Nucleotide substitution rates for the full set of mitochondrial protein coding genes in Coleoptera

Joan Pons[1], Ignacio Ribera[2], Jaume Bertranpetit[2,3] and Michael Balke[4]

[1]Departament de Biodiversitat i Conservació, Institut Mediterrani d'Estudis Avançats IMEDEA (CSIC-UIB), Miquel Marqués, 21, Esporles, 07190 Illes Balears, Spain

[2]Institute of Evolutionary Biology (CSIC-UPF), Passeig Marítim de la Barceloneta 37, 08003 Barcelona, Spain

[3] Unitat de Biologia Evolutiva, Universitat Pompeu Fabra, Doctor Aiguader 88, Barcelona, 08003, Spain

[4]Zoological State Collection, Muenchhausenstrasse 21, 81247 Munich, Germany

Email: Joan Pons*–jpons@imedea.uib-csic.es; Ignacio Ribera–ignacio.ribera@ibe.upf-csic.es; Jaume Bertranpetit–jaume.bertranpetit@upf.edu; Michael Balke–Coleoptera-ZSM@zsm.mwn.de.

*Corresponding author

Key words: nucleotide substitution rate, Coleoptera suborders phylogeny, mitochondrial protein-coding genes, codon partitioning

Abstract

The ages of cladogenetic events in Coleoptera are frequently estimated with mitochondrial protein coding genes (MPCGs) and the "standard" mitochondrial nucleotide substitution rate for arthropods. This rate has been used for different mitochondrial gene combinations and time scales despite it was estimated on short mitochondrial sequences from few comparisons of close related species. These shortcomings may cause greater impact at deep phylogenetic levels as errors in rates and ages increase with branch lengths. We use the full set of MPCGs of 15 species of beetles (two of them newly sequenced here) to estimate the nucleotide evolutionary rates in a reconstructed phylogeny among suborders, paying special attention to the effect of data partitioning and model choices on these estimations. The optimal strategy for nucleotide data, as measured with Bayes factors, was partitioning by codon position. This retrieved Adephaga as a sister group to Myxophaga with strong support (expected likelihood weights test 0.94–1) and both sisters to Polyphaga, in contradiction with the most currently accepted views. The hypothesis of Archostermata being sister to the remaining Coleoptera, which is in agreement with morphology, was increasingly supported when third codon sites were recoded or completely removed, sequences were analyzed as AA, and heterogeneous models were implemented but the support levels remained low. Nucleotide substitution rates were strongly affected by the choice of data partitioning (codon position *versus* individual genes), with up to sixfold levels of variation, whereas differences in the molecular clock algorithm produced changes of only about 20%. The global mitochondrial protein coding rate using codon partitioning and an estimated age of 250 million years (MY) for the origin of the Coleoptera was 1.34% per branch per MY, which closely matches the 'standard' clock of 1.15% per MY. The estimation of the rates on alternative topologies gave similar results. Using local molecular clocks, the evolutionary rate in the Polyphaga and Archostemata was estimated to be nearly twice as fast as in the Adephaga and Myxophaga (1.03% *versus* 0.53% per MY). Rates across individual genes varied from 0.55% to 8.61% per MY. Our results suggest that *cox1* might not be an optimal gene for implementing molecular clocks in deep phylogenies for beetles because

it shows relatively slow rates at first and second codon positions but very fast rates at third ones. In contrast, *nad5*, *nad4* and *nad2* perform better, as they exhibit more homogeneous rates among codon positions.

Introduction

The Coleoptera forms the most diverse group of extant Metazoa, with *c*. 380,000 described species and an estimated total number of species ranging from one to more than three millions (Hammond, 1994; Ødegaard, 2000). Beetles are virtually everywhere and they are extremely diverse ecologically and morphologically. Their size ranges from tiny animals of *c*. 0.3 mm body length (Ptiliidae) to 'giants' of almost 18 cm (*Titanus giganteus*, the giant Amazonian longhorn beetle). This extraordinary diversity has long attracted the attention of evolutionary biologists and systematists (Hutchinson, 1959; Crowson, 1960, 1981; Farrell, 1998; Hunt et al., 2007), who have frequently built phylogenies of Coleoptera with mitochondrial (mt) DNA sequences for estimating the dates of evolutionary events. However, there is rarely enough fossil, geological or biogeographical evidence for node calibration, and many studies rely on molecular clocks with the 'standard' arthropod nucleotide substitution rate of 1.15% substitutions per million years (MY) or 2.3% sequence divergences per MY between species (Brower, 1994). This rate has several shortcomings: (1) it was estimated from eight arthropod examples only; (2) comparisons were only made at the intraspecific level or between closely related species; (3) the sequences used were short, representing only a small fraction of the mitochondrial genome (partial *cox1*, *cox2*, *rrnS* and *rrnL* sequences) and (4) genetic distances were calculated from restriction site polymorphisms, DNA–DNA hybridization, and only three were derived from pairwise comparisons of partial *cox1–cox2* DNA sequences assuming simple evolutionary models (Brower, 1994). However, this 'standard' clock has been used for multiple organisms, mitochondrial gene combinations, time scales (obviating saturation) and reconstruction methods. This is of special concern in the context of deep level phylogenies, as errors in branch length estimation—and hence in rates and ages—increase with the length of the branches (Buckley et al., 2001; Lemmon and Moriarty, 2004).

Here, we have used a phylogeny of the Coleoptera constructed using the full set of mitochondrial protein coding genes (MPCGs) of 15 species (two of them newly sequenced here) of the four extant suborders to estimate the rate of nucleotide substitution at the order and suborder

levels. The analysis of full mitochondrial genomes has been established as a powerful approach to elucidate deeper-level relationships among vertebrates (e.g., Zardoya and Meyer, 1996; Meyer and Zardoya, 2003; Murata et al., 2003) and also among Arthropods (e.g., Nardi et al., 2003; Masta et al., 2009). Recent studies have explored the utility of applying mitochondrial genome data to resolve phylogenetic relationships at the intraordinal level of insects with promising results, as for the Diptera (Cameron et al., 2007), Orthoptera (Fenn et al., 2008) and Hymenoptera (Dowton et al., 2009). Because the phylogenetic relationship among the four beetle suborders is still disputed (see below), we also explore the phylogenetic reconstruction of the four suborders of Coleoptera based on full mitochondrial sequences. We paid special attention to model and partition choice in both resolving the phylogeny and estimating the nucleotide substitution rates. Our general aim was to provide rates with more general applicability for evolutionary studies in the Coleoptera and especially for deeper level phylogenetics from the family to the subordinal level.

Background on the deep phylogeny of Coleoptera

The bulk of the diversity of Coleoptera is concentrated in two of the four currently recognized suborders, the Adephaga and Polyphaga, with *c*. 40,000 and 340,000 described species respectively (Beutel and Leschen, 2005). Of the two smaller suborders, the Myxophaga contains *c*. 100 predominantly aquatic beetle species, with a body length of less than 3 mm. The Archostemata is composed of only about 40 terrestrial species, ranging in length from 1.3 to 27 mm (Beutel and Leschen, 2005). The latter are rarely collected: three of the four families are known from single species only and two of them known only from the types. The earliest undisputed beetle fossils date back to the Upper Permian, about 250 MY ago (Ponomarenko, 1969) and belong to the Archostemata, strongly resembling even older stem-line Coleoptera such as the Permocupedidae (Ponomarenko, 1969; Beutel and Leschen, 2005; Grimaldi and Engel, 2005).

