends on the absolute size of the trees being compared, so that two trees with the same shape (topology and relative branch length) but different global rates will give rise to a very high BLD (Kuhner and Felsenstein, 1994), which may be unwanted.

In our method, we introduce a factor, K, to scale tree T' so that both trees, T and T', have a similar global divergence. Thus, we are interested in calculating BLD after scaling T' with a factor K:

$$BLD(K) = \sqrt{\sum_{i=1}^{N} (b_i - Kb_i')^2} \quad (1)$$

To obtain the value of K that minimizes BLD we differentiate Equation (1). It can be shown that the value of K that makes this derivative zero is:

$$K = \frac{\sum_{i=1}^{N} \left(b_i b_i'\right)}{\sum_{i=1}^{N} b_i'^2} \quad (2)$$

*To whom correspondence should be addressed.

We then substitute this value of K in Equation (1) and obtain the minimum branch length distance or K tree score. It should be taken into account that the K tree score is not symmetric, that is, the result from T to T' may not be the same than from T' to T, and, in consequence, the K score does not have the mathematical properties of a distance. Thus, this score is generally not useful to compare only two trees (although the K factor of Equation (2) can be very valuable for scaling purposes; see below). The K tree score is most useful when there is a tree that serves as reference (T) and several other trees (T') that will be scaled and compared to T. In such cases, trees T' that are similar in shape to T will receive a low K tree score whereas those that are very different will get a relatively higher K score, regardless of their overall rates.

The method that calculates the K tree score (as well as other tree comparison measures) is implemented in a Perl program called Ktreedist.

3 APPLICATIONS

There are several applications of the K tree score. First, it can be used to evaluate the quality of phylogenetic reconstructions in simulated alignments by comparing the true tree to the trees obtained with different phylogenetic methods. For example, the reference tree shown in Figure 1A was used to simulate with SeqGen (Rambaut and Grassly, 1997) 100 alignments of 1000 positions with a GTR model and gamma rate heterogeneity ($\alpha = 1.5$). We then constructed maximum-likelihood (ML) trees from such simulations using Phyml (Guindon and Gascuel, 2003) with two different conditions: without and with rate heterogeneity. To facilitate the comparison between both phylogenetic methods we imposed the topology of the reference tree during the ML reconstructions. After averaging the branch lengths of the 100 reconstructed trees, we obtained one tree for each phylogenetic method. Both trees differed in their overall rates (with the nonrate heterogeneity tree not capturing all substitutions, leading to a K scale factor $\gg 1$) but, importantly, they also differed in their shapes: see, for example, the relative lengths of sp3, sp4 and sp5. The differences in shape were reflected in the K scores: 0.197 for the average tree without rate heterogeneity and 0.030 for the average tree calculated with rate heterogeneity, indicating the better performance of the latter method. (Differences also appeared after averaging the K score from the 100 trees obtained with each method although, in this case, the magnitude of the difference was smaller.) Thus, the K tree score can be used to quantify the different quality in branch length reconstruction of different phylogenetic methods. The K score can also be used with trees that have different topologies. In such cases, nonshared branches that are relatively long will contribute to the K score much more than small conflicting branches. This is different from the symmetric difference (Robinson and Foulds, 1981), in which all topological differences count the same.

In a second example, we show how the K tree score can be used to make an accurate selection of orthologous genes. Orthologs should reflect the same topology of the species tree but they should also give rise, in principle, to a similar tree shape. We extracted from the ENSEMBL database (Hubbard *et al.*, 2007) the tables of pairs of orthologous genes of seven

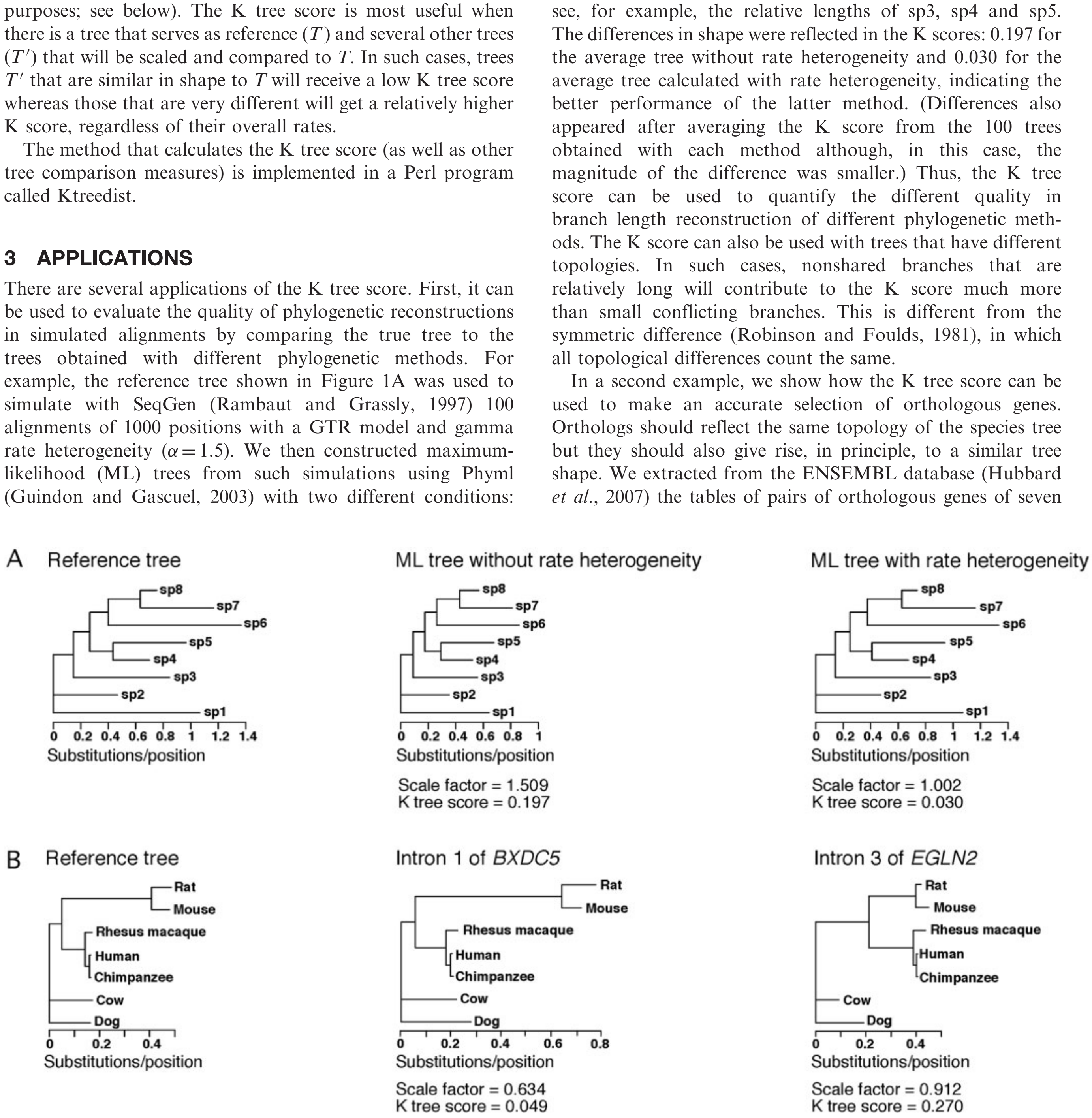


Fig. 1. (**A**) Reference tree used to simulate 100 alignments and the average reconstructions obtained by ML without and with rate heterogeneity. (**B**) Trees obtained with 472 concatenated introns (reference tree) and with two individual introns (intron 1 of *BXDC5* and intron 3 of *EGLN2*).

mammalian species. By matching the pairwise orthology tables, we constructed a set of one-to-one orthologs, and we downloaded the corresponding genes. We then extracted the introns and, after applying several filters (elimination of very long introns, those with problematic alignments, etc.), we obtained a set of 472 putative orthologous introns. Some of these introns produced ML phylogenetic trees that were of unusual shape, which could be due to different rates of evolution in different lineages (heterotachy) or could indicate that they do not come from orthologous genes (hidden paralogy). We then constructed a reference tree (Fig. 1B) with the concatenated alignment of the 472 introns using the RAxML program (Stamatakis, 2006), which can handle very long alignments, with a GTR model of evolution and four rate categories. This tree should reflect the average divergence of the seven genomes and, as expected, rodents showed a higher acceleration in their branches. We then calculated the K score of the trees of all individual introns with respect to the reference tree. We show in Figure 1B the trees of two putative orthologous introns. Intron 1 of *BXDC5*, despite having a high global rate, produced a phylogeny with the same topology and a very similar tree shape to the reference tree. This was reflected in a low K score: 0.049, smaller than the mean of the distribution of K scores of all individual introns (0.104), which is indicative of a very likely ortholog. (The K score would also be low in a similar tree but with a topological conflict affecting a small branch, which would not affect the high probability of orthology.) Intron 3 of *EGLN2* also reproduced the reference topology. However, this tree showed a relatively long basal branch in primates as well as a long branch connecting Euarchontoglires and Laurasiatherians. In consequence, the K score for this tree with respect to the reference is much higher: 0.270. In fact, this value is a clear outlier in the distribution of K scores. Although heterotachy cannot be discarded, the chances that the latter gene contains hidden paralogs in some species are higher than in the first gene. Thus, the K score can be used to establish a certain threshold and make a more accurate selection of orthologous genes.

If orthology is ensured for a set of genes, then a high K tree score with respect to a given reference will be indicative of trees with very fast-evolving species or with a significant amount of other types of heterotachy. These trees are of more difficult reconstruction, and thus the K tree score can be used to select (in a similar way as above) a set of the most reliable genes for estimating species phylogenies.

On a more practical side, the K scale factor [Equation (2)] can be used in instances where it is necessary to scale trees to have equivalent divergences. For example, the linearization of trees by means of a method like nonparametric rate smoothing produces trees with an arbitrary scale when no dates are known for the tree nodes (Sanderson, 1997). In such cases, one can make use of the K scale factor obtained from the comparison between the linearized tree and the original (reference) tree: the scaling of the linearized tree with this K factor will re-establish a genetic distance scale equivalent to that of the original tree.

ACKNOWLEDGEMENTS

J.C. is supported by grant number CGL2005-01341/BOS from the Plan Nacional I+D+I of the MEC (Spain), cofinanced with FEDER funds.

Conflict of Interest: none declared.

REFERENCES

Castresana,J. (2007) Topological variation in single-gene phylogenetic trees. *Genome Biol.*, **8**, 216.

Felsenstein,J. (2004) *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts, pp. 531–533.

Guindon,S. and Gascuel,O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52**, 696–704.

Hall,B.G. (2005) Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences. *Mol. Biol. Evol.*, **22**, 792–802.

Hillis,D.M. *et al.* (2005) Analysis and visualization of tree space. *Syst. Biol.*, **54**, 471–482.

Hubbard,T.J.P. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610–D617.

Huerta-Cepas,J. *et al.* (2007) The human phylome. *Genome Biol.*, **8**, R109.

Kuhner,M.K. and Felsenstein,J. (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.*, **11**, 459–468.

Rambaut,A. and Grassly,N.C. (1997) Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.*, **13**, 235–238.

Robinson,D.F. and Foulds,L.R. (1981) Comparison of phylogenetic trees. *Math. Biosci.*, **53**, 131–147.

Sanderson,M.J. (1997) A nonparametric approach to estimating divergence times in the absence of rate constancy. *Mol. Biol. Evol.*, **14**, 1218–1231.

Stamatakis,A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.

Journal of Biogeography (*J. Biogeogr.*) (2011)

ORIGINAL ARTICLE

Tracing the origin of disjunct distributions: a case of biogeographical convergence in *Pyrgus* butterflies

Juan L. Hernández-Roldán[1,2], Cesc Múrria[3], Helena Romo[2], Gerard Talavera[1], Evgeny Zakharov[4], Paul D. N. Hebert[4] and Roger Vila[1,5]*

[1]Institut de Biologia Evolutiva (CSIC-UPF), Passeig Marítim de la Barceloneta, 37-49, E-08003 Barcelona, Spain, [2]Departamento de Biología (Zoología), Facultad de Ciencias de la Universidad Autónoma de Madrid, C/Darwin, 2, E-28049 Madrid, Spain, [3]Department of Entomology, The Natural History Museum, London SW7 5BD, UK, [4]Biodiversity Institute of Ontario, University of Guelph, Guelph, ON, Canada N1G 2W1, [5]Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, E-08010 Barcelona, Spain

*Correspondence: Roger Vila, Institut de Biologia Evolutiva (CSIC-UPF), Passeig Marítim de la Barceloneta, 37-49, E-08003 Barcelona, Spain.
E-mail: roger.vila@ibe.upf-csic.es

ABSTRACT

Aim To study the biogeographical factors responsible for the current disjunct distributions of two closely related species of butterflies (*Pyrgus cinarae* and *Pyrgus sidae*, Lepidoptera: Hesperioidea). Both species have small populations in the Iberian Peninsula that are isolated by more than 1000 km from their nearest conspecifics. Because these species possess similar ecological preferences and geographical distributions, they are excellent candidates for congruent biogeographical histories.

Location The Palaearctic region, with a special focus on the Mediterranean peninsulas as glacial refugia.

Methods We integrated phylogeography and population genetic analyses with ecological niche modelling. The mitochondrial gene cytochrome *c* oxidase subunit 1 (COI) and the non-coding nuclear marker internal transcribed spacer 2 (ITS2) were analysed for 62 specimens of *P. cinarae* and for 80 of *P. sidae* to infer phylogeography and to date the origin of disjunct distributions. Current and ancestral [Last Glacial Maximum using MIROC (Model for Interdisciplinary Research on Climate) and CCSM (Community Climate System Model) circulation models] distribution models were calculated with Maxent. Using present climatic conditions, we delimited the ecological space for each species.

Results The genetic structure and potential ancestral distribution of the two species were markedly different. While the Iberian population of *P. cinarae* had an old origin (*c.* 1 Ma), that of *P. sidae* was closely related to French and Italian lineages (which jointly diverged from eastern populations *c.* 0.27 Ma). Ecological niche modelling showed that minor differences in the ecological preferences of the two species seem to account for their drastically different distributional response to the last glacial to post-glacial environmental conditions. Although the potential distribution of *P. cinarae* was largely unaffected by climate change, suitable habitat for *P. sidae* strongly shifted in both elevation and latitude. This result might explain the early origin of the disjunct distribution of *P. cinarae*, in contrast to the more recent disjunction of *P. sidae*.

Main conclusions We show that convergent biogeographical patterns can be analysed with a combination of genetic and ecological niche modelling data. The results demonstrate that species with similar distributional patterns and ecology may still have different biogeographical histories, highlighting the importance of including the temporal dimension when studying biogeographical patterns.

Keywords

Biogeography, COI, disjunct distribution, ecology, ITS2, Lepidoptera, niche modelling, Palaearctic region, palaeoclimate, phylogeography.

http://wileyonlinelibrary.com/journal/jbi
doi:10.1111/j.1365-2699.2011.02547.x

INTRODUCTION

Disjunct distributions (i.e. populations separated by a wide area where the species does not occur) represent extreme biogeographical patterns characterized by unusual evolutionary histories (Schmitt & Hewitt, 2003; Schmitt *et al.*, 2006; Garcia Collevatti *et al.*, 2009). Such distributions may reflect range fragmentation of a formerly widely distributed species due to changes in environmental conditions that affected suitable habitat distribution (Cox & Moore, 2005), as in the case of isolation of populations in refugia from glacial periods (e.g. Stehlik *et al.*, 2000). Alternatively, they can arise by extraordinary long-distance dispersal events, which result in the colonization of new suitable habitats (Davis & Shaw, 2001; Cox & Moore, 2005; Garcia Collevatti *et al.*, 2009). Determining how historical and present-day factors interact to initiate and maintain disjunct distributions in different evolutionary scenarios is a challenge for biogeographers and evolutionary biologists.

The distribution of organisms in the Palaearctic region has been strongly influenced by historical and current factors, especially the cyclic glacial and interglacial periods during the Pleistocene. These events affected large geographical areas and produced cycles of demographic contraction and expansion in species with low dispersal capacity, or recurrent shifts in the distributions of species with good dispersal abilities (Hewitt, 2000, 2004; Schmitt, 2007). Present-day factors have an important role in the maintenance of disjunct distributions by reducing connectivity among isolated populations. As a result, the current geographical distribution of genetic structure within species may reflect the impacts of Pleistocene climatic oscillations on refugial location, the level of gene flow between refugia during interglacial periods and varying connectivity linked to dispersal capacity (Avise, 2000; Hewitt, 2004). As a consequence, it is anticipated that most current disjunct distributions in the Palaearctic region arose from populations isolated in glacial refugia where varied species found suitable conditions for their survival during glacial maxima and that there has been low or no gene flow among these habitats.

In Europe, the main glacial refugia were located in the Mediterranean area: the Iberian, Italian and Balkan peninsulas (Hewitt, 1996). Phylogeographical studies based on species variability have shown a different molecular biogeographical pattern for this area compared with Continental and Arctic European regions (Schmitt, 2007). This pattern is characterized by one or more genetic lineages that began to diverge in Mediterranean refugia. In many species of animals and plants (Taberlet *et al.*, 1998; Hewitt, 1999, 2000) gene flow between populations in these former refugia is at present absent or very limited, and they apparently evolved independently (e.g. Schmitt & Seitz, 2002; Schmitt & Krauss, 2004; Habel *et al.*, 2005). Studying gene flow during interglacial periods, which depends on dispersal traits and habitat connectivity, is thus critical for understanding the biogeographical history of species and to reveal the origin of disjunct distributions.

Molecular genetic techniques are now used widely in phylogeographical studies of both animals and plants (e.g. Hewitt, 2004; Garcia Collevatti *et al.*, 2009). Recently, new methodologies based on environmental variables have been developed to estimate the potential geographical distributions of species (Guisan & Zimmerman, 2000; Elith *et al.*, 2006; Phillips *et al.*, 2006). Among them, the program Maxent has been shown to perform better than other methods [for example GARP (Stockwell & Peters, 1999) and BIOCLIM (Nix, 1986)] in predicting species distributions from presence-only data (Phillips *et al.*, 2004, 2006; Elith *et al.*, 2006; Wisz *et al.*, 2008). Maxent is based on maximum entropy modelling of species geographical distributions, and computes a probability distribution of habitat suitability over the geographical area of the units considered. The integration of phylogeographical and distribution modelling seems to be a promising way to unravel the biogeographical history behind disjunct distributions (Weaver *et al.*, 2006; Jakob *et al.*, 2007; Alsos *et al.*, 2009).

In this study we trace the genesis of disjunct distributions in two members of the genus *Pyrgus* (Lepidoptera: Hesperioidea: Hesperiidae: Pyrginae). *Pyrgus cinarae* (Rambur, 1839) and *Pyrgus sidae* (Esper, 1784) have very similar distributions: in Europe they are limited to the north Mediterranean, extending deeply into central Asia, and both display a disjunct distribution with small isolated populations in the Iberian Peninsula (Kudrna, 2002; García-Barros *et al.*, 2004). The extremely localized Iberian population is restricted to central Spain, *c.* 1800 km (*P. cinarae*) and *c.* 1000 km (*P. sidae*) from the nearest conspecific populations (see Fig. 1). Despite the apparent presence of suitable habitats and host plants (*Potentilla recta* and *Potentilla hirta*) between the Iberian and non-Iberian populations, no studies have suggested their connectivity. *Pyrgus cinarae* and *P. sidae* display similar ecological preferences (they frequently co-occur in the same habitats and share host plants), and both have low dispersal ability (Hernández-Roldán *et al.*, 2009; Wagner, 2009), especially across water surfaces (they are not present on islands or in North Africa), a trait shared by all members of the genus *Pyrgus*.

Species with similar current distributions and ecological traits are good candidates for having congruent biogeographical histories, i.e. similar responses to the same environmental changes or geographical events (Bocxlaer *et al.*, 2006; Noonan & Chippindale, 2006). Recent calls to integrate the temporal and spatial dimensions in biogeographical studies, especially when studying patterns across taxa, have been made (Hunn & Upchurch, 2001; Donoghue & Moore, 2003). In this regard, it is worth noting that two species could display spatially congruent but temporally incongruent biogeographical histories (Loader *et al.*, 2007). In this paper, by combining molecular data (both phylogeography and population genetics) with ecological niche modelling, we study two closely related species that are apparently equivalent in most regards to test whether they share parallel evolutionary histories.

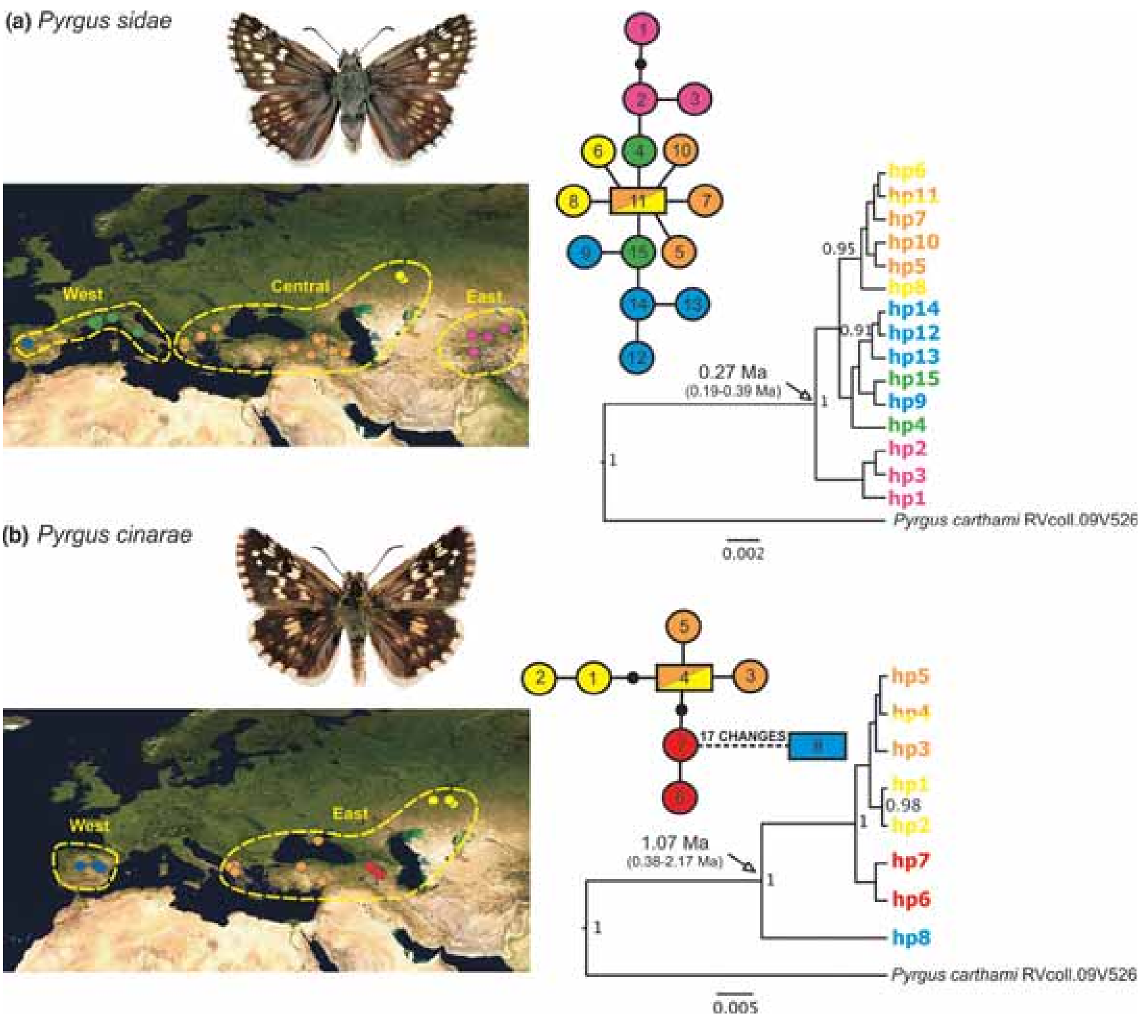


Figure 1 Maps showing localities of studied specimens, COI haplotype networks and phylogenetic trees for (a) *Pyrgus sidae* and (b) *P. cinarae*. TCS v.121 with a 95% connection limit was used to reconstruct the haplotype networks. In the maps, the three main haplotype clades for *P. sidae* and the two for *P. cinarae* are indicated by discontinuous lines. Colours indicate five populations for *P. sidae* (magenta, Kyrgyzstan; yellow, south Urals; orange, Caucasus–Black Sea–Balkans; green, France–Italy; blue, Iberian Peninsula) and four populations for *P. cinarae* (yellow, south Urals; red, Caucasus; orange, Black Sea–Balkans; blue, Iberian Peninsula). BEAST v.1.5.8 under a coalescent model was used for Bayesian tree inference. Scale bars in the trees show divergence in substitutions/site and only posterior probabilities > 0.9 are shown in the nodes. The divergence time of the oldest split within each species calculated by MDIV models is indicated, with the confidence interval in the form of 95% highest posterior densities in parentheses.

MATERIALS AND METHODS

Sampling and gathering of the molecular data set

Samples

Sixty-two specimens of *P. cinarae* and 80 specimens of *P. sidae* were collected from 18 and 32 localities, respectively, covering the known ranges of both species (Fig. 1). The 48 sampling sites were partitioned into five populations for *P. sidae* (Kyrgyzstan, south Urals, Caucasus–Black Sea–Balkans, France–Italy, Iberian Peninsula) and into four populations for *P. cinarae* (south Urals, Caucasus, Black Sea–Balkans, Iberian Peninsula) that were separated by more than 1000 km without any record for the species (Fig. 1). Samples were preserved in 100% ethanol for molecular analysis. Identification codes and collection localities for the samples used are listed in Appendix S1 in Supporting Information. Voucher specimens were deposited in the collection of the Institut de Biologia Evolutiva (CSIC-UPF), Barcelona, Spain.

Cytochrome c oxidase subunit I (COI) amplification

DNA was extracted using a glass fibre protocol (Ivanova *et al.*, 2006) from a single leg of each specimen. A 658-bp fragment of the mitochondrial gene cytochrome *c* oxidase subunit I (COI) was targeted for polymerase chain reaction (PCR) amplification using the primers LepF (5′-ATTCAACCAATCATAAAGA TATTGG-3′) and LepR (5′-TAAACTTCTGGATGTCCAAAA AATCA-3′) (Hajibabaei *et al.*, 2005; deWaard *et al.*, 2008).

Samples that did not produce a PCR product with the primers LepF and LepR were reamplified with the primers LepF and Enh_LepR (5′-CTCCWCCAGCAGGATCAAAA-3′), which amplify a 609-bp fragment of COI. All specimens were successfully amplified for this marker. One specimen of *Pyrgus carthami* (Hübner, 1813) was amplified and used as an outgroup. One COI sequence of *Pyrgus communis* (Grote, 1872) obtained from GenBank (accession number AF170857) was also added to the dataset as an outgroup. Sequences were obtained with an ABI 3730 × 1 sequencer (Applied Biosystems, Carlsbad, CA, USA) following the manufacturer's recommendations.

Internal transcribed spacer 2 (ITS2) amplification

A total of 23 *P. sidae* and 12 *P. cinarae* samples representing all the COI haplotype clades, plus a specimen of *Pyrgus armoricanus* (Oberthür, 1910) used as an outgroup, were sequenced for the non-coding nuclear marker internal transcribed spacer 2 (ITS2). Total genomic DNA was extracted using Chelex 100 resin, 100–200 mesh, sodium form (Bio-Rad, Richmond, CA, USA), under the following protocol: one leg was removed and introduced into a tube with 100 μL of Chelex 10% solution and 5 μL of Proteinase K (20 mg mL^{-1}). The samples were incubated overnight at 55 °C and then incubated at 100 °C for 15 min. The samples were subsequently centrifuged for 10 s at 1500 *g* and the supernatant was used for PCR amplification. A 684-bp fragment at the 5′-end of ITS2 was amplified using the primers ITS3 (5′-GCATCGATGAAGAACGCAGC-3′) and ITS4 (5′-TCCTCCGCTTATTGATATGC-3′) (White *et al.*, 1990). Double-stranded DNA was amplified in 25 μL volume reactions: 16.7 μL ultra pure (high-performance liquid chromatography quality) water, 2.5 μL 10× buffer, 1 μL 100 mm $MgCl_2$, 0.25 L 100 mm dNTP, 1.2 μL of each primer (10 mm), 0.15 μL Taq DNA Polymerase (Bioron GmbH, Ludwigshafen, Germany) and 2 μL of extracted DNA. The typical thermal cycling profile was: 95 °C for 45 s, 47 °C for 60 s and 72 °C for 60 s, for 40 cycles. Finally, ITS2 PCR products were purified and sequenced by Macrogen Inc. (Seoul, Korea).

Sequences, specimen photographs and associated data for the COI sequences are available in the 'Butterflies of Spain' project Barcode of Life Data Systems (http://www.barcodinglife.org, downloaded 2 November 2010). COI and ITS2 sequences are also available in GenBank (see Appendix S1 for accession numbers).

Data analyses

Phylogenetic inference

COI and ITS2 sequences were edited and aligned using Geneious Pro v.4.8.3 (Biomatters Ltd., 2009; http://www.geneious.com/). These analyses resulted in four final alignments: (1) 654 bp and 80 specimens for *P. sidae* COI, (2) 654 bp and 62 specimens for *P. cinarae* COI, (3) 611 bp and 23 specimens for *P. sidae* ITS2, and (4) 636 bp and 12 specimens for *P. cinarae* ITS2. As we were not estimating population sizes, a selection of unambiguous haplotypes using tcs v.1.21 (Clement *et al.*, 2000) was used for COI phylogenetic inference resulting in 15 haplotypes for *P. sidae* and 8 haplotypes for *P. cinarae*. *Pyrgus carthami* and *P. armoricanus* were used as outgroups for COI and ITS2, respectively. A coalescent approach was used to reconstruct Bayesian trees for all the alignments using beast v.1.5.3 (Drummond & Rambaut, 2007). For COI trees, HKY + I and GTR + I models of nucleotide substitution were used for *P. sidae* and *P. cinarae*, respectively, and HKY for both ITS2 alignments, according to jModelTest v.0.1 (Posada, 2008) suggestions for the Akaike information criterion (AIC). Parameters were estimated using two independent runs of 10 million generations each (with a pre-run burn-in of 100,000 generations) to ensure convergence.

Estimate of evolutionary entities

Genetic clusters representing evolutionary entities were established using the generalized mixed Yule-coalescent (GMYC) model (Pons *et al.*, 2006; Fontaneto *et al.*, 2007). This model tests for a change in branching rates at the species boundary to classify the observed intervals of genetic divergence to either inter-specific ('diversification') or intra-specific ('coalescent') processes to delimit 'independently evolving' mitochondrial DNA (mtDNA) clusters. All individual haplotypes of COI sequences for *P. sidae* and *P. cinarae*, as well as the outgroups *P. carthami* and *P. communis*, were used to perform the GMYC analysis to detect populations that are evolving independently within the study species. A maximum likelihood phylogeny was obtained with RAxML v.7.0.4 (Stamatakis, 2006) under a GTR + γ substitution model. The resulting topology was made ultrametric using a penalized likelihood as implemented in r8s v.1.7 (Sanderson, 2003). The GMYC analysis was conducted using 'Splits' from the R package (http://www.r-project.org/) with the 'single threshold' option.

Population genetics, demographic and genetic divergence analyses

To describe genetic diversity for each species, we calculated polymorphic sites and nucleotide diversity π_T (i.e. the average number of nucleotide differences per site between two sequences) and its variance (Nei, 1987). To visualize relationships among haplotypes, a statistical parsimony haplotype network was constructed with a 95% connection limit using tcs v.1.21 (Clement *et al.*, 2000), and these haplotypes were hierarchically nested in clades following Templeton's rules (Templeton & Sing, 1993). Closely related haplotypes were generally distributed in geographical proximity, and the tcs analysis resulted in three haplotype clades for *P. sidae* (West, Central and East) and two haplotype clades for *P. cinarae* (West and East) (Fig. 1, see Results). These haplotype clades included more than one population (defined in the sampling design) and both geographical levels (clades and populations) were used in the following analyses of genetic structure (Fig. 1).

To study genetic divergence and current gene flow among populations, we calculated pairwise Φ_{ST} values (an analogue to F_{ST} that incorporates genetic divergence between sequences) between them and their genetic divergence D_{xy} (i.e. the average number of nucleotide substitutions per site between populations; Nei, 1987). In order to assess the level of gene flow among populations we used S_{nn} statistics (nearest-neighbour statistics; Hudson, 2000), excluding the Iberian population of *P. cinarae*, where only one haplotype was observed. In addition, to assess at which spatial scale the genetic variability was structured, i.e. to detect if gene flow was more effective at the level of populations or haplotype clades, we used analysis of molecular variance (AMOVA) with hierarchical partitioning (Excoffier *et al.*, 2005). AMOVA was carried out by estimating Φ_{ST} at three hierarchical levels using 10,000 random permutations. Hierarchical level tests included the following: among haplotype clades; among populations in each haplotype clades; and within populations. A Mantel test was used to detect associations between genetic (i.e. pairwise Φ_{ST} between populations) and geographical distances, which were measured as the shortest distance between the centroids of two populations. We used the software DnaSP v.4.10 (Rozas *et al.*, 2003) to calculate the genetic diversity parameters and S_{nn} statistics. All other analyses were carried out using Arlequin v.3.1 (Excoffier *et al.*, 2005).

In order to detect evidence of population expansion and infer the demographic history of each species, we calculated the statistics Tajima's *D*, Fu's F_S, Fu and Li's *D* and *F* (Ramos-Onsins & Rozas, 2002) using DnaSP v.4.10 (Rozas *et al.*, 2003). Tajima's *D* tested neutrality (i.e. populations evolved under neutrality) against non-random processes (i.e. population expansion or natural selection). To distinguish between the non-random processes, we tested Fu and Li's *D* and *F* and Fu's F_S. The time of divergence between haplotype clades and time to most recent common ancestor (TMRCA) were inferred using the program mdiv (Nielsen & Wakeley, 2001) under the finite sites model. We reported the mode and 95% highest posterior densities (HPD) for each parameter (2,000,000 Markov chain interactions; burn-in = 500,000; $M = 0$; $T_{max} = 5$; θ = auto-initialize). In the absence of a specific mutation rate for *Pyrgus*, we assumed a generation time of 1 year based on the evidence that these species are univoltine (Hernández-Roldán *et al.*, 2009; Wagner, 2009), and a 1.5% Myr^{-1} divergence rate for arthropod COI (Quek *et al.*, 2004) in all historical demographic analyses.