This disparate distribution of taxonomic diversity implies the possibility of multiple scenarios for explaining its origin, entirely depending on the inferred phylogenetic relationships

among the suborders. Virtually all possible scenarios have been suggested (see Beutel, 2005; Beutel and Hass, 2000 and Friedrich et al., 2009 for reviews). Two alternative hypotheses are supported by strong evidence as follows.

(1) (Archostemata + (Adephaga + (Myxophaga + Polyphaga))) (Crowson, 1960; Klausnitzer, 1975; Beutel, 1997; Beutel and Hass, 2000; Beutel, 2005; Hughes et al. ,2006; Friedrich et al., 2009). This hypothesis is supported by the most comprehensive morphological analyses to date (Beutel and Hass, 2000; Friedrich et al., 2009) and by the analysis of a group of ribosomal protein sequences (Hughes et al., 2006), although the latter gives low support.

(2) (Archostemata + (Myxophaga + (Adephaga +Polyphaga))). This hypothesis is supported by the analysis of full-length *SSU* (18S rRNA) sequences, although with low support (Shull et al., 2001; Caterino et al., 2002; Ribera et al., 2002; Vogler, 2005). The sister relationship of Adephaga and Polyphaga was also supported by the analysis of *SSU, rrnL* and *cox1* sequences for nearly 1900 species (Hunt et al., 2007).

In any case, the Archostemata is generally accepted to be a sister group to the remaining Coleoptera, which is in agreement with the fossil evidence (Ponomarenko, 1995; Beutel and Leschen, 2005; Grimaldi and Engel, 2005; Beutel and Pohl, 2006; Beutel et al., 2008; Friedrich et al., 2009). The second hypothesis above suggests a low-diversity stem Coleopteran lineage, with a single shift to a greatly increased diversification rate at the base of the Polyphaga + Adephaga grouping. Under the first hypothesis, several scenarios are possible, with at least two independent shifts in diversification rate, e.g., independent increases at the bases of the Polyphaga and Adephaga, or a single increase at the stem lineage sister group to the Archostemata with a subsequent decrease in the Myxophaga.

Materials and methods

Samples and DNA extraction

Aspidytes niobe Ribera et al., 2002 (Coleoptera, Adephaga, Aspidytidae) was collected in the Republic of South Africa (Mitchell's Pass, Ceres, ix.2002, D. T. Bilton leg.) and *Hydroscapha granulum* Motschulsky, 1855 (Coleoptera, Myxophaga, Hydroscaphidae) in Italy (Móngia, Piamonte, V. Móngia, 31.vii.2005, 720 m, 44° 17' 35.8" N; 7° 58' 34.5" E; I. Ribera and A. Cieslak leg.). DNA samples and voucher specimens are kept in the Natural History Museum London, Department of Entomology (urn:lsid:biocol.org:col:1009) (*A. niobe*, DNA ref. MB 302/BMNH(E) 703115) and the Museo Nacional de Ciencias Naturales, Madrid, Spain (MNCN) (urn:lsid:biocol.org:col:33867) (*H. granulum*, DNA ref. MNCN-JP1). Amplification and sequencing protocols of the mitochondrial genomes are supplied as additional information.

Other Coleopteran and outgroup sequences

GenBank featured 13 complete Coleopteran mitochondrial genomes as of November 2008: (1) one Adephaga: Carabidae: *Trachypachus holmbergi* (NC_011329); (2) one Archostemata: Ommatidae: *Tetraphalerus bruchi* (NC_011328); (3) one Myxophaga: Sphaeriusidae: *Sphaerius* sp. (NC_011322) and (4) ten Polyphaga: Chrysomelidae (*Crioceris duodecimpunctata*, NC_003372), Lampyridae (*Pyrocoelia rufa* , NC_003970), Tenebrionidae (*Tribolium castaneum*, NC_003081), Cerambycidae (*Anoplophora glabripennis* , NC_008221), Phloeostichidae (*Priasilpha obscura*, NC_011326), Melyridae (*Chaetosoma scaritides*, NC_011324), Phengodidae (*Rhagophthalmus lufengensis*, NC_010969, and *Rhagophthalmus ohbai* NC_010964), Elateridae (*Pyrophorus divergens*, NC_009964) and Scirtidae (*Cyphon* sp., NC_011320). The Coleoptera belongs to the Holometabola (Grimaldi and Engel, 2005). As there was no mitochondrial genome available for the Neuropteroidea, the assumed sister group of Coleoptera, we downloaded two Diptera and three Lepidoptera genomes to be used as outgroups because they are the closest relatives to Coleoptera among the available mitochondrial genomes (Nardi et al., 2003; Grimaldi and Engel, 2005;

Wiegmann et al., 2009). These were *Aedes albopictus* (NC_006817), *Ceratitis capitata* (NC_000857), *Ostrinia furnacalis* (NC_003368), *Antheraea pernyi* (NC_004622) and *Bombyx mori* (NC_002355). To root the tree we used a Hemimetabolan sequence of a group assumed to be close to the Holometabola (Psocoptera; Grimaldi and Engel, 2005): '*lepidopsocid* sp. RS-2001' (NC_004816). No sequence was available for a small fragment at the 3' end of *cox3* of *Bombyx mori*.

Nucleotide and amino acid composition and sequence alignment

Nucleotide sequences of the 13 MPCGs were translated into protein and each gene was individually aligned using the MAFFT 4.0 software (Katoh et al., 2005) using default parameters (BLOSUM62 matrix, open penalty 1.53 and extension penalty 0.123). Nucleotide sequences were subsequently aligned matching the AA alignment; i.e., aligned as triplets, using the tranalign tool (http://mobyle.pasteur.fr/cgi-bin/MobylePortal/portal.py). The 13 alignments were concatenated into a single matrix with 11,478 nucleotide sites (3826 AA positions) that is available upon request from JP. We checked for the presence of compositional biases, since they are expected to introduce artifacts in tree topology and branch lengths (Hassanin, 2006; Phillips, 2009). Nucleotide and AA compositions and relative synonymous codon usage (RSCU) were estimated using MEGA v4.0.2 (Tamura et al., 2007). The effective number of codons (ENC) was determined according to INCA v1.20 (Supek and Vlahovicek, 2004). Nucleotide compositional heterogeneity across species was estimated by self-organizing clustering and analysis of heterogeneity (Jermiin et al., 2004), as implemented in SeqVis v1.4 (http://www.bio.usyd.edu.au/jermiin/SeqVis/). Intrastrand equimolarity between A and T nucleotides and between G and C, or relative skew between complementary nucleotides within the same strand (Lobry, 1995), were calculated as follows: AT skew = (A – T)/(A + T) and GC skew = (G – C)/(G + C). Nucleotide and AA compositional deviations of each taxon were measured in PhyloBayes (Lartillot and Philippe, 2004) using the 'ppred –comp' command. The deviation was measured as the sum over the 20 AAs of the absolute differences between the taxon-specific and global empirical frequencies. We considered that a taxon had a deviated

composition if the z-score was > 2 (2 standard deviations above the mean, $p < 0.025$; Lartillot et al., 2009).

Phylogenetic analyses

We selected the model to analyze each partition in jModelTest (Posada, 2008) using the Bayesian Information Criterion. The best model of evolution for all partitions and genes was a general time-reversible model with gamma distribution plus a proportion of invariant sites (GTR+I+G) except for *atp*8 (GTR+I) and *nad*4L (GTR+G). Notwithstanding this, we also applied a GTR+I+G model to the later two genes to ease analyses. Phylogenetic Bayesian analyses were conducted in the parallel version of MrBayes 3.1.2 (Huelsenbeck and Ronquist, 2001) and run over the eight nodes on a Macintosh MacPro computer with 2 × 2.8 GHz Quad-Core Intel Xeon processors (Apple Inc., Cupertino, CA, USA). Each Bayesian search performed two independent runs starting with default prior values, random trees, and three heated and one cold Markov chains that ran for two million generations (3–10 million for protein analyses) sampled at intervals of 1000 generations. Burn-in and convergence of runs was assessed by examining the plot of generations against likelihood scores using the *sump* command in MrBayes. The convergence of all parameters of the two independent runs was also assessed using the program Tracer v1.4 (Rambaut and Drummond, 2007). We estimated the effective sample sizes for all parameters in the final set to test if they were greater than 100, indicating that the sampled generations were uncorrelated and the posterior distribution of the parameters was long and accurate (Rambaut and Drummond, 2007). Convergence of posterior clade probabilities in a single run (cumulative) and for independent ones (compare) was assessed with the software AWTY (Wilgenbusch et al., 2004). Trees from the two independent runs (once burn-in samples were discarded) were combined in a single majority consensus topology using the *sumt* command in MrBayes and the frequencies of the nodes in a majority rule tree were taken as *a posteriori* probabilities (Huelsenbeck and Ronquist, 2001).

Bayesian analyses were performed at the AA level with the mtArt+I+G model (Abascal et

al., 2007), selected by Protest 10.2 (Abascal et al., 2005) as optimal. We also implemented the CAT model in PhyloBayes (Lartillot et al., 2009), which estimates the distribution of site-specific effects underlying each dataset by combining infinite K categories of site-specific rates (Huelsenbeck and Suchard, 2007) and site-specific profiles over the 20 AA frequencies (Lartillot and Philippe, 2004). The global exchange rates was fixed to flat values (the CAT-Poisson) or inferred from the data (CAT–GTR settings). Four independent analyses were run in different cores until they converged. Convergence of split frequencies was assessed with the 'bpcomp' command and effective sample size for all parameters with the 'tracecomp' command. The first 1000 samples were discarded as burn-in and then sampled every 10 generations. We considered that independent runs converged when the maximum split frequency was < 0.1 and effective sample size was > 100 (Lartillot et al., 2009). To further reduce the compositional bias, we recoded the AA residues in six categories ('recode dayhoff6' command in PhyloBayes). Finally, we implemented the CAT+BP model in nh_PhyloBayes, which is heterogeneous across sites (CAT component) and nonstationary over time (BP component) (Blanchart and Lartillot, 2008). The convergence of the two independent runs was assessed using the command 'compchain' after 6000 generations.

We implemented different evolutionary models, data partitioning strategies, tree reconstruction and clock estimation methods to assess their effects on tree topology, branch lengths and rates since they are prone to several error sources that aggravate at deep phylogenetic levels (Roger and Hug, 2006; Philips, 2009 and references therein). We explored five different partitioning strategies: (1) a single partition with the 13 MPCGs concatenated; (2) two partitions, sites of first and second codons *versus* positions of third codon; (3) three partitions, by codon (first, second and third); (4) codon-based models that explicitly incorporate information on the genetic code (i.e., AA replacement rates) but have codons rather than nucleotides as states (Goldman and Yang, 1994); and finally (5) we applied 13 partitions, one for each protein-coding gene. We implemented the simplest codon model (omega=equal) since they are computationally expensive and they need long runs to converge (Shapiro et al., 2006). A preliminary test with six independent runs of one million

generations did not converge. We then set two independent runs of three million generations for 200 h on eight nodes (the maximum time allowed for the cluster), which converged in the last 200,000 generations. This was insufficient to guarantee that our result was not a local minimum.

Competing partitioning strategies were compared using Bayes factors, as they allow one to contrast non-nested models with a distinct number of parameters (Brown and Lemmon, 2007; Miller et al., 2009). Bayes factors are the ratio of the marginal likelihoods of two alternative hypotheses and are calculated as the difference between the natural logarithms of the harmonic means (Kass and Raftery, 1995). Marginal likelihoods and harmonic means were estimated in Tracer v1.4 using the Newton and Raftery approach with the modifications proposed by Suchard et al. (2001). A Bayes factor larger than 150 was considered decisive in favor of the tree with the higher likelihood score (Kass and Raftery, 1995). Note that Bayes factors estimated with marginal likelihoods seem to not penalize for overparameterization and hence favor highly dimensional models (Lartillot and Philippe, 2006). This could be because of the bias towards sampling high-likelihood regions of the harmonic mean estimator using MCMC, which results in the underestimation of the dimensional penalty. However, other studies suggest that Bayes factors do not select overparameterized models when the (admittedly subjective) cut off of 10 is used (Brown and Lemmon, 2007; Miller et al., 2009). To try to avoid overparameterization, we also implemented the PM factor suggested by Miller et al. (2009) ($PM = \Delta LnL/\Delta p$, where p = number of free parameters). This factor is based on the suggestion by Pagel and Meade (2004) that to accept an extra GTR matrix in a model the likelihood should be improved with at least 70–80 log-likelihood units, which would be equivalent to a PM factor of 10 or greater. The PM factor is identical to the double of the 'relative Bayes Factor' of Castoe et al. (2005) and hence the threshold suggested by Miller et al. (2009) would be equivalent to a 'relative Bayes Factor' of >20. We also estimate Akaike Information Critetion ($AIC = -2\log(L)+2k$, where k is the number of free parameters), and the increment of AIC ($\Delta AIC = AICi\_model - AIC\_bestmodel$), (Akaike, 1974; Posada and Buckley, 2004).

We did not test the model while splitting each individual gene by codon (39 partitions in total), as it would result in an overparameterization and an increase in stochastic error because of the small number of sites sampled, especially for the shorter genes (*atp8*, *nad4L* and *nad3*). Lemmon and Moriarty (2004) observed that overparameterization led to biased bipartition posterior probabilities that were more pronounced in short sequences.

In some analyses, we recoded nucleotides at first and third codons as purines (R) or pyrimidines (Y) (the only net transversion model) as it typically reduces composition bias, phylogenetic signal decay (saturation) and other artifacts associated with integrating a model incorporating a single rate of heterogeneity across multiple substitution types (Philips, 2009). Moreover, this model should improve parameter estimation and hence lead to more accurate branch lengths (Hassanin, 2006). Partitions recoded as R/Y were analyzed with a F81+I+G model (Felsenstein, 1981). Finally, we deleted third codon positions because generally such codon sites are highly saturated (Hassanin, 2006), with sites of the first codon recoded as R/Y.