Current and ancestral distribution modelling

The latitude and longitude of the centroid of each 10 × 10 km Universal Transverse Mercator (UTM) square, measured as *x* and *y* coordinates in metres, were selected as spatial variables, considering 50 different distribution points in *P. cinarae* and 154 in *P. sidae* based on the current known distribution of both species in Europe (Abadjiev, 2001; Kudrna, 2002; García-Barros *et al.*, 2004; J. G. Coutsis, Athens, Greece, pers. comm.).

The 19 WorldClim variables (http://www.worldclim.org/, described by Hijmans *et al.*, 2005) were considered. As WorldClim variables generally show a high collinearity that can distort the results obtained, a subselection of variables was employed. Taking into account the study area considered for the distribution modelling analyses (Europe), 10,000 randomly generated points were chosen to extract the values of the WorldClim variables. Using these values, the level of correlation between pairs of variables was analysed. When two variables shared a Pearson correlation coefficient of 0.8 or higher (Rissler & Apodaca, 2007), we selected the biologically most meaningful variable according to the physiological requirements of the *Pyrgus* species (usually that related to the activity of the more sensitive adult stage) or the variable that was easier to interpret (that encompassing a wider temporal range). In this way, 8 out of 19 variables were retained: Bio1 (annual mean temperature), Bio2 (mean diurnal range), Bio7 (temperature annual range), Bio8 (mean temperature of the wettest quarter), Bio12 (annual precipitation), Bio13 (precipitation of the wettest period), Bio15 (precipitation seasonality) and Bio18 (precipitation of the warmest quarter).

To predict the potential distribution models we employed Maxent v.3.3.2 (http://www.cs.princeton.edu/~schapire/maxent/), which uses a machine-learning algorithm to identify the areas in which the environmental conditions are suitable for the species considered in the model. For each species, starting from a uniform distribution, the program performs a number of iterations, each of which increases the probability of the sample locations for the species. The probability is displayed in terms of 'gain', and this gain increases iteration by iteration, until the change from one iteration to the next falls below a specified threshold, or a maximum number of iterations have been performed (Phillips *et al.*, 2006).

The default parameter settings were used (maximum number of background points 10,000; regularization multiplier 1; auto features; maximum iterations 500; convergence threshold 0.00001) as suggested by Phillips *et al.* (2006). Each model was run with 100 replicates and cross-validation, using 25% of the presence data to test the model and 75% to train the model (as suggested and used by other authors, e.g. Moffett *et al.*, 2007; Pawar *et al.*, 2007; Alba-Sánchez *et al.*, 2010). The logistic output was selected due to the easier interpretation of the results (interpreted as probability of presence of the species) compared to raw and cumulative output formats, and the results were presented on a linear scale (Phillips, 2008; Phillips & Dudík, 2008). Values of 0.5 indicate typical presence data points, and the most suitable sites are those whose logistic values are close to 1. The 95% confidence output files were chosen to represent the results, adjusting the threshold of maximum probability of presence to the closest one representing the real known distribution of the species.

To test the accuracy of the models, the area under the receiver operating characteristic curve (AUC) and the following set of 11 binomial tests were performed (Phillips *et al.*, 2006; Moffett *et al.*, 2007; Pawar *et al.*, 2007): (1) fixed cumulative value 1, (2) fixed cumulative value 5, (3) fixed cumulative value 10, (4) minimum training presence, (5) tenth percentile training presence, (6) equal training sensitivity and specificity, (7) maximum training sensitivity plus specificity, (8) equal test sensitivity and specificity, (9) maximum test sensitivity plus specificity, (10) balance training omission, predicted area and threshold value, and (11) equate entropy of thresholded and original distributions. All 11 binomial tests were required to be significant at $P < 0.01$. If the predictions yielded by the model were not better than random, the AUC value would be equal to 0.5. Values of AUC higher than 0.7 (Pearce & Ferrier, 2000; Elith, 2002; Newbold *et al*., 2009) or 0.85 (Newbold, 2009) are considered acceptable.

After calibrating the models for their current distributions in relation to the present climate, we modelled the distribution onto Last Glacial Maximum (LGM) WorldClim data, with two general atmospheric circulation models: CCSM (Community Climate System Model) and MIROC (Model for Interdisciplinary Research on Climate) models from the Paleoclimate Modelling Intercomparison Project Phase II (PMIP2) through WorldClim, using the same eight variables considered in the present models. The original variables (2.5 arcmin) were transformed according to the scale used with DIVA-GIS software v.7.1.7 (Hijmans *et al*., 2005). Only those areas that were recovered as suitable according to both models were considered.

A comparison between the predicted areas in present and past times was made with STATISTICA v.7 (StatSoft, Inc., 2004), representing the elevation and latitude values for both species, in order to test their relationships.

Measurement of the ecological niche

To estimate the ecological niche of each species, we performed a principal components analysis (PCA) on the eight log-transformed WorldClim selected variables (see above). We plotted values on PCA axes 1 and 2 of 178 localities from Europe for *P. sidae*, 51 localities from Europe (except the Iberian Peninsula) for *P. cinarae* and 12 localities of *P. cinarae* from the Iberian Peninsula and assumed these limits as demarcating the ecological niches of each taxon. Additionally, we calculated Pearson correlations between species scores on PCA axes and the environmental data. We tested niche overlap between species conducting a multivariate analysis of variance (MANOVA) and *post hoc* Newman–Keuls tests using PCA species scores as dependent variables, and species as factors. Statistical analyses were computed using the 'ade4' of the R package (http://www.r-project.org/) and STATISTICA v. 7 software (StatSoft, Inc., 2004).

RESULTS

Phylogenetic analyses

The Bayesian phylogenetic tree for *P. sidae* COI haplotypes displayed three main clades with differing geographical origins: one clade included populations from the western part of the distribution, another from the central area, while the third included the eastern populations (Fig. 1a; see localities and haplotypes in Appendix S1). However, only the clade from the central region was statistically supported. For *P. cinarae*, the COI tree showed two strongly supported clades, one including the Iberian Peninsula populations and the other the rest (Fig. 1b). Interestingly, the divergence between these two clades (17 substitutions, 2.60%) was much deeper than those between clades of *P. sidae* (1 substitution, 0.15%), suggesting a much older origin for the disjunct distribution in *P. cinarae*.

The nuclear marker ITS2 resulted in a tree with low intraspecific divergences for *P. sidae* that failed to recover its eastern, central and western clades (Fig. 2a). However, it did recover the same two clades revealed by COI for *P. cinarae*, with monophyly of the western clade being statistically

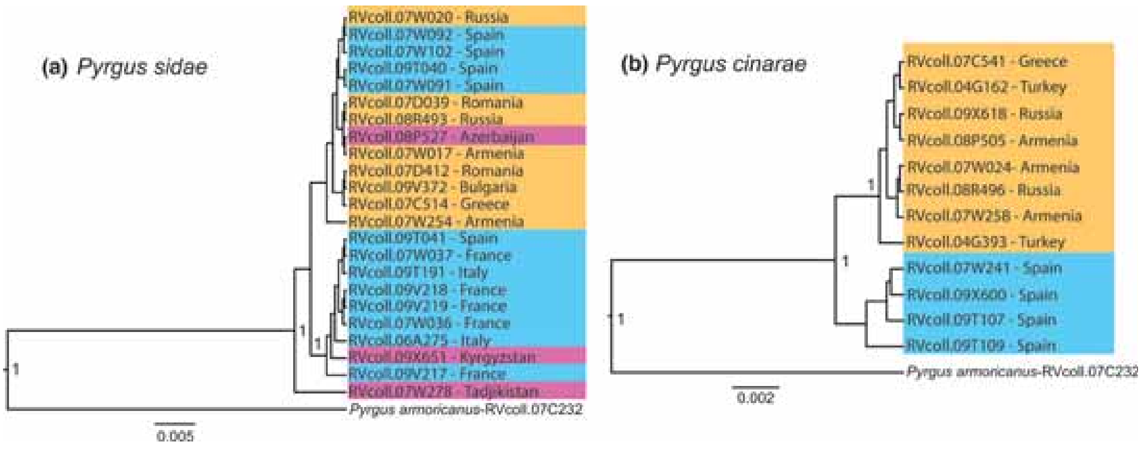


Figure 2 Internal transcribed spacer 2 (ITS2) gene trees for (a) *Pyrgus sidae* and (b) *P. cinarae*. BEAST v.1.5.8 under a coalescent model was used for Bayesian tree inference. Scale bars show divergence in substitutions per site. Only posterior probabilities > 0.9 are shown in the nodes. Colours correspond to the main regions determined by cytochrome *c* oxidase subunit 1 (COI) haplotype clades, as shown on the maps in Fig. 1.

significant (Fig. 2b). GMYC analysis grouped haplotypes into five evolutionarily independent entities (Appendix S2), namely the external groups *P. communis* and *P. carthami*, as well as *P. sidae* and two lineages of *P. cinarae,* one composed exclusively of the specimens from the Iberian Peninsula and the other of the remaining populations.

Population genetics and demographic analyses

From the COI alignments, 10 polymorphic sites (1.52%) with four singleton variable sites and nucleotide diversity (π) of 0.00408 ± 0.00048 were detected for *P. sidae*, whereas *P. cinarae* had 21 polymorphic sites (3.21%) with 15 singleton variable sites and a nucleotide diversity (π) of 0.01032 ± 0.00444. The parsimony haplotype network showed no haplotypes of any species present in all populations. On the contrary, all haplotypes were restricted to single populations, except haplotype 11 for *P. sidae* and haplotype 4 for *P. cinarae*, which were shared by two populations and are possibly ancestral haplotypes (Fig. 1, Table 1). Remarkably, TCS was unable to link the single Iberian haplotype (haplotype 8) of *P. cinarae* to the rest because of very high divergence (17 changes separate this haplotype from the closest). The ancestral haplotypes suggested by TCS for both *P. sidae* and non-Iberian *P. cinarae* were located in the Balkans–Caucasus and south Urals. In the haplotype network, haplotypes were grouped into clades strongly related to their geographical distribution. *Pyrgus sidae* showed three haplotype clades – Eastern (Kyrgyzstan and Tajikistan), Central (south Urals, Caucasus, Black Sea and Balkans) and Western (France–Italy and the Iberian Peninsula) clades. *Pyrgus cinarae* showed similar genetic groups, with two main independent haplotype clades – Iberia and the rest (south Urals, Caucasus, Black Sea and the Balkans), but this species is not present in Kyrgyzstan or Tajikistan. For *P. cinarae* there was one missing haplotype between the ancestral Balkans–Black Sea populations and south Urals or Caucasus, suggesting reduced historical gene flow among them despite the presence of the ancestral haplotype 4 in the south Urals. This is not the case for *P. sidae* because there were no missing haplotypes between populations.

All pairwise Φ_{ST} comparisons among populations were significant for both species, suggesting limited current gene flow among populations (Table 2). Both D_{xy} and Φ_{ST} values between populations were similar, with the notable exception of the *P. cinarae* Iberian population, which was shown to be genetically unique. S_{nn} tests of genetic structure showed differences among populations for *P. sidae* (S_{nn} = 0.46250, P = 0.0010), whereas *P. cinarae* (excluding the Iberian population) showed non-significant differences among populations (S_{nn} = 0.625, P = 0.0620). The absence of genetic structure for *P. cinarae* seems related to the low genetic intra-populational diversity detected for this species. This is most acute in the Iberian Peninsula where only one haplotype was detected among 12 individuals. The AMOVA analyses showed similar patterns for both species, with the highest genetic structure among haplotype clades (50.34% of genetic variation detected for *P. sidae*, 86.03% for *P. cinarae*) (Table 3). In both cases, the genetic differentiation at this spatial scale was non-significant ($P > 0.05$), probably due to the lack of statistical power (low replication on statistical inference based on permutations) because of the low number

Table 1 Cytochrome *c* oxidase subunit 1 (COI) haplotype composition of the sampled populations for *Pyrgus sidae* and *P. cinarae*: populations, nucleotide diversity (π), sampled sites (see Appendix S1 for site description), haplotypes present (*n* = number of individuals) and number of specimens sequenced (*N*). The ancestral haplotype for each species is highlighted.

Species	Haplotype clades	Populations	π	Sites	COI haplotype (*n*)	*N*
P. sidae	East	Kyrgyzstan	0.00306	Kyrgyzstan	2 (17), 3 (1)	18
				Tajikistan	1 (2),	2
	Central	South Urals	0.00204	Russia	6 (4), 8 (1), **11** (3)	8
		Caucasus–Black Sea–Balkans	0.00229	Azerbaijan	5 (2)	2
				Armenia	**11** (12)	12
				Turkey	**11** (2),	2
				Bulgaria	5 (2)	2
				Romania	5 (4), 7 (3)	7
				Greece	5 (2), 10 (1)	3
	West	France–Italy	0.00153	Italy	15 (2)	2
				France	4 (1), 15 (10)	11
		Iberian Peninsula	0.00306	Spain	9 (1), 12 (2), 13 (2), 14 (6)	11
P. cinarae	East	South Urals	0.00306	Russia	1 (19), 2 (1), **4** (1)	21
		Caucasus	0.00153	Armenia	6 (11), 7(2)	13
		Black Sea–Balkans	0.00204	Ukraine	3 (1), **4** (2)	3
				Turkey	**4** (5)	5
				Greece	**4** (7), 5 (1)	8
	West	Iberian Peninsula	0	Spain	8 (12)	12

Table 2 Average number of nucleotide substitutions per site (D_{xy}; Nei, 1987) between populations (above the diagonal), and values of Φ_{ST} from pairwise population comparisons (below the diagonal) for cytochrome *c* oxidase subunit 1 (COI) haplotypes of *Pyrgus sidae* and *P. cinarae*.

P. sidae	Kyrgyzstan	South Urals	Balkans	France	Spain
Kyrgyzstan		0.005	0.005	0.003	0.006
South Urals	0.796*		0.002	0.002	0.004
Balkans	0.770*	0.311*		0.002	0.004
France	0.839*	0.742*	0.647*		0.002
Spain	0.790*	0.665*	0.671*	0.603*	

P. cinarae	South Urals	Caucasus	Balkans	Spain
South Urals		0.005	0.003	0.0295
Caucasus	0.91*		0.004	0.0267
Balkans	0.869*	0.911*		0.029
Spain	0.99*	0.991*	0.992*	

*$P < 0.05$.

Table 3 Analysis of molecular variance (AMOVA) among *Pyrgus sidae* and *P. cinarae* populations grouped by main cytochrome *c* oxidase subunit 1 (COI) haplotype clades (see Fig. 1). Values for the variance components, the percentage of variation at each hierarchical level (%), *F*-statistics and *P*-values are shown.

	Variance components	Percentage of variation	Fixation indices*	*P*-value
P. sidae				
Among haplotype clades	0.615	50.34	$F_{CT} = 0.503$	0.074
Among populations within haplotype clades	0.293	23.99	$F_{SC} = 0.743$	0
Within populations	0.313	25.68	$F_{ST} = 0.743$	0
P. cinarae				
Among haplotype clades	8.086	86.03	$F_{CT} = 0.860$	0.245
Among populations within haplotype clades	1.202	12.79	$F_{SC} = 0.916$	0
Within populations	0.111	1.18	$F_{ST} = 0.988$	0

*Fixation indices for AMOVA measured the degree of genetic differentiation at different organizational levels: F_{CT}, differentiation among haplotype clades relative to total; F_{SC}, differentiation among populations relative to haplotypes clades; F_{ST}, differentiation within populations relative to total populations.

of populations per haplotype clade (Fitzpatrick, 2009). We calculated that the minimum expected *P*-value for the rejection of the null hypothesis for *P. sidae* (two groups of two populations each and one group of one population) was $\alpha = 0.066$, whereas for *P. cinarae* (one group of three populations and one group of one population) it was $\alpha = 0.25$. Thus, the genetic structure among haplotype clades was significant for *P. cinarae* ($P = 0.245$) (Table 3), and close to significant for *P. sidae* ($P = 0.074$). Genetic structure was significant among populations within haplotype clades and within populations for both species, especially for *P. sidae* (Table 3). Non-significant relationships were found between genetic and geographical distances in the Mantel test (*P. sidae* $r = 0.48$, $P = 0.06$; *P. cinarae* $r = 0.74$, $P = 0.25$), which discarded a pure isolation-by-distance model.

Non-random genetic divergence was detected for both species by the negative values of Tajima's *D* ($D = -0.816$ for *P. sidae* and $D = -0.869$ for *P. cinarae*, $P > 0.1$). The results obtained for Fu and Li's *D* ($D = -0.149$ for *P. sidae* and $D = -0.988$ for *P. cinarae*, $P > 0.1$) and Fu and Li's *F* ($F = -0.380$ for *P. sidae* and $F = -1.068$ for *P. cinarae*, $P > 0.1$) indicated that the populations have expanded or are under selection. Population expansion is confirmed by Fu's F_S ($F_S = -17.433$, $P = 0.001$ for *P. sidae* and $F_S = -3.088$, $P = 0.044$ for *P. cinarae*). Therefore, the demographic history of both species suggests the existence of a historical bottleneck and current population expansion. Times of divergence between clades calculated by MDIV models were different for the two species. For *P. sidae* the time of divergence between west and central clades was 0.25 Ma (95% HPD = 0.18–0.29 Ma), and between central and east clades was 0.27 Ma (95% HPD = 0.19–0.39 Ma). For *P. cinarae* the time of divergence between the Iberian clade and the rest was 1.07 Ma (95% HPD = 0.38–2.17 Ma). Estimates for the TMRCA were also different for the two species: TMRCA for *P. sidae* was 0.51 Ma (95% HPD = 0.37–0.72 Ma) for west and central clades and 0.52 Ma (95% HPD = 0.31–0.71 Ma) for east and central clades, while TMRCA for *P. cinarae* was 1.90 Ma (95% HPD = 1.33–2.48 Ma) for the two main clades.

Current and ancestral distribution modelling

The models obtained showed high mean AUC scores (averaged across all 100 runs) in both species (*P. cinarae* 0.989, SD 0.006; and *P. sidae* 0.956, SD 0.011) according to the evaluation test provided by Maxent software. These high AUC values demonstrated a good model performance. Besides, predictions for *P. sidae* and *P. cinarae* were significantly different from random because all 11 binomial omission test thresholds proved significant (*P*-value < 0.01) across all 100 runs.

The predicted distribution for both species was quite similar to the actual one, although there were some differences. In the case of *P. sidae* (Fig. 3a) the areas with higher probability of presence are located in different mountain chains, such as the Iberian System and Pyrenees in the Iberian Peninsula, the Italian Alps, and the Balkans. The Hungarian populations of *P. sidae* seem to be located in a rather unfavourable area (i.e. with lower probability of presence). On the other hand, areas favourable for *P. cinarae* (Fig. 3b) were mainly in the Iberian System and the Balkans. There is a favourable area in the Iberian Peninsula from where *P. cinarae* has never been recorded that deserves deeper inspection in future faunistic surveys.

A heuristic estimate of relative contributions of the environmental variables to the Maxent model is shown in Appendix S3. Variables related to mean temperature and precipitation in the summer (such as precipitation of the warmest quarter, annual temperature range or mean diurnal range) had a prominent role in the model estimated by Maxent and thus seem to represent environmental factors affecting the distribution of both *Pyrgus* species. However, *P. sidae* was more influenced by the annual temperature, and *P. cinarae* by the availability of water in summer dry periods. Jackknife tests of variable importance indicated that annual range of temperature in the case of *P. sidae* and precipitation in the warmest quarter for *P. cinarae* had the highest gain when

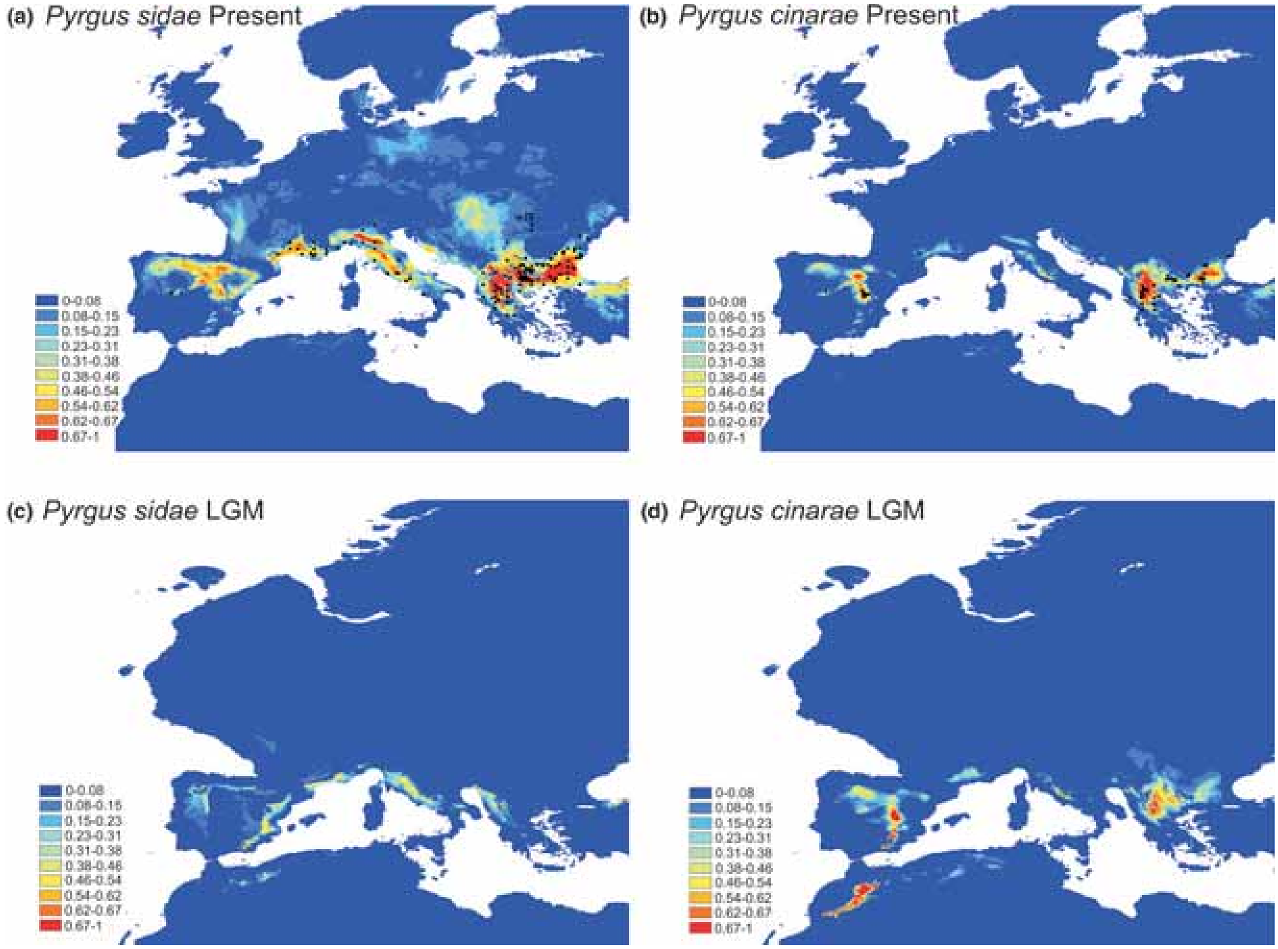


Figure 3 Potential distributions of *Pyrgus sidae* and *P. cinarae* in Europe obtained with Maxent based on current climatic data (a, b), and on the Last Glacial Maximum palaeoclimatic data (c, d). The most probable areas are shown in warm colours. The current known distribution of both species is represented in black.

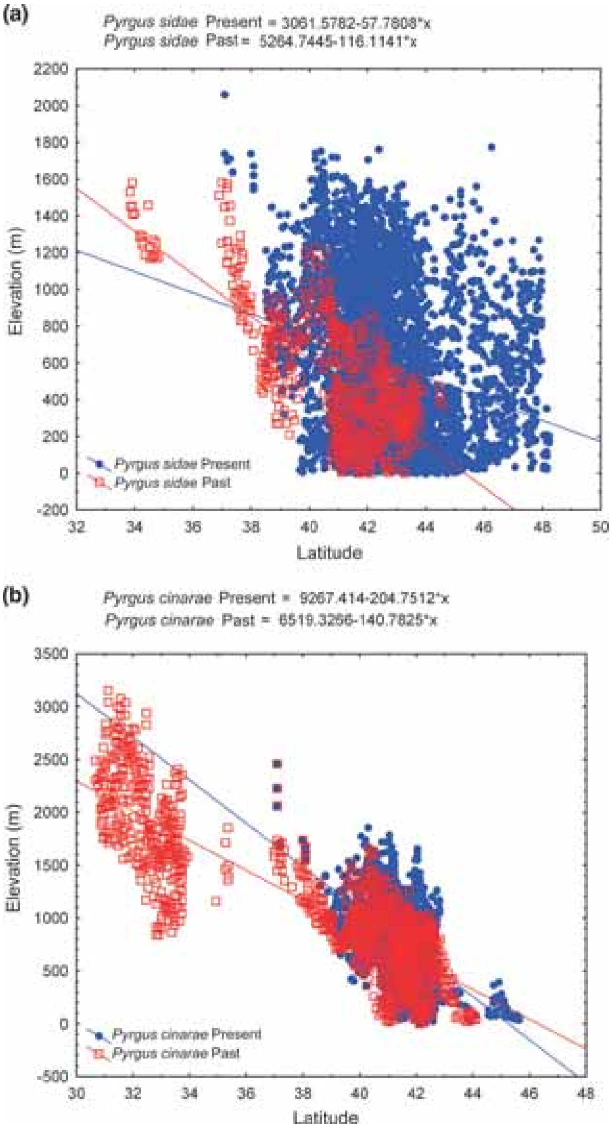


Figure 4 Elevation and latitude values of the predicted areas obtained with MAXENT at present (in blue) and during the Last Glacial Maximum (in red) for (a) *Pyrgus sidae* and (b) *P. cinarae*.

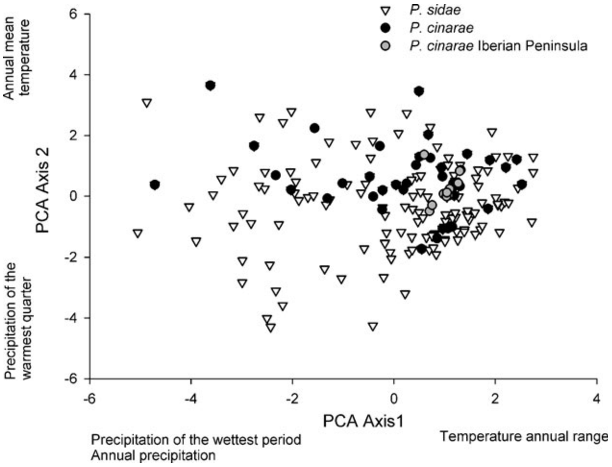


Figure 5 Results of the principal components analysis on environmental descriptors for *Pyrgus sidae* and the two main clades of *P. cinarae*. Environmental descriptors with the highest positive and negative correlations with species scores are indicated on the axes (see details of correlations in Appendix S4).

used in isolation, suggesting that these variables contained the most useful information, and also decreased the gain the most when omitted, suggesting that they contained the most information not present in the other variables.

A comparison of modern to LGM distributions (considering only minimum areas under both the CCSM and MIROC climatic models), revealed that climatically suitable areas have increased for *P. sidae* since the LGM. Its distribution now extends into the Balkan Peninsula (Fig. 3a,c), while suitable areas were previously much smaller, restricted to low elevations and latitudes (Fig. 4a). By contrast, *P. cinarae* encountered similarly favourable areas in both periods (Fig. 3b,d). Therefore, the distribution of *P. sidae* apparently increased in the interglacial periods while *P. cinarae* has been generally stable. Interestingly, connectivity between suitable areas remained quite stable for both species, at least along the coast in the case of *P. sidae*. In both cases there was no continuous suitable habitat connecting present-day populations, although connectivity was always much lower in *P. cinarae*. For both species, the existence of suitable habitat in North African mountains during the LGM was predicted, although these areas are not suitable for the species at present (Fig. 4). Thus, North Africa could have represented a glacial refugium for both species, although it is unlikely that they colonized this region because of their aversion to dispersal across water.

In the PCA, axes 1 and 2 explained 36.72% and 24.11%, respectively, of the total ecological niche variation. Axis 1 was correlated positively with annual temperature range and negatively with precipitation of the wettest period and annual precipitation (Appendix S4). Axis 2 was correlated with average annual temperature (positive values) and with rain in the warmest quarter (negative values). The two-dimensional niche defined by the PCA axes was significantly different among species (Wilks' $\lambda = 0.932$, d.f. = 474, $P < 0.05$). *Pyrgus sidae* occupied the largest ecological space and *P. cinarae*'s preferences were slightly more restricted, although the two species broadly overlapped. The Iberian *P. cinarae* populations displayed a rather small ecological space, clearly nested within that of the rest of *P. cinarae* populations (Fig. 5). Newman–Keuls post hoc tests only indicated niche differentiation for PCA axis 1 for Iberian specimens of *P. cinarae* with *P. sidae* ($P = 0.03$), but not between the remaining populations of *P. cinarae* with *P. sidae* ($P = 0.27$) or between the two main *P. cinarae* clades ($P = 0.14$). Any comparison showed non-significant niche differentiation on PCA axis 2 ($P > 0.05$).

DISCUSSION

Molecular data, both COI and ITS2 sequences, point to an old origin for the disjunct distribution of *P. cinarae*. This is suggested by the topology of the phylogenetic trees and haplotype networks (Fig. 1b), which recover the Iberian population as sister to the rest, by the high divergences separating these two main clades (17 substitutions, 2.6% in COI; 2 substitutions, 0.03% in ITS2), and by the higher pairwise D_{xy} and Φ_{ST} values for *P. cinarae* than for *P. sidae* (Table 2). Age estimates based on MDIV analysis situate the origin of this disjunct distribution at 1.1 Ma (0.4–2.2 Ma), with a TMRCA of 1.9 Ma (1.3–2.5 Ma). Despite the potential error involved in coalescent age estimates, we can safely state that the disjunct distribution in *P. cinarae* is much older than that of *P. sidae*, and that it dates back to the initial Pleistocene glaciations. We can thus infer that during the latest several glacial and interglacial periods there has been no gene flow between the two disjunct groups of populations of *P. cinarae*, or at least it has left no signal in current genetic structure. Indeed, the GMYC analysis (Appendix S2) confirms that the *P. cinarae* Iberian isolate is an independent evolutionary lineage. This scenario is corroborated by distribution modelling results, which show that during the latest glacial to interglacial period the distribution of *P. cinarae* in Europe has been largely unaffected, without major elevational or latitudinal shifts (Fig. 4b). Such a distributional stasis (no important changes in distribution along time) at a general scale suggests that *P. cinarae*, unlike *P. sidae*, does not have the capacity to expand its distribution range to new habitats during interglacial periods. Interestingly, the potential present-day distribution shows that no substantial suitable habitat exists along the 1800 km separating Iberian and Balkan populations (Fig. 3b). The origin of the *P. cinarae* disjunct distribution is probably related to more general climatic trends that occurred during the last several million years, and has been maintained by distributional stasis during more recent glacial to interglacial fluctuations. Nevertheless, at a local scale, glacial cycles possibly had variable demographic and distributional effects on populations of this species. This seems true at least for the Iberian population, which is distributed in a very limited area and displays an extremely low genetic variability (a single COI haplotype and three haplotypes of ITS2 were detected). These particularities, coupled with evidence for demographic expansion in the species as a whole provided by the coalescence analysis, suggest that a recent bottleneck occurred in the Iberian Peninsula. Interestingly, the Iberian Peninsula is the only region where a substantial mismatch between potential and realized distribution is observed, suggesting that the capacity of these populations to expand may be hindered by current low genetic variability and population densities.

Genetic data indicate a much more recent origin for the observed distributional pattern of *P. sidae* compared with that of *P. cinarae*. All COI analyses recovered three main clades (west, central and east) that have low divergences between them (they only differ by single changes). The nuclear marker ITS2 was less informative and did not recover any clear pattern consistent with geographical distance for *P. sidae*, which could be interpreted as a lack of resolution of this marker or as the existence of recent gene flow between populations. The Iberian *P. sidae* population is embedded within the western clade, which also includes south France and central Italy. The MDIV age estimate for the divergence between the western and central clades is *c.* 0.27 Ma (0.18–0.29 Ma). Thus, we can safely assume that the 1000 km disjunct Iberian population originated earlier than this date, most probably during one of the latest glacial events. Given the COI divergences between Iberian and non-Iberian populations of *P. sidae* (one substitution, 0.15%) and of *P. cinarae* (17 substitutions, 2.60%), we believe that the difference in age estimates obtained by MDIV (less than 0.18–0.29 Ma and between 0.38 and 2.17 Ma, respectively) are reasonable despite uncertainty involved in coalescent age estimates.

Isolation since the LGM has been invoked as the cause for many cases of disjunct distributions in Europe and elsewhere (Avise, 2000; Hewitt, 2000). More generally, this has also been suggested to be the cause for patterns of population genetic structure, regardless of the existence of distributional discontinuity at present (Schmitt & Seitz, 2001, 2002; Schmitt & Krauss, 2004; Schmitt *et al.*, 2005; Schmitt, 2007). As the term ‘disjunct distribution’ can be applied at very different spatial scales, it has been used from rather local studies (Williams, 1980; Flinn *et al.*, 2010) to discontinuities of thousands of kilometres (Wahlberg & Saccheri, 2007; Garcia Collevatti *et al.*, 2009; Cagnon & Turgeon, 2010; Peña *et al.*, 2010). We could say that the case of *P. sidae* fits the conventional model of a species that suffers a marked distributional shift southwards and to lower elevations, a demographic reduction, and retains viable populations only in Mediterranean peninsulas that act as refugia during the glacial periods. During the interglacial periods, this species would expand to the European mainland, with the possibility of re-establishing gene flow between isolates, given enough time. In this case, the disjunct distribution is the product of the LGM because gene flow was not re-established. However, our results for *P. cinarae* show that the origin of long-distance disjunct distributions might not always be a direct consequence of one of the latest glacial periods, even in the case of a current distribution in known glacial refugia such as the Mediterranean peninsulas. This species, despite apparent similarity to *P. sidae*, displays unique particularities. One of them is an apparent stasis in the morphology and ecology after *c.* 1 million years of isolation of the Iberian population. *Pyrgus cinarae* also displays a surprising stasis in its potential distribution despite climatic oscillations. Being restricted to habitats not strongly affected by climatic oscillations could be the product of an unusually limited dispersal capability, which would account for the inability to have expanded during interglacial periods, producing a long-term disjunct distribution. If this hypothesis is correct, the predicted dispersal capability of *P. cinarae* should be substantially lower than that of *P. sidae*, which is difficult to imagine because a capture–mark–recapture study on Iberian *P. sidae* showed that this is already a typical sedentary species with low dispersal capabilities (Hernández-Roldán *et al.*, 2009).