Maximum likelihood analyses were performed using RAxML v7.0.4, implementing a fast bootstrapping algorithm (Stamatakis et al., 2005). We implemented a GTR model with CAT approximation to incorporate rate heterogeneity across sites, although the final likelihood value and branch lengths were optimized according to GTR+I+G (GTRMIXI model in RAxML). Analyses with AA sequences used an mtArt+I+G model. Finally, we tested the statistical significance of alternative topologies of the Coleopteran suborders at both DNA and protein levels using the Shimodaira–Hasegawa (SH; Shimodaira and Hasegawa, 1999; Goldman et al., 2000) and Expected-Likelihood Weight (ELW; Strimmer and Rambaut, 2002) tests with 500 bootstrap replicates as implemented in RAxML v7.0.4.

Estimation of rates of nucleotide evolution

We estimated the mean rate of nucleotide substitution using BEAST v1.4.7 (Drummond and Rambaut, 2007), enforcing a relaxed molecular clock with an uncorrelated log–normal distribution

and a Yule speciation model. Relaxed uncorrelated clock models assume independent rates on different branches as there is no *a priori* correlation between one particular lineage's rate and that of its ancestor. We enforced a fixed topology (excluding outgroups) in all BEAST analyses, allowing the optimization of all other parameters. The crown age of Coleoptera was set with a normal prior mean of 250 MY (Ponomarenko, 1969; Grimaldi and Engel, 2005) and a standard deviation of 25 MY. We did not attempt to estimate the ages for the different suborders, as the species included here represent just a small sample of beetle diversity and do not include some of the most divergent lineages within each of the suborders. The BEAST analyses were run for 20 million generations, sampling every 1000 generations, except for individual genes, which were run for 40 million generations. The output was analyzed using Tracer v1.4 after discarding the first 2–4 million generations. We also estimated rates of MPCGs using the semiparametric penalized likelihood in r8s (Sanderson, 2002). The optimal smoothing value was estimated by cross-validation, and we tested 32 smoothing values ranging from 1 to 57 millions. The branch lengths for the fixed topology were estimated in RAxML using an independent GTR+I+G model and nucleotide frequencies set for each codon partition. To estimate possible differences in evolutionary rates across suborders we implemented two to five local clocks in r8s.

Results

Mitochondrial genomes of Aspidytes niobe and Hydroscapha granulum

We obtained the complete mitochondrial genome of two species of the suborders Adephaga (*Aspidytes niobe*, family Aspidytidae) and Myxophaga (*Hydroscapha granulum*, family Hydroscaphidae), except for part of the control region of the former. The annotation of the two mitochondrial genomes (AM_493667 and AM_493668 for *Hydroscapha* and *Aspidytes*, respectively) is supplied as additional information (see Suppl. Material).

Nucleotide composition of the MPCGs

We analyzed MPCG sequences to detect any bias in the nucleotide composition across species, genes or codon positions, which could introduce phylogenetic artifacts (Hassanin, 2006; Sheffield et al., 2009). First, we analyzed MPCGs by codon sites because they generally show different nucleotide compositions (Beard et al., 1993; Sheffield et al., 2008, 2009). First codon positions had slightly higher frequencies of G (18.0 ± 1.4%) than C (12.3 ± 2.4%) and similar frequencies of A and T (34.4 ± 1.5% and 35.3 ± 2.5, respectively). In contrast, second codon positions were biased towards C (C 18.4 ± 0.9% and G 13.8 ± 0.5%) and T (A 20.4 ± 0.8% and T 47.4 ± 0.8%). Third codon positions showed similar frequencies of complementary nucleotides (G 5.3%, C 7.6%, A 42.6%, and T 44.5%) but were about 20% A+T richer (87.1 ± 8.0%) than first and second positions. The strong bias found on third positions was confirmed by the analysis of codon usage (RSCU and ENC) because A+T rich codons were used preferentially over other synonymous codons (e.g., UUU was ~10 times more frequent than UUC; AUU was three times more frequent than AUC and UCU, and UCA was six times more frequent than UCC and UCG). This bias was also reflected at the protein level, because sequences were mainly composed of AA coded by A+T rich codons: Phe, Ile, Leu and Ser (> 9% each) and Gly, Met and Asn (> 5% each). The six AA coded by G+C rich codons only accounted for about 10% of the total sites (Cys, Asp, Glu, His, Gln and Arg, < 2% each). Individual genes had a similar pattern irrespective of the coding strand (data not shown).

Next, we analyzed biases across species based on self-organizing clustering and analysis of heterogeneity. Nucleotide composition was homogeneous across the studied Coleoptera for all 13 MPCGs combined as well as for first and second codon positions. However, third codon positions of *T. bruchi*, *P. divergens* and *T. castaneum* were dissimilar from the other beetle sequences (lower A+T richness at 71.4%, 74.0% and 76.4%, respectively). Interestingly, these three species were also the most divergent in first and second positions, although the differences were not statistically significant. When the genes of these three species were analyzed individually, some had

significantly different nucleotide compositions: *atp6*, *cox2* and *nad1* (without codon partitioning); *atp8*, *nad2*, *nad4L* and *nad6* (only first codon positions) and *atp8* (only second codon positions). Composition across genes was most variable on third codon sites and this was explained statistically by three to five clusters, depending on the gene, with *T. bruchi*, *P. divergens*, *T. castaneum* and *Cyphon* sp. being the most dissimilar. When compared with the z-score test, only three species did not deviate at the nucleotide and four at the AA level (i.e., $z < 2$).

Finally, for the MPCGs we also estimated the relative skew between complementary nucleotides within the same strand. Skewness varied very little across species but was highly dependent on the coding strand (Fig. 1). MPCGs on the minus strand showed skew towards G, whereas those coded on the plus strand had the opposite skew towards C. All genes showed a skew towards T irrespective of the coding strand, but those on the minus strand showed higher negative values (stronger T skew) than those coded on the plus strand. This was because the first and second codon sites of genes on the plus strand generally had a slight skew towards A.

Phylogenetic analyses of Coleopteran suborders

The results of the Bayesian analyses of the 13 MPCGs implementing the five partition strategies (see Materials and Methods) were compared using Bayes factors (Table 1). The optimal model was found to be partition by codon position. This model was also preferred when using PM-factors (Table 1), which correct for differences in the harmonic means (Bayes factors) by accounting for the increase in the number of free parameters and thus avoid overparameterization. Similar results were found using ΔAIC which also takes into account for the increment of parameters. The tree resulting from the Bayesian analysis implementing the optimal model gave strong support (Bayesian posterior probabilities, BPP = 1.0) for the monophyly of Coleoptera and of the three suborders with more than one species. It also supported the sister group relationship between Adephaga and Myxophaga, with the family Scirtidae (*Cyphon*) being a sister group of the remaining Polyphaga, and the Cucujiformia and Elateriformia being monophyletic and sister