In conclusion, we show that *P. cinarae* and *P. sidae* display different population genetic structures and ancestral potential distributions, which reveal that they have undergone different biogeographical histories. Indeed, the effect on these two species of the last glacial to post-glacial period was radically different according to our results: while the distribution of *P. cinarae* was not substantially affected in Europe, that of *P. sidae* greatly changed in both latitude and elevation. Remarkably, the two dissimilar biogeographical histories resulted in very similar distributions across the Palaearctic at present, including a disjunct distribution with isolated populations on the Iberian Peninsula. This could thus be considered a case of convergence in biogeography, even more so when the two species involved belong to the same genus, frequently coexist in the same habitat, share host plants, are univoltine and are usually synchronic. Why then have they not followed parallel biogeographical histories? We demonstrate general ecological similarities, because their ecological niche space broadly overlaps in a PCA analysis. However, differences exist that are not readily obvious without conducting an ecological niche modelling exercise followed by PCA. We show that the rather subtle particularities of each species in their niche preferences result in a characteristic response to environmental change that subsequently determines population genetic structure. This case illustrates how minor ecological differences may lead to very different biogeographical histories and highlights the important role that ecology has played during the evolutionary history of species. By combining molecular data and ecological niche modelling it is possible to reconstruct and understand the history of a species to a fine level. Interpreting biogeographical patterns is nevertheless a complex exercise and the existence of convergence in biogeography should be a warning to avoid generalizations and not to extrapolate results for a taxon to other species, even if they are apparently equivalent from an ecological perspective. Finally, our results highlight the importance of integrating the spatial and temporal dimensions in biogeography.

ACKNOWLEDGEMENTS

We thank Ángel Blázquez, Ernst Brockmann, Vlad Dinca, Zdravko Kolev, Tristan Lafranchis, Vladimir Lukhtanov, Santi Viader, Juan Carlos Vicente and Wolfgang Wagner for help in collecting material. We are also grateful to John G. Coutsis for providing distribution records for both species in Greece, to Joan García-Porta for his very insightful comments regarding ecological niche modelling, and to Blanca Huertas for access to the collection of *Pyrgus* in the Natural History Museum, London. Support for this research was provided by the Spanish Ministerio de Ciencia e Innovación (projects CGL2007-60516/BOS and CGL2010-21226/BOS to J.L.H.-R., G.T. and R.V.; and fellowship BES-2008-002054 to G.T.) and by Genome Canada through the Ontario Genomics Institute to P.D.N.H.

REFERENCES

Abadjiev, S. (2001) *An atlas of the distribution of the butterflies in Bulgaria (Lepidoptera: Hesperioidea & Papilionoidea)*. Pensoft Publishers, Sofia-Moscow.

Alba-Sánchez, F., López-Sáez, J.A., Benito-de Pando, B., Linares, J.C., Nieto-Lugilde, D. & López-Merino, L. (2010) Past and present potential distribution of the Iberian *Abies* species: a phytogeographic approach using fossil pollen data and species distribution models. *Diversity and Distributions*, **16,** 214–228.

Alsos, I.G., Alm, T., Normand, S. & Brochmann, C. (2009) Past and future range shifts and loss of diversity in dwarf willow (*Salix herbacea* L.) inferred from genetics, fossils and modelling. *Global Ecology and Biogeography*, **18,** 223–239.

Avise, J.C. (2000) *Phylogeography: the history and formation of species*. Harvard University Press, Cambridge, MA.

Bocxlaer, I.V., Roelants, K., Biju, S.D., Nagaraju, J. & Bossuyt, F. (2006) Late Cretaceous vicariance in Gondwana amphibians. *PLoS ONE*, **1,** e74.

Cagnon, M.C. & Turgeon, J. (2010) Disjunct distribution in *Gerris* species (Insecta: Hemiptera: Gerridae): an analysis based on spatial and taxonomic patterns of genetic diversity. *Journal of Biogeography*, **37,** 170–178.

Clement, M., Posada, D. & Crandall, K. (2000) TCS: a computer program to estimate gene genealogies. *Molecular Ecology*, **9,** 1657–1660.

Cox, C.B. & Moore, P.D. (2005) *Biogeography: an ecological and evolutionary approach*, 7th edn. Blackwell Publishing, Oxford.

Davis, M.B. & Shaw, R.G. (2001) Range shifts and adaptive responses to Quaternary climate change. *Science*, **292,** 673–679.

Donoghue, M.J. & Moore, B.R. (2003) Toward an integrative historical biogeography. *Integrative and Comparative Biology*, **43,** 261–270.

Drummond, A.J. & Rambaut, A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, **7,** 214.

Elith, J. (2002) Quantitative methods for modeling species habitat: comparative performance and an application to Australian plants. *Quantitative methods for conservation biology* (ed. by S. Ferson and M. Burgman), pp. 39–58. Springer, New York.

Elith, J., Graham, C.H., Anderson, R. *et al.* (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29,** 129–151.

Excoffier, L., Laval, G. & Schneider, S. (2005) Arlequin ver. 3.0: an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online*, **1,** 47–50.

Fitzpatrick, B.M. (2009) Power and sample size for nested analysis of molecular variance. *Molecular Ecology*, **18,** 3961–3966.

Flinn, K.M., Waterway, M.J. & Lechowicz, M.J. (2010) Disjunct performance and distribution in the sedge *Carex prasina*. *Oecologia*, **163,** 119–126.

Fontaneto, D., Herniou, E.A., Boschetti, C., Caprioli, M., Melone, G., Ricci, C. & Barraclough, T.G. (2007) Independently evolving species in asexual bdelloid rotifers. *PLoS Biology*, **5,** 914–921.

Garcia Collevatti, R., Gonçalves Rabelo, S. & Vieira, R.F. (2009) Phylogeography and disjunct distribution in *Lychnophora ericoides* (Asteraceae), an endangered cerrado shrub species. *Annals of Botany*, **104,** 655–664.

García-Barros, E., Munguira, M.L., Martín, J., Romo, H., García-Pereira, P. & Maravalhas, E.S. (2004) *Atlas de las mariposas diurnas de la Península Ibérica y Baleares (Lepidoptera: Papilionoidea & Hesperioidea)*. Monografías SEA, Vol. 11. Zaragoza.

Guisan, A. & Zimmerman, N.E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135,** 147–186.

Habel, J.C., Schmitt, T. & Müller, P. (2005) The fourth paradigm pattern of postglacial range expansion of European terrestrial species: the phylogeography of the marbled white butterfly (Satyrinae, Lepidoptera). *Journal of Biogeography*, **32,** 1489–1497.

Hajibabaei, M., deWaard, J.R., Ivanova, N.V., Ratnasingham, S., Dooh, R.T., Kirk, S.L., Mackie, P.M. & Hebert, P.D.N. (2005) Critical factors for assembling a high volume of DNA barcodes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **360,** 1959–1967.

Hernández-Roldán, J.L., Munguira, M.L. & Martín, J. (2009) Ecology of a relict population of the vulnerable butterfly *Pyrgus sidae* on the Iberian Peninsula (Lepidoptera: Hesperiidae). *European Journal of Entomology*, **106,** 611–618.

Hewitt, G.M. (1996) Some genetic consequences of ice ages, and their role in divergence and speciation. *Biological Journal of the Linnean Society*, **58,** 247–276.

Hewitt, G.M. (1999) Post-glacial re-colonization of European biota. *Molecular Genetics in Animal Ecology*, **68,** 87–112.

Hewitt, G.M. (2000) The genetic legacy of the Quaternary ice ages. *Nature*, **405,** 907–913.

Hewitt, G.M. (2004) Genetic consequences of climatic oscillations in the Quaternary. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **359,** 183–195.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25,** 1965–1978.

Hudson, R. (2000) A new statistic for detecting genetic differentiation. *Genetics*, **155,** 2011–2014.

Hunn, C.A. & Upchurch, P. (2001) The importance of time/space in diagnosing the causality of phylogenetic events: towards a 'chronobiogeographical' paradigm? *Systematic Biology*, **50,** 1–17.

Ivanova, N.V., deWaard, J.R. & Hebert, P.D.N. (2006) An inexpensive, automation-friendly protocol for recovering high-quality DNA. *Molecular Ecology Notes*, **6,** 998–1002.

Jakob, S.S., Ihlow, A. & Blattner, F.R. (2007) Combined ecological niche modelling and molecular phylogeography revealed the evolutionary history of *Hordeum marinum* (Poaceae) – niche differentiation, loss of genetic diversity, and speciation in Mediterranean Quaternary refugia. *Molecular Ecology*, **16,** 1713–1727.

Kudrna, O. (2002) The distribution atlas of European butterflies. *Oedippus*, **20,** 1–342.

Loader, S.P., Pisani, D., Cotton, J.A., Gower, D.J., Day, J.J. & Wilkinson, M. (2007) Relative timescales reveal multiple origins of parallel disjunct distributions of African caecilian amphibians. *Biology Letters*, **3,** 505–508.

Moffett, A., Shackelford, N. & Sarkar, S. (2007) Malaria in Africa: vector species' niche models and relative risk maps. *PLoS ONE*, **2,** e824.

Nei, M. (1987) *Molecular evolutionary genetics*. Columbia University Press, New York.

Newbold, T. (2009) *The value of species distribution models as a tool for conservation and ecology in Egypt and Britain*. PhD Thesis, University of Nottingham, UK.

Newbold, T., Gilbert, F., Zalat, S., El-Gabbas, A. & Reader, T. (2009) Climate-based models of spatial patterns of species richness in Egypt's butterfly and mammal fauna. *Journal of Biogeography*, **36,** 2085–2095.

Nielsen, R. & Wakeley, J. (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, **158,** 885–896.

Nix, H.A. (1986) A biogeographic analysis of Australian elapid snakes. *Atlas of elapid snakes of Australia* (ed. by R. Longmore), pp. 4–15. Australian Government Publishing Service, Canberra.

Noonan, B.P. & Chippindale, P.T. (2006) Vicariant origin of Malagasy reptiles supports Late Cretaceous Antarctic land bridge. *The American Naturalist*, **168,** 730–741.

Pawar, S., Koo, M.S., Kelley, C., Ahmed, F.M., Choudhury, S. & Sarkar, S. (2007) Conservation assessment and prioritization of areas in northeast India: priorities for amphibians and reptiles. *Biological Conservation*, **136,** 346–361.

Pearce, J. & Ferrier, S. (2000) Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, **133,** 225–245.

Peña, C., Nylin, S., Freitas, A.V.L. & Wahlberg, N. (2010) Biogeographic history of the butterfly subtribe *Euptychiina* (Lepidoptera, Nymphalidae, Satyrinae). *Zoologica Scripta*, **39,** 243–258.

Phillips, S.J. (2008) Transferability, sample selection bias and background data in presence-only modelling: a response to Peterson *et al.* (2007). *Ecography*, **31,** 272–278.

Phillips, S.J. & Dudík, M. (2008) Modelling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, **31,** 161–175.

Phillips, S.J., Dudík, M. & Schapire, R.E. (2004) A maximum entropy approach to species distribution modelling. *Proceedings of the 21st International Conference on Machine Learning*, pp. 655–662. ACM Press, New York.

Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006) Maximum entropy modelling of species geographic distributions. *Ecological Modelling*, **190,** 231–259.

Pons, J., Barraclough, T.G., Gomez-Zurita, J., Cardoso, A., Duran, D.P., Hazell, S., Kamoun, S., Sumlin, W.D. & Vogler, A.P. (2006) Sequence-based species delimitation for the

DNA taxonomy of undescribed insects. *Systematic Biology*, **55,** 595–609.

Posada, D. (2008) jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution*, **25,** 1253–1256.

Quek, S.P., Davies, S.J., Itino, T. & Pierce, N.E. (2004) Codiversification in an ant-plant mutualism: stem texture and the evolution of host use in *Crematogaster* (Formicidae: Myrmicinae) inhabitants of macaranga (Euphorbiaceae). *Evolution*, **58,** 554–570.

Ramos-Onsins, S.E. & Rozas, J. (2002) Statistical properties of new neutrality test against population growth. *Molecular Biology and Evolution*, **19,** 2092–2100.

Rissler, L.J. & Apodaca, J.J. (2007) Adding more ecology into species delimitation: ecological niche models and phylogeography help define cryptic species in the black salamander (*Aneides flavipunctatus*). *Systematic Biology*, **56,** 924–942.

Rozas, J., Sánchez-DelBarrio, J.C., Messeguer, X. & Rozas, R. (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*, **19,** 2496–2497.

Sanderson, M.J. (2003) r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, **19,** 301–302.

Schmitt, T. (2007) Molecular biogeography of Europe: Pleistocene cycles and postglacial trends. *Frontiers in Zoology*, **4,** 11.

Schmitt, T. & Hewitt, G.M. (2003) Molecular biogeography of the arctic-alpine disjunct burnet moth species *Zygaena exulans* (Zygaenidae, Lepidoptera) in the Pyrenees and Alps. *Journal of Biogeography*, **31,** 885–893.

Schmitt, T. & Krauss, J. (2004) Reconstruction of the colonization route from glacial refugium to the northern distribution range of the European butterfly *Polyommatus coridon* (Lepidoptera: Lycaenidae). *Diversity and Distributions*, **10,** 271–274.

Schmitt, T. & Seitz, A. (2001) Allozyme variation in *Polyommatus coridon* (Lepidoptera: Lycaenidae): identification of ice-age refugia and reconstruction of post-glacial expansion. *Journal of Biogeography*, **28,** 1129–1136.

Schmitt, T. & Seitz, A. (2002) Postglacial distribution area expansion of *Polyommatus coridon* (Lepidoptera: Lycaenidae) from its Ponto-Mediterranean glacial refugium. *Heredity*, **89,** 20–26.

Schmitt, T., Varga, Z. & Seitz, A. (2005) Are *Polyommatus hispana* and *Polyommatus slovacus* bivoltine *Polyommatus coridon* (Lepidoptera: Lycaenidae)? The discriminatory value of genetics in taxonomy. *Organisms Diversity & Evolution*, **5,** 297–307.

Schmitt, T., Hewitt, G.M. & Müller, P. (2006) Disjunct distributions during glacial and interglacial periods in mountain butterflies: *Erebia epiphron* as an example. *Journal of Evolutionary Biology*, **19,** 108–113.

Stamatakis, A. (2006) Raxml-vi-hpc: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22,** 2688–2690.

StatSoft, Inc. (2004) *STATISTICA (data analysis software system), version 7*. StatSoft, Inc., Tulsa, OK. Available at: http://www.statsoft.com.

Stehlik, I., Holderegger, R., Schneller, J.J., Abbott, R.J. & Bachmann, K. (2000) Molecular biogeography and population genetics of alpine plant species. *Bulletin of the Geobotanical Institute ETH*, **66,** 47–59.

Stockwell, D.R.B. & Peters, D. (1999) The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographical information Science*, **13,** 143–158.

Taberlet, P., Fumagalli, L., Wust-Saucy, A.G. & Cosson, J.E. (1998) Comparative phylogeography and postglacial colonization routes in Europe. *Molecular Evolution*, **7,** 453–464.

Templeton, A. & Sing, C. (1993) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analyses with cladogram uncertainty and recombination. *Genetics*, **134,** 659–669.

deWaard, J.R., Ivanova, N.V., Hajibabaei, M. & Hebert, P.D.N. (2008) Assembling DNA barcodes: analytical protocols. *Methods in molecular biology: environmental genetics* (ed. by C. Martin), pp. 275–293. Humana Press, Totowa, NJ.

Wagner, W. (2009) Zur Ökologie von *Pyrgus cinarae* (Rambur 1839) nebst Beobachtungen zu *Spialia phlomidis* (Herrich-Schäffer 1845) – Larvalhabitat, Präimaginalstadien und Entwicklungszyklus (Lepidoptera: Hesperiidae). *Nachrichten des Entomologischen Vereins Apollo*, **29,** 199–204.

Wahlberg, N. & Saccheri, I. (2007) The effects of Pleistocene glaciations on the phylogeography of *Melitaea cinxia* (Lepidoptera: Nymphalidae). *European Journal of Entomology*, **104,** 675–684.

Weaver, K.F., Anderson, T. & Guralnick, R. (2006) Combining phylogenetic and ecological niche modeling approaches to determine distribution and historical biogeography of Black Hills mountain snails (Oreohelicidae). *Diversity and Distributions*, **12,** 756–766.

White, T.J., Bruns, T., Lee, S. & Taylor, J. (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. *PCR protocols: a guide to methods and applications* (ed. by M.A. Innis, D.H. Gelfand, J.J. Sninsky and T.J. White), pp. 315–322. Academic Press, New York.

Williams, E.H. (1980) Disjunct distributions of two aquatic predators. *Limnology and Oceanography*, **25,** 999–1006.

Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H. & Guisan, A. (2008) Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, **14,** 763–773.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Details of *Pyrgus* spp. and sequences used in the present study.
Appendix S2 GMYC model applied to a maximum likelihood tree of cytochrome *c* oxidase subunit 1 (COI) haplotypes for *Pyrgus sidae* and *P. cinarae*.
Appendix S3 Heuristic estimate of relative contributions of the environmental variables to the Maxent model in *Pyrgus sidae* and *P. cinarae*.
Appendix S4 Pearson correlations between species scores on principal components analysis (PCA) axes and environmental descriptors.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

BIOSKETCH

The Butterfly Diversity and Evolution Lab at the Institut de Biologia Evolutiva (CSIC-UPF) combines molecular, morphological and ecological data to study the biodiversity, biogeography and evolution of Lepidoptera. The Biodiversity Institute of Ontario (University of Guelph) is involved in the oversight of a large-scale programme on DNA barcoding, which employs this approach to probe biodiversity in varied groups of eukaryotes.

Author contributions: J.L.H.-R. and R.V. designed the research and obtained the specimens; J.L.H.-R., G.T. and E.Z. obtained the molecular data; G.T. performed phylogenetic analyses; C.M. performed population genetic, GMYC and principal components analyses; H.R. performed ecological niche modelling analyses based on distribution data compiled by J.L.H.-R. All authors contributed to discussing the results and to writing the paper.

Editor: Brett Riddle

BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Unprecedented within-species chromosome number cline in the Wood White butterfly *Leptidea sinapis* and its significance for karyotype evolution and speciation

Vladimir A Lukhtanov[1,2*†], Vlad Dincă[3,4†], Gerard Talavera[3,4] and Roger Vila[3,5*]

Abstract

Background: Species generally have a fixed number of chromosomes in the cell nuclei while between-species differences are common and often pronounced. These differences could have evolved through multiple speciation events, each involving the fixation of a single chromosomal rearrangement. Alternatively, marked changes in the karyotype may be the consequence of within-species accumulation of multiple chromosomal fissions/fusions, resulting in highly polymorphic systems with the subsequent extinction of intermediate karyomorphs. Although this mechanism of chromosome number evolution is possible in theory, it has not been well documented.

Results: We present the discovery of exceptional intraspecific variability in the karyotype of the widespread Eurasian butterfly *Leptidea sinapis*. We show that within this species the diploid chromosome number gradually decreases from 2n = 106 in Spain to 2n = 56 in eastern Kazakhstan, resulting in a 6000 km-wide cline that originated recently (8,500 to 31,000 years ago). Remarkably, intrapopulational chromosome number polymorphism exists, the chromosome number range overlaps between some populations separated by hundreds of kilometers, and chromosomal heterozygotes are abundant. We demonstrate that this karyotypic variability is intraspecific because in *L. sinapis* a broad geographical distribution is coupled with a homogenous morphological and genetic structure.

Conclusions: The discovered system represents the first clearly documented case of explosive chromosome number evolution through intraspecific and intrapopulation accumulation of multiple chromosomal changes. *Leptidea sinapis* may be used as a model system for studying speciation by means of chromosomally-based suppressed recombination mechanisms, as well as clinal speciation, a process that is theoretically possible but difficult to document. The discovered cline seems to represent a narrow time-window of the very first steps of species formation linked to multiple chromosomal changes that have occurred explosively. This case offers a rare opportunity to study this process before drift, dispersal, selection, extinction and speciation erase the traces of microevolutionary events and just leave the final picture of a pronounced interspecific chromosomal difference.

* Correspondence: lukhtanov@mail.ru; roger.vila@ibe.upf-csic.es
† Contributed equally
[1] Department of Karyosystematics, Zoological Institute of Russian Academy of Science, Universitetskaya nab. 1, 199034 St. Petersburg, Russia
[3] Institut de Biologia Evolutiva (CSIC-UPF), Passeig Marítim de la Barceloneta 37-49, 08003 Barcelona, Spain
Full list of author information is available at the end of the article

Background

Despite the fundamental role of chromosomal change in eukaryotic evolution, the mechanisms related to this process are still poorly known. Main karyotypic features of organisms, such as the number of chromosomes, are usually stable within species [1,2]. This stability is in good correspondence with the fact that new chromosomal rearrangements usually originate as heterozygotes and are often - although not always - associated with heterozygote disadvantage (=negative heterosis; =underdominance). Therefore, their spread to fixation within a large population has low probability [2]. At the same time, differences in karyotype characters between species, including diploid chromosome number (2n), are extremely common. Numerous cases of extraordinary differences in chromosome number, especially in plants, are due to polyploidy [3]. Even when excluding polyploidy, interspecific variation remains very frequent, and many closely related species often have substantially different chromosome numbers. In metazoan animals, the greatest range of within-genus karyotype variation not related to polyploidy is found in *Agrodiaetus* blue butterflies, where diploid chromosome number ranges between species from 2n = 20 to 2n = 268 in spite of morphological similarity and very recent time of species divergence [4]. Interestingly, *Agrodiaetus* also tends to demonstrate the greatest karyotype difference between very closely related species, e.g. sister species *A. biruni* and *A. posthumus* have 2n = 20 and 2n = 180 respectively with no intermediates between them.

In vertebrates, the range of chromosome number variation between closely related species is smaller, yet still impressive. For example, the analysis of 11 species of the catfish genus *Corydoras* revealed that they have karyotypes ranging from 2n = 44 to 2n = 102 [5]. The tuco-tucos, South American rodents of the genus *Ctenomys*, show chromosomal variation with diploid numbers varying from 2n = 10 to 2n = 70 among the 60 species described [6]. The deer genus *Muntiacus* includes species with different karyotypes, ranging from 2n = 6 to 2n = 46 [7]. In plants, the greatest range of within-genus karyotype variation not related to polyploidy is found in *Carex*, where diploid chromosome number ranges from 2n = 12 to 2n = 132 [8].

The discrepancy between intra- and interspecific variability in chromosome numbers poses a serious evolutionary problem. How can numerous species with extremely diverse karyotypes evolve in a relatively short period of time, if major chromosomal rearrangements changing the number of chromosomes are mostly underdominant and, consequently, intraspecific variations are rare and their range is limited?

One possible explanation is that extremely different chromosome numbers evolve gradually through multiple speciation/raciation events, each involving the fixation of a single (or few) chromosomal rearrangement(s), and followed by the subsequent extinction of species or races with intermediate karyotypes. This step-by-step mechanism of karyotype evolution seems to be common in nature, and its initial phase can be observed in some chromosomally polymorphic organisms such as the mouse *Mus musculus domesticus* and the shrew *Sorex araneus* [9-13]. It has been recently demonstrated that the reduction in fertility of hybrids between the house mouse races separated by fixed monobrachial differences is not so pronounced as previously supposed [14]. Nevertheless, this study generally supported the chromosomally-based monobrachial speciation model as a process that accelerates the acquisition of reproductive isolation in the house mouse [14]. In the step-by-step process, the transitional forms are expected to demonstrate a chromosomal fusion/fission polymorphism and, accordingly, numerous examples are known where single or few chromosomal fusions exist in the polymorphic phase, e.g., Robertsonian fusions in *Drosophila americana* [15], melanopline grasshoppers [16] and rodents of the genus *Ctenomys* [6,17].

An alternative hypothesis is that dramatic changes in chromosome number appear as a consequence of a within-species accumulation of numerous chromosomal rearrangements, resulting in highly polymorphic systems with the subsequent extinction of intermediate karyomorphs. A necessary precondition for this mechanism is that major chromosomal rearrangements changing the number of chromosomes are not strongly underdominant. This seems to hold true for different groups such as butterflies, flies, grasshoppers, spiders, fishes and mammals [6,15-25].

While the within-species mechanism of explosive chromosome number evolution is possible in theory, it has been less well documented compared to the evolution through multiple speciation/raciation events. In practice, it is difficult to record such an extensive within-species accumulation for two reasons. First, the transition from one chromosomal form to another may be very fast compared to the species lifespan. The only exception is the chromosomal evolution operated by balancing selection. However, this mechanism seems to be rare, except in the case of inversions [[26,27], but see [28]]. Second, even if polymorphism for multiple chromosomal rearrangements is found, it may be difficult to distinguish between a polymorphic system primarily evolved within a species and a polymorphism resulting from hybridization between different, chromosomally diverged species. For example, in the hybridization zones between low and high chromosome number species of the rodent genus *Ellobius*, there is a so-called "chromosomal fan" including all chromosome numbers from 2n = 31 to 2n = 54 [29]. In fact, this case does not represent evidence for within-species accumulation of chromosomal changes, but simply represents the outcome of secondary parapatry by previously isolated chromosomal races.

Furthermore, the clinal geographical distribution of chromosomal races observed in some organisms [1,2], apparently compatible with gradual within-species accumulation of chromosomal changes, may be better explained by the multiple speciation mechanism. For example, in butterflies of the *Erebia tyndarus* complex there are several geographically isolate chromosomal races (chromosome numbers ranging from 2n = 16 to 2n = 102) [30], and in fossorial mole rats of the *Spalax ehrenbergi* complex four linearly distributed chromosomal races exist (from 2n = 52 to 2n = 60) [31]. In these cases, intrapopulation chromosomal polymorphism is absent and differences between neighbouring chromosomal races, although minor, are fixed. Detailed molecular and morphological studies provide evidence for non-conspecificity of the *E. tyndarus* and *S. ehrenbergi* forms, and several distinct species were identified and formally described [32,33].

In this study we describe a chromosomal cline in the Wood White butterfly, *Leptidea sinapis* (Insecta, Lepidoptera, Pieridae) that provides strong evidence for rapid and extensive within-species chromosome number evolution through accumulation of multiple chromosomal changes. This cline is exceptional in the geographic area that it covers (6000 km) and in its range of within-species chromosome number variation (2n = 56-106). Excluding polyploidy, this is the widest known within-species chromosome number range for any animal or plant, and it is comparable with the highest known level of within-genus karyotype variability.

Results and Discussion

We analyzed the karyotype, mitochondrial and nuclear genetic markers, and the morphology of the Wood White butterfly *L. sinapis*. This is a common species widely distributed from Portugal and Spain in the west to Siberia in the east [34]. From this territory different chromosome numbers have been reported in literature ranging from n = 28 to n = 41 [35]. However, these results are impossible to interpret in practice because of the discovery in 1993 of a cryptic sympatric species (*L. reali*) in Europe and Asia [36]. As all karyotype data for *L. sinapis* were published before this date, it is unclear whether reported chromosome numbers reflect inter- or intraspecific variability.

Our study covers populations from different parts of the *L. sinapis* distribution (Figures 1, 2), as well as the closely related species *L. reali* and *L. morsei* as comparison. We discovered that diploid chromosome number ranges in *L. sinapis* from 2n = 106 in Spain to 2n = 56 in eastern Kazakhstan in a longitudinal cline (Figure 1a; for more details, see Additional file 1). These findings are based on the examination of 209 male specimens, with metaphase plates observed in 35 individuals, out of which 23 had unambiguous chromosome number counts (Spain - 4, France - 2, Italy - 2, Romania - 8, Kazakhstan - 7). We also found that chromosome numbers are not stable within some populations from Italy, Romania and Kazakhstan. Specimens with different chromosome numbers were found within each of these populations, and the great majority of the individuals were chromosomal heterozygotes displaying from one to six multivalents in metaphase I of meiosis (Additional file 1, Figure S1). In the heterozygotes, we observed no abnormalities in the anaphase I stage of meiosis, and the first division of meiosis resulted in normal haploid metaphase II cells where, as expected, two types of metaphase plates with different chromosome numbers were observed. Therefore we conclude that chromosomal rearrangements are not fixed in several of the populations studied, and there seems to be no strong selection against chromosomal heterozygotes. Interestingly, chromosome number range overlaps between some studied populations separated by hundreds of kilometers, e.g. in Kazakhstan between the population from Landman (2n = 56-61) and the population from Saur (2n = 56-64).

In certain species, variation in chromosome number may be caused by the presence of so-called B-chromosomes (=additional chromosomes, =supernumerary chromosomes) [37]. B-chromosomes consist mainly of repetitive DNA and can be usually found in low numbers (one to five) in a percentage of the individuals of a given population. Although they are dispensable, they can sometimes accumulate through processes of mitotic or meiotic drive [38]. B-chromosomes can be distinguished from normal A-chromosomes because they are usually smaller and can be seen as additional chromosomes present in only some of the individuals in a population. The best diagnostic feature is their identity at meiosis, where they may be found as univalents, or in various pairing configurations (bivalents or multivalents), but never pairing with A-chromosomes. Thus, meiotic analysis is critical to distinguish between B-chromosomes and normal A-chromosomes [37,38]. Although we cannot totally exclude that B-chromosomes can be found in *L. sinapis*, especially taking into account that they are known in other genera of the family Pieridae [39], there is good evidence that B-chromosomes are not a valid explanation for the chromosome number cline found in *L. sinapis*. This is due to the fact that in the Spanish population, where the number of chromosomes is maximal (and correspondingly where the highest number of B-chromosomes would be expected), they seem to be completely

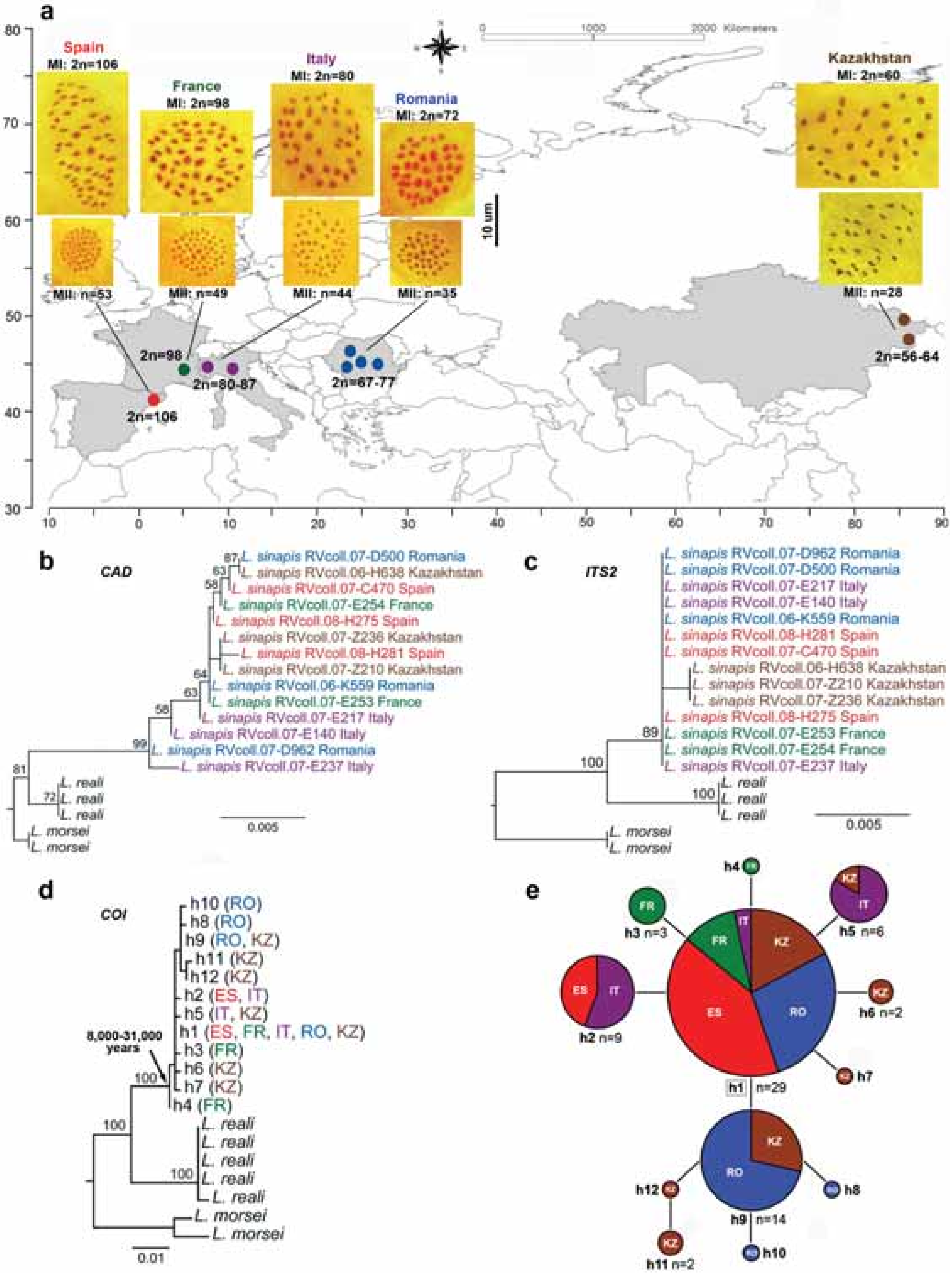

a
Spain
MI: 2n=106
France
MI: 2n=98
Italy
MI: 2n=80
Romania
MI: 2n=72
Kazakhstan
MI: 2n=60
MII: n=53
MII: n=49
MII: n=44
MII: n=35
MII: n=28
10 um
0 1000 2000 Kilometers
2n=98
2n=80-87
2n=67-77
2n=106
2n=56-64
80
70
60
50
40
30
10
0
10
20
30
40
50
60
70
80
90
b
CAD
87 L. sinapis RVcoll.07-D500 Romania
63 L. sinapis RVcoll.06-H638 Kazakhstan
58 L. sinapis RVcoll.07-C470 Spain
L. sinapis RVcoll.07-E254 France
L. sinapis RVcoll.08-H275 Spain
L. sinapis RVcoll.07-Z236 Kazakhstan
L. sinapis RVcoll.08-H281 Spain
L. sinapis RVcoll.07-Z210 Kazakhstan
64 L. sinapis RVcoll.06-K559 Romania
63 L. sinapis RVcoll.07-E253 France
58 L. sinapis RVcoll.07-E217 Italy
L. sinapis RVcoll.07-E140 Italy
99 L. sinapis RVcoll.07-D962 Romania
L. sinapis RVcoll.07-E237 Italy
81
L. reali
72 L. reali
L. reali
L. morsei
L. morsei
0.005
c
ITS2
L. sinapis RVcoll.07-D962 Romania
L. sinapis RVcoll.07-D500 Romania
L. sinapis RVcoll.07-E217 Italy
L. sinapis RVcoll.07-E140 Italy
L. sinapis RVcoll.06-K559 Romania
L. sinapis RVcoll.08-H281 Spain
L. sinapis RVcoll.07-C470 Spain
L. sinapis RVcoll.06-H638 Kazakhstan
L. sinapis RVcoll.07-Z210 Kazakhstan
L. sinapis RVcoll.07-Z236 Kazakhstan
L. sinapis RVcoll.08-H275 Spain
89 L. sinapis RVcoll.07-E253 France
L. sinapis RVcoll.07-E254 France
100 L. sinapis RVcoll.07-E237 Italy
100 L. reali
L. reali
L. reali
L. morsei
L. morsei
0.005
d
COI
h10 (RO)
h8 (RO)
h9 (RO, KZ)
h11 (KZ)
h12 (KZ)
h2 (ES, IT)
h5 (IT, KZ)
h1 (ES, FR, IT, RO, KZ)
8,000-31,000 years
h3 (FR)
h6 (KZ)
h7 (KZ)
h4 (FR)
100
100
L. reali
L. reali
L. reali
100 L. reali
L. reali
L. morsei
L. morsei
0.01
e
h4 FR
FR
h3 n=3
KZ
IT
h5 n=6
ES
IT
h2 n=9
FR
IT
KZ
ES
RO
KZ
h6 n=2
KZ h7
h1 n=29
KZ
RO
h12 KZ
RO h8
h9 n=14
KZ
h11 n=2
RO h10

Figure 1 Chromosomal cline in *Leptidea sinapis* across the Palaearctic region. **a**. Sampling sites and karyotype results. Metaphase plates were observed in 35 individuals, out of which 23 had unambiguous chromosome number counts: Spain - 4, France - 2, Italy - 2, Romania - 8, Kazakhstan - 7. Top row of microphotographs: examples of diploid chromosome number (2n) counted in metaphase I of meiosis (MI). Bottom row of microphotographs: examples of haploid chromosome number (n) counted in metaphase II of meiosis (MII). Maximum likelihood trees for **b**. *CAD*, **c**. *ITS2* and **d**. *COI*. Bootstrap supports (>50%) are shown for each node. **e**. Most parsimonious *COI* haplotype network. Colours refer to each studied region as indicated in the map. ES - Spain, FR - France, IT - Italy, RO - Romania, KZ - Kazakhstan.

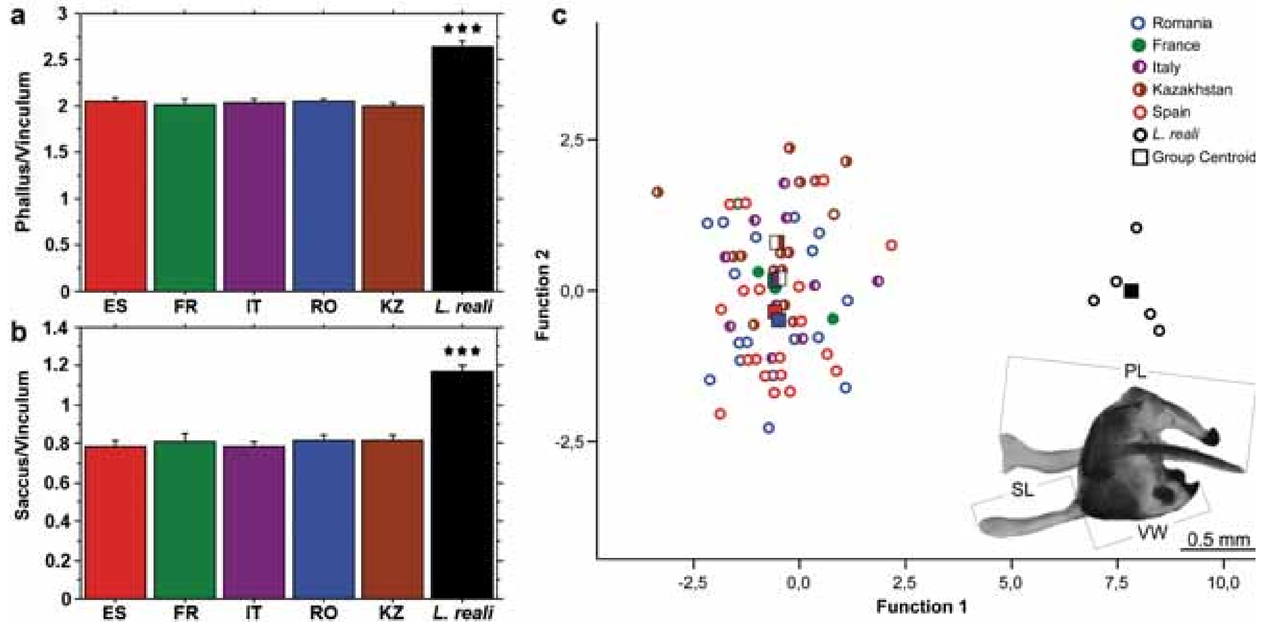


Figure 2 Male genitalia morphology of *Leptidea sinapis* reveals no significant intraspecific differences. One-way ANOVA for **a**. phallus length/vinculum width and **b**. saccus length/vinculum width. The sibling species *L. reali* is included as positive control. Only *L. reali* versus all *L. sinapis* groups is significantly different (p < 0.0001 for both analyses). The bars represent two standard errors. **c**. Canonical discriminant analysis based on phallus length, saccus length and vinculum width.

absent: the chromosome number is stable within as well as between individuals, and no univalents have been observed during meiosis. Moreover, no univalents have been observed during meiosis in any of the other populations studied. Additionally, the following clear pattern was observed: the higher the chromosome numbers in a population, the smaller the size of chromosomes, and vice versa (Figure 1; Additional file 1, Figure S1). This regularity indicates that chromosomal fusions/fissions (but not B-chromosomes) were the main mechanism of karyotype evolution.

Leptidea sinapis can be distinguished from its closest relative *L. reali* by the length of the phallus, saccus and vinculum (in male genitalia) or of the ductus bursae (in female genitalia) [36,40] as well as by molecular markers [41,42]. Therefore, to exclude the possibility of cryptic species involved in the formation of the extraordinarily high chromosomal variability and to demonstrate the conspecificity of the populations studied, we performed morphological and molecular analysis of each studied individual.

The measured variables of the male genitalia showed no significant difference or apparent trend between chromosomal races according to one-way ANOVA (Figure 2a, b) and to discriminant analysis (DA) (Figure 2c). 100% of the *L. reali* were correctly classified to species with the DA, but within *L. sinapis*, between 0 (France and Italy) and 62.5% (Kazakhstan) of specimens were correctly assigned to region (Additional file 1, Table S1).

The mitochondrial *Cytochrome Oxidase I* (*COI*) and nuclear *carbamoyl-phosphate synthetase 2/aspartate transcarbamylase/dihydroorotase* (*CAD*) and *internal transcribed spacer 2* (*ITS2*) markers analyzed did not reveal deep intraspecific levels of divergence (maximum uncorrected p distance of 0.61% for *COI*, 0.7% for *CAD* and 0.16% for *ITS2*) suggesting the absence of cryptic species (Figures 1b-d and Figure 3). The *COI* haplotype network (Figure 1e) shows that the maximum

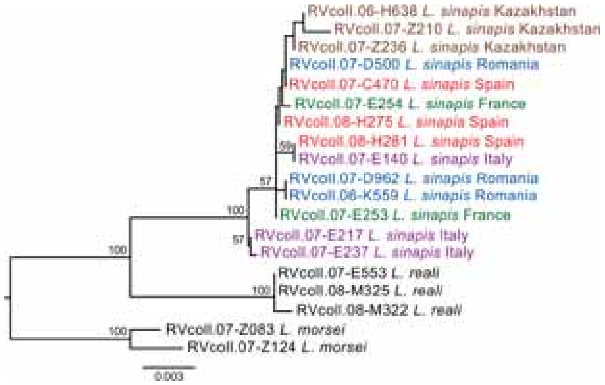

RVcoll.06-H638 L. sinapis Kazakhstan
RVcoll.07-Z210 L. sinapis Kazakhstan
RVcoll.07-Z236 L. sinapis Kazakhstan
RVcoll.07-D500 L. sinapis Romania
RVcoll.07-C470 L. sinapis Spain
RVcoll.07-E254 L. sinapis France
RVcoll.06-H275 L. sinapis Spain
59
RVcoll.06-H281 L. sinapis Spain
RVcoll.07-E140 L. sinapis Italy
57
RVcoll.07-D962 L. sinapis Romania
RVcoll.06-K559 L. sinapis Romania
100
RVcoll.07-E253 L. sinapis France
57
RVcoll.07-E217 L. sinapis Italy
RVcoll.07-E237 L. sinapis Italy
100
RVcoll.07-E553 L. reali
100
RVcoll.08-M325 L. reali
RVcoll.08-M322 L. reali
100
RVcoll.07-Z083 L. morsei
RVcoll.07-Z124 L. morsei
0.003

Figure 3 Maximum Likelihood tree of *Leptidea sinapis* based on the combined analysis of mitochondrial *COI* and nuclear *CAD* and *ITS2* according to the HKY model (log likelihood score = -3159.19036) and 100 bootstrap replicates. The scale bar represents 0.003 substitutions/position.

connection steps are only four, and that the most common haplotype is found in all the studied regions. The observed genetic variability is rather low for an almost pan-Palaearctic species (e.g. [42,43]), even more so since *L. sinapis* is considered a non-migratory poor flyer. The fact that the same low variability is shown by several independent markers rejects a recent mitochondrial genetic sweep and strongly suggests a very recent geographic expansion. Coalescence-based dating with each marker and with all the markers combined estimates that the time to the most recent common ancestor of all the populations is only 8,500 to 31,000 years. Thus, we conclude that there is no evidence for multiple species involved in the formation of the discovered cline, and that its origin is very recent.

It is known that in some systems, variation in chromosome number may be a result of ongoing hybridization between different, chromosomally diverged species [29]. Therefore, the chromosome number variability discovered may be a consequence of hybridization between *L. sinapis* and its sibling species *L. reali*. This explanation may seem possible given that the presence of putative F_1 hybrids between *L. sinapis* and *L. reali* was suggested [44]. However, these results [44] were based on some apparent mismatches between DNA-based identifications (which were congruent for RAPD markers and *COI*) and morphometry of the male genitalia. The classification of the sequenced specimens based on their genitalia was made by employing a bivariate plot, which took into account only the lengths of the phallus and saccus. A recent comprehensive morphometrical study on *L. sinapis* and *L. reali* from Central Italy [40] highlighted the limitation of the “phallus and saccus” approach, which can lead to ambiguous classifications. The same study showed that this limitation can be corrected when using additional genitalic characters (especially the vinculum width) and performing multivariate analyses. Therefore, the report of possible hybrids between *L. reali* and *L. sinapis* requires confirmation since it may actually represent an artifact caused by the interpretation of insufficient morphological traits. Moreover, in case of interspecific hybridization we can expect that some individuals would be heterozygous for species-specific nuclear molecular markers and specimens with intermediate morphology of genitalia should be found. None of the specimens studied in our work has shown these characteristics (see above). Due to genitalic morphological constraints between the two species, introgression is likely to be unidirectional with female *L. sinapis* potentially inseminated by male *L. reali* [36,44]. Finally, several studies dealing with the mating behaviour of *L. sinapis* and *L. reali* reported that females of both species exclusively mated with conspecific males, suggesting the presence of strong precopulatory barriers [36,45,46]. Therefore, we can conclude that interspecific hybridization is an unlikely explanation for the origin of the discovered chromosomal cline.

The clinal distribution of chromosome numbers in *L. sinapis* is statistically significant ($p < 0.0001$) and it is very unlikely to have arisen by chance (Figure 4). Interestingly, the cline is longitudinally oriented (Figure 1a), indicating either the direction of selective pressure involved in its formation, or the direction of population dispersal, or both of these processes. According to our dating, the moment of this dispersal would correspond to the upper Pleistocene and the Holocene, a period characterized by a strong glaciation in northern Europe and the Alps [47]. Thus, our estimates indicate that the dispersal of *L. sinapis* could have occurred before or after the last glacial maximum (24,000 to 17,000 years ago).

Several other cases of broad intraspecific chromosomal polymorphism have been described in animals [6,18-21,23,24,48-56] and plants [8,57]. However, all these cases differ from the cline found in *L. sinapis* by the essentially smaller range of karyotype variability and by the possible existence of two or more cryptic species involved in the formation of the polymorphic chromosomal system. In order to demonstrate the intraspecific nature of karyotype variability, the following three criteria should be met simultaneously: 1) segregating chromosomal polymorphism within a population should be demonstrated, 2) molecular markers should not suggest the presence of potential cryptic species, and 3) species-

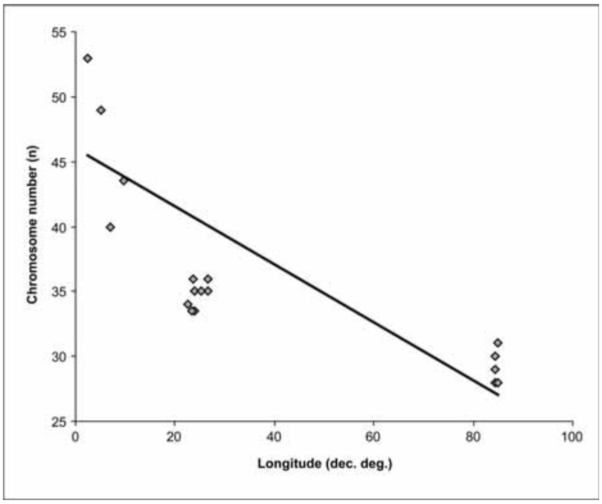


Figure 4 Variation of *L. sinapis* chromosome number across geographical longitude. Chromosome number is inversely correlated with longitude according to a linear function (r = 0.826; $p < 0.0001$). Results based on 23 specimens with unambiguous chromosome number counts (Spain - 4, France - 2, Italy - 2, Romania - 8, Kazakhstan - 7).

diagnostic morphological differences should be lacking. To our knowledge, only studies on the common shrew and the house mouse have met all these criteria, but chromosomal races within these mammals have essentially smaller differences in chromosome number and apparently evolved through a step-by-step accumulation of single chromosomal rearrangements [9-13] rather than through wide intraspecific and intrapopulation chromosome number polymorphism.

Conclusions

Given that (a) chromosomal races of *L. sinapis* belong to the same species, (b) intrapopulation chromosome number polymorphism exists, (c) the chromosome number range overlaps between some populations separated by hundreds of kilometers, (d) the species has broad ecological preferences and is widely distributed, (e) the species has a rather homogenous genetic structure, and (f) chromosomal heterozygotes are abundant, this represents a clearly documented case of rapid and massive within-species accumulation of multiple chromosomal rearrangements affecting the number of chromosomes.

The chromosomal rearrangements discovered in our investigation display segregating polymorphism that seems not to strongly affect reproductive fitness within the populations studied. However, these rearrangements are not necessarily irrelevant to the process of formation of reproductive isolation (i.e. to speciation). It is well known that Robertsonian rearrangements (i.e. nonreciprocal translocations involving fission and fusion at or near a centromere), have the potential to limit gene flow and drive speciation [58,59]. The Wood White butterfly, like other Lepidoptera and some other insects, has holokinetic chromosomes in which the centromere is not localized and centromeric activity is distributed along the length of the chromosome [35,60-62]. It has been recently demonstrated that fusions/fissions of holokinetic chromosomes restrict gene flow too, and that this effect is cumulative (i.e. increases proportionally with the level of chromosomal differences) [57]. In the case of *L. sinapis* all evidence suggests that neighbour populations with relatively low differences in chromosome number are reproductively compatible. We cannot exclude that geographically distant and chromosomally divergent populations would display reduced fertility if crossed, although they are connected by a chain of compatible populations that should allow gene flow. Therefore, the discovered system opens the possibility to study clinal speciation, a process that is theoretically possible but difficult to document [[63], pages 113-123].

Chromosomal rearrangements are known to limit introgression in parapatry or sympatry with regard to isolation genes, thus facilitating the maintenance of incipient species boundaries [64,65], and serving as regions where isolation genes can accumulate [15,27,66-68]. The preservation and/or accumulation of isolation genes protected by chromosomal rearrangements could represent a prerequisite for speciation by means of suppressed-recombination mechanisms [15,27,64-68].

In conclusion, the *L. sinapis* chromosomal cline seems to represent a narrow time-window of the very first steps of species formation linked to multiple chromosomal changes that have occurred explosively. This case offers a rare opportunity to study this process before drift, dispersal, selection, extinction and speciation erase the traces of microevolutionary events and just leave the final picture of a pronounced interspecific chromosomal difference.

Methods

Note: During the publication process of this paper it has been shown that the Romanian specimens of *L. reali* used here as outgroup actually belong to a new cryptic species named *Leptidea juvernica* [69].

Sample collecting

Fresh male *Leptidea* specimens (Additional file 1, Table S2) were collected with the insect net and were kept alive in glassine envelopes. In the laboratory, butterflies were killed by pressing the thorax and testes were removed from the abdomen and immediately placed into a 0.5 ml vial with freshly prepared Carnoy fixative (ethanol and glacial acetic acid, 3:1). Bodies were placed into a 2 ml plastic vial with 100% ethanol for DNA analysis and wings were stored in glassine envelopes. Each sample has been assigned a unique sample ID. All the samples are stored in Roger Vila's DNA and Tissues Collection in Barcelona, Spain.

Genitalia preparation and morphometric analyses

Male genitalia were prepared according to the following protocol: maceration for 15 minutes at 95°C in 10% potassium hydroxide, dissection and cleaning under a stereomicroscope and storage in tubes with glycerin. Genitalia were photographed laterally (Figure 2c), without being pressed, in a thin layer of distilled water under a Carl Zeiss Stemi 2000-C stereomicroscope equipped with a DeltaPix Invenio 3S digital camera. Measurements were performed based on the digital photographs by using AxioVision software. A total of 73 specimens of *L. sinapis* were included in the morphometrical analyses (Additional file 1, Table S3). These included 35 of the karyotyped samples, and 38 individuals collected in the same locality and moment for which the cytogenetic studies did not produce results. In addition, five specimens of the sibling *L. reali* were added as outgroup. Three elements of the male genitalia were measured: phallus, saccus and vinculum width.

These are the best diagnostic characters to separate *L. sinapis* from *L. reali* [40]. The vinculum width was used to normalize the size of the specimen.

StatView 5.0.1 (SAS Institute Inc., 1992-1998) was used to perform one-way ANOVA in order to test for differences in the length of the phallus and saccus, each normalized by the width of the vinculum, between regions for *L. sinapis*, and between *L. sinapis* and *L. reali*. All variables were normally distributed (Kolmogorov-Smirnov Test, p > 0.05). The software SPSS 14.0 was used to perform a discriminant analysis by employing the stepwise method. The Box's M test was used to evaluate the homogeneity of covariance assumption (p > 0.05). The variables were selected with the Wilks' lambda statistic, which measures how each function separates cases into groups. In order to test the obtained classification a cross validation was carried out.

Karyotype analyses

Gonads were stored in Carnoy fixative (ethanol and glacial acetic acid, 3:1) for 2-6 months at 4ï,°C and then stained with 2% acetic orcein for 30 days at 20ï,°C as it was previously described [70,71].

Chromosomes of butterflies (Lepidoptera) are small, numerous and uniform in both shape and size [35]. They lack a distinct primary constriction (the centromere) and are regarded as holokinetic with kinetochores extended over a large portion of the chromosome surface [60]. The uniformity of lepidopteran chromosomes, the absence of morphological markers such as the centromeres and the lack of convenient differential banding techniques [61] make difficult the identification of individual chromosomes by standard cytogenetic methods. Although new approaches to individual identification of the Lepidoptera chromosomes based on the fluorescent *in situ* hybridization (FISH) technique have been recently elaborated [72-74], they are applicable only for studying species bred in the laboratory. For this reason, the chromosome number remains the most commonly used karyotypic character in Lepidoptera cytogenetics and karyosystematics. In our study we counted the diploid chromosome numbers (2n) in mitotic spermatogonial cells and the haploid chromosome numbers (n) in metaphase II of male meiosis. We also counted the number of chromosomal elements (n) (bivalents + multivalents) in metaphase I of male meiosis. In the last case, the number of chromosomal elements was equal to the haploid number (n) if all the elements were represented by bivalents, or less if some elements were represented by multivalents. To distinguish between bivalents and multivalents, we used a special method [75]. Briefly, by varying the pressure on the coverslip, we were able to manipulate chromosomes, e.g. change their position and orientation in intact (not squashed) spermatocyte cells, and consequently to analyze the structure of the bivalents and multivalents.

In total, preparations from 209 males were analyzed. As cell divisions are extremely rare in *Leptidea* during imago stage [76], metaphase plates were observed in only 35 individuals (Additional file 1, Table S2). These individuals were also used for morphological and molecular analysis.

Geographical longitude vs. chromosome number

Pearson correlation coefficients were used to assess the degree of association between haploid karyotype and geographical longitude. Longitude was measured in decimal degrees and only 23 samples with unambiguous chromosome number counts were included (see Additional file 1, Table S4). If the specimen showed different chromosome numbers in different cells, the average between the different chromosome numbers was used.

Specimen sequencing

Total genomic DNA was extracted using Chelex 100 resin, 100-200 mesh, sodium form (Biorad), under the following protocol: one leg was removed and introduced into 100 μl of Chelex 10% and 5 μl of Proteinase K (20 mg/ml) were added. The samples were incubated overnight at 55°C, afterwards were incubated at 100°C for 15 minutes and were subsequently centrifuged for 10 seconds at 3000 rpm.

A 676 bp fragment at the 5' end of the mitochondrial gene cytochrome oxidase subunit I (*COI*) was amplified by polymerase chain reaction using the primers LCO 1490 (5'-GGTCAACAAATCATAAAGATATTGG-3') [77] and Nancy (5'-CCCGGTAAAATTAAAATA-TAAACTTC-3') [78]. When these primers failed, we used the primers LepF1 (5'-ATTCAACCAATCATAAA-GATATTGG-3') and LepR1 (5'-TAAACTTCTG-GATGTCCAAAAAATCA-3') [79], which amplified a 658 bp fragment of *COI*. Double-stranded DNA was amplified in 25 μl volume reactions: 13.22 μl ultra pure (HPLC quality) water, 2.5 μl 10× buffer, 4.5 μl 25 mM MgCl2, 0.25 μl 100 mM dNTP, 1.2 μl of each primer (10 mM), 0.13 μl *Taq* DNA Gold Polymerase (Qiagen) and 2 μl of extracted DNA. The typical thermal cycling profile was: 95°C for 60 seconds, 44°C for 60 seconds and 72°C for 90 seconds, for 40 cycles. A total of 70 *L. sinapis* samples were successfully sequenced for this marker. These included 34 of the karyotyped samples, and 36 individuals collected in the same locality as the karyotyped samples. Five *L. reali* and two *L. morsei* specimens were also sequenced and used as outgroup.

A 571 bp fragment at the 5' end of the nuclear gene *CAD* was amplified by polymerase chain reaction using the primers CADFa (5'-GDATGGTYGATGAAAATGT-TAA-3') and CADRa (5'- CTCATRTCGTAATCYG-TRCT-3') (designed by A. Kaliszewska). Double-stranded

DNA was amplified in 25 μl volume reactions: 16.65 μl ultra pure (HPLC quality) water, 2.5 μl 10× buffer, 1 μl 100 mM MgCl2, 0.25 μl 100 mM dNTP, 1.2 μl of each primer (10 mM), 0.2 μl *Taq* DNA Polymerase (Bioron, GmbH) and 2 μl of extracted DNA. The typical thermal cycling profile was: 95°C for 60 seconds, 48°C for 60 seconds and 72°C for 90 seconds, for 40 cycles. A total of 14 samples (all karyotyped) were sequenced for this marker. Three *L. reali* and two *L. morsei* specimens were also sequenced and used as outgroup.

A 684 bp fragment at the 5' end of the nuclear internal transcribed spacer 2 (*ITS2*) was amplified by polymerase chain reaction using the primers ITS3 (5'-GCATCGATGAAGAACGCAGC-3') and ITS4 (5'-TCCTCCGCTTATTGATATGC-3') [80]. Double-stranded DNA was amplified in 25 μl volume reactions: 16.7 μl ultra pure (HPLC quality) water, 2.5 μl 10× buffer, 1 μl 100 mM MgCl2, 0.25 μl 100 mM dNTP, 1.2 μl of each primer (10 mM), 0.15 μl *Taq* DNA Polymerase (Bioron, GmbH) and 2 μl of extracted DNA. The typical thermal cycling profile was: 95°C for 45 seconds, 47°C for 60 seconds and 72°C for 60 seconds, for 40 cycles. A total of 14 samples (all karyotyped) were sequenced for this marker. Three *L. reali* and two *L. morsei* specimens were also sequenced and used as outgroup. PCR products were purified and sequenced by Macrogen Inc. (Seoul, Korea). Sequences obtained specifically for this study were deposited in GenBank (accession numbers indicated in Additional file 1, Table S2).

Sequence alignment and phylogenetic inference

COI, *ITS2* and *CAD* sequences were edited and aligned using Geneious Pro 4.7.5 [81]. These resulted in three final alignments of 658 bp and 77 specimens for *COI*, 571 bp and 19 specimens for *CAD*, and 684 bp and 19 specimens for *ITS2*. For *COI*, duplicate haplotypes were removed using Collapse 1.2 [82]. Maximum Likelihood (ML) phylogenetic trees were inferred for *CAD*, *ITS2* and *COI* using Phyml 2.4.4 [83], with the nucleotide substitution model HKY [84] for nuclear markers and HKY+I for *COI*, as suggested by jModeltest 0.1 [85], and 100 bootstrap replicates.

Haplotype network

In order to examine relationships among haplotypes, a maximum parsimony haplotype network was constructed using TCS 1.21 [86]. The haplotype network was built with a 99% parsimony connection limit. The network presented one loop, which was broken according to frequency and geographic criteria [87].

Estimation of TMRCA

Time to the most recent common ancestor (TMRCA) of *L. sinapis* was inferred with BEAST v.1.5.3 [88] independently for *COI*, *ITS2* and *CAD* haplotypes under a Coalescent model with constant population size. Duplicate haplotypes were removed from the matrix using Collapse 1.2 [82]. A lognormal distribution (Mean = 0.15, Stdev = 0.798) was used assuming a maximum possible limit of 405000 years as the 95% HPD of the distribution, trying to let the maximum exploratory space to MCMC runs. To estimate this prior, we used the maximum *COI* intraspecific divergence for *L. sinapis* under a rather slow invertebrate mitochondrial substitution rate: 1.5% uncorrected pairwise distance per million years [89]. Since substitution rates are known to overestimate ages for recent lineages still under the coalescence process, we are certain that 405000 years is a good maximum estimate for the TMRCA of this species. The dataset was analyzed using the HKY model and applying a strict molecular clock along the branches. Base frequencies were estimated and a randomly generated initial tree was used. Parameters were estimated using two independent runs of 10 million generations each (with a pre-run burn-in of 100,000 generations) to ensure convergence, checked with the program Tracer v1.4.

A multi-locus approach with *BEAST [90] was also employed to check the results with a smaller set of 12 samples, including those with most divergent *COI* haplotypes. In order to study the effect of the outgroup, *COI* and multilocus analyses were conducted by both including and excluding *L. reali* haplotypes (Additional file 1, Table S5).

Additional material

Additional file 1: Additional Text, Figures and Tables. a) Additional results of chromosomal analyses. b) Figure S1. Karyotypes of *Leptidea sinapis*. c) Table S1. Discriminant analysis classification results for chromosomal races of *L. sinapis* and *L. reali*. d) Table S2. List of specimens included in this study. e) Table S3. Results of morphometric analysis of the male genitalia. f) Table S4. List of the specimens included in the analysis of geographical longitude vs. chromosome number. g) Table S5. Estimation of TMRCA of *L. sinapis* under a coalescent model.

Acknowledgements

We thank L. Dapporto, A.-M. Hereş, J. Martínez-Vilalta and C. Sañudo for advice on the statistical analyses and to I.Y. Bakloushinskaya and V.G. Kuznetsova for suggestions on the manuscript. A. Kaliszewska designed the primers for CAD. Support for this research was provided by the Russian Foundation for Basic Research (grants RFFI 09-04-01234, 11-04-00734 and 11-04-01119), by grant NSH-3332.2010.4 (Leading Scientific Schools) and by the programs of the Presidium of Russian Academy of Science 'Gene Pools and Genetic diversity' and 'Origin of biosphere and evolution of geo-biological systems' to V.A.L.; by the Spanish Ministerio de Ciencia e Innovación (project CGL2007-60516/BOS to V.D., G.T. and R.V., and predoctoral fellowship BES-2008-002054 to G.T.); and by a predoctoral fellowship from Universitat Autònoma de Barcelona to V.D.

Author details

[1]Department of Karyosystematics, Zoological Institute of Russian Academy of Science, Universitetskaya nab. 1, 199034 St. Petersburg, Russia. [2]Department

of Entomology, St. Petersburg State University, Universitetskaya nab. 7/9, 199034 St. Petersburg, Russia. [3]Institut de Biologia Evolutiva (CSIC-UPF), Passeig Marítim de la Barceloneta 37-49, 08003 Barcelona, Spain. [4]Departament de Genètica i Microbiologia, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain. [5]Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Spain.

Authors' contributions

VAL, VD, GT and RV designed the experiments and analyzed the data, VAL performed the analysis of karyotypes, VD and GT performed phylogenetic analyses, VD performed morphological analyses, VAL, VD and RV wrote the paper, and all co-authors contributed in the form of discussion and suggestions. All authors read and approved the final manuscript.