groups, respectively (Fig. 2). Archostemata was retrieved as a sister group to the Polyphaga (BPP = 1.00) despite most accepted hypotheses suggesting it to hold a basal position within the Coleoptera. To investigate further whether the position of Archostemata was driven by its different nucleotide composition (see above), we recoded first and third codon positions as purines and pyrimidines (R/Y), which reduced compositional biases and homoplasy (Hassanin, 2006). The R/Y recoding of first and third codon sites retrieved a similar overall topology, with three alternative positions for the Archostemata: as a sister group to the Polyphaga (BPP = 0.55), to the Coleoptera (BPP = 0.35) or to the (Adephaga + Myxophaga) (BPP = 0.10). Complete removal of the third codon position resulted in very similar topologies for the Archostemata (BPP = 0.51, 0.29 and 0.20 for the same three alternatives). The analysis of the AA sequences with the mtArt+G+I model selected by Protest (not shown) reduced the support of the sister relationship between Archostemata and the rest of Coleoptera (0.21 BPP), whereas the sister relationship to Adephaga had the highest posterior probability (0.74 BPP). Maximum likelihood analyses using RAxML v7.0.4 showed similar results to those retrieved by MrBayes at both the DNA and protein levels, with all nodes well supported except for the position of Archostemata (Fig. 2). Since the AA composition was not stationary (see above), we performed phylogenetic searches using the CAT (Poisson, GTR, and recoding with Dayhoff matrix) and CAT+BP models. Analyses gave alternative relationships among the four suborders, always with low supports, but most of them placing Archostemata as sister to the other Coleoptera (not shown).

Finally, we repeated the phylogenetic analyses omitting the outgroups, as their distant relationships with the Coleoptera could have produced long branch attractions within the ingroup (Rota-Stabellia and Telford, 2008). To speed up searches we reduced the number of species in Polyphaga to three (one Cucujiformia, *A. glabripennis*; one Elateriformia, *R. ohbai*, and *Cyphon*). All analyses with this reduced dataset of eight Coleopteran species resulted in topologies very similar to those found with the full dataset of 21 species (results not shown) although alternative topologies (Adephaga + Polyphaga or Myxophaga + Polyphaga) despite being rare had slightly

higher probabilities. For instance, when third codon sites were completely removed from the analysis, the probability of the Adephaga being a sister group to the Polyphaga increased to BPP = 0.23.

In all analyses and under all analytical conditions the most likely topology was of the Adephaga being a sister group to the Myxophaga. To further assess the support of this relationship we compared the three alternative topologies (leaving the rest of the nodes of the tree unchanged) with a SH test. The hypothesis of (Polyphaga + Myxophaga) *versus* the preferred (Adephaga + Myxophaga) was rejected by the SH test ($p < 0.05$) for both protein and nucleotide data (the latter analyzed as a single partition or by two or three independent codon partitions; see Materials and Methods). The grouping of (Polyphaga + Adephaga) was rejected at the DNA level but not at the AA level ($p > 0.05$). The ELW test strongly supported the hypothesis of (Adephaga + Myxophaga) with probabilities ranging from 0.94 to 1.0 at both protein and nucleotide levels, with the latter using different partition schemes. Alternative topologies had very small probabilities: (Adephaga + Poyphaga) gave p = 0.05–0.001 and for (Myxophaga + Polyphaga) p = 0.008–0. The SH tests were run using the reduced eight species dataset with similar results. The ELW test also showed high support for the preferred hypothesis of (Adephaga + Myxophaga) giving p = 0.99 at the DNA level using three codon partitions. The probabilities for the alternative hypothesis increased when third codon positions were removed or when the AA sequence was analyzed: (Adephaga + Polyphaga) gave p = 0.29 and p = 0.34 for DNA without third codon positions and AA, respectively, and (Myxophaga + Polyphaga) gave p = 0.11 and p = 0.08, respectively.

Estimation of nucleotide substitution rates

We estimated the rates of nucleotide substitution in MPCGs in beetles using the 15 available mitochondrial genomes and assuming the topology of the tree in Figure 2 as being fixed but with the Archostemata *T. bruchi* constrained as a sister group to the remaining Coleoptera, in agreement with previous morphological analyses (Ponomarenko, 1969; Beutel and Leschen, 2005; Grimaldi

and Engel, 2005; Friedrich et al., 2009). We kept Adephaga as sister to Myxophaga despite being against the currently most accepted hypothesis, as this was strongly supported in the different analyses. However, we also estimated the rates for the alternative relationships (Polyphaga + Myxophag or Polyphaga + Adephaga) to assess the impact of the topology on the rates (see below). The phylogeny of the Coleoptera based on the 13 MPCGs, partitioned by codon position (first, second, third) and with an independent GTR+I+G model and nucleotide frequencies for each partition, rejected a constant molecular clock based on the likelihood ratio test ($p < 0.01$). Hence, we used different algorithms to account for rate variations across trees: (1) a Bayesian relaxed clock with a uncorrelated log normal distribution in BEAST (Drummond et al., 2006); (2) a relaxed clock with semiparametric penalized likelihood in r8s (Sanderson, 2002) and (3) local clocks on different parts of the tree in r8s (Sanderson, 2002).

The overall rate using a relaxed clock with a log normal distribution in BEAST for the combined analysis of the 13 MPCGs partitioned by codon position was 0.0134 nucleotide substitutions per site per MY per lineage (subs/s/my/l), which is equivalent to 2.68% of pairwise sequence divergence (Table 2). Effective sample sizes for all parameters were greater than 100, indicating that the sampled generations were uncorrelated and that the posterior distributions of the parameters were long and accurate (Rambaut and Drummond, 2007). The 95% confidence interval for the rate values ranged from 0.0088 to 0.0189 subs/s/my/l, with a median of 0.0130. Analysis of the same dataset with two partitions (i.e., merging first and second codon sites) reduced the values slightly to 0.0111 subs/s/my/l. This rate was within the 95% confidence intervals of the value estimated in the three codon partition scheme, although it was significantly worse (Bayes factor of 374). For the analyses with penalized likelihood using r8s, we estimated the initial branch lengths under a maximum likelihood criterion in RAxML with an identical three codon partition strategy and models. The rates obtained were slightly lower (0.0090 subs/s/my/l), although this was also within the lower range of the 95% confidence intervals of the rate estimated using BEAST with three codon partitions.

In all phylogenetic analyses, the species within the Polyphaga and Archostemata had longer branches than did those of the Adephaga and Myxophaga for both DNA and protein data (see Fig. 2). In attempting to account for this difference, we implemented two local clocks in r8s: one for the Polyphaga and Archostemata and another for the Adephaga and Myxophaga. The estimated rate in the Polyphaga and Archostemata (0.0103 subs/s/my/l) was nearly twice as fast as in the Adephaga and Myxophaga (0.0053 subs/s/my/l). Results were similar when five local clocks were implemented: two in the Polyphaga (Cucujiformia and Elateriformia) and one each for the Adephaga, Myxophaga and Archostemata (Table 3). Faster rates in the Polyphaga and Archostemata were also observed in the BEAST analyses when partitioning the data according to the three codon positions (Table 3). To take into account possible alternative relationships among suborders, we also estimated rates constraining Polyphaga as a sister group to the Myxophaga, and Polyphaga as a sister group to the Adephaga. Both analyses also retrieved rates about twice as fast in the Polyphaga and Archostemata (0.0181–0.0192 subs/s/my/l) than in the Myxophaga (0.0097–0.0078 subs/s/my/l) or the Adephaga (0.0058–0.0067 subs/s/my/l). The overall rates for Coleoptera for the alternative relationships were also similar to that found for the best topology (Adephaga + Myxophaga) at 0.0134 subs/s/my/l: (Polyphaga + Adephaga) at 0.0141 subs/s/my/l and (Polyphaga + Myxophaga) 0.0136 subs/s/my/l.

Finally, we estimated the rate of each individual MPCG across the Coleoptera (Table 4). For those analyses, we merged first and second codon sites in a single partition because stochastic error is expected to be larger in short sequences (200–1800 bp; Felsenstein, 2004); particularly if the number of substitutions is low, as expected for such codon sites. Rates across genes varied up to 15 fold, ranging from 0.0861 to 0.0055 subs/s/my/l, as estimated with Bayesian probabilities in BEAST (Table 4). Generally, genes coded on the plus strand showed faster rates than those coded on the minus strand (Table 4), with the only exception being *nad6* (plus strand). The rates estimated with the semiparametric penalized likelihood in r8s were about 34% slower on average than the values obtained with BEAST, although they showed a similar trend (Table 4). When the estimation

was done without partitioning the gene by codon site, the rates were much lower, ranging from 0.002 to 0.005 subs/s/my/l except for *atp8* (0.0151 subs/s/my/l) and *nad3* (0.0074 subs/s/my/l). In this case, there were no apparent differences between genes coded on the plus or minus mtDNA strands.

Discussion

Phylogenetic reconstruction of Coleopteran suborders

Estimation of the evolutionary rates of MPCGs in Coleoptera requires an accurate reconstruction of the topology and the branch lengths of the relationships among the main lineages. Both in turn are highly dependent on the evolutionary models used in the phylogenetic analyses, especially when studying a reduced dataset of highly divergent sequences (Sullivan and Joyce, 2005; Sheffield et al., 2009). In our case, most of the nodes were stable and highly supported, regardless of the data source (nucleotide or AA), phylogenetic reconstruction method (Bayesian probabilities or ML fast algorithms), partition scheme (by gene or by codon position) or evolutionary model used (GTR+I+G, CAT or CAT-BP). The results emphasize the monophyly of Coleoptera and the monophyly of the three suborders for which there are more than one example (Myxophaga, Adephaga and Polyphaga). They also imply the sister group relationship between the Adephaga and Myxophaga, between the Scirtoidea (*Cyphons* sp.) and the rest of Polyphaga and the respective monophyly and sister group relationship between Elateriformia and Cucujiformia within the Polyphaga. However, the placement of the supposedly archaic Archostemata (Friedrich et al., 2009) remained ambiguous in our analyses. The low support for the phylogenetic position of this rarely collected suborder could arise from the significantly different nucleotide composition in *T. bruchi* with respect to all other Coleoptera taxa, as already found by Sheffield et al. (2008, 2009). In the latter study, the authors used three assumed relationships to assess the performance of different methods to account for nucleotide biases and nonstationarity: the monophyly of Cucujiformia and Elateriformia and the sister group relationship of Archostemata with the rest of Coleoptera

(Sheffield et al., 2009). In our case, some of these nodes were not recovered only when nonoptimal models were implemented (results not shown). Most of the relationships of the optimal trees (Fig. 2) conform to current knowledge of beetle phylogenetics (see Hunt et al., 2007 for a recent review), including the placement of species with biased nucleotide compositions such as *Pyrophorus divergens*, *Tribolium castaneum* and *Cyphon* sp. Of these, the first two species were considered to produce phylogenetic artifacts because of their nucleotide composition (Sheffield et al., 2009). The most important differences found in nucleotide composition were at the codon level. Therefore, the phylogenetic signal of MPCGs would be better recovered when first, second and third codon sites were split in three different partitions with an independent model of evolution (Shapiro et al., 2006), as previously implemented in other insect studies (Cameron et al., 2007; Fenn et al., 2008; Dowton et al., 2009).

Interestingly, the RY recoding or removal of third codon positions and some analyses of AA increased the probability of Archostemata being placed as a sister group to the other Coleoptera, as expected from morphological and fossil data (Friedrich et al., 2009), but the support levels remained low. This was probably because of reduced saturation and nucleotide composition biases (Cameron et al., 2007; Fenn et al., 2008; Dowton et al., 2009; Miller et al., 2009; Philips, 2009). However and despite the increased accuracy of the model, the removal or recoding of third codon sites had very limited impact on the tree topology as observed in Hymenoptera, Orthoptera or Diptera (Cameron et al., 2007; Fenn et al., 2008; Dowton et al., 2009). Hence, we suggest that most of the likelihood improvement arose from a more accurate estimate of the branch lengths.

The implementation of models taking into account nonstationary compositions (CAT Lartillot and Philippe, 2004; and CAT-BP Blanchart and Lartillot, 2008), did not resolve the ambiguous position of the Archostemata with strong support. This is a similar result to that obtained by Sheffield et al. (2009) with a more limited sampling (only suborders Polyphaga and Archostemata and without the inclusion of *Cyphon* sp.). The tree obtained with these models and those retrieved with analyses of the reduced dataset without outgroups which reduce the long

branch attraction caused by the compositional bias and the inclusion of distant outgorups shows a reduction of the support of Adephaga being sister to Myxophaga. These results suggest that the relationship between Adephaga and Myxophaga could be driven by long attraction effect caused by the low nucleotide substitution rates shown by both lineages. However, those heterogeneous models failed to retrieve a fully supported topology, indicating that the phylogenetic signal of the MPCG is hard to retrieve even for complex models of evolution.

The use of marginal likelihoods and harmonic means to estimate Bayes factors has been criticized for not penalizing overparameterization (Lartillot and Philippe, 2006). However, the use of PM-factors (Miller et al., 2009), which takes into account the number of parameters, did not change the selection of the models. This was in agreement with the findings of Brown and Lemmon (2007), suggesting that Bayes factors might not be particularly sensitive to overparameterization. Finally, the analysis of MPCGs using full codon models (those explicitly incorporating information on the genetic code as AA–codon replacement rates; Goldman and Young, 1994) was not found by Bayes and PM-factors to be better than the analysis with three codon partitions (first, second and third). This was contrary to the results of Shapiro et al. (2006) for viruses. It must be noted that full codon models are computationally very costly (Shapiro et al., 2006) and the runs did not reach a good degree of convergence even when analyzing small datasets of eight taxa with six independent runs for more than six million generations.