Received: 7 December 2010 Accepted: 20 April 2011
Published: 20 April 2011

References

1. White MJD: *Animal Cytology and Evolution* Cambridge: Cambridge University Press; 1973.
2. King M: *Species Evolution: The Role of Chromosomal Change* Cambridge: Cambridge University Press; 1993.
3. Coghlan A, Eichler EE, Oliver SG, Paterson AH, Stein L: **Chromosome evolution in eukaryotes: a multi-kingdom perspective.** *Trends Genet* 2005, **21**:673-682.
4. Lukhtanov VA, Kandul NP, Plotkin JB, Dantchenko AV, Haig D, Pierce NE: **Reinforcement of pre-zygotic isolation and karyotype evolution in *Agrodiaetus* butterflies.** *Nature* 2005, **436**:385-389.
5. Shimabukuro-Dias CK, Oliveira C, Foresti F: **Karyotype variability in eleven species of the catfish genus *Corydoras* (Siluriformes: Callichthyidae).** *Ichtyiol Explor Freshwaters* 2004, **15**:135-146.
6. Giménez MD, Mirol PM, Bidau CJ, Searle JB: **Molecular analysis of populations of *Ctenomys* (Caviomorpha, Rodentia) with high karyotypic variability.** *Cytogenet Genome Res* 2002, **96**:130-136.
7. Huang L, Wang J, Nie W, Su W, Yang F: **Tandem chromosome fusions in karyotypic evolution of *Muntiacus*: evidence from *M. feae* and *M. gongshanensis*.** *Chrom Res* 2006, **14**:637-647.
8. Hipp AL: **Nonuniform processes of chromosome evolution in sedges (*Carex*: Cyperaceae).** *Evolution* 2007, **61**:2175-2194.
9. Britton-Davidian J, Catalan J, Ramalhinho MD, Ganem G, Auffray JC, Capela R, Biscoito M, Searle JB, Mathias MD: **Rapid chromosomal evolution in island mice.** *Nature* 2000, **403**:158.
10. Pialek J, Hauffe HC, Rodriguez-Clark KM, Searle JB: **Raciation and speciation in house mice from the Alps: the role of chromosomes.** *Mol Ecol* 2001, **10**:613-625.
11. Castiglia R, Annesi F, Capanna E: **Geographical pattern of genetic variation in the Robertsonian system of *Mus musculus domesticus* in central Italy.** *Biol J Linn Soc* 2005, **84**:395-405.
12. White TA, Bordewich M, Searle JB: **A network approach to study karyotypic evolution: the chromosomal races of the common shrew (*Sorex araneus*) and house mouse (*Mus musculus*) as model systems.** *Syst Biol* 2010, **59**:262-276.
13. Franchini P, Castiglia R, Capanna E: **Reproductive isolation between chromosomal races of the house mouse *Mus musculus domesticus* in a parapatric contact area revealed by an analysis of multiple unlinked loci.** *J Evol Biol* 2008, **21**:502-513.
14. Nunes AC, Catalan J, Lopez J, da Graça Ramalhinho M, da Luz Mathias M, Britton-Davidian J: **Fertility assessment in hybrids between monobrachially homologous Rb races of the house mouse from the island of Madeira: implications for modes of chromosomal evolution.** *Heredity* 2011, **106**:348-356.
15. McAllister BF, Sheeley SL, Mena PA, Evans AL, Schlötterer C: **Clinal distribution of a chromosomal rearrangement: a precursor to chromosomal speciation?** *Evolution* 2008, **62**:1852-1865.
16. Bidau CJ, Mirol PM: **Orientation and segregation of Robertsonian trivalents in *Dichroplus pratensis* (Acrididae).** *Genome* 1988, **30**:947-955.
17. Fornel R, Cordeiro-Estrela P, De Freitas TRO: **Skull shape and size variation in *Ctenomys minutus* (Rodentia: Ctenomyidae) in geographical, chromosomal polymorphism, and environmental contexts.** *Biol J Linn Soc* 2010, **101**:705-720.
18. Sharp HE, Rowell DM: **Unprecedented chromosomal diversity and behaviour modify linkage patterns and speciation potential: structural heterozygosity in an Australian spider.** *J Evol Biol* 2007, **20**:2427-2439.
19. Qumsiyeh MB, Coate JL, Peppers JA, Kennedy PK, Kennedy ML: **Robertsonian chromosomal rearrangements in the short-tailed shrew, *Blarina carolinensis*, in western Tennessee.** *Cytogenet Cell Genet* 1997, **76**:153-158.
20. Qumsiyeh MB, Barker S, Dover S, Kennedy PK, Kennedy MP: **A potential model for early stages of chromosomal evolution via concentric Robertsonian fans: A large area of polymorphism in southern short-tailed shrews (*Blarina carolinensis*).** *Cytogen Cell Genet* 1999, **87**:27-31.
21. Völker M, Sonnenberg R, Kullmann RPH: **Karyotype differentiation in *Chromaphyosemion* killifishes (Cyprinodontiformes, Nothobranchiidae). III: Extensive karyotypic variability associated with low mitochondrial haplotype differentiation in *C. bivittatum*.** *Cytogen Genome Res* 2007, **116**:116-126.
22. Nagaraju J, Jolly MS: **Interspecific hybrids of *Antheraea roylei* and *A. pernyi* - A cytogenetic reassessment.** *Theor App Genet* 1986, **72**:269-273.
23. Nachman MW, Myers P: **Exceptional chromosomal mutations in a rodent population are not strongly underdominant.** *Proc Natl Acad Sci USA* 1989, **86**:6666-6670.
24. Kerridge DC, Baker RJ: **Genetic variation and origin of the most chromosomally polymorphic natural mammalian population.** *Cytogen. Cell Genet* 1990, **53**:5-7.
25. Nogueira CDA, Fagundes V: ***Akodon cursor* Winge, 1887 (Rodentia: Sigmodontinae): one or two species? New evidences based on molecular data.** *Zootaxa* 2008, **1768**:41-51.
26. Dobzhansky T: *Genetics of the evolutionary process* New York: Columbia University Press; 1970.
27. Kirkpatrick M: **How and why chromosome inversions evolve.** *PLoS Biol* 2010, **8**:e1000501.
28. Baker RJ, Chesser RK, Koop BF, Hoyt RA: **Adaptive nature of chromosomal rearrangement differential fitness in pocket gophers *Geomys bursarius*.** *Genetica* 1983, **61**:161-164.
29. Lyapunova EA, Bakloushinskaya IY, Saidov AS, Saidov KK: **Dynamics of chromosome variation in mole voles *Ellobius tancrei* (Mammalia, Rodentia) in Pamiro-Alai in the period from 1982 to 2008.** *Russian J Genet* 2010, **46**:566-571.
30. Lorković Z: **Some peculiarities of spatially and sexually restricted gene exchange in the *Erebia tyndarus* group.** *Cold Spring Harb Symp Quant Biol* 1958, **23**:319-325.
31. Nevo E, Cleve H: **Genetic differentiation during speciation.** *Nature* 1978, **275**:125-126.
32. Albre J, Gers C, Legal L: **Molecular phylogeny of the *Erebia tyndarus* (Lepidoptera, Rhopalocera, Nymphalidae, Satyrinae) species group combining CoxII and ND5 mitochondrial genes: A case study of a recent radiation.** *Mol Phyl Evol* 2008, **47**:196-210.
33. Karanth KP, Avivi A, Beharav A, Nevo E: **Microsatellite diversity in populations of blind subterranean mole rats (*Spalax ehrenbergi* superspecies) in Israel: speciation and adaptation.** *Biol J Linn Soc* 2004, **83**:229-241.
34. Gorbunov PY: *The butterflies of Russia (Lepidoptera: Hesperioidea and Papilionoidea): classification, genitalia, keys for identification* Ekaterinburg: Thesis; 2001.
35. Robinson R: *Lepidoptera Genetics* Pergamon Press; 1971.
36. Lorković Z: ***Leptidea reali* Reissinger, 1989 (= *lorkovicii* Real, 1988), a new European species (Lepid., Pieridae).** *Natura Croatica* 1993, **2**:1-26.
37. Camacho JPM, Sharbel TF, Beukeboom LW: **B-chromosome evolution.** *Phil Trans R Soc Lond B* 2000, **355**:163-178.
38. Jones RN, Gonzalez-Sanchez M, Gonzalez-Garcia M, Vega JM, Puertas MJ: **Chromosomes with a life of their own.** *Cytogenet Genome Res* 2008, **120**:265-280.
39. Lukhtanov VA: **Evolution of the karyotype and system of higher taxa of the Pieridae (Lepidoptera) of the world fauna.** *Entomol Obozr* 1991, **70**:619-641.
40. Fumi M: **Distinguishing between *Leptidea sinapis* and *L. reali* (Lepidoptera: Pieridae) using a morphometric approach: impact of measurement error on the discriminative characters.** *Zootaxa* 2008, **1819**:40-54.
41. Martin J, Gilles A, Descimon H: **Species concepts and sibling species: the case of *Leptidea sinapis* and *Leptidea reali*.** In *Butterflies: Ecology and*

Evolution Taking Flight. Edited by: Boggs CL, Watt WB, Ehrlich PR. Chicago: Chicago University Press; 2003:459-476.

42. Dincă V, Zakharov EV, Hebert PDN, Vila R: **Complete DNA barcode reference library for a country's butterfly fauna reveals high performance for temperate Europe.** *Proc R Soc B* 2011, **278**:347-355.
43. Wahlberg N, Saccheri I: **The effects of Pleistocene glaciations on the phylogeography of *Melitaea cinxia* (Lepidoptera: Nymphalidae).** *Eur J Entomol* 2007, **104**:675-684.
44. Verovnik R, Glogovčan P: **Morphological and molecular evidence of a possible hybrid zone of *Leptidea sinapis* and *L. reali* (Lepidoptera: Pieridae).** *Eur J Entomol* 2007, **104**:667-674.
45. Freese A, Fiedler K: **Experimental evidence for specific distinctness of the two wood white butterfly taxa, *Leptidea sinapis* and *L. reali* (Pieridae).** *Nota lepid* 2002, **25**:39-59.
46. Friberg M, Vongvanich N, Borg-Karlson AK, Kemp DJ, Merilaita S, Wiklund C: **Female mate choice determines reproductive isolation between sympatric butterflies.** *Behav Ecol Sociobiol* 2008, **62**:873-886.
47. Simakova A, Puzachenko A: **The vegetation during the last glacial maximum (LGM) (24.0 - 17.0 kyr BP).** In *Evolution of European ecosystems during Pleistocene - Holocene transition (24.0 - 8.0 kyr BP).* Edited by: Markova AK, van Kolfschoten T. Moscow: KMK Scientific Press; 2008:315-341.
48. Freitas TRO, Mattevi MS, Oliveira LFB, Souza MJ, Yonenagayassuda Y, Salzano FM: **Chromosome relationships in 3 representatives of the genus *Holochilus* (Rodentia, Cricetidae) from Brazil.** *Genetica* 1983, **61**:13-20.
49. Koop BF, Baker RJ, Genoways HH: **Numerous chromosomal polymorphisms in a natural population of rice rats *Oryzomys* (Cricetidae).** *Cytogenet Cell Genet* 1983, **35**:131-135.
50. Yonenaga-Yassuda Y, Doprado RC, Mello DA: **Supernumerary chromosomes in *Holochilus brasiliensis* and comparative cytogenetic analysis with *nectomys -squamipes* (Cricetidae, Rodentia).** *Rev Bras Genet* 1987, **10**:209-220.
51. Angines N, Guilera M: **Chromosome polymorphism in *Holochilus venezuelae* (Rodentia, Cricetidae) - C-bands and G-bands.** *Genome* 1991, **34**:13-18.
52. Nachman MW: **Geographic patterns of chromosomal variation in South American marsh rats, *Holochilus brasiliensis* and *H. vulpinus*.** *Cytogen Cell Genet* 1992, **61**:10-16.
53. Volobuev VT, Aniskin VM: **Comparative chromosome banding analysis of three South American species of rice rat of the genus *Oryzomys* (Rodentia, Sigmodontidae).** *Chrom Res* 2000, **8**:295-304.
54. Andrades-Miranda J, Zanchin NIT, Oliveira LFB, Langguth AR, Mattevi MS: **Cytogenetic studies in nine taxa of the genus *Oryzomys* (Rodentia, Sigmodontinae) from Brazil.** *Mammalia* 2001, **65**:461-472.
55. Brant SV, Ortí G: **Molecular phylogeny of short-tailed shrews, *Blarina* (Insectivora: Soricidae).** *Mol Phyl Evol* 2002, **22**:163-173.
56. Silva MJJ, Yonenaga-Yassuda Y: **B chromosomes in Brazilian rodents.** *Cytogenet Genome Res* 2004, **106**:257-263.
57. Hipp AL, Rothrock PE, Whitkus R, Weber JA: **Chromosomes tell half of the story: the correlation between karyotype rearrangements and genetic diversity in sedges, a group with holocentric chromosomes.** *Mol Ecol* 2010, **19**:3124-3138.
58. Baker RJ, Bickham : **Speciation by monobrachial centric fusion.** *Proc Natl Acad Sci USA* 1986, **83**:8245-8248.
59. Basset P, Yannic G, Brunner H, Hausser J: **Restricted gene flow at specific parts of the shrew genome in chromosomal hybrid zones.** *Evolution* 2006, **60**:1718-1730.
60. Wolf KW: **The structure of condensed chromosomes in mitosis and meiosis of insects.** *Int J Insect Morphol Embryol* 1996, **25**:37-62.
61. Lukhtanov VA, Kuznetsova VG: **Molecular and cytogenetic approaches to species diagnostics, systematics, and phylogenetics.** *Zh Obshch Biol* 2009, **70**:415-437.
62. Lukhtanov VA, Kuznetsova VG: **What genes and chromosomes say about the origin and evolution of insects and other arthropods.** *Russian J Genet* 2010, **46**:1115-1121.
63. Coyne JA, Orr AH: *Speciation* Sunderland, MA: Sinauer; 2004.
64. Noor M, Grams KL, Bertucci LA, Reiland J: **Chromosomal inversions and the reproductive isolation of species.** *Proc Natl Acad Sci USA* 2001, **98**:12084-12088.
65. Rieseberg LH: **Chromosomal rearrangements and speciation.** *Trends Ecol Evol* 2001, **16**:351-358.
66. Faria R, Navarro A: **Chromosomal speciation revisited: rearranging theory with pieces of evidence.** *Trends Ecol Evol* 2010, **25**:660-669.
67. Lowry DB, Willis JH: **A widespread chromosomal inversion polymorphism contributes to a najor life-history transition, local adaptation, and reproductive isolation.** *PLoS Biol* 2010, **8(9)**:e1000500.
68. Ayala FJ, Coluzzi M: **Chromosome speciation: humans, Drosophila, and mosquitoes.** *Proc Natl Acad Sci USA* 2005, **102(suppl 1)**:6535-6542.
69. Dincă V, Lukhtanov VA, Talavera G, Vila R: **Unexpected layers of cryptic diversity inwood white *Leptidea* butterflies.** *Nature Comm* 2011, **2**:324.
70. Lukhtanov VA, Vila R, Kandul NP: **Rearrangement of the *Agrodiaetus dolus* species group (Lepidoptera, Lycaenidae) using a new cytological approach and molecular data.** *Insect Syst Evol* 2006, **37**:325-334.
71. Vershinina AO, Lukhtanov VA: **Geographical distribution of the cryptic species *Agrodiaetus alcestis alcestis*, *A. alcestis karacetinae* and *A. demavendi* (Lepidoptera, Lycaenidae) revealed by cytogenetic analysis.** *Comparative Cytogenetics* 2010, **4**:1-11.
72. Marec F, Sahara K, Traut W: **Rise and fall of the W chromosome in Lepidoptera.** In *Molecular Biology and Genetics of the Lepidoptera.* Edited by: Marec F, Goldsmith MR. London-New York: CRC Press; 2010:49-63.
73. Traut W, Sahara K, Otto TD, Marec F: **Molecular differentiation of sex chromosomes probed by comparative genomic hybridization.** *Chromosoma* 1999, **108**:173-180.
74. Yoshido A, Bando H, Yasukochi Y, Sahara K: **The *Bombyx mori* karyotype and the assignment of linkage groups.** *Genetics* 2005, **170**:675-685.
75. Lukhtanov VA, Dantchenko AV: **Principles of highly ordered metaphase I bivalent arrangement in spermatocytes of *Agrodiaetus* (Lepidoptera).** *Chrom Res* 2002, **10**:5-20.
76. Lorković Z: **The butterfly chromosomes and their application in systematics and phylogeny.** In *Butterflies of Europe. Volume 2.* Edited by: Kudrna O. Wiesbaden: Aula-Verlag; 1990:332-396.
77. Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R: **DNA primers for amplification of mitochondrial Cytochrome C oxidase subunit I from diverse metazoan invertebrates.** *Mol Mar Biol Biotech* 1994, **3**:294-299.
78. Simons C, Frati R, Beckenbach A, Crespit B, Liu H, Floors P: **Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers.** *Ann Ent Soc Am* 1994, **87**:651-701.
79. Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W: **Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*.** *Proc Natl Acad Sci USA* 2004, **101**:14812-14817.
80. White TJ, Bruns T, Lee S, Taylor J: **Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics.** In *PCR Protocols: a guide to methods and applications.* Edited by: Innis MA, Gelfand DH, Sninsky JJ, White TJ. San Diego: Academic Press; 1990:315-322.
81. **Biomatters Ltd. 2009 Geneious v.4.8.3.** [http://www.geneious.com/].
82. Posada D: *Collapse: Describing haplotypes from sequence alignments* Vigo (Spain): University of Vigo; 2004 [http://darwin.uvigo.es/software/collapse.html].
83. Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Syst Biol* 2003, **52**:696-704.
84. Hasegawa M, Kishino H, Yano TA: **Dating of the human ape splitting by a molecular clock of mitochondrial DNA.** *J Mol Evol* 1985, **22**:160-174.
85. Posada D: **jModelTest: phylogenetic model averaging.** *Mol Biol Evol* 2008, **25**:1253-1256.
86. Clement M, Posada D, Crandall K: **TCS: a computer program to estimate gene genealogies.** *Mol Ecol* 2000, **9**:1657-1660.
87. Excoffier L, Langaney A: **Origin and differentiation of human mitochondrial DNA.** *Am J Human Gen* 1989, **44**:73-85.
88. Drummond AJ, Rambaut A: **BEAST: Bayesian evolutionary analysis by sampling trees.** *BMC Evol Biol* 2007, **7**:214.
89. Quek SP, Davies SJ, Itino T, Pierce NE: **Codiversification in an ant-plant mutualism: Stem texture and the evolution of host use in Crematogaster (Formicidae: Myrmicinae) inhabitants of Macaranga (Euphorbiaceae).** *Evolution* 2004, **58**:554-570.
90. Heled J, Drummond AJ: **Bayesian inference of species trees from multilocus data.** *Mol Biol Evol* 2010, **27**:570-580.

doi:10.1186/1471-2148-11-109
Cite this article as: Lukhtanov *et al.*: Unprecedented within-species chromosome number cline in the Wood White butterfly *Leptidea sinapis* and its significance for karyotype evolution and speciation. *BMC Evolutionary Biology* 2011 **11**:109.

nature COMMUNICATIONS

ARTICLE

Received 12 Jan 2011 | Accepted 28 Apr 2011 | Published 24 May 2011

DOI: 10.1038/ncomms1329

Unexpected layers of cryptic diversity in wood white *Leptidea* butterflies

Vlad Dincă[1,2], Vladimir A. Lukhtanov[3,4], Gerard Talavera[1,2] & Roger Vila[1,5]

Uncovering cryptic biodiversity is essential for understanding evolutionary processes and patterns of ecosystem functioning, as well as for nature conservation. As European butterflies are arguably the best-studied group of invertebrates in the world, the discovery of a cryptic species, twenty years ago, within the common wood white *Leptidea sinapis* was a significant event, and these butterflies have become a model to study speciation. Here we show that the so-called 'sibling' *Leptidea* actually consist of three species. The new species can be discriminated on the basis of either DNA or karyological data. Such an unexpected discovery challenges our current knowledge on biodiversity, exemplifying how a widespread species can remain unnoticed even within an intensely studied natural model system for speciation.

[1] Institut de Biologia Evolutiva (CSIC-UPF), Passeig Marítim de la Barceloneta 37–49, 08003 Barcelona, Spain. [2] Departament de Genètica i Microbiologia, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain. [3] Department of Karyosystematics, Zoological Institute of Russian Academy of Science, Universitetskaya nab. 1, 199034 St Petersburg, Russia. [4] Department of Entomology, St Petersburg State University, Universitetskaya nab. 7/9, 199034 St Petersburg, Russia. [5] Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Spain. Correspondence and requests for materials should be addressed to R.V. (email: roger.vila@ibe.upf-csic.es).

Given the global biodiversity crisis[1–3], cataloguing the earth's species has become a race against time. Several studies have highlighted the presence and importance of cryptic biodiversity, which might represent a substantial proportion of Earth's natural capital. An estimate of cryptic species diversity is important to better understand evolutionary processes and patterns of ecosystem functioning, while also having deep implications for nature conservation[4,5]. The recent increase in the number of reported cryptic species is, in large part, owing to an increasing amount of studies incorporating DNA-based techniques, including large-scale approaches such as DNA barcoding[6], which often provide resolution beyond the boundaries of morphological information. However, documenting cryptic diversity based on DNA data alone is generally not sufficient, prompting calls for integrative morphological, ecological and molecular approaches[7,8]. Recent estimates on the distribution of cryptic diversity are contradictory, and are based on a thin empirical foundation[4,9]. In any case, it is to be expected that the highest number of yet-to-be-discovered cryptic species is concentrated in the most biodiverse and least explored regions of our planet (that is, tropical areas). In temperate regions such as Europe, it is assumed that the level of unrecognized diversity is low, not only because of lower species richness, but also because taxonomic research has been intense for many groups of organisms. Such a case is represented by butterflies, probably the best-studied group of invertebrates, which have become a flagship for insect conservation efforts in Europe[10,11].

The discovery of a new European species of wood white (*Leptidea* sp.) at the end of the twentieth century was an important event in butterfly systematics. *Leptidea sinapis* (Linnaeus, 1758), a common butterfly with Palaearctic distribution was found to 'hide' a cryptic species, *Leptidea reali* (Reissinger, 1989)[12,13]. After the two species were shown to be separable based on their genitalia—but not on their wing morphology[13]—several studies revealed that *L. reali* is often sympatric with *L. sinapis* and that its distribution is almost as wide as that of *L. sinapis*[14,15]. Molecular data (allozyme markers and mitochondrial DNA) also supports the specific distinctness of *L. reali*[16]. Moreover, much attention has been paid to behavioural and ecological aspects of the species pair *L. sinapis*–*L. reali*, to the point that they have become a model for studying speciation in cryptic species. Such studies revealed that: a premating reproductive barrier exists (females only accept conspecific males)[17,18]; the two species display only limited ecological differentiation (larval food plant preference and performance)[17,19]; niche separation between the two species (forests or meadows) is not caused by fixed between-species differences[20]; differences in phenology and voltinism are mostly the result of environmentally induced pleiotropic effects[21]; larval diapause is determined by information from the host plant[22]; and behavioural polyphenism has been documented in female propensity to mate[23].

In this paper, we integrate molecular (mitochondrial and nuclear DNA markers), cytological (chromosome number) and morphological data (male genitalia morphometry) to study the species pair *L. sinapis*–*L. reali*. We found an unexpected pattern showing that *L. reali* actually comprises two synmorphic, yet genetically and karyotypically distinct, groups, with the new cryptic entity being sister to *L. sinapis*+*L. reali*, producing what may be called nested cryptic species. Therefore, the so-called 'twin species' *L. sinapis*–*L. reali* are actually a triplet of cryptic species, a result that asks for a reconsideration of previous knowledge and exemplifies the advantages of an integrative approach when studying closely related taxa.

Results

Analysis of mitochondrial and nuclear DNA markers. The Bayesian and maximum likelihood gene genealogies estimated for each of the mitochondrial (cytochrome oxidase subunit I (*COI*), NADH dehydrogenase subunit 1 (*ND1*)) and nuclear loci (internal transcribed spacer 2 (*ITS2*), wingless (*Wg*), carbamoyl-phosphate synthetase 2/aspartate transcarbamylase/dihydroorotase (*CAD*)) gave largely congruent results for the species pair *L. sinapis* and *L. reali*. Depending on their degree of variability, the markers had different resolving power, but all suggested that specimens that are morphologically attributable to *L. reali* (based on their genitalia) are not monophyletic. Moreover, the more variable genes *COI*, *ND1* and *ITS2* showed that *L. reali* formed two clades and was paraphyletic with respect to *L. sinapis* (Supplementary Figs S1–S5).

The topology of the partitioned Bayesian, maximum likelihood and maximum parsimony multi-gene trees revealed three major well-resolved clades within the *L. sinapis*–*L. reali* group (Fig. 1). Whereas *L. sinapis* was recovered as monophyletic, specimens morphologically attributable to *L. reali* (based on their genitalia) formed two strongly supported clades and were paraphyletic with respect to *L. sinapis*. One of these clades was sister to *L. sinapis* and included all specimens from Spain and Italy, as well as several from southern France (Fig. 1). This clade is certainly attributable to genuine *L. reali*, as the type locality of this species lies in the French Pyrenees[12]. The other clade of specimens with *reali*-like morphology consisted of samples from several countries ranging from Ireland and France in the west, to eastern Kazakhstan in the east, and was recovered as sister to *L. sinapis* plus genuine *L. reali* with good support (Fig. 1). This pattern was recovered by the Bayesian coalescent-based species tree estimation as well, confirming the topological relationships of the three lineages (Fig. 2). The species-tree approach is less prone to misleading results than combining data by partitions, because it incorporates uncertainty associated with gene trees (probability of unsorted ancestral polymorphism), nucleotide substitution model parameters, and the coalescent process. These results, together with the karyotypic data, strongly suggest that the non-Mediterranean clade of *L. reali* represents a different species. The oldest available name that we could assign to the new species is *juvernica* (Williams, 1946), described as a subspecies for Irish populations with *reali*-like morphology[14]. Therefore, in accordance with the International Code of Zoological Nomenclature, we hereafter refer to the new species as *Leptidea juvernica* stat. nov.

Our sampling revealed that *L. reali* and *L. juvernica* stat. nov. display non-overlapping geographical distributions, but some populations are parapatric —at least in southeastern France, where they are separated by only 87 kilometres (Fig. 3, Supplementary Table S1). It is worth noting that we did not find any case of introgression between these two species in the parapatry zone or elsewhere.

Karyotype analysis. Diploid chromosome numbers $2n=52$, $2n=53$ and $2n=54$ were found in *L. reali*. The individuals with $2n=52$ and $2n=54$ presented 26 and 27 bivalents during first meiotic division (MI), and 26 and 27 chromosomes during second meiotic division (MII), respectively. Individuals with $2n=53$ were heterozygous for one chromosomal fusion/fission and demonstrated 25 bivalents and one trivalent in MI (Supplementary Note 1). Thus, we established the chromosome number of *L. reali* is not fixed and ranges between $2n=52–54$.

Leptidea juvernica stat. nov. displayed clearly higher chromosome numbers and, at the same time, a higher level of chromosome number variation than *L. reali*. We have found in mitotic cells, or have reconstructed based on meiotic cells, the following numbers: $2n=80$, $2n=82$ and $2n=84$, $2n=$ ca. 81–84, $2n=$ ca. 83–85. Some of the individuals studied were chromosomal heterozygotes displaying up to six multivalents in metaphase I of meiosis (Supplementary Note 1). Given the karyotypes observed in MI and MII cells and taking into account all possible combinations of gametes, we concluded that chromosome numbers ranging from $2n=76$ to $2n=88$ are expected to be found in *L. juvernica* stat. nov. Our results show that *L. reali* and *L. juvernica* stat. nov. are differentiated by at least 11 chromosomal fusions/fissions.

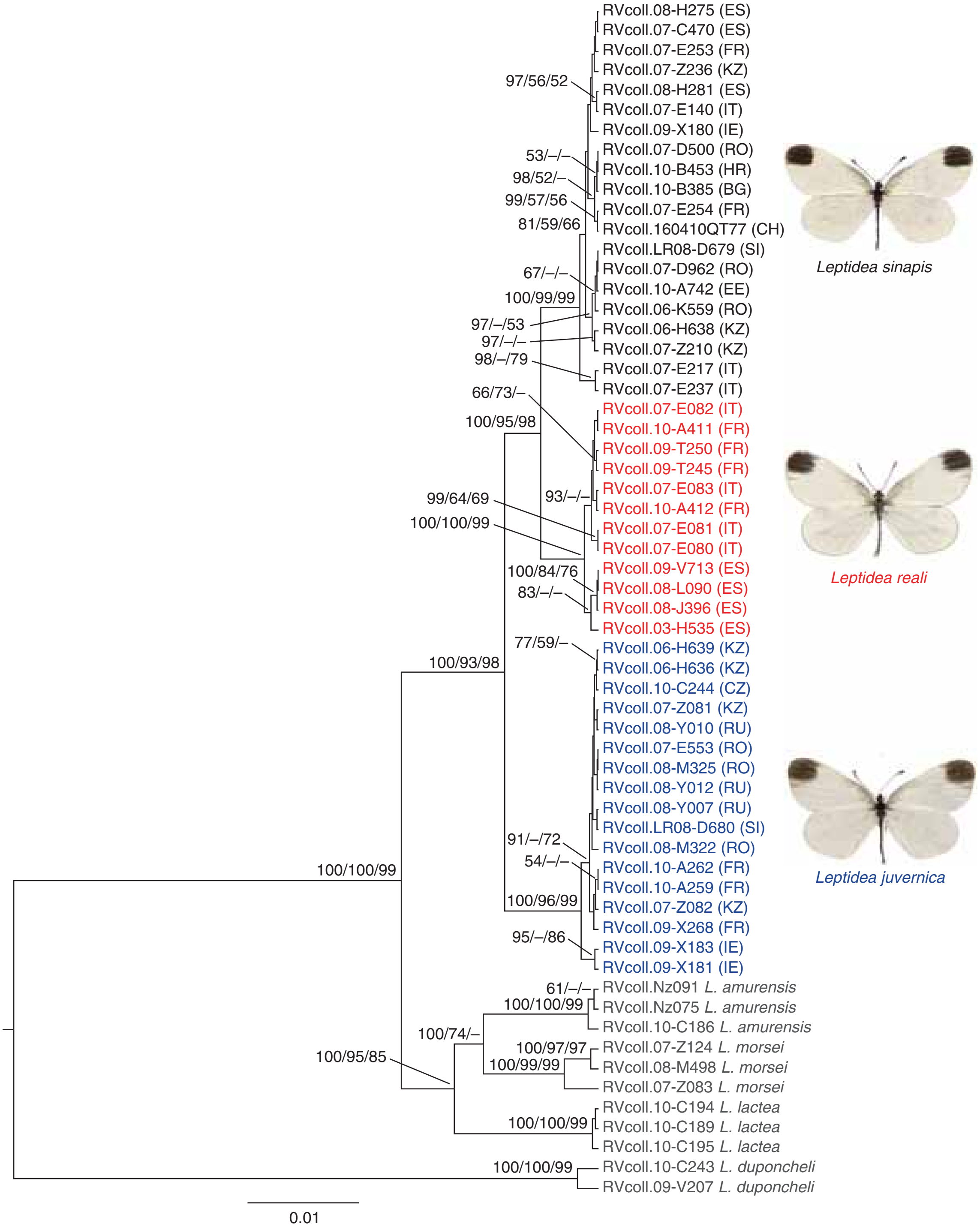


Figure 1 | Leptidea molecular phylogeny. Bayesian ultrametric tree based on the combined analysis of *COI*, *ND1*, *ITS2*, *Wg* and *CAD*. *Leptidea juvernica* stat. nov. is monophyletic and sister to *L. sinapis* + *L. reali*. Bayesian posterior probabilities, maximum likelihood and maximum parsimony bootstrap values (>50%) are shown above recovered branches. IE, Ireland; ES, Spain; FR, France; IT, Italy; CH, Switzerland; SI, Slovenia; HR, Croatia; RO, Romania; BG, Bulgaria; EE, Estonia; RU, Russia; KZ, Kazakhstan.

Morphological analysis. The Shapiro-Wilk test supported normal distributions for the five measured variables (phallus length (PL), saccus length (SL), vinculum width (VW), genital capsule length and uncus length (UL)) (P>0.05). For the discriminant analysis, the variables included in the prediction equation with the stepwise method and using Wilks' Lambda were PL, VW and SL. The first two canonical discriminant functions explained 100% of the variance and were used in the analysis. The first function alone accounted for 99.4% of the variance displaying a strong canonical correlation of 0.951 and a highly significant Wilks' Lambda (0.091, P<0.001). The second function explained 0.6% of the variance, displayed a canonical correlation of 0.227 and had a significant Wilks' Lambda (0.949, P=0.032). The structure matrix that was obtained (Supplementary Table S2) showed the canonical weight of each variable which is an indicator of its discriminatory power.

All specimens attributed to *L. sinapis* in the molecular analysis were correctly classified by the discriminant analysis, supporting previous results indicating that male genitalia allow for the separation of *L. sinapis* and *L. reali sensu lato*. On the other hand, classification was much less accurate for *L. reali* and *L. juvernica* stat. nov. (61.5% for *L. reali* and 62.5% for *L. juvernica* with 53.8% and 58.3%, respectively, after cross validation) (Fig. 4, Supplementary Table S3), indicating that they cannot be reliably identified based on the parameters involved. To further test these results, another discriminant analysis was run including only *L. reali* and *L. juvernica* stat. nov. Two variables were introduced in the prediction equation: SL and genital capsule length (GL). The first function explained 100% of the variance and was used in the analysis. This function displayed a moderate canonical correlation of 0.357 and a significant Wilks' Lambda of 0.873 ($P=0.003$). Classification results were similar to the previous analysis, with 56.4% of *L. reali* and 64.6% of *L. juvernica* stat. nov. correctly identified (53.8% and 62.5%, respectively, after cross-validation). This confirmed that, although there seemed to be a slight tendency of larger genitalia for *L. reali* specimens (Fig. 4), identification was unreliable based on male genitalia.

Female genitalia of a few specimens corresponding to *L. reali* and *L. juvernica* stat. nov. were also examined. Although our sample was too small to permit statistical analyses, we did not notice any apparent difference in the length of the ductus bursae, the most useful character to discriminate females of *L. sinapis* from *L. reali*[13,24].

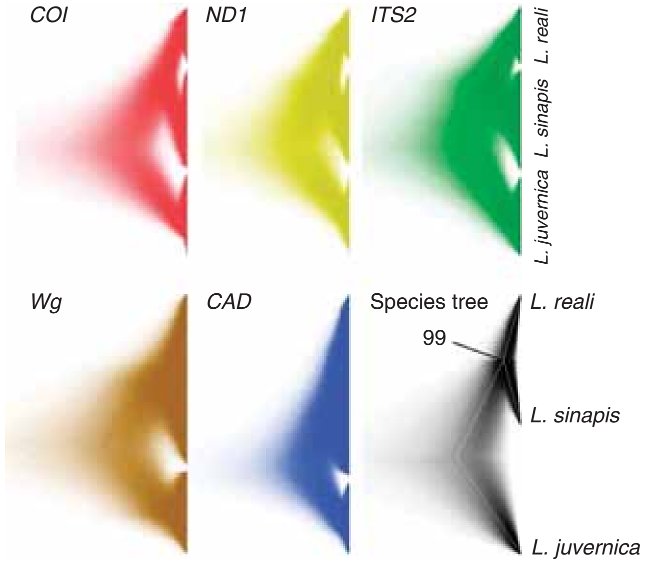


Figure 2 | Co-estimation of five gene trees embedded in a shared species tree. Two mitochondrial (*COI* and *ND1*) and three nuclear loci (*ITS2*, *Wg* and *CAD*) were used to estimate a species tree (in black) for *L. sinapis*, *L. reali* and *L. juvernica* stat. nov. using *BEAST. The consensus is represented in white within the species tree and the posterior probability for the *L. reali*—*L. sinapis* clade is shown. Trees are figured with DensiTree[55] displaying all trees of the Markov chain Monte Carlo chain with a burn-in of 5,000 trees. Higher levels of uncertainty are represented by lower densities.

Discussion

This study presents strong evidence for the existence of a previously unnoticed, widespread species of *Leptidea*. This is clearly supported by our combined molecular phylogeny based on two mitochondrial and three nuclear markers, as well as by the coalescent-based species tree reconstruction, which showed that the new species *L. juvernica* stat. nov. is sister to the species pair *L. sinapis* and *L. reali*. No topological discordance in the relationships among the three species was detected in the single-gene trees (Fig. 2), except for *CAD* and *Wg*, which mixed specimens of *L. sinapis* and *L. reali*. The slower mutation rate and/or coalescent process of these two nuclear markers is most probably the cause, but it is worth noting that they recovered the new species as a dif-

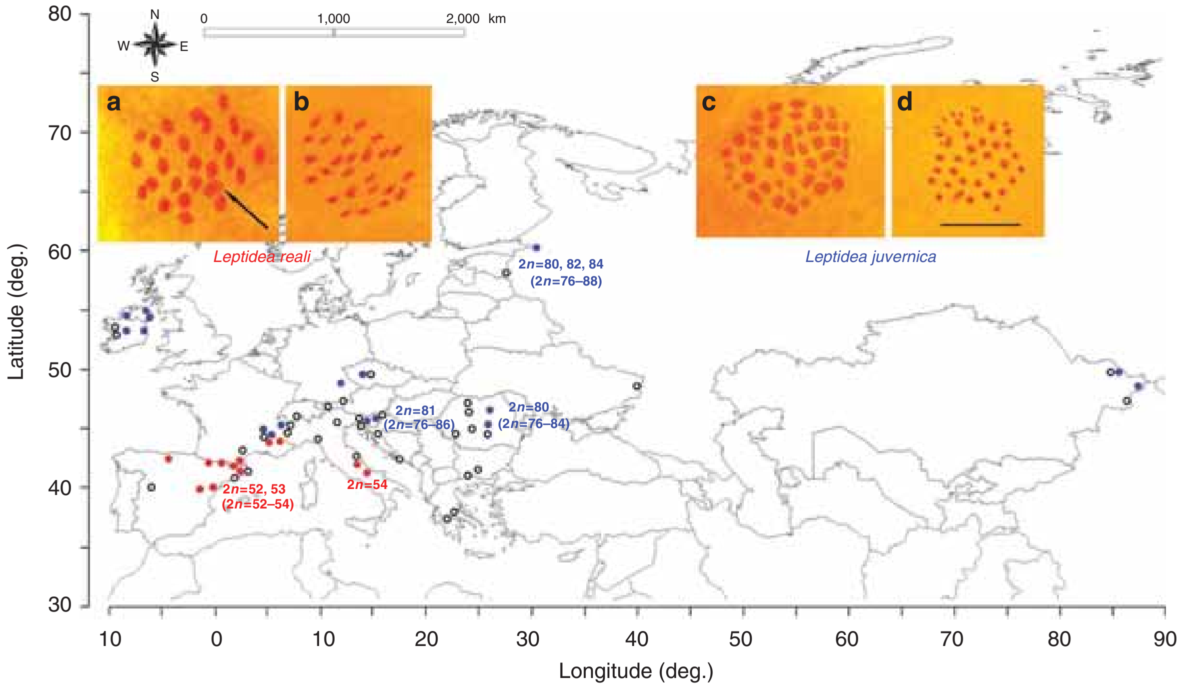


Figure 3 | Chromosome number results and sampling localities. *Leptidea reali* (red dots), *L. juvernica* stat. nov. (blue dots) and *L. sinapis* (empty circles). Although *L. reali* and *L. juvernica* display non-overlapping distributions, they come into close contact in southeastern France and are differentiated by at least 11 fixed chromosome fusions/fissions. (**a**) Karyotype of *L. reali*, Spain, RVcoll.03-H535: MI cell demonstrating 25 bivalents and one multivalent (most likely a trivalent, indicated by an arrow). (**b**) *L. reali*, Spain, RVcoll.07-F514: MII cell demonstrating 27 chromosomes. (**c**) *L. juvernica*, Russia, RVcoll.08-Y012: MI cell demonstrating 42 bivalents. (**d**) *L. juvernica*, Russia, RVcoll.08-Y010: MII cell demonstrating 41 chromosomes. Scale bar corresponds to 10 μm in all figures.