The most unexpected result of our phylogenies was the relationship between suborders (Adephaga + Myxophaga), which is against the two most accepted views of the Myxophaga being a sister group to the Polyphaga (Friedrich et al., 2009) and of the Adephaga being a sister group to the Polyphaga (Vogler, 2005 and see Introduction). However, the sister group relationship between the Adephaga and Myxophaga had been proposed by Ponomarenko (1969, 1973), based on the system of wing venation and the folding mechanism of some fossil groups, and by Forbes (1926) and Kukalová-Peck and Lawrence (1993, 2004), based on the same type of characters but studied in extant taxa. The same similarities were noted by Hammond (1979), who suggested the possibility

that they were related to the reduced size of the Myxophaga (as in the Clambidae, a miniaturized Polyphagan family). In any case, our results strongly contradict the possibility of a sister group relationship between the Myxophaga and Polyphaga, as suggested by morphology (Crowson, 1960; Klausnitzer, 1975; Baehr, 1979; Beutel, 1997; Beutel and Hass, 2000; Beutel et al., 2008; Friedrich et al., 2009) and by the analysis of some molecular datasets (Hughes et al., 2006). The second alternative tested (Adephaga + Polyphaga), despite not being statistically rejected by SH testing at the protein level, was rejected at the DNA level and had small probabilities based on the ELW test. Analyses without outgroups that could cause long branch attractions within the ingroup (Rota-Stabellia and Telford, 2008) retrieved similar results when the trees were rooted in the Archostemata, although the clade (Adephaga + Polyphaga) increased its probabilities in both the SH and ELW tests (up to $p = 0.23$). However, many of these analyses also found the Polyphaga to be paraphyletic, especially if the taxon sampling was reduced to eight, which clearly shows the importance of taxon sampling. It may thus be necessary to include more species of a wider diversity of extant lineages of the Archostemata, Myxophaga and Adephaga, together with a closer range of outgroups, to obtain well-resolved and supported relationships among the four beetle suborders.

The prevalent view of evolution of the Coleoptera is that three main lines were derived from an original stock with detritivorous or fungicolous and subcortical habits: the Archostemata, with wood-boring habits, the predatory Adephaga and the (Polyphaga + Myxophaga) group, which kept plesiomorphic habits in their stem lineage (Crowson, 1960). The alternative hypothesis, supported by molecular data (see above), requires a single shift in diversification rates at the base of the (Adephaga + Polyphaga) group. Unfortunately, the internal phylogenies of the Adephaga and Polyphaga are still largely unresolved (Beutel and Leschen, 2005; Hunt et al., 2007). Therefore, at present it is not possible to elaborate more detailed hypotheses on the origin of the two megadiverse lineages of Coleoptera, the family Carabidae in Adephaga (> 35,000 species, Arndt et al., 2005) or the Polyphaga with the exclusion of some species-poor clades (Scirtoidea and Derodontidae; Lawrence, 2001; Hunt et al., 2007; and this paper). However, our results open the possibility that

these two radiations are the product of independent colonizations of a fully terrestrial environment from a stem lineage strongly associated with aquatic or semiaquatic habits, as seen in the Myxophaga, and all Adephagan families except the Caraboidea and Scirtoidea.

Nucleotide substitution rates and ages

The nucleotide evolutionary rates estimated here for the combined 13 MCPGs, partitioning the data by codon position and using an estimated age of the Coleoptera of 250 MY, closely match the standard mitochondrial arthropod clock of 0.0115 subs/s/my/l reported by Brower (1994). A similar rate was also found in other studies comparing more closely related species of Coleoptera and using different combinations of mitochondrial genes, both ribosomal and protein-coding (e.g., Leys et al., 2003; Pons and Vogler, 2005; Pons et al., 2006; Balke et al., 2009; Ribera et al., 2010). These rates also closely match estimates for humans using complete mitochondrial genomes of 0.0126 subs/s/my/l (Mishamar et al., 2003) or 0.0166–0.0171 subs/s/my/l (Soares et al., 2009) and the 'standard mitochondrial clock rate' of 0.01 subs/s/my/l for vertebrates estimated from restriction fragment length polymorphisms (Brown et al., 1979; Wilson et al., 1985). However, we found strong rate variations depending on the partition scheme and evolutionary model used. Thus, the analysis of the data as a single partition with a single GTR+I+G model greatly reduced the rates using both BEAST (0.0022 subs/s/my/l) and r8s (0.0021 subs/s/my/l). These values were similar to the substitution rate estimated in r8s using branch lengths calculated in RAxML with the mtArt+I+G model (0.0030 subs/s/my/l). Our results show that differences in the estimated rate of nucleotide evolution arising from clock model selection (Bayesian relaxed clocks in BEAST or penalized likelihood in r8s) were much smaller (about 20%) than differences caused by partitioning strategy and branch length estimation (up to sixfold for unpartitioned data). Thus, model misspecification might have a larger impact on the rate estimations than differences in how to model rate change across the tree (see also Philips, 2009). Phylogenetic studies applying a predefined substitution rate for estimating branch lengths might obtain widely different results, depending on the partition scheme they apply (codon position or single partition), with a stronger

effect on older nodes (Buckley et al., 2001; Lemmon and Moriarty, 2004). When a partition by codon position was implemented, the rate estimates for third codon sites (0.0242 subs/s/my/l) were about 15–20 times faster than the rate estimates for first (0.0017 subs/s/my/l) and second (0.0008 subs/s/my/l) codon sites. However, when rates were estimated based on pairwise sequence divergence with a GTR model, third codon positions were only three to four times faster than second ones and about twice as fast as using first positions. This implies that model misspecification affects third more than first and second codon sites, as expected from the high degree of saturation of third codon sites in MPCGs.

In all phylogenetic analyses at both mtDNA and protein levels, the Adephagan and Myxophagan species showed shorter branches than did the Polyphagan and Archostematan species. More precisely, the Polyphaga and Archostemata had longer branches, as the Adephagan and Myxophagan branches were of similar lengths to those from the outgroups (Diptera and Lepidoptera). This translated into large differences in the rate estimates when local clocks were applied (Table 3), which is something to be taken into account when applying *a priori* rates for specific taxa. Possibly, these rate differences might have introduced some artifacts in the topologies obtained with mitochondrial genes (a ‘short branch attraction’; Philippe et al., 2005), as in the Hymenoptera (Dowton et al., 2009).

Analyses of the individual genes showed again the strong influence of model and partition choice in the estimations of evolutionary rates. As expected, mean rates estimated for first and second sites were more conserved across genes than those estimated for third codon sites, with genes coded on the plus mtDNA strand having the highest values except for *nad2* and *nad6* (Table 4). Similar differences were found in the analysis of the mitochondrial genomes of 48 vertebrates, with large rate variation across lineages and genes (Pereira and Baker, 2006), although rate values were more similar to those estimated without partitioning. Of special interest were the results obtained for the *cox1* gene, commonly used for species-level phylogenies in the Coleoptera (e.g., Hunt et al., 2007), as well as a universal ‘barcode’, or identification tag, for animals (e.g., Hebert et