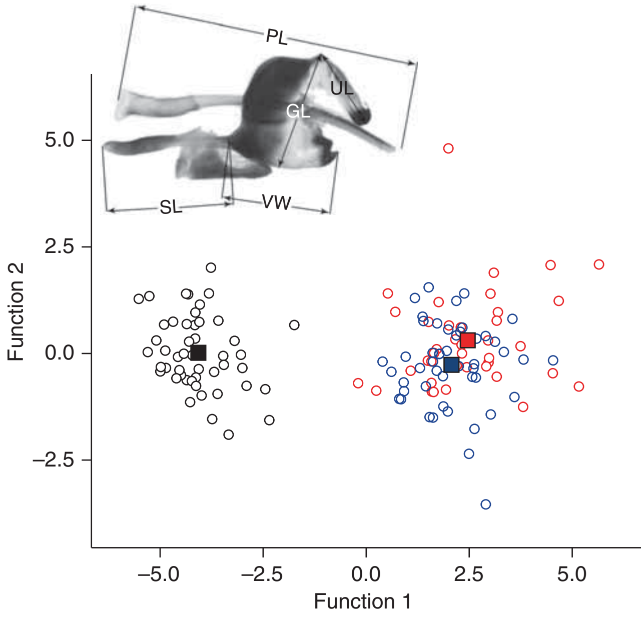


Figure 4 | Discriminant analysis based on male genitalia morphometry. *L. sinapis* (black) is identifiable based on male genitalia, but there is broad overlap between *L. reali* (red) and *L. juvernica* stat. nov. (blue). Circles represent individuals and squares represent centroids for each species. Elements of the male genitalia measured were PL, SL, VW, GL and UL. The discriminant variables were PL and SL for function 1 and SL and VW for function 2. The upper left corner image indicates the variables measured for *Leptidea* male genitalia.

ferent entity. When combining the three nuclear loci, the monophyly of all species was strongly supported (Supplementary Fig. S6).

Our conclusions based on molecular data were also supported by karyotype analyses, which revealed different chromosome numbers between *L. reali* ($2n=52$–54) and *L. juvernica* stat. nov. ($2n=80$–84). Although the chromosome number is not fixed in these species, intraspecific variation is limited and the interspecific gap is pronounced (at least 11 chromosomal rearrangements). It is relevant that karyotype characteristics (chromosome numbers and level of chromosome number variability) are nearly identical in geographically distant populations within each species, whereas *L. reali* from Italy and *L. juvernica* stat. nov. from Slovenia are drastically different, despite geographical proximity.

Mating between species with different karyotypes is known to produce hybrids that are heterozygous for chromosomal rearrangements fixed between parental species. These hybrids typically have reduced fertility due to partial missegregation of homologous chromosomes during the MI[25]. Although different kinds of chromosomal rearrangements have various effects on the fertility of heterozygous hybrids[26], hybrid fertility is generally negatively correlated with the extent of karyotypic divergence between parental taxa[27,28], and multiple chromosome fusion/fissions, such as those we detected in *L. reali* and *L. juvernica* stat. nov., can strongly contribute to postzygotic reproductive isolation. Although we have no direct data on the degree of postzygotic isolation, the chromosomal differentiation between *L. reali* and *L. juvernica* stat. nov. is high and can be considered as additional independent evidence that there are two distinct species.

Morphometry results showed that specimens of *L. reali* and *L. juvernica* cannot be reliably distinguished, whereas *L. sinapis* was clearly differentiated based on genitalic measurements. Wing and preimaginal stage morphology did not appear to be useful for identification either, as already shown by several studies comparing *L. sinapis* and *L. reali sensu lato*[13]. Therefore, *L. reali* and *L. juvernica* stat. nov. seem to represent the plesiomorphic state and to have remained in morphological stasis, wheras *L. sinapis* evolved genitalic differences. The fact that the new cryptic species reported here is apparently fully synmorphic to *L. reali* explains why it remained unnoticed for such a long time despite intensive research. We propose the common name 'cryptic wood white' for *L. juvernica* stat. nov.

The relationships between the three studied species suggest that the common ancestor of the triplet of species (ancestor A) (Fig. 5a,b) probably originated in central or western Asia and subsequently spread over western Europe. The hypothesis of an eastern origin is also supported by the exclusively eastern distribution of the closest relatives (*L. amurensis*, *L. morsei* and *L. lactea*) to the triplet of cryptic species (Fig. 1). About 270,000 years ago (Supplementary Table S4), probably in southwestern Europe, ancestor A speciated producing the common ancestor of *L. sinapis* and *L. reali* (ancestor B), and the *L. juvernica* stat. nov. lineage that established across temperate Europe and Asia (Fig. 5c). About 120,000 years ago, ancestor B diverged into *L. sinapis* and *L. reali* (Fig. 5d). Later on (ca. 27,000 years ago), *L. sinapis* expanded north and east into the territory of *L. juvernica* stat. nov. (Fig. 5e). On the basis of our sampling, *L. reali* and *L. juvernica* stat. nov. are most likely parapatric, with *L. reali* confined to southwestern Europe and *L. juvernica* stat. nov. spread across temperate Europe and Asia. This provides a totally new view on *L. reali* which is actually a west Mediterranean species and not a widely distributed taxon as concluded before. Our sampling suggests a potential contact zone in southeastern France, where populations of the two species are separated by less than 90 kilometres (Fig. 3).

To know the causes behind the apparent inability of *L. reali* and *L. juvernica* stat. nov. to coexist will require further studies. It has been shown that, besides differences in the genitalia, behavioural aspects related to mate choice maintain reproductive isolation between *L. sinapis* and *L. reali sensu lato*[17,18]. Previous data on the biology and ecology of *L. reali sensu lato*, as well as our view of the speciation processes undergone by *Leptidea*, need to be extensively revised in light of these results[17–23,29]. Our observations revealed that both *L. reali* and *L. juvernica* stat. nov. can use *Lathyrus pratensis* as a larval food plant (oviposition observed in Spain for *L. reali* and in Romania for *L. juvernica* stat. nov. and adults obtained from these eggs by rearing in the laboratory). It is thus to be expected that ecological differentiation between the two species is minimal.

In this study, we show that assessing cryptic diversity is a challenging task even in well-studied groups of organisms. What has been formerly called the cryptic species pair, *L. sinapis*–*L. reali* comprises a triplet of species, and new research is needed to clarify their distribution, ecology and conservation status. Our findings exemplify that cryptic biodiversity may consist of finely nested layers and highlight the importance of using an array of techniques when dealing with closely related species.

Methods

Specimen sequencing. The mitochondrial marker *COI* was sequenced in 166 specimens, the mitochondrial *ND1* in 85 specimens, the nuclear *ITS2* in 91 specimens, the nuclear wingless (*Wg*) in 67 specimens and the nuclear *CAD* in 43 specimens.

Thirteen GenBank *COI* sequences of *L. sinapis* from Spain[30], France[31], Slovenia[32], Greece[21] and Kazakhstan[33], seven sequences of *L. juvernica* stat. nov. from Slovenia[32], three sequences of *L. amurensis*[33] from Russia and two sequences of *L. morsei*[33] from Kazakhstan were also added to the dataset. Additionally, one sequence of *L. sinapis* from Austria and three sequences of *L. juvernica* stat. nov. from Germany were included from the publicly available project 'Fauna Bavarica—Lepidopera Rhopalocera' included in the Barcode of Life Data System at http:\\www.barcodinglife.org. Four *ND1 Leptidea* GenBank sequences (two *L. reali* and two *L. sinapis*)[16] were also added to the dataset. All novel sequences obtained in this study have been deposited to GenBank under accession codes JF512569 to JF513007 (for details see Supplementary Table S1).

Total genomic DNA was extracted using Chelex 100 resin, 100–200 mesh, sodium form (Bio-rad), under the following protocol: one leg was removed and introduced into 100 μl of Chelex 10% and 5 μl of Proteinase K (20 mg ml^{-1}) were added. The samples were incubated overnight at 55 °C and were subsequently incubated at 100 °C for 15 min. Afterwards they were centrifuged for 10 s at 3,000 r.p.m.

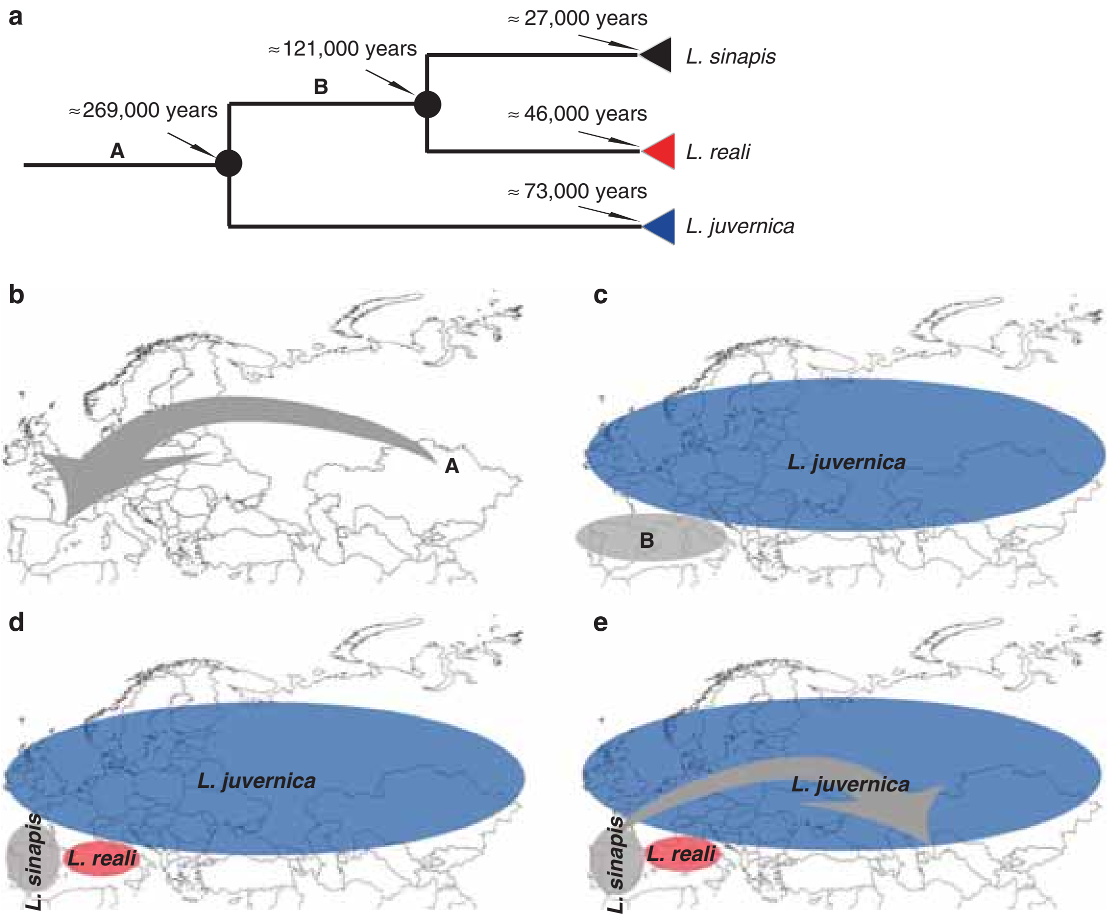


Figure 5 | Phylogenetic relationships and proposed speciation scenario. (**a**) *L. juvernica* stat. nov. is sister to *L. sinapis* + *L. reali*. Age estimations are indicated for each node. (**b**) The common ancestor of the whole group (ancestor A) probably originated in central or western Asia and subsequently colonized western Europe. (**c**) Ancestor A split into *L. juvernica* in temperate Europe and Asia and the common ancestor of *L. sinapis* and *L. reali* (ancestor B) in southwestern Europe (**d**) Ancestor B speciated into *L. sinapis* and *L. reali*. (**e**) Subsequently, *L. sinapis* rapidly spread north and east into the territory of *L. juvernica*.

The primers used were: for *COI* (676 bp) LCO 1490 (5′-GGTCAACAAATCATAAAGATATTGG-3′)[34] and Nancy (5′-CCCGGTAAAATTAAAATATAAACTTC-3′)[35], or (658 bp) LepF1 (5′-ATTCAACCAATCATAAAGATATTGG-3) and LepR1 (5′-TAAACTTCTGGATGTCCAAAAAATCA-3′)[36]; for *ND1* (790–794 bp) 5′- CTGTTCGATCATTAAAATCTTAC-3′ (forward)[37] and 5′-ATCAAAAGGAGCTCGATTAGTTTC-3′ (reverse)[38]; for *ITS2* (684 bp) ITS3 (5′-GCATCGATGAAGAACGCAGC-3′) and ITS4 (5′-TCCTCCGCTTATTGATATGC-3′)[39]; for *Wg* (403 bp) Wg1 (5′-GARTGYAARTGYCAYGGYATGTCTGG-3′) and Wg2 (5′-ACTICGCRCACCARTGGAATGTRCA-3′)[40]; for *CAD* (571 bp) CADFa (5′-GDATGGTYGATGAAAATGTTAA-3′) and CADRa (5′-CTCATRTCGTAATCYGTRCT-3′).

Double-stranded DNA was amplified in 25 µl volume reactions: 13.22 µl ultra pure (HPLC quality) water, 2.5 µl 10× buffer, 4.5 µl 25 mM $MgCl_2$, 0.25 µl 100 mM dNTP, 1.2 µl of each primer (10 mM), 0.13 µl Taq DNA Gold Polymerase (Qiagen) and 2 µl of extracted DNA. The typical thermal cycling profile for *COI* was 95 °C for 60 s, 44 °C for 60 s and 72 °C for 90 s, for 40 cycles. The annealing temperature varied according to marker: 48 for *ND1*, 47 for *ITS2*, 51 for *Wg*, and 48 for *CAD*.

PCR products were purified and sequenced by Macrogen Inc. All the samples are stored in the Institut de Biologia Evolutiva collection in Barcelona, Spain, and are available upon request.

Phylogenetic analyses and species tree estimation. *COI*, *ND1*, *ITS2*, *Wg* and *CAD* sequences were edited and aligned using Geneious Pro 4.7.5[41]. These resulted in five alignments of 676 bp and 195 specimens for *COI*, 794 bp and 89 specimens for *ND1*, 715 bp and 91 specimens for *ITS2*, 403 bp and 67 specimens for *Wg*, and 571 bp and 43 specimens for *CAD*. For *COI*, duplicate haplotypes (excluding outgroups) were removed using TCS 1.21[42].

Individual Bayesian and ML phylogenetic trees were inferred using *COI*, *ND1*, *ITS2*, *Wg*, and *CAD* with BEAST 1.6.0[43] and GARLI 1.0[44]. Relationships based on the combined dataset were estimated using partitioned Bayesian and ML analyses using BEAST 1.6.0 and GARLI-PART v. 0.97[44] with substitution models by markers according to the suggestions of jModeltest 0.1[45]. The models employed for the partitioned ML analysis were TPMuf + I + G for *COI*, HKY + I for *CAD* and *ND1*, TVM for *ITS2* and TPM2 for *Wg*. For the partitioned BI, GTR + I + G was used for *COI*, HKY + I for *CAD* and *ND1*, GTR for *ITS2* and HKY for *Wg*.

Branch support was assessed by 100 bootstrap replicates for maximum likelihood, and Markov chain Monte Carlo convergence was checked after two independent runs of 10 million generations each (with a pre-run burn in of 100,000 generations) for Bayesian inference. A multilocus coalescent-based Bayesian species tree was estimated with *BEAST[46]. *L. sinapis*, *L. reali* and *L. juvernica* stat. nov. specimens were defined as three taxonomic units in accordance with clades previously inferred by single-gene and five-loci combined trees. A relaxed clock with uncorrelated lognormal distribution[47] and a Yule speciation process as tree prior were used. The length of the Markov chain Monte Carlo chain was set at 50 million generations sampling every 1,000 runs with a burn-in set to the first 500,000 generations. A maximum parsimony tree based on the five markers combined was inferred with MEGA4[48] and branch supports were assessed by 100 bootstrap replicates.

Dating divergence events. Node ages were inferred with BEAST 1.6.0[43] using the *COI* haplotype dataset under a coalescent model with constant population size. We calibrated the phylogeny at two selected nodes: the *L. sinapis* common ancestor node as an example of a very recent clade supposedly under a coalescent process, and the root of the tree as a clearly coalesced node. For the age of the root node, we used a normally distributed prior ranging between 2.2 and 4 MYA based on slow and fast published invertebrate mitochondrial rates of 1.3 and 2.3% uncorrected pairwise distance per million years[49,50]. The prior range assumed for the common ancestor of *L. sinapis* was a normal distribution between 8,500–31,000 years, as previously inferred[51]. The dataset was analysed using the GTR + I + G model and applying an uncorrelated lognormal relaxed molecular clock[47] along the branches. Base frequencies were estimated, six gamma rate categories were selected and a randomly generated initial tree was used. Parameters were estimated using two independent runs of 10 million generations each (with a pre-run burn in of 100,000 generations) to ensure convergence, and were checked with the program Tracer v1.5.

Karyotype analyses. Gonads were stored in Carnoy fixative (ethanol and glacial acetic acid, 3:1) for 2–6 months at 4 °C and then stained with 2% acetic orcein for 30 days at 20 °C. Cytogenetic analysis was conducted using a two-phase method of chromosome analysis[52].

In our study, we have counted the diploid chromosome numbers (2n) in mitotic spermatogonial cells and the haploid chromosome numbers (n) in metaphase II of male meiosis. We also counted the number of chromosomal elements (n) (bivalents + multivalents) in metaphase I of male meiosis. In the last case, the number of chromosomal elements was equal to the haploid number (n), if all the elements were represented by bivalents, or less if some elements were represented by multivalents. To distinguish between bivalents and multivalents, we used a special method[53]. Briefly, by varying the pressure on the coverslip, we were able to manipulate chromosomes, for example, change their position and orientation in intact (not squashed) spermatocyte cells, and consequently to analyse the structure of the bivalents and multivalents.

In total, preparations from 68 males were analysed. As cell divisions are extremely rare in *Leptidea* during imago stage[54], metaphase plates were observed in only 14 individuals (Supplementary Table S1). These individuals have also been used for morphological and molecular analysis.

Genitalia preparation and morphometrics. Male genitalia were prepared according to the following protocol: maceration for 15 min at 95 °C in 10% potassium hydroxide, dissection and cleaning under a stereomicroscope and storage in tubes with glycerine.

Genitalia were photographed in a thin layer of distilled water (without being pressed under a cover slip) under a Carl Zeiss Stemi 2000-C stereomicroscope equipped with a DeltaPix Invenio 3S digital camera. Measurements were performed based on the digital photographs by using AxioVision software. A total of 39 specimens of *L. reali*, 48 of *L. juvernica* stat. nov. and 48 of *L. sinapis* were included in the morphometrical analyses (Supplementary Table S5). Five elements of the male genitalia were measured: PL, SL, VW, GL (measured from the ventral edge of the vinculum to the uncus apex) and UL. The first three elements combined were reported to be the best to discriminate between *L. sinapis* and *L. reali*[13,16,24].

Statistical analyses were carried out using the software SPSS 14.0 for Windows. The first batch of analyses was run by including three groups: *Leptidea reali*, *L. juvernica* stat. nov. and *L. sinapis*. Subsequently the analyses were repeated including only *L. reali* and *L. juvernica* stat. nov. A Shapiro-Wilk normality test was employed. Subsequently, a discriminant analysis was performed by employing the stepwise method. In order to test the obtained classification, a cross validation was carried out ('leave-one-out' method).

References

1. Gaston, K. J. & Fuller, R. A. Biodiversity and extinction: losing the common and the widespread. *Prog, Phys, Geogr.* **31,** 213–225 (2007).
2. Thomas, J. A. *et al.* Comparative losses of British butterflies, birds, and plants and the global extinction crisis. *Science* **303,** 1879–1881 (2004).
3. Brooks, T. M. *et al.* Global biodiversity conservation priorities. *Science* **313,** 58–61 (2006).
4. Bickford, D. *et al.* Cryptic species as a window on diversity and conservation. *Trends Ecol. Evol.* **22,** 148–155 (2006).
5. Esteban, G. F. & Finlay, B. J. Conservation work is incomplete without cryptic biodiversity. *Nature* **463,** 293 (2010).
6. Hebert, P. D. N., Cywinska, A., Ball, S. L. & deWaard, J. R. Biological identifications through DNA barcodes. *Proc. R. Soc. B* **270,** 313–321 (2003).
7. Schlick-Steiner, B. C., Seifert, B., Stauffer, C., Christian, E., Crozier, R. H. & Steiner, F. M. Without morphology, cryptic species stay in taxonomic crypsis following discovery. *Trends Ecol. Evol.* **22,** 391–392 (2007).
8. Beheregaray, L. B. & Caccone, A. Cryptic biodiversity in a changing world. *J. Biol.* **6,** 9 (2007).
9. Pfenninger, M. & Schwenk, K. Cryptic animal species are homogeneously distributed among taxa and biogeographical regions. *BMC Evol. Biol.* **7,** 121 (2007).
10. New, T. R., Pyle, R. M., Thomas, J. A., Thomas, C. D. & Hammond, P. C. Butterfly conservation management. *Annu. Rev. Entomol.* **40,** 57–83 (1995).
11. Thomas, J. A. in *Ecology and Conservation of Butterflies* (ed. Pullin, A. S.) 180–197 (Chapman & Hall, 1995).
12. Réal, P. Lépidoptères nouveaux principalement jurassiens. *Mém. Comité de Liaison Rech. Ecofaunist. Jura* **4,** 1–28 (1988).
13. Lorković, Z. *Leptidea reali* Reissinger, 1989 (=*lorkovicii* Real 1988), a new European species (Lepid., Pieridae). *Nat. Croatica* **2,** 1–26 (1993).
14. Mazel, R. *Leptidea sinapis* L., 1758—*L. reali* Reissinger, 1989, le point de la situation (Lepidoptea: Pieridae, Dismorphiinae). *Linneana Belgica* **18,** 199–202 (2001).
15. Mazel, R. & Eitschberger, U. Répartition géographique de *Leptidea sinapis* (L., 1758) et *L. reali* Reissinger, 1989 au nord de l'Europe, en Russie et dans quelques pays d'Asie (Lepidoptera: Pieridae, Dismorphiinae). *Linneana Belgica* **18,** 373–376 (2002).
16. Martin, J., Gilles, A. & Descimon, H. in *Butterflies: Ecology and Evolution Taking Flight* (eds Boggs, C. L., Watt, W. B. & Ehrlich, P. R.) 459–476 (Chicago University Press, 2003).
17. Freese, A. & Fiedler, K. Experimental evidence for specific distinctness of the two wood white butterfly taxa, *Leptidea sinapis* and *L. reali* (Pieridae). *Nota Lepid.* **25,** 39–59 (2002).
18. Friberg, M., Vongvanich, N., Borg-Karlson, A.- K., Kemp, D. J., Merilaita, S. & Wiklund, C. Female mate choice determines reproductive isolation between sympatric butterflies. *Behav. Ecol. Sociobiol.* **62,** 873–886 (2008).
19. Friberg, M. & Wiklund, C. Host plant preference and performance of the sibling species of butterflies *Leptidea sinapis* and *Leptidea reali*: a test of the trade-off hypothesis for food specialisation. *Oecologia* **159,** 127–137 (2009).
20. Friberg, M., Olofsson, M., Berger, D., Karlsson, B. & Wiklund, C. Habitat choice precedes host plant choice—niche separation in a species pair of a generalist and a specialist butterfly. *Oikos* **117,** 1337–1344 (2008).
21. Friberg, M., Bergman, M., Kullberg, J., Wahlberg, N. & Wiklund, C. Niche separation in space and time between two sympatric sister species—a case of ecological pleiotropy. *Evol. Ecol.* **22,** 1–18 (2008).
22. Friberg, M. & Wiklund, C. Host-plant-induced larval decision-making in a habitat/host-plant generalist butterfly. *Ecology* **91,** 15–21 (2010).
23. Friberg, M. & Wiklund, C. Generation dependent female choice: behavioral polyphenism in a bivoltine butterfly. *Behav. Ecol.* **18,** 758–763 (2007).
24. Fumi, M. Distinguishing between *Leptidea sinapis* and *L. reali* (Lepidoptera: Pieridae) using a morphometric approach: impact of measurement error on the discriminative characters. *Zootaxa* **1819,** 40–54 (2008).
25. Kandul, N. P., Lukhtanov, V. A. & Pierce, N. E. Karyotypic diversity and speciation in *Agrodiaetus* butterflies. *Evolution* **61,** 546–559 (2007).
26. King, M. *Species Evolution* (Cambridge University Press, 1993).
27. White, M. J. D. *Animal Cytology and Evolution* (Cambridge University Press, 1973).
28. Gropp, A. H., Winking, H. & Redi, C. in *Genetic Control of Gamete Production and Function* (eds Crosignani, P. G., Rubin, B. L. & Franccaro, M.) 115–134 (Academic Press, 1982).
29. Beneš, J., Konvička, M., Vrabec, V. & Zámečník, J. Do the sibling species of small whites, *Leptidea sinapis* and *L. reali* (Lepidoptera, Pieridae) differ in habitat preferences? *Biologia* **58,** 943–951 (2003).
30. Braby, M. F., Vila, R. & Pierce, N. E. Molecular phylogeny and systematics of the Pieridae (Lepidoptera: Papilionoidea): higher classification and biogeography. *Zool J. Linn. Soc.* **147,** 239–275 (2006).
31. Mutanen, M., Wahlberg, N. & Kaila, L. Comprehensive gene and taxon coverage elucidates radiation patterns in moths and butterflies. *Proc. R. Soc. B* **277,** 2839–2848 (2010).
32. Verovnik, R. & Glogovčan, P. Morphological and molecular evidence of a possible hybrid zone of *Leptidea sinapis* and *L. reali* (Lepidoptera: Pieridae). *Eur. J. Entomol.* **104,** 667–674 (2007).
33. Lukhtanov, V. A., Sourakov, A., Zakharov, E. V. & Hebert, P. D. N. DNA barcoding Central Asian butterflies: increasing geographical dimension does not significantly reduce the success of species identification. *Mol. Ecol. Resour.* **9,** 1302–1310 (2009).
34. Folmer, O. *et al.* DNA primers for amplification of mitochondrial Cytochrome C oxidase subunit I from diverse metazoan invertebrates. *Mol. Mar. Biol. Biotech.* **3,** 294–299 (1994).
35. Simon, C. *et al.* Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. *Ann. Entomol. Soc. Am.* **87,** 651–701 (1994).
36. Hebert, P. D. N. *et al.* Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proc. Natl Acad. Sci. USA* **101,** 14812–14817 (2004).
37. Aubert, J., Barascud, B., Descimon, H. & Michel, F. Ecology and genetics of interspecific hybridization in the swallowtails, *Papilio hospiton* Gene and *P. machaon* L., in Corsica (Lepidoptera: Papilionidae). *Biol. J. Linn. Soc.* **60,** 467–492 (1997).
38. Aubert, J., Barascud, B., Descimon, H. & Michel, F. Systematique moleculaire des Argynnes (Lepidoptera: Nymphalidae). *Comptes rendus de l'Academie des Sciences. Serie 3. Sciences de la Vie* **319,** 647–651 (1996).
39. White, T. J. *et al.* in *PCR Protocols: A Guide to Methods and Applications* (eds Innis, M. A. *et al.*) 315–322 (Academic Press, 1990).
40. Brower, A. V. Z. & DeSalle, R. Mitochondrial vs. nuclear DNA sequence evolution among nymphalid butterflies: the utility of Wingless as a source of characters for phylogenetic inference. *Insect Mol. Biol.* **7,** 1–10 (1998).
41. Drummond, A. J. *et al.* Geneious v4.7. Available from http://www.geneious.com/ (2009).
42. Clement, M., Posada, D. & Crandall, K. Tcs: a computer program to estimate gene genealogies. *Mol. Ecol.* **9,** 1657–1660 (2000).
43. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7,** 214 (2007).
44. Zwickl, D. J. *Genetic Algorithm Approaches for the Phylogenetic Analysis of Large Biological Sequence Datasets Under the Maximum Likelihood Criterion*. PhD dissertation, The University of Texas at Austin (2006). Available from: https://www.nescent.org/wg_garli/Main_Page.
45. Posada, D. jModelTest: phylogenetic model averaging. *Mol. Biol. Evol.* **25,** 1253–1256 (2008).

46. Heled, J. & Drummond, A. J. Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.* **27,** 570–580 (2010).
47. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4,** e88 (2006).
48. Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24,** 1596–1599 (2007).
49. Quek, S. P. *et al.* Codiversification in an ant-plant mutualism: stem texture and the evolution of host use in *Crematogaster* (Formicidae: Myrmicinae) inhabitants of *Macaranga* (Euphorbiaceae). *Evolution* **58,** 554–570 (2004).
50. Brower, A. V. Z. Rapid morphological radiation and convergence among races of the butterfly *Heliconius erato* inferred from patterns of mitochondrial-DNA evolution. *Proc. Natl Acad. Sci. USA* **91,** 6491–6495 (1994).
51. Lukhtanov, V. A., Dincă, V., Talavera, G. & Vila, R. Unprecedented within-species chromosome number cline in the Wood White butterfly and its significance for karyotype evolution and speciation. *BMC Evol. Biol.* **11,** 109 (2011).
52. Lukhtanov, V. A., Vila, R. & Kandul, N. P. Rearrangement of the *Agrodiaetus dolus* species group (Lepidoptera, Lycaenidae) using a new cytological approach and molecular data. *Insect Syst. Evol.* **37,** 325–334 (2006).
53. Lukhtanov, V. A. & Dantchenko, A. V. Principles of highly ordered metaphase I bivalent arrangement in spermatocytes of *Agrodiaetus* (Lepidoptera). *Chromosome Res.* **10,** 5–20 (2002).
54. Lorković, Z. in *Butterflies of Europe*, Vol 2 (ed. Kudrna, O.) 332–396 (Aula, 1990).
55. Bouckaert, R. R. DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics* **26,** 1372–1373 (2010).

Acknowledgments

We are grateful to B. Aldwell, K. Bond, S. Cuvelier, M. Fumi, J. Harding, J. Haslett, J. Hernández-Roldán, M. Hughes, G. Jeffcoate, M. Marín, X. Merit, S. Montagud, E. Regan, I. Rippey, V. Tshikolovets, N. Shapoval, S. Viader and M. Warren for providing samples. Special thanks to B. Nelson for organizing a specimen loan from National Museums Northern Ireland, and to G. Lamas for comments on the manuscript. Support for this research was provided by the Russian Foundation for Basic Research (grants RFFI 09-04-01234, 11-04-00734 and 11-04-00076), by grant NSH-3332.2010.4 (Leading Scientific Schools) and by the programs of the Presidium of Russian Academy of Science 'Gene Pools and Genetic diversity' and 'Origin of biosphere and evolution of geo-biological systems' to V.A.L.; by the Spanish MICINN (projects CGL2007-60516/BOS and CGL2010-21226/BOS to V.D., G.T. and R.V., and predoctoral fellowship BES-2008-002054 to G.T.); and by a predoctoral fellowship from UAB to V.D.

Author contributions

All authors conceived and designed the research. G.T. and V.D. performed phylogenetic analyses. V.D. performed morphometrical analyses. V.A.L. performed karyological analyses. V.D. and R.V. wrote the manuscript, and all authors discussed the results and commented on the manuscript.

Additional information

Accession Codes: All novel sequences obtained in this study have been deposited to GenBank under accession codes JF512569 to JF513007.

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Dincă, V. *et al.* Unexpected layers of cryptic diversity in wood white *Leptidea* butterflies. *Nat. Commun.* 2:324 doi: 10.1038/ncomms1329 (2011).

References

References

Acinas, S.G., Klepac-Ceraj, V., Hunt, D.E., Pharino, C., Ceraj, I., Distel, D.L., Polz, M.F. **2004**. Fine-scale phylogenetic architecture of a complex bacterial community. *Nature*. 430(6999):551-4.

Ackermann, M., Achatz, M., Weigend, M. **2008**. Hybridization and crossability in Caiophora (Loasaceae subfam. Loasoideae): are interfertile species and inbred populations results of a recent radiation? *American Journal of Botany* 95:1109–1121.

Aguileta, G., Bielawski, J.P., Yang, Z. **2006**. Evolutionary rate variation among vertebrate [beta] globin genes: Implications for dating gene family duplication events. *Gene* 380(1):21-29.

Andersson, L. **1990**. The driving force: Species concepts and ecology. *Taxon* 39:375–382.

Ané, C., Larget, B., Baum, D.A., Smith, S.D, Rokas, A. **2007**. Bayesian estimation of concordance among gene trees. *Molecular Biology and Evolution* 24(2):412-26.

Anisimova, M., Gascuel, O. **2006**. Approximate likelihood ratio test for branchs: A fast, accurate and powerful alternative. *Systematic Biology*, 55(4), 539-552, 2006.

Atz, J.W. **1970**. The application of the idea of homology to behavior. In *Development and Evolution of Behavior.* Aronson, L.R., Tolbach, E., Lehrman, D.S., Rosenblatt, J.S., [Eds]. Freeman, San Francisco, CA.

Avise, J.C., Arnold, J., Ball Jr, R.M., Bermingham, E., Lamb, T., Neigel, J.E., Reeb, C.A., Saunders, N.C. **1987**. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics* 18:489-522.

Avise, J.C. **1992**. Molecular population structure and the biogeographic history of a regional fauna: a case history with lessons for conservation biology. *Oikos* 63:62-76.

Avise, J.C. **2000**. Phylogeography: The History and Formation of Species. Harvard University Press, Cambridge, MA.

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A., Johnson, E.A. **2008**. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* 3(10):e3376.

Barraclough, T.G., Barclay, M.V., Vogler, A.P. **1998**. Species richness: does flower power explain beetle-mania? *Current Biology* 8:843-845.

Barrett, M., Donoghue, M.J., Sober, E. **1991**. Against consensus. *Systematic Zoology* 40:486–493.

Baum, D. A., Shaw, K.L. **1995**. Genealogical perspectives on the species problem. In *Experimental and molecular approaches to plant biosystematics*. Hoch, P.C., Stephenson, A.G., [Eds]. Missouri Botanical Garden, St. Louis.

Baum, D.A., Smith, S.D., Donovan, S.S. **2005**. The tree-thinking challenge. *Science* 310 (5750):979-989.

Baum, D.A., Smith, S.D. **2012**. Tree thinking: an introduction to phylogenetic biology. Roberts and Company, CO, USA.

Baurain, D., Brinkmann, H., and Philippe, H. **2007**. Lack of resolution in the animal phylogeny: closely spaced cladogeneses or undetected systematic errors? *Molecular Biology and Evolution* 24, 6–9.

Beaumont, M. A., Nielsen, R., Robert, C., Hey, J., Gaggiotti, J., Knowles, L., Estoup, A., Panchal, M., Corander, J., Hickerson, M., Sisson, S.A., Fagundes, N., Chikhi, L., Beerli, P., Vitalis, R., Cornuet, J.M., Huelsenbeck, J., Foll, M., Yang, Z.H., Rousset, F., Bal, D. **2010**. In defence of model-based inference in phylogeography. *Molecular Ecology* 19:436-446.

Beaumont, M.A. **2010**. Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics* 41: 379-406.

Beaumont, M.A., Panchal. **2008**. On the validity of nested clade phylogeographical analysis. *Molecular Ecology* 17(11):2563-2565.

Beaumont, M.A., Zhang, W., Balding, D.J. 2002. Approximate Bayesian computation in population genetics. *Genetics* 162(4):2025-35.

Beerli, P., Felsenstein, J. **2001**. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences USA* 98(8):4563-8.

Benton, M.J., Donoghue P.C.J. **2007**. Paleontological evidence to date the tree of life. *Molecular Biology and Evolution* 24(1):26-53.

Bergsten, J. **2005**. A review of long-branch attraction. *Cladistics* 21(2):163-193.

Bessey, C.E. **1915**. The phylogenetic taxonomy of flowering plants. *Annals of the Missouri Botanical Garden* 2(1/2):109-164.

Beutel, R.G., Pohl, H. **2006**. Endopterygote systematics - where do we stand and what is the goal (Hexapoda, Arthropoda)? *Systematic Entomology* 31(2):202-219.

Bininda-Emonds, O.R.P., Gittleman, J.L., Steel, M.A. **2002**. The (super)tree of life: procedures, problems, and prospects. *Annual Reviews of Ecology and Systematics* 33: 265–289

Bininda-Emonds, O.R.P. **2003**. Novel versus unsupported clades: assessing the qualitative support for clades in MRP supertrees. *Systematic Biology* 52: 839–848.

Bininda-Emonds, O.R.P. **2004**. The evolution of supertrees. *Trends in Ecology and Evolution* 19: 315–322.

Bininda-Emonds, O.R.P. **2007**. Fast genes and slow clades: comparative rates of molecular evolution in mammals. Evolutionary Bioinformatics Online, 3, 59-85.

Bininda-Emonds, O.R.P. **2011**. Inferring the Tree of Life: chopping a phylogenomic problem down to size? *BMC Evolutionary Biology* 9:59.

Bond, J.E., Stockman, A.K. **2008**. An integrative method for delimiting cohesion species: finding the population-species interface in a group of Californian trapdoor spiders with extreme genetic divergence and geographic structuring. *Systematic Biology* 57:628–646.

Bos, D., Posada, D. **2005**. Using models of nucleotide evolution to build phylogenetic trees. *Developmental and Comparative Immunology* 29: 211-227.

Bousquet, J., Strauss, S.H., Doerksen, A.H., Price, R.A. **1992**. Extensive variation in evolutionary rate of rbcL gene sequences among seed plants. *Proceedings of the National Academy of Sciences USA* 89(16):7844 -7848.

Braby, M.F., Eastwood, R., Murray, N. **2012**. The subspecies concept in butterflies: has its application in taxonomy and conservation biology outlived its usefulness? *Biological Journal of the Linnean Society* 106(4):699-716.

Bradley, T.J., Briscoe, A.D., Brady, S.G., Contreras, H.L., Danforth, B.N., Dudley, R., Grimaldi, R., Harrison, J.F., Kaiser, A., Merlin, C., Reppert, S.M., VandenBrooks, J.M., Yanoviak, S.P. **2009**. Episodes in insect evolution. *Integrative and Comparative Biology* 49(5):590-606.

Bremer, K., **1988**. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. *Evolution* 42, 795–803.

Bremer, K. **1992**. Ancestral areas: A cladistic reinterpretation of the center of origin concept. *Systematic Biology* 41:436–445.

Bremer, K. **1995**. Ancestral areas: Optimization and probability. *Systematic Biology* 44:255–259.

Brinkmann, H., Van der Giezen, M., Zhou, Y., Poncelin de Raucourt, G., Philippe, H. **2005**. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Systematic Biology* 54(5):743-757.

Brooks, D.R., McLennan, D.A. **1991**. Phylogeny, Ecology, and Behavior: a Research Program in Comparative Biology. Chicago: University of Chicago Press.

Brooks, T.M., Mittermeier, R.A., da Fonseca, G.A., Gerlach, J., Hoffmann, M., Lamoreux, J.F., Mittermeier, C.G., Pilgrim, J.D., Rodrigues, A.S. **2006**. Global biodiversity conservation priorities. *Science* 313(5783):58-61.

Brower, A.V.Z. **1999**. Delimitation of phylogenetic species with DNA sequences: A critique of Davis and Nixon's population aggregation analysis. *Systematic Biology* 48: 199-213.

Brower, A.V.Z. **2006**. Problems with DNA barcodes for species delimitation:'ten species' of Astraptes fulgerator reassessed (Lepidoptera: Hesperiidae). *Systematics and Biodiversity* 4(2):127-132.

Buckley, T. R., Cunningham, C.W. **2002**. The effects of nucleotide substitution model assumptions on estimates of nonparametric bootstrap support. *Molecular Biology and Evolution*, 19: 394-402.

Buckley, T.R., Cordeiro, M., Marshall, D.C., Simon, C. **2006**. Differentiating between hypotheses of lineage sorting and introgression in New Zealand alpine cicadas (Maoricicada Dugdale). *Systematic Biology* 55(3):411-25.

Burger, G., Gray, M.W., Lang, B.F. **2003**. Mitochondrial genomes: anything goes. *Trends in Genetics* 19(12):709-716.

Butler, M.A., King, A. **2004**. Phylogenetic Comparative Analysis: A modeling approach for adaptive evolution. *American Naturalist* 164:683-695.

Cadena, C.D., Cuervo, A.M. **2010**. Molecules, ecology, morphology and songs in concert: Can we know how many species is *Arremon torquatus* (Aves, Emberizidae)? *Biological Journal of the Linnean Society* 99:152-176.

Cap, H., Deleporte, P., Joachim, J., Reby, D. **2008**. Male vocal behavior and phylogeny in deer. *Cladistics* 24:1-15.

Carstens BC, Knowles LL. **2007**. Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from Melanoplus grasshoppers. *Systematic Biology* 56(3):400-11.

Carstens, B.C., Dewey, T.A. **2010**. Species delimitation using a combined coalescent and information theoretic approach: An example from North American Myotis bats. *Systematic Biology* 59, 400-414.

Castresana, J. **2000**. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution* 17:540–552.

Castro, L.R., Austin, A.D., Dowton, M. **2002**. Contrasting rates of mitochondrial molecular evolution in parasitic diptera and hymenoptera. *Molecular Biology and Evolution* 19(7):1100-1113.

Cavalli-Sforza, L.L., Edwards, A.W.F. **1967**. Phylogenetic analysis: Models and estimation procedures. *Evolution* 21(3): 550-570.

Chapman, A.D. **2009**. Numbers of Living Species in Australia and the World. 2nd edition. Australian Biodiversity Information Services, Toowoomba, Australia.

Charleston, M.A. **2002**. Principles of cophylogeny maps. In *Biological Evolution and Statistical Physics.* Lässig, M., Valleriani, A., [Eds]. Springer-Verlag.

Chen, F.-C., Li, W.-H. **2001**. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *American Journal of Human Genetics* 68, 444–456.

Chen, F., Mackey, A.J., Vermunt, J.K., Roos, D.S. **2007**. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE* 2: e383.

Chenuil, A., McKey, D.B. **1996**. Molecular phylogenetic study of a myrmecophyte symbiosis: Did Leonardoxa ant associations diversify via cospeciation? *Molecular Phylogenetics and Evolution* 6: 270–286.

Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B., Bork, P. **2006**. Toward automatic reconstruction of a highly resolved tree of life. *Science*. 311(5765):1283-7.

Clark, J.R., Ree, R.H., Alfaro, M.E., King, M.G., Wagner, W.L., Roalson, E.H. **2008**. A Comparative Study in

Ancestral Range Reconstruction Methods: Retracing the Uncertain Histories of Insular Lineages. *Systematic Biology* 57:693–707.

Cooper, A., Lalueza-Fox, C., Anderson, S., Rambaut, A., Austin, J., Ward, R. **2001**. Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution. *Nature* 409(6821):704-7.

Coyne, J.A., Orr, A.H. **2004**. Speciation. Sunderland, MA: Sinauer.

Cracraft, J. **1983**. Species concepts and speciation analysis. *Current Ornithology* 1:159–187.

Creevey, C.J., Muller, J., Doerks, T., Thompson, J.D., Arendt, D., Bork, P. **2011**. Identifying Single Copy Orthologs in Metazoa. *Plos Computational Biology* 7(12):e1002269.

Cummings, M.P., Handley, S.A., Myers, D.S., Reed, D.L., Rokas, A., Winka, K. **2003**. Comparing bootstrap and posterior probability values in the four-taxon case. *Systematic Biology*. 52:477-487.

Cummings, M.P., Neel, M.C., Shaw, K.L. **2008**. A genealogical approach to quantifying lineage divergence. *Evolution* 62(9):2411-2422.

Darwin, C. **1859**. On the origin of species by means of natural selection. London: John Murray.

Darwin, E. **1794**. Zoonomia. London: J. Johnson.

Davis, J.I., Nixon, K.C. **1992**. Populations, genetic variation, and the delimitation of phylogenetic species. *Systematic Biology* 41:421–435.

Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C. **1978**. A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, vol 5, suppl. 3, ed. Dayhoff, M.O., pp. 345-352. Washington DC, USA: National Biomedical Research Foundation.

de Queiroz, A., Gatesy J. **2007**. The supermatrix approach to systematics. *Trends in Ecology and Evolution* 22:34–41.

de Queiroz, K. **1998**. The general lineage concept of species, species criteria, and the process of speciation: A conceptual unification and terminological recommendations. In *Endless forms: Species and speciation* (D. J. Howard, D.J., Berlocher, S.H., [Eds]. Oxford University Press, New York.

de Queiroz, K. **2005**. Different species problems and their resolution. *BioEssays* 27:1263–1269.

de Queiroz, K. **2007**. Species concepts and species delimitation. *Systematic Biology* 56:879–886.

Deffontaine, V., Libois, R., Kotlík, P., Sommer, R., Nieberding, C., Paradis, E., Searle, J.B., Michaux, J.R. **2005**. Beyond the Mediterranean peninsulas: evidence of central European glacial refugia for a temperate forest mammal species, the bank vole (Clethrionomys glareolus). *Molecular Ecology* 14(6):1727-39.

Delsuc, F., Brinkmann, H., Philippe, H. **2005**. Phylogenomics and the reconstruction of the tree of life. *Nature Review Genetics* 6(5):361-75.

Desalle, R. **2006**. Species discovery versus species identification in DNA barcoding efforts: response to Rubinoff. *Conservation Biology* 20(5):1545–7.

Dincă, V., Dapporto, L., Vila, R. **2011a**. A combined genetic-morphometric analysis unravels the complex biogeographical history of Polyommatus icarus and Polyommatus celina common blue butterflies. *Molecular Ecology* 20(18):3921-35.

Dincă, V., Lukhtanov, V.A., Talavera, G., Vila, R. **2011b**. Unexpected layers of cryptic diversity in wood white Leptidea butterflies. *Nature Communications* 2:234.

Dobzhansky, T. **1950**. Mendelian populations and their evolution. *American Naturalist* 84:401–418.

Donoghue, M.J. **1985**. A critique of the biological species concept and recommendations for a phylogenetic alternative. *Bryologist* 88:172– 181.

Drew, L.W. **2011**. Are we losing the science of taxonomy? *BioScience* 61:942–946

Drummond, A.J., Ho S.Y.W., Phillips, M.J., Rambaut, A. **2006**. Relaxed phylogenetics and dating with confidence. *PLoS Biology*, 4(5):e88.

Dunn, C.W., Hejnol, A., Matus, D.Q., Pang, K., Browne, W.E., Smith SA, Seaver, E., Rouse, G.W., Obst, M., Edgecombe, G.D., Sørensen, M.V., Haddock, S.H., Schmidt-Rhaesa, A., Okusu, A., Kristensen, R.M., Wheeler, W.C., Martindale, M.Q., Giribet, G. **2008**. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* 452: 745–749.

Ebach, M.C., Holdrege. C. **2005**. DNA barcoding is no substitute for taxonomy. *Nature* 434:697.

Efron, B., Halloran, E., Holmes, S. **1996**. Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences USA* 93:7085-7090

Eisen, J.A., Fraser, C.M. **2003**. Phylogenomics: intersection of evolution and genomics. *Science* 300(5626):1706-7

Elango, N., Lee, J., Peng, Z., Loh, Y.-H.E., Yi, S.V. **2009**. Evolutionary rate variation in Old World monkeys. *Biology Letters* 5(3):405-408.

Erwin, T.L. **1993**. Biodiversity at its utmost: tropical forest beetles. In *Biodiversity II. Understanding and Protecting our Biological Resources.* Reaka-Kudla, M.L., Wilson, D.E., Wilson, E.O., [Eds]. Joseph Henry Press, Washington, DC.

Erwin, T.L. **2004**. The biodiversity question: how many species of terrestrial arthropods are there?. In *Forest Canopies*. Lowman, M., Brinker, B., [Eds]. Academic Press London, UK.

Faircloth, B.C., McCormack, J.E., Crawford, N.G., Brumfield, R.T., Glenn, T.C. **2012**. Ultraconserved elements anchor thousands of genetic markers for target enrichment spanning multiple evolutionary timescales. *Systematic Biology* 61:717–726.

Farrell, B.D., Mitter, C. **1990**. Phylogenesis of insect/plant interactions: have Phyllobrotica leaf beetles (Chrysomelidae) and the Lamiales diversified in parallel? *Evolution* 44:1389–1403.

Farrell, B. D. **1998**. "Inordinate fondness" explained: Why are there so many beetles? *Science* 281:553-557.

Farris, J.S. **1970**. Estimating phylogenetic trees from distance matrixes. *American Nature* 106:645-668.

Felsenstein, J. **1978**. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* 27(4):401-410.

Felsenstein, J. **1985**a. Confidence limits on phylogenies: An approach using the bootstrap.

Evolution 39:783-791.

Felsenstein, J. **1985**b. Phylogenies and the comparative method. *American Naturalist* 125: 115.

Felsenstein, J. **2004**. Inferring phylogenies. Sinauer Associates, Sunderland, MA, USA.

Ferrier, S., Guisan, A. **2006**. Spatial modelling of biodiversity at the community level. *Journal of Applied Ecology* 43(3):393-404.

Fitch, W.M. **1970**. Distinguishing homologous from analogous proteins. *Systematic Zoology,* 19: 99–113.

Fitch, W.M. **1971**. Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology* 20:406-416.

Fitch, W.M., Margoliash, E. **1967**. Construction of phylogenetic trees. *Science* 155(760): 279-284.

Foottit, R.G., Adler, P.H. **2009**. Insect Biodiversity: Science and Society*.* Canada: Wiley - Blackwell.

Gadagkar, S.R., Rosenberg, M.S., Kumar, S. **2005**. Disparity index: a simple statistic to measure and test the homogeneity of substitution patterns between molecular sequences. *J Exp Zool B Mol Dev Evol* 304(1):64.

Garrick, R.C., Dyer, R.J., Beheregaray, L.B., Sunnucks, P. **2008**. Babies and bathwater: a comment on the premature obituary for nested clade phylogeographic analysis. *Molecular Ecology* 17:1401–1403.

Gaston, K.J. **1991**. The magnitude of global insect species richness. *Conservation Biology* 5:283–296.

Gatesy, J., Matthee, C., Desalle, R., Hayashi, C. **2002**. Resolution of a supertree/supermatrix paradox. *Systematic Biology*, 51: 652–664.

Gatesy, J., Baker, R.H., Hayashi, C. **2004**. Inconsistencies in arguments for the supertree approach: supermatrices versus supertrees of Crocodylia. *Systematic Biology*, 53: 342–355.

Gaunt, M.W., Miles, M.A. **2002**. An insect molecular clock dates the origin of the insects and accords with palaeontological and biogeographic landmarks. *Molecular Biology and Evolution* 19(5):748-61.

Giribet, G., Edgecombe, G. D., Wheeler, W. C., Babbitt, C. **2002**. Phylogeny and systematic position of Opiliones: A combined analysis of Chelicerate relationships using morphological and molecular data. *Cladistics 18*:5-70.

Goldman, N. **1993**. Statistical tests of models of DNA substitution. *Journal of Molecular Evolution* 36:182-198

Gompert, Z., Forister, M.L., Fordyce, J.A., Nice, C.C., Williamson, R., Buerkle, C.A. **2010a**. Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of *Lycaeides* butterflies. *Molecular Ecology* 19:2455-2473.

Gompert, Z., Lucas, L.K., Fordyce, J.A., Forister, M.L., Nice, C.C. **2010b**. Secondary contact between *Lycaeides idas* and *L. melissa* in the Rocky Mountains: extensive introgression and a patchy hybrid zone. *Molecular Ecology* 19:3171-3192.

Gonnet, G.H., Cohen, M.A., Benner, S.A. **1992**. Exhaustive matching of the entire protein sequence database. *Science* 256(5062):1443-5.

Gonzalez-Rodriguez, A., Bain, J.F., Golden, J.L., Oyama, K. **2004**. Chloroplast DNA variation in the Quercus affinis-Q. Laurina complex in Mexico: geographical structure and associations with nuclear and morphological variation. *Molecular Ecology* 13:3467–3476.

Goodman, M., Olson, C.B., Beeber, J.E., Czelusniak, J. **1982**. New perspectives in the molecular biological analysis of mammalian phylogeny. *Acta Zoologica Fennica* 169: 19–35.

Górecki P, Tiuryn J. **2007**. Inferring phylogeny from whole genomes. *Bioinformatics* 23(2):e116-22.

Grandcolas, P., Deleporte, P., Desutter-Grandcolas, L., Daugeron, C., **2001**. Phylogenetics and ecology: as many characters as possible should be included in the cladistic analysis. *Cladistics* 17, 104–110.

Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H., Hansen, N.F., Durand, E.Y., Malaspinas, A.S., Jensen, J.D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H.A., Good, J.M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E.S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Z., Gusic, I., Doronichev, V.B., Golovanova, L.V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R.W., Johnson, P.L., Eichler, E.E., Falush, D., Birney, E., Mullikin, J.C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M,, Reich, D., Pääbo, S. **2010**. A draft sequence of the Neandertal genome. *Science* 328(5979):710-22.

Gribaldo, S., Philippe, H. Ancient phylogenetic relationships. **2002**. *Theoretical Population Biology* 61(4):391-408.

Grimaldi, D.A., Engel, M. **2005**. The Evolution of Insects. Cambridge University Press, Cambridge.

Gullan, P.J., Cranston, P.S. **2010**. Insects: An Outline of Entomology, 4th edition. Wiley-Blackwell.

Hajibabaei, M., Singer, G.A., Hebert, P.D., Hickey, D.A. **2007**. DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends in Genetics* 23:167-172.

Hammond, P.M. **1992**. Species inventory. In *Global Biodiversity: status of the Earth's living resources*. Groombridge, B., [Ed]. Chapman and Hall, London, UK.

Harvey, P.H., Pagel, M.D. **1991**. The Comparative Method in Evolutionary Biology. New York: Oxford University Press.

Harvey, P.H., May, R.M., Nee, S. **1994**. Phylogenies without fossils. *Evolution* 48(3):523-529.

Hassanin, A., Leger, N., Deutsch, J. **2005**. Evidence for multiple reversals of asymmetric mutational constraints during the evolution of the mitochondrial genome of Metazoa, and consequences for phylogenetic inferences. *Systematic Biology* 54(2):277-298.

Hausdorf, B. **1998**. Weighted ancestral area analysis and a solution of the redundant distribution problem. *Systematic Biology* 47:445–456.

Hawksworth, D.L., Kalin-Arroyo, M.T. **1995**. Magnitude and distribution of biodiversity. In *Global biodiversity assessment*. Heywood, V.H., [Ed]. Cambridge University Press, Cambridge.

Hebert P.D.N., Cywinska A., Ball S.L. **2003a.** Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences* 270:313–321.

Hebert, P.D.N., Ratnasingham, S., deWaard, J.R. **2003b**. Barcoding animal life: cytochrome c oxidase subunit 1 diverges among closely related species. *Proceedings of the Royal Society B: Biological Sciences (Suppl.)* 270: 96– 99.

Hebert, P.D.N., Penton, E.H., Burns, J.M., Janzen, D.H., Hallwachs, W. **2004**. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator. *Proceedings of the National Academy of Sciences USA* 101(41):14812-7.

Hebert, P.D.N., Gregory, T.R. **2005**. The promise of DNA barcoding for taxonomy. *Systematic Biology* 54:852-859.

Heled J, Drummond AJ. **2008**. Bayesian inference of population size history from multiple loci. *BMC Evolutionary Biology* 8:289.

Heled, J., Drummond, A. **2010.** Bayesian Inference of Species Trees from Multilocus Data. *Molecular Biology and Evolution* 27: 570-580.

Henikoff, S., Hnekoff, J.G. **1992**. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences USA* 89(22):10915-10919.

Hennig, W. **1966**. Phylogenetic systematics. Urbana (IL): University of Illinois Press.

Hennig, W. **1966**. Phylogenetic systematics. Urbana (IL):University of Illinois Press.

Henz, S.R., Huson, D.H., Auch, A.F., Nieselt-Struwe, K., Schuster, S.C. **2005**. Whole-genome prokaryotic phylogeny. *Bioinformatics* 21(10):2329-35.

Hewitt, G.M. **2001**. Speciation, hybrid zones and phylogeography or seeing genes in space and time. *Molecular Ecology* 10:537–549.

Hewitt, G.M. **2004.** Genetic consequences of climatic oscillations in the Quaternary. *Philos Trans R Soc Lond B Biol Sci*. 359(1442):183-95.

Hey, J. **1992**. Using phylogenetic trees to study speciation and extinction. *Evolution* 46(3):627-640.

Hey, J. **2001.** The Mind of the Species Problem. *Trends in Ecology and Evolution* 16: 326-329.

Hey, J. **2006a**. On the failure of modern species concepts. *Trends in Ecology and Evolution* 21:447-450.

Hey, J. **2006b**. Recent advances in assessing gene flow between diverging populations and species. *Current Opinion in Genetics & Development* 16:592-596.

Hey, J., Nielsen, R. **2004**. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* 167(2):747-60.

Hey, J., Nielsen, R. **2007**. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences USA* 104(8):2785-90

Hickerson, M.J., Carstens, B.C., Cavender-Bares, J., Crandall, K.A., Graham, C.H., Johnson, J.B., Rissler, L., Victoriano, P.F., Yoder, A.D. **2010**. Phylogeography's past, present, and future: 10 years after Avise, 2000. *Molecular Phylogenetics and Evolution* 54(1):291-301.

Hickerson, M.J., Meyer, C.P, Moritz, C. **2006**. DNA barcoding will often fail to discover new animal species over broad parameter space. *Systematic Biology* 55:729–739.

Hillis, D.M. **1996**. Inferring complex phylogenies. *Nature* 383:130.

Hillis, D.M., Moritz, C., Mable, B.K. **1996**. Molecular Systematics. Sunderland, MA: Sinauer Associates.

Hodkinson, I.D., Casson, D. **1991**. A lesser predilection for bugs: Hemiptera (Insecta) diversity in tropical rain forests. *Biological Journal of the Linnean Society* 43:101-109.

Hubbard, T.J.P., Aken, B.L., Beal, K., Ballester, B., Caccamo, M., et al. **2007**. Ensembl 2007. *Nucleic Acids Research* 35: D610–D617.

Hudson, R.R., Coyne, J.A. **2002**. Mathematical consequences of the genealogical species concept. *Evolution* 56, 1557–1565.

Huelsenbeck, J.P., Crandall, K.A. **1997**. Phylogeny estimation and hypothesis testing using maximum likelihood. *Annual Review of Ecology and Systematics* 28:437-466.

Huelsenbeck, J.P., Rannala, B, Yang, Z. **1997**. Statistical tests of host-parasite cospeciation. *Evolution* 51: 410-419.

Huelsenbeck, J.P., Rannala, B. **1997**. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science* 276:227–232.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R., Bollback, J.P. **2001**. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294:2310-2314.

Huelsenbeck, J.P., Nielsen, R., Bollback, J.P. **2003**. Stochastic Mapping of Morphological Characters. *Systematic Biology* 52(2):131-158.

Huelsenbeck, J. P., Rannala, B. **2004**. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Systematic Biology* 53(6):904-913.

Hull, D.L. **1988**. Science as a process: An evolutionary account of the social and conceptual development of science. Chicago: University of Chicago Press.

Hulsen, T., Huynen, M.A., de Vlieg, J., Groenen, P.M. **2006**. Benchmarking ortholog identification methods using functional genomics data. *Genome Biology* 7:31.

Hunt, T., Bergsten, J., Levkanicova, Z., Papadopoulou, A., St John, O., Wild, R., Hammond, P.M., Ahrens, D., Balke, M., Caterino, M.S., Gómez-Zurita, J.,Ribera, I., Barraclough, T.G., Bocakova, M., Bocak, L., Vogler, A.P. **2007**. A comprehensive phylogeny of beetles reveals the evolutionary origins of a super-radiation. *Science* 318:1913-1916.

Huynen, L., Millar, C.D., Lambert, D.M. **2012**. Resurrecting ancient animal genomes: the extinct moa and more. *Bioessays* 34(8):661-9.

Ishiwata, K., Sasaki, G., Ogawa, J., Miyata, T., Su, Z. **2010**. Phylogenetic relationships among insect orders based on three nuclear protein coding gene sequences. *Molecular Phylogenetics and Evolution* 58:169–80.

IUCN. **2011**. IUCN Red List of Threatened Species. Version 2011.1. <www.iucnredlist.org>.

Jabot, F., Chave, J. **2009**. Inferring the parameters of the neutral theory of biodiversity using phylogenetic information and implications for tropical forests. *Ecology Letters* 12:239–248.

Jackson, A.P. **2004**. A reconciliation analysis of host switching in plant-fungal symbioses. *Evolution* 58(9):1909-23.

Janda, M., Folkova, D., Zrzavy, J. **2004**. Phylogeny of *Lasius* ants based on mitochondrial DNA and morphology, and the evolution of social parasitism in the Lasiini (Hymenoptera: Formicidae). *Molecular Phylogenetics and Evolution* 33:595-614.

Janzen, D.H., Hajibabaei, M., Burns, J.M., Hallwachs, W., Remigio, E., **Hebert,** P.D. **2005**. Wedding biodiversity inventory of a large and complex Lepidoptera fauna with DNA barcoding. *Philosophical Transactions of the Royal Society London B: Biological Sciences* 360(1462):1835-45.

Jeffroy, O., Brinkmann, H., Delsuc, F., Philippe, H. **2006**. Phylogenomics: the beginning of incongruence? *Trends in Genetics* 22(4):225-31.

Johnson, G.B. **2003**. The living world. New York: McGraw Hill.

Johnson, K.P., Yoshizawa, K., Smith, V.S. Multiple origins of parasitism in lice. **2004.** *Proceedings of the Royal Society B: Biological Sciences* 271(1550):1771-1776.

Joly, S., McLenachan, P. A., Lockhart, P. J. **2009.** A statistical approach for distinguishing hybridization and incomplete lineage sorting. *American Naturalist*, 174, e54–e70.

Jones, D.T., Taylor, W.R., Thornton, J.M. **1992**. The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences* 8:275-282.

Joost, S., Kalbermaten, M., Bonin, A. **2008**. Spatial analysis method (SAM): a software tool combining molecular and environmental data to identify candidate loci for selection. *Molecular Ecology Resources* 8:957.

Joost, S.A.B., Bruford, M.W., Després, L., Conord, C., Erhardt, G., Taberlet, P. **2007**. A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Molecular Ecology* 16:3955–3969.

Joseph L, Moritz C, Hugall A. **1995**. Molecular support for vicariance as a source of diversity in rainforest. *Proceedings of the Royal Society B: Biological Sciences* 260(1358):177-82.

Jurado-Rivera, J., Vogler, A.P., Reid, C.A.M., Petitpierre, E., Gómez-Zurita, J. **2009**. DNA barcoding insect-host plant associations. *Proceedings of the Royal Society B: Biological Sciences* 276: 639–648.

Kandul, N.P., Lukhtanov, V.A., Dantchenko, A.V., Coleman, J.W.S, Sekercioglu, C.H., Haig, D., Pierce, N.E. **2004**. Phylogeny of *Agrodiaetus* Hübner, 1822 (Lepidoptera: Lycaenidae) inferred from mtDNA sequences of COI and COII and nuclear sequences of EF1-a: karyotype diversification and species radiation. *Systematic Biology* 53: 278–298.

Kandul, N.P., Lukhtanov, V.A., Pierce, N.E. **2007**. Karyotypic diversity and speciation in *Agrodiaetus* butterflies. *Evolution* 61: 546–559.

Kawakita, A., Takimura, A., Terachi, T., Sota, T., Kato, M. **2004**. Cospeciation analysis of an obligate pollination mutualism: have Glochidion trees (Euphorbiaceae) and pollinating Epicephala moths (Gracillariidae) diversified in parallel? *Evolution* 58: 2201–2214.

Kim, K. C., Byrne, L.B. **2006**. Biodiversity loss and the taxonomic bottleneck: emerging biodiversity science. *Ecological Research* 21:794–810.

Kingman, J.F.C. **1982**. The coalescent. *Stochastic processes and their applications* 13:235-248.

Kishino, H., Thorne, J.L., Bruno, W.J. **2001**. Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Molecular Biology and Evolution* 18(3):352-61.

Kluge, A.G. **1998**. Total evidence or taxonomic congruence: cladistics or consensus classification. *Cladistics*. 14:151–158.

Kluge, A.G., Farris, J.S. **1969**. Quantitative phyletics and the evolution of anurans. *Systematic Zoology*. 19:1-32.

Knowles, L.L., Maddison, W.P. **2002**. Statistical phylogeography. *Molecular Ecology* 11:2623-2635.

Knowles, L.L. **2004**. The burgeoning field of statistical phylogeography. *Journal of Evolutionary Biology* 17:1-10.

Knowles, L.L., Carstens, B.C. **2007**. Delimiting species without monophyletic gene trees. *Systematic Biology* 56:887–895.

Knowles, L.L. **2009**. Statistical phylogeography. *Annual Review of Ecology and Systematics* 40:593-612.

Kolaczkowski, B., Thornton, J.W. **2004**. Performance of maximum parsimony and maximum likelihood phylogenetics when evolution is heterogeneous. *Nature* 431: 980–984.

Krause, J., Fu, Q., Good, J.M., Viola, B., Shunkov, M.V., Derevianko, A.P., Pääbo, S. **2010**. The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature* 64(7290):894-7.

Kubatko, L., Degnan, J. **2007**. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology* 56, 17–24.

Kubatko, L.S., Carstens, B.C., Knowles, L.L. **2009**. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics* 25(7):971-3.

Kubo, T., Iwasa, Y. **1995**. Inferring the rates of branching and extinction from molecular phylogenies. *Evolution* 49(4):694-704.

Kuhner, M.K. **2006**. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics* 22(6):768-70.

Kuhner, M.K., Yamato, J., Felsenstein, J. **1998**. Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics* 149(1):429-34.

Kumar, S. & S. Subramanian. **2002**. Mutation rates in mammalian genomes. *Proceedings of the National Academy of Sciences USA* 99(2),803 -808.

Kumar, S. **2005**. Molecular clocks: four decades of evolution. *Nature Reviews Genetics* 6(8):654- 662.

Lamm, K., Redelings, B.D. **2009**. Reconstructing ancestral ranges in historical biogeography: properties and prospects. *Journal of Systematics and Evolution*, 47(5):369-382.

Lamoreux, J.F., Morrison, J.C., Ricketts, T.H., Olson, D.M., Dinerstein, E., McKnight, M.W., Shugart, H.H. **2006.** Global tests of biodiversity concordance and the importance of endemism. *Nature* 440(7081):212-4.

Larget, B., Kotha, S.K., Dewey, C.N., Ané, C. **2010**. BUCKy: Gene tree / species tree reconciliation with the Bayesian concordance analysis. *Bioinformatics* 26(22): 2910-2911.

Lartillot, N., Philippe, H. **2004**. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology Evolution* 21(6):1095-109.

Le, S.Q., Gascuel, O. **2008**. An improved general amino acid replacement matrix. *Molecular Biology and Evolution* 25(7):1307-20.

Le, S.Q., Lartillot, N., Gascuel, O. **2008**. Phylogenetic mixture models for proteins. *Philosofical Transactions of the Royal Society of London B Biological Sciences* 363(1512):3965-76.

Leaché, A.D., Koo, M.S., Spencer, C.L., Papenfuss, T.J., Fisher, R.N., McGuire, J.A. **2009**. Quantifying ecological, morphological, and genetic variation to delimit species in the coast horned lizard species complex (*Phrynosoma*). *Proceedings of the National Academy of Sciences USA* 106:12418-12423.

Leavitt, D.H., Bezy, R.L., Crandall, K.A., Sites Jr., J.W. **2007**. Multi-locus DNA sequence data reveal a history of deep cryptic vicariance and habitat-driven convergence in the desert night lizard Xantusia vigilis species complex (Squamata: Xantusiidae). *Molecular Ecology* 16:4455–4481.

Leigh, J.W., Lapointe, F.J., Lopez, P., Bapteste, E. **2011**. Evaluating phylogenetic congruence in the post-genomic era. *Genome Biology and Evolution* 3:571-87.

Lemmon, A.R., Emme, S.A., Lemmon, E.M. **2012**. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology* 61(5):727-44.

Lewis, P.O. **1988.** A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data. *Molecular Biology and Evolution* 5(3):277-83.

Lewis, P. O., Holder, M. T., Holsinger, K. E. **2005**. Polytomies and Bayesian phylogenetic inference. *Systematic Biology* 54(2): 241-253.

Li, W.H. **1997**. Molecular Evolution. Sunderland, MA: Sinauer Associates.

Lin, C., Danforth, B. **2004**. How do insect nuclear and mitochondrial gene substitution patterns differ? Insights from Bayesian analyses of combined datasets. *Molecular Phylogenetics and Evolution* 30(3):686-702.

Linnæus, C. **1735**. Systema naturæ, sive regna tria naturæ systematice proposita per classes, ordines, genera, & species. pp 1–12. Lugduni Batavorum.

Liò, P., Goldman, N. **1998**. Models of molecular evolution and phylogeny. *Genome Research*, 8(12), 1233-1244.

Liu, L., Pearl, D.K. **2007**. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Systematic Biology* 56(3):504-14.

Lukhtanov, V.A., Kandul, N.P., Plotkin, J.B., Dantchenko, A.V., Haig, D., Pierce, N.E. **2005**. Reinforcement of pre-zygotic isolation and karyotype evolution in *Agrodiaetus* butterflies. *Nature* 436, 385–389.

Lyell, C. **1832**. Principles of geology. Vol II. London: John Murray.

Maddison, W.P. **1990**. A method for testing the correlated evolution of two binary characters: are gains or losses concentrated on certain branches of a phylogenetic tree. *Evolution* 44:539-557.

Maddison, W.P., Knowles, L.L. **2006**. Inferring phylogeny despite incomplete_lineage sorting. *Systematic Biology* 55(1):21-30.

Maddison, W.P., Maddison, D.R. **1992.** MacClade: Analysis of phylogeny and character evolution. Version 3. Sinauer Associates, Sunderland, Massachusetts.

Mallet, J. **1995**. A species definition for the modern synthesis. *Trends in Ecology and Evolution* 10:294–299.

Mallet, J. **2001**. Species, concepts of. In *Encyclopedia of Biodiversity* Volume 5. Levin, S., [Eds]. Academic Press. Elsevier, Oxford.

Mallet, J., Willmott, K. **2003**. Taxonomy: Renaissance or Tower of Babel?. *Trends in Ecology and Evolution* 18(2):57-59.

Masters, B.C., Fan, V, Ross, H.A. **2011**. Species Delimitation - a Geneious plugin for the exploration of species boundaries. *Molecular Ecology Resources* 11:154-157.

May, R.M. **1990**. How many species? *Philosophical Transactions of the Royal Society. B. Biological Sciences* 330:293-304.

May, R.M. **1998**. The dimensions of life on Earth. Pp. 30–45. In *Nature and Human Society: The Quest for a Sustainable World.* Raven, P.H., [Ed]. National Academy Press, Washington, DC.

May, R.M. **2000**. The dimensions of life on Earth. In *Nature and Human Society: The Quest for a Sustainable World.* Raven, P.H., Williams, T., [Eds]. National Academies Press, Washington, D.C.

Mayden, R.L. **1997**. A hierarchy of species concepts: the denouement in the saga of the species problem. Pp. 381–424. In *Species: The Units of Biodiversity*. Claridge, M.F., Dawah, H.A., Wilson, M.R., [Eds]. Chapman and Hall, London.

Mayr, E. **1942**. Systematics and the origin of species. Columbia University Press, New York.

Mayr, E. **1982**. The growth of biological thought: diversity, evolution, and inheritance. Harvard University Press, Cambridge, MA.

Mayr, E., Ashlock, P.D. **1991**. *Principles of Systematic Zoology*. 2nd edition. McGraw-Hill College, New York.

McCormack, J.E., Faircloth, B.C., Crawford, N.G., Gowaty, P.A., Brumfield, B.T., Glenn, T.C. **2012**. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species tree analysis. *Genome Research* 22:746–754.

McCulloch, G.A., Wallis, G.P., Waters, J.M. **2010**. Onset of glaciation drove simultaneous vicariant isolation of Alpine insects in New Zealand. *Evolution* 64(7):2033-43.

McKenna, D, Sequeira, A.S., Marvaldi, A.E., Farrell, B.D. **2009**. Temporal lags and overlap in the diversification of weevils and flowering plant. *Proceedings of the National Academy of Sciences USA* 106:7083-7088.

Medina, M., Collins, A.G., Silberman, J.D., Sogin, M.L. **2001**. Evaluating hypotheses of basal animal phylogeny using complete sequences of large and small subunit rRNA. *Proc. Natl. Acad. Sci. USA* 98, 9707–9712.

Meier, R., Shiyang, K., Vaidya, G., Ng, P.K.L. **2006**. DNA barcoding and taxonomy in Diptera: A tale of high intraspecific variability and low identification success. *Systematic Biology* 55:715–728.

Meng, C., Kubatko, L.S. **2009**. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. *Theoretical Population Biology* 75:35-45.

Meusemann, K., von Reumont, B.M., Simon, S., Roeding, F., Strauss, S., Kück, P., Ebersberger, I., Walzl, M., Pass, G., Breuers, S., Achter, V., von Haeseler, A., Burmester, T., Hadrys, H., Wägele, J.W., Misof, B. **2010**. A phylogenomic approach to resolve the arthropod tree of life *Molecular Biology and Evolution* 27:2451–64.

Meyer, C.P., Paulay, G. **2005**. DNA Barcoding: Error Rates Based on Comprehensive Sampling. *PLoS Biology* 3:e422.

Miller, J. S., and J. W. Wenzel. **1995**. Ecological characters and phylogeny. *Annual Review of Entomology* 40: 389-415.

Miller, W., Drautz, D.I., Ratan, A., Pusey, B., Qi, J., Lesk, A.M., Tomsho, L.P., Packard, M.D., Zhao, F., Sher, A., Tikhonov, A., Raney, B., Patterson, N., Lindblad-Toh, K., Lander, E.S., Knight, J.R., Irzyk, G.P., Fredrikson, K.M., Harkins, T.T., Sheridan, S., Pringle, T., Schuster, S.C. **2008**. Sequencing the nuclear genome of the extinct woolly mammoth. *Nature* 456(7220):387-90.

Moreau, C.S., Bell, C.D., Vila, R., Archibald, S.B., Pierce, N.E. **2006.** Phylogeny of the Ants: Diversification in the Age of Angiosperms. *Science* 312(5770):101-104.

Moritz, C., Cicero, C. **2004**. DNA Barcoding: Promise and Pitfalls. *PLoS Biology* 2(10):1529–31.

Mossel, E., Vigoda, E. **2005**. Phylogenetic MCMC algorithms are misleading on mixtures of trees. *Science*, 309:2207–2209.

Nabholz, B., Glemin, S., Galtier, N. **2008**. Strong variations of mitochondrial mutation rate across mammals-the longevity hypothesis. *Molecular Biology and Evolution* 25(1):120-130.

Nabholz, B., Glemin, S., Galtier, N. **2009**. The erratic mitochondrial clock: variations of mutation rate, not population size, affect mtDNA diversity across birds and mammals. *BMC Evolutionary Biology*, 9(1):54.

Nee, S., Holmes, E.C., May, R.M., Harvey, P.H. **1994**a. Extinction rates can be estimated from molecular phylogenies. *Philosophical Transactions of the Royal Society B: Biological Sciences* 344(1307):77-82.

Nee, S., May, R.M., Harvey, P.H. **1994**b. The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society B: Biological Sciences* 344(1309):305-11.

Nee, S., Mooers, A., Harvey, P. **1992**. Tempo and mode of evolution revealed from molecular phylogenies. *Proceedings of the National Academy of Sciences USA* 89(17):8322-8326.

Nielsen, R., Beaumont, M.A. **2009**. Statistical inferences in phylogeography. *Molecular Ecology* 18:1034-1047.

Nielsen, R., Matz, M. **2006**. Statistical approaches for DNA barcoding. *Systematic Biology* 55:162–169.

Nielsen, R., Wakeley, J. **2001**.Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* 158(2):885-96.

Nosil, P., Parchman, T.L., Feder, J.L., Gompert, Z. **2012**. Do highly divergent loci reside in genomic regions affecting speciation? A test using next-generation sequence data in *Timema* stick insects. *BMC Evolutionary Biology* 12:164.

Nylander, J.A.A., Ronquist, F., Huelsenbeck, J.P., Nieves-Aldrey, J.L. **2004**. Bayesian phylogenetic analysis of combined data. *Systematic Biology* 53:47–67.

Ogden, T. H., Rosenberg, M.S. **2006.** Multiple sequence alignment accuracy and phylogenetic inference. *Systematic Biology* 55:314-328.

Omland, K.E., Cook, L.G., Crisp, M.D. 2008. Tree thinking for all biology: The problem with reading phylogenies as ladders of progress. *Bioessays* 30 (9): 854-867.

Ord, T.J., Martins, E.P. **2010**. The evolution of behavior: phylogeny and the origin of present-day diversity. In *Evolutionary Behavioral Ecology*. Westneat, D.F., Fox C.W., [Eds]. Oxford University Press, New York.

Orme, C.D.L., Davies, R.G., Burgess, M., Eigenbrod, F., Pickup, N., Olson, V.A., Webster, A.J., Ding, T.S., Rasmussen, P.C., Ridgely, R.S., Stattersfield, A.J., Bennett, P.M., Blackburn, T.M., Gaston, K.J., Owens, I.P.F. **2005**. Global hotspots of species richness are not congruent with endemism or threat. *Nature* 436:1016-1019.

Page, R.D.M., Holmes, E.C. **1998**. Molecular Evolution: a phylogenetic approach. Oxford, UK: Blackwell Science.

Pagel, M. **1994**. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society, London, Series B* 255:37-45.

Pagel, M. **1999**. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Systematic Biology*, 48, 612-622.

Pagel, M. **1999**b. Inferring the historical patterns of biological evolution. *Nature* 401: 877-884.

Pagel, M., Meade, A. **2008**. Modelling heterotachy in phylogenetic inference by reversible-jump Markov chain Monte Carlo. *Philosophical Transactions of the Royal Society B-Biological Sciences* 363(1512):3955-3964.

Panchal, M., Beaumont, M.A. **2007**. The automation and evaluation of nested clade phylogeographic analysis. *Evolution* 61:1466-1480.

Paradis, E. **1997**. Assessing temporal variations in diversification rates from phylogenies: estimation and hypothesis testing. *Proceedings of the Royal Society B: Biological Sciences* 264(1385):1141-1147.

Patterson, C., Williams, D.M., Humphries, C.J. **1993**. Congruence between molecular and morphological phylogenies. *Annual Review of Ecology and Systematics* 24: 153-188.

Penn, O., Privman, E., Ashkenazy, H., Landan, G., Graur, D., Pupko, T. **2010**. GUIDANCE: a web server for assessing alignment confidence scores. *Nucleic Acids Research* 38(2):W23-28.

Penn, O., Privman, E., Landan, G., Graur, D., Pupko, T. **2010**. An alignment confidence score capturing robustness to guide tree uncertainty. *Molecular Biology and Evolution* 27(8): 1759-1767.

Percy, D.M., Page, R.D.M., Cronk, Q.C.B. **2004**. Plant-insect interactions: double-dating associated insect and plant lineages reveals asynchronous radiations. *Systematic Biology* 53:120–127.

Pereira, S.L., Baker, A.J. **2006**. A mitogenomic timescale for birds detects variable phylogenetic rates of molecular evolution and refutes the standard molecular clock. *Molecular Biology and Evolution* 23(9):1731-1740.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. **2012**. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One* 7(5):e37135.

Peterson, A.T., Soberon, J., Sanchez-Cordero, V. **2002**. Distributional prediction based on ecological niche modeling of primary occurrence data. In: Scott, J.M.,

Heglund, P.J., Morrison, M., Haufler, J.B., Raphael, M.G., Wall, W.A., Samson, F.B., [Eds]. **2002**. *Predicting Species Occurrences: Issues of Scale and Accuracy*. Island Press, Washington, DC.

Petit, R. J. **2008**. The coup de grâce for nested clade phylogeographic analysis? *Molecular Ecology* 17:516–518.

Petit, R.J., Grivet. D. **2002**. Optimal randomization strategies when testing the existence of a phylogeographic structure. *Genetics* 161, 469–471.

Philippe, H., Chenuil, A., Adoutte, A. **1994**. Can the Cambrian explosion be inferred through molecular phylogeny? *Development* 120: S15–S25.

Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N., Delsuc, F. **2005a**. Heterotachy and long-branch attraction in phylogenetics. *BMC Evolutionary Biology* 5:50.

Philippe, H., Lartillot, N., Brinkmann, H. **2005b**. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Molecular Biology and Evolution* 22, 1246–1253.

Philippe, H., Derelle, R., Lopez, P., Pick, K., Borchiellini, C., Boury-Esnault, N., Vacelet, J., Renard, E., Houliston, E., Quéinnec, E., Da Silva, C., Wincker, P, Le Guyader, H., Leys, S., Jackson, D.J., Schreiber, F., Erpenbeck, D., Morgenstern, B., Wörheide, G., Manuel, M. **2009**. Phylogenomics revives traditional views on deep animal relationships. *Current Biology* 19: 706–712

Philippe, H., Brinkmann, H., Lavrov, D.V., Littlewood, D.T.J., Manuel, M., Wörheide, G., Baurain, D. **2011**. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biology* 9(3): e1000602.

Pierce, N.E., Braby, M.F., Heath, A., Lohman, D.J, Mathew, J, Rand, D.B., Travassos M.A. **2002**. The ecology and evolution of ant association in the Lycaenidae (Lepidoptera). *Annual Review of Entomology* 47: 733-771.

Pons, J., Barraclough, T., Gomez-Zurita, J., Cardoso, A., Duran, D., Hazell, S., Kamoun, S., Sumlin, W., Vogler, A. **2006**. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology* 55:595–610.

Pons, J., Ribera, I, Bertranpetit, J., Balke, M. **2010**. Nucleotide substitution rates for the full set of mitochondrial protein-coding genes in Coleoptera. *Molecular Phylogenetics and Evolution* 56(2):796-807.

Posada, D., Crandall, K.A. **2001**. Intraspecific phylogenetics: Trees grafting into networks. *Trends in Ecology and Evolution* 16(1):37-45.

Posada, D. **2009**. Selecting Models of Molecular Evolution. In *The Phylogenetic Handbook*, 2nd Edition. Vandamme, A.M., Salemi, M., Lemey, P., [Eds]. Cambridge University Press.

Posada, D. **2012**. Reconstrucción de árboles filogenéticos. In *El árbol de la vida: sistemática y evolución de los seres vivos.* Vargas, P., Zardoya, R., [Eds]. Madrid.

Puillandre, N., Lambert, A., Brouillet, S., Achaz, G. **2012.** ABGD, Automatic barcode gap discovery for primary species delimitation. *Molecular Ecology* 21(8):1864-77.

Puorto, G., Salomao, M.G, Theakston, R.D.G., Thorpe, R.S., Warrell, D.A, Wuster, W. **2001**. Combining mitochondrial DNA sequences and morphological data to infer species boundaries: Phylogeography of lanceheaded pitvipers in the Brazilian Atlantic forest, and the status of *Bothrops pradoi* (Squamata: Serpentes: Viperidae). *Journal of Evolutionary Biology* 14:527–538.

Quang, le S., Gascuel, O., Lartillot, N. **2008**. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* 24(20):2317-23.

Rannala, B., Yang, Z. **1996**. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *Journal of Molecular Evolution* 43(3), 304-311.

Rannala, B., Yang, Z. **2003**. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* 164(4):1645-56.

Rannala B, Yang Z. **2007**. Inferring speciation times under an episodic molecular clock. Systematic Biology 56(3):453-66.

Rannala B, Yang Z. **2008**. Phylogenetic inference using whole genomes. *Annual Review of Genomics and Human Genetics* 9:217-31.

Ratnasingham, S., Hebert, P.D.N. **2007**. BOLD: the barcode of life data system (www.Barcodinglife.org). *Molecular Ecology Notes* 7:355-364.

Rausher, M.D., Miller R.E., Tiffin, P. **1999**. Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. *Molecular Biology and Evolution*, 16(2), 266 -274.

Raxworthy, C., Ingram, C., Rabibisoa, N., Pearson, R. **2007**. Applications of ecological niche modeling for species delimitation: a review and empirical evaluation using day geckos (Phelsuma) from Madagascar. *Systematic Biology* 56:907–923.

Ree, R.H., Moore, B.R., Webb, C.O., Donoghue, M.J. **2005**. A likelihood framework for inferring the evolution of geographic range on phylogenetic trees. *Evolution* 59(11):2299-2311.

Ree, R. H. and S. A. Smith. **2008**. Maximum-likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. *Systematic Biology* 57(1):4-414.

Regier, J.C., Shultz, J.W., Ganley, A.R., Hussey, A., Shi, D., Ball, B., Zwick, A., Stajich, J.E., Cummings, M.P., Martin, J.W., Cunningham, C.W. **2008**. Resolving arthropod phylogeny: exploring phylogenetic signal within 41 kb of protein-coding nuclear gene sequence. *Systematic Biology* 57:920–38.

Remm, M., Storm, C.E., Sonnhammer, E.L. **2001**. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology* 314: 1041–1052.

Rieppel, O. **2005**. The philosophy of total evidence and its relevance for phylogenetic inference. *Papéis Avulsos de Zoologia (São Paulo)* 45:1–31.

Rodriguez-Ezpeleta, N., Brinkmann, H., Roure, B., Lartillot, N., Lang, B.F., Philippe, H. **2007**. Detecting and overcoming systematic errors in genome-scale phylogenies. *Systematic Biology*, 56(3):389-399.

Rokas, A., Holland, P.W. Rare genomic changes as a tool for phylogenetics. **2000**. *Trends in Ecology and Evolution* 15(11):454-459.

Rokas, A., Carroll, S.B. **2005**. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. *Molecular Biology and Evolution* 22(5):1337-44.

Rokas, A., Carroll, S. B. **2006**. Bushes in the Tree of Life. *PLoS Biology*: 4: e352.

Rokas, A., Williams, B.L., King, N., Carroll, S.B. **2003**. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425:798-804.

Ronquist, F. **1997**. Dispersal-vicariance analysis: a new approach to the quantification of historical biogeography. *Systematic Biology* 46:195-203.

Ronquist, F., Deans, A.R. **2010**. Bayesian phylogenetics and its influence on insect systematics. *Annual Review of Entomology* 55:189–206.

Rosen, D.E. **1979**. Fishes from the uplands and intermontane basins of Guatemala: Revisionary studies and comparative geography. *Bulletin of the American Museum of Natural History*162:267–376.

Rosenberg, N.A., Nordborg, M. **2002**. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Review Genetics*. 3(5):380-90.

Rota-Stabelli, O., Campbell, L., Brinkmann, H., Edgecombe, G.D., Longhorn, S.J., Peterson, K.J., Pisani, D., Philippe, H., Telford, M.J. **2011**. A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. *Proceedings of the Royal Society B: Biological Sciences* 278(1703):298-306.

Rubin, B.E., Ree, R.H., Moreau, C.S. 2012. Inferring phylogenies from RAD sequence data. *PLoS One* 7(4):e33394.

Rubinoff D., Holland B.S. **2005**. Between two extremes: mitochondrial DNA is neither the panacea nor the nemesis of phylogenetic and taxonomic inference. *Systematic Biology* 54:952–961.

Rubinoff, D. **2006**. Utility of mitochondrial DNA barcodes in species conservation. *Conservation Biology* 20(4):1026–33.

Rzhetsky, A., Nei, M. **1992**. A simple method for estimating and testing minimum-evolution trees. *Molecular Biology and Evolution* 9(5):945-967.

Saitou, N., Nei, M. **1986**. The number of nucleotides required to determine the branching order of three species, with special reference to the human-chimpanzee-gorilla divergence. *Journal of Molecular Evolution* 24: 189–204.

Saitou, N., Nei, N. **1987**. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4(4): 406-425.

Savard, J., Tautz, D., Richards, S., Weinstock, G., Gibbs, R., Werren, J., Tettelin, H., Lercher, M. **2006**. Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of Holometabolous insects. *Genome Research*

16(11):1334-1338.

Schierwater B, Eitel M, Jakob W, Osigus HJ, Hadrys H, Dellaporta S.L., Kolokotronis, S.O., Desalle, R. **2009**. Concatenated analysis sheds light on early metazoan evolution and fuels a modern ''urmetazoon'' hypothesis. *PLoS Biology* 7(1): e20.

Schluter, D. **2009**. Evidence for ecological speciation and its alternative. *Science* 323: 737–741.

Schneider, C.J., Cunningham, M., Moritz, C. **1998**. Comparative phylogeography and the hystory of endemic vertebrates in the wet tropics rainforests of Australia. *Molecular Ecology* 7(4):487-498.

Shapiro, B., Hofreiter, M., [Eds]. **2012**. Ancient DNA: Methods and protocols. Humana Press Inc, NY, USA.

Simmons, M.P., Pickett, K.M., Miya, M. **2004**. How meaningful are bayesian Support Values?. *Molecular Biology and Evolution* 21(1):188-199.

Simpson, G.G. **1951**. The species concept. *Evolution* 5:285–298.

Simpson, G.G. **1964**. Organisms and molecules in evolution. *Science* 146(3651): 1535-1538.

Sims, G.E., Jun, S.R., Wu, G.A., Kim, S.H. **2009**. Whole-genome phylogeny of mammals: evolutionary information in genic and nongenic regions. *Proceedings of the National Academy of Sciences USA* 106(40):17077-82

Sites, J.W., Marshall, J.C. **2003**. Delimiting species: A renaissance issue in systematic biology. *Trends in Ecology and Evolution* 18:462–470.

Sites, J.W., Marshall, J.C. **2004**. Operational criteria for delimiting species. *Annual Review of Ecology, Evolution and Systematics* 35:199–227.

Smith, S.A., Donoghue, M.J. **2008**. Rates of molecular evolution are linked to life history in flowering plants. *Science*, 322(5898): 86-89.

Smith, V.S. **2005**. DNA barcoding: perspectives from a "Partnership for enhancing expertise in taxonomy" (PEET) debate. *Systematic Biology* 54:841-844.

Sneath, P.H.A., Sokal, R.R. **1973**. Numerical Taxonomy. San Francisco: W.H. Freeman.

Sokal, R., Michener, C. **1958**. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409-1438.

Sokal, R.R., Sneath, P.H.A. **1963**. Principles of Numerical Taxonomy. San Francisco: W.H. Freeman

Soria-Carrasco, V., Castresana, J. **2008**. Estimation of phylogenetic inconsistencies in the three domains of life. *Molecular Biology and Evolution* 25, 2319-2329.

Steel, M.A., Lockhart, P.J., Penny, D. **1993**. Confidence in evolutionary trees from biological sequence data. *Nature* 364, 440-442.

Steel, M.A., Lockhart, P.J., Penny, D. **1995**. A frequency-dependent significance test for parsimony. *Molecular Phylogenetics and Evolution* 4:64-71.

Stockman A.K., Bond J.E. **2007**. Delimiting cohesion species: extreme population structuring and the role of ecological interchangeability. *Molecular Ecology* 16:3374–3392.

Stork, N.E. **1988**. Insect diversity: facts, fiction and speculation. *Biological Journal of the Linnean Society* 35:321-337.

Stork, N.E. **1993**. How many species are there? *Biodiversity and Conservation* 2:215-232.

Sullivan, J., Joyce, P. **2005**. Model selection in phylogenetics. *Annual Review of Ecology, Evolution and Systematics*, 36:445-466.

Swenson, N.G., Howard, D.J. **2005**. Clustering of contact zones, hybrid zones, and phylogeographic breaks in North America. *American Naturalist* 166:581–591.

Taberlet, P., Cheddadi, R. **2002**. Quaternary refugia and persistence of biodiversity. *Science* 297(5589):2009-2010.

Talavera, G., Castresana, J. **2007**. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology* 56, 564–577.

Tautz, D., Arctander, P., Minelli, A., Thomas R.H., Vogler, A.P. **2003**. A plea for DNA taxonomy. *Trends in Ecology and Evolution* 18:70–74.

Templeton, A.R. **1992**. Human origins and analysis of mitochondrial DNA sequences. *Science* 255:737

Templeton, A.R., Routman, E., Phillips, C. **1995**. Separating population structure from population history: a cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the Tiger Salamander, Ambystoma tigrinum. *Genetics* 140:767-782.

Templeton, A.R. **1998**. Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history. *Molecular Ecology* 7:381-397.

Templeton, A.R. **2008**. Nested clade analysis: an extensively validated method for strong phylogeographic inference. *Molecular Ecology* 17:1877-1880.

Templeton, A.R. **2009**a. Statistical hypothesis testing in intraspecific phylogeography: nested clade phylogeographical analysis vs. approximate Bayesian computation. *Molecular Ecology* 18(2):319-331.

Templeton, A.R. **2009**b. Why does a method that fails continue to be used? The answer. *Evolution* 63:807-812.

Thomas, J.A., Welch, J.J., Woolfit, M., Bromham, L. **2006**. There is no universal molecular clock for invertebrates, but rate variation does not scale with body size. *Proceedings of the National Academy of Sciences USA* 103(19):7366 -7371.

Thomsen, P.F., Elias, S., Gilbert, M.T., Haile, J., Munch, K., Kuzmina, S., Froese, D.G., Sher, A., Holdaway, R.N., Willerslev, E. **2009**. Non-destructive sampling of ancient insect DNA. *PLoS ONE* 4(4):e5048.

Thorne, J.L., Kishino, H.,Painter, I.S. **1998**. Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution* 15(12):1647-1657.

Tilmon, K.J., [Ed]. **2008**. Specialization, Speciation and Radiation: The Evolutionary Biology of Herbivorous Insects. University of California Press, Berkeley, CA.

Trautwein, M.D., Wiegmann, B.M., Beutel, R., Kjer, K.M., Yeates, D.K. **2012**. Advances in insect phylogeny at the dawn of the postgenomic era. *Annual Review of Entomology* 57:449-68.

Van Valen, L. **1976**. Ecological species, multispecies, and oaks. *Taxon* 25:233–239.

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A.,

Wu, D., Paulsen, I., Nelson, K.E., Nelson, W., Fouts, D.E., Levy, S., Knap, A.H., Lomas, M.W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.H., Smith, H.O. **2004** Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304, 66–74.

von Reumont, B.M., Jenner, R.A., Wills, M.A., Dell'ampio, E., Pass,G., Ebersberger, I., Meyer, B., Koenemann, S., Iliffe, T.M., Stamatakis, A., Niehuis, O., Meusemann, K., Misof, B. **2012**. Pancrustacean Phylogeny in the Light of New Phylogenomic Data: Support for Remipedia as the Possible Sister Group of Hexapoda. *Molecular Biology and Evolution* 29(3): 1031-1045.

Wahlberg, N., Braby, M.F., Brower AVZ, de Jong, R., Lee, M., Nylin, S., Pierce, N.E., Sperling, F.A.H., Vila, R., Warren, A.D., Zakharov, E. **2005**. Synergistic effects of combining morphological and molecular data in resolving the phylogeny of butterflies and skippers. *Proc. R. Soc. Series B.* 272: 1577-1586.

Wakeley, J. **2008**. Coalescent theory an introduction. Greenwood Village (CO): Roberts and Company.

Waltari, E., Hijmans, R.J., Peterson, A.T., Nyári, A.S., Perkins, S.L., Guralnick, R.P. **2007**. Locating pleistocene refugia: comparing phylogeographic and ecological niche model predictions. *PLoS One* 2:e563.

Weiblen, G.D., Bush, G.L. **2002**. Speciation in fig pollinators and parasites. *Molecular Ecology* 11(8):1573-1578.

Weir, J.T., Schluter, D. **2008**. Calibrating the avian molecular clock. *Molecular Ecology* 17(10):2321-2328.

Wenzel, J. W. **1992**. Behavioral homology and phylogeny. *Annual Review of Ecology and Systematics* 23: 361-381.

Wheeler, Q.D., Valdecasas, A.G. **2007**. Taxonomy: Myths and Misconceptions. *Anales del Jardín Botánico de Madrid* 64(2):237-241.

Whelan S, Goldman N. **2001**. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution* 18(5):691-9.

Whitfield, J.B., Kjer, K.M. **2008**. Ancient rapid radiations of insects: challenges for phylogenetic analysis. *Annual Review of Entomology* 53:449–72.

Whitworth, T.L., Dawson, R.D., Magalon, H., Baudry, E. **2007**. DNA barcoding cannot reliably identify species of the blowfly genus Protocalliphora (Diptera: Calliphoridae). *Proceedings of the Royal Society B: Biological Sciences* 274(1619):1731-9.

Wiegmann, B.M., Mitter, C., Farrell, B. **1993**. Diversification of parasitic insects: extraordinary radiation or specialized dead end?. *American Naturalist* 142: 737-754.

Wiegmann, B., Trautwein, M., Kim, J.W., Cassel, B., Bertone, M., Winterton, S.L., Yeates, D.K. **2009**. Single-copy nuclear genes resolve the phylogeny of the holometabolous insects. *BMC Biology* 7:34.

Wiens, J.J., Servedio, M.R. **2000**. Species delimitation in systematics: inferring diagnostic differences between species. *Proceedings of the Royal Society of London, Series B* 267:631–636.

Wiens J.J., Penkrot T.A. **2002**. Delimiting species using DNA and morphological variation and discordant species limits in spiny lizards (Sceloporus). *Systematic Biology* 51:69–91.

Wiens, J. J. **2004**. The role of morphological data in phylogeny reconstruction. *Systematic Biology* 53:653-661.

Wiens, J.J. Graham, C.H. **2005**. Niche conservatism: integrating evolution, ecology, and conservation biology. *Annual Review of Ecology, Evolution, and Systematics* 36:519–539.

Wiens J. **2007**. Species delimitation: new approaches for discovering diversity. *Systematic Biology* 56:875–879.

Wiley, E.O. **1978**. The evolutionary species concept reconsidered. *Systematic Zoology* 27:17–26.

Will K.W., Mishler B.D., Wheeler Q.D. **2005**. The perils of DNA barcoding and the need for integrative taxonomy. *Systematic Biology* 54:844–851.

Will, K.W., Rubinoff, D. **2004**. Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics* 20:47–55.

Willerslev, E., Gilbert, M.T, Binladen, J., Ho, S.Y., Campos, P.F., Ratan, A., Tomsho, L.P., da Fonseca, R.R., Sher, A., Kuznetsova, T.V., Nowak-Kemp, M., Roth, T.L., Miller, W., Schuster, S.C. **2009**. Analysis of complete mitochondrial genomes from extinct and extant rhinoceroses reveals lack of phylogenetic resolution. *BMC Evolutionary Biology* 9:95.

Williams, J.W., Jackson, S.T., Kutzbach, J.E. **2007**. Projecte distributions of novel and disappearing climates by 2100 AD. *Proceedings of the National Academy of Sciences USA* 104(14):5738-5742.

Willis, K.J., Rudner, E., Sumegi. P. **2000**. The Full-Glacial Forests of Central and Southeastern Europe. *Quaternary Research* 53(2):203-213.

Wilson, E.O. **1955**. A monographic revision of the ant genus *Lasius*. *Bulletin of the Musuem of Comparative Zoology* 113:3-205.

Wilson, E.O. **1992**. The Diversity of Life. Belknap Press of Harvard University Press, Cambridge, Massachusetts.

Wolf, Y.I., Rogozin, I.B., Grishin, N.V., Koonin, E.V. **2002**. Genome trees and the Tree of Life. *Trends in Genetics* 18(2):472-9.

Wortley, A.H., Scotland, R.W. **2006**. The Effect of Combining Molecular and Morphological Data in Published Phylogenetic Analyses. *Systematic Biology* 55(4): 677-685.

Wright, S. **1940**. The statistical consequences of Mendelian heredity in relation to speciation. In *The new systematics*. Huxley, J. [Ed]. Oxford University Press, London.

Wu, M., Eisen, J.A. **2008**. A simple, fast, and accurate method of phylogenomic inference. *Genome Biology* 9(10):151.

Wu, M., Chatterji, S., Eisen, J. A. **2012**. Accounting for alignment uncertainty in phylogenomics. *PLoS ONE* 7(1):e30288.

Xia, X., Xie, Z. **2001**. DAMBE: data analysis in molecular biology and evolution. *Journal of Heredity* 92:371-373.

Xia, X.H., Xie, Z., Salemi, M., Chen, L., Wang, Y. **2003**. An index of substitution saturation and its application. *Molecular Phylogenetics and Evolution* 26:1-7.

Yang, Z., Nielsen, R. **1998**. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of Molecular Evolution* 46(4):409-418.

Yang, Z., Rannala, B. **2005**. Branch-length prior influences Bayesian posterior probability of phylogeny. *Systematic Biology* 54:455-470.

Yang, Z. **2006**. Computational molecular evolution. Oxford University Press, Oxford.

Yang, Z., Rannala, B. **2010**. Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences USA* 107:9264-9269.

Yoder, A.D., Yang, Z. **2000**. Estimation of Primate speciation dates using local molecular clocks. *Molecular Biology and Evolution*, 17(7):1081-1090.

Yule, G.U. **1925**. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. *Philosophical Transactions of the Royal Society B: Biological Sciences* 213:21-87.

Zapata, F., Jiménez, I. **2012**. Species Delimitation: Inferring gaps in morphology across geography. *Systematic Biology* 62(2):179-94.

Zuckerkandl, E., Pauling, L. **1962**. Molecular disease, evolution, and genetic heterogeneity. In *Horizons in Biochemistry*. Kasha, M., Pullman, M.B., [Eds]. New York: Academic Press.

Universitat Autònoma
de Barcelona