al., 2003). When the estimation was applied using partitioning by codon position (merging first and second positions), the *cox1* gene showed the lowest rates for first and second codons but the fastest for third codon sites. The low rates for first and second codon positions were expected, because alignments at both mtDNA and protein level were very conserved, but the extremely high estimated rate for third codon positions (0.2566 subs/s/my/l) was unexpected. The overall rate (0.0861 subs/s/my/l) reflects this disproportionately fast rate (Table 4). In a study with closely related species of Adephaga, Pons et al. (2005) found slower rates for *cox1* (0.0167 *versus* 0.0861 subs/s/my/l) but similar rates to those estimated here for *cob* (0.0211 *versus* 0.0171 subs/site/my/l, Table 4). Ribera et al. (2010) also found slower *cox1* rates (0.02 subs/site/my/l) for a group of Polyphagan cave beetles in a time interval of ~40 MY, estimated with a single GTR+I+G model. A similar study performed on complete mitochondrial genomes of 27 salamander species, and using partitioning by codon sites showed a similar trend (Mueller, 2006). Cytochrome genes were faster rates than *nad* genes, and *cox1* had the fastest one that was nearly twice as fast than the second one. Mueller (2006) found that the slowest evolutionary rate of *cox1* at the amino acid level was coupled to the fastest rates at the nucleotide level (including all codon positions). He suggested that the relatively higher number of (mainly synonymous) substitutions should occur at the third codon positions of this gene, and our estimations demonstrate that the overall faster rate of *cox1* in beetles is due to the extremely high rates on those third codon sites. This pattern suggest functional constraints on the *cox1* sequence that cause its rates of synonymous substitution to be very sensitive to changes on both amino acid and nucleotide compositions (Mueller, 2006). These strong differences suggest that extreme caution should be exerted in the choice of evolutionary model and partition scheme when using *a priori* rates with individual genes. In particular, *cox1* might be a problematic gene to be used as a phylogenetic and/or molecular clock marker for deep level phylogenies. In contrast, *nad5*, *nad4* and *nad2* could be better markers, as they are long and have rates that are more similar across codon positions, therefore being less likely to be affected by methodological artifacts. The superiority of *nad4* and *nad5* to build deep level phylogenies of

vertebrates over the extensively used *cox1*, *cox2* and *cob* genes was already suggested elsewhere (Russo et al. 1996; Mueller 2006). The genes *atp6*, *atp8* and *nad6* could also be good candidates for estimating ages, but they are very short and thus offer limited information content. In general, our results suggest that the *atp* and *cox* genes had more evolutionary constraints at the protein level than did *nad* genes.

Authors' contributions

JP, IR and MB outlined the project, MB (*Aspidytes*) and JP (*Hydroscapha*) performed the sequencing, JP did the primary analyses of the sequences, JP, IR and MB performed the phylogenetic analyses and drafted the paper. All authors contributed to the discussion of results and conclusions and approved the final version of the paper.

Acknowledgments

We thank D. Bilton (Plymouth) for providing specimens of *Aspidytes niobe*, A. Stamatakis for helping with RAxML, J. Bergsten for helping with the estimation of BF and PM factors, S. Blanquart for providing the software nh_PhyloBayes, and the Centro de Supercomputación de Galicia for access to their computer clusters. We also thank J. Castresana, D. San Mauro, and an anonymous reviewer for their comments and suggestions that greatly improved the manuscript. Funding was provided by projects CGL2007-61665/BOS to IR; DFG 2152/3-1, 3-2 and European Union (EU) Commission SYNTHESYS ES-TAF 193 and 2197 to MB; and project CGL2006-01365 of the Spanish Ministry of Science and Innovation (MICINN) and EU Commission FEDER funds, Ramón y Cajal Fellowships from MICINN and EU Commission SYNTHESYS grant GB-TAF-4237 to JP.

References

Abascal, F., Posada, D., Zardoya, R., 2007. MtArt: a new model of amino acid replacement for Arthropoda. Mol. Biol. Evol. 24 (1), 1–5.

Abascal, F.,. Zardoya, R.,. Posada, D., 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21 (9), 2104–2105.

Akaike, H,. 1974. A new look at statistical model identification. IEEE Trans. Automatic Control 19(6), 716–723.

Arnd, E., Beutel, R.G., Will, K., 2005. Carabidae Latreille 1802. In: Beutel, Leschen (Eds). Handbook of Zoology Vol IV. Arthropoda, Insecta. Part 38, Coleoptera, Volume Vol 1. Morphology and Systematics (Archostemata Adephaga Myxophaga Polyphaga (partim). De Gruyter, Berlin and New York. pp. 119–146.

Baehr, M., 1979. Vergleichende Untersuchungen am Skelett und an der Coxalmuskulatur des Prothorax der Coleoptera Ein Beitrag zur Klärung der phylogenetischen Beziehungen der Adephaga (Coleoptera Insecta). Zoologica 44 (4), 1–76.

Balke, M., Ribera, I., Miller, M., Hendrich, L., Sagata, K., Posman, A., Vogler, A.P., Meier, R., 2009. New Guinea highland origin of a widespread arthropod supertramp. Proc. Roy. Soc. B 276 (1666), 2359–2367.

Beard, C.B., Hamm, D.M., Collins, F.H., 1993. The mitochondrial genome of the mosquito *Anopheles gambiae* DNA sequence genome organization and comparisons with mitochondrial sequences of other insects. Insect. Mol. Biol. 2 (2), 103–124.

Beutel, R.G., 1997. Über Phylogenese und Evolution der Coleoptera (Insecta) insbesondere der Adephaga. Verh. Naturwiss. Ver. Hamburg (NS) 31, 1–164.

Beutel, R.G., 2005. Systematic position basal branching pattern and early evolution. In: Beutel, Leschen (Eds). Handbook of Zoology Vol IV. Arthropoda, Insecta. Part 38, Coleoptera, Volume Vol 1. Morphology and Systematics (Archostemata Adephaga Myxophaga Polyphaga (partim). De Gruyter, Berlin and New York. pp. 1–9.

Beutel, R.G., Haas, F., 2000. Phylogenetic relationships of the suborders of Coleoptera (Insecta). Cladistics 16 (1), 103–141.

Beutel, R.G., Leschen, R.A.B., 2005. Handbook of Zoology Vol IV. Arthropoda, Insecta. Part 38, Coleoptera, Vol 1. Morphology and Systematics (Archostemata Adephaga Myxophaga Polyphaga (partim). Berlin and New York, De Gruyter. 453 p.

Beutel, R.G., Pohl, H., 2006. Endopterygote systematics – where do we stand and what is the goal (Hexapoda Arthropoda)? Syst. Entomol. 31 (2), 202–219.

Beutel, R.G., Hörnschemeyer, T., Ge, S.Q., 2008. On the head morphology of *Tetraphalerus* the phylogeny of Archostemata and the basal branching events in Coleoptera. Cladistics 24 (3), 270–98.

Blanquart, S., Lartillot, N., 2008. A site- and time-heterogeneous model of amino acid replacement. Mol. Biol. Evol. 25 (5), 842-858.

Brower, A.V.Z., 1994. 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 (14), 6491–6495.

Brown, J.M., Lemmon, A.R., 2007. The importance of data partitioning and the utility of Bayes factors in Bayesian phylogenetics. Syst. Bio. 56 (4), 643–655.

Brown, W.M, George, M., Wilson, A.C., 1979. Rapid evolution of animal mitochondrial DNA. Proc. Natl. Acad. Sci. USA 76 (4), 1967–1971.

Buckley, T.R., Simon, C., Chambers, G.K., 2001. Exploring among–site rate variation models in maximum likelihood framework using empirical data. Effects of model assumptions on estimates of topology branch lengths and bootstrap support. Syst. Bio. 50 (1), 67–86.

Cameron, S.L., Lambkin, C.L., Barker, S.C., Whiting, M.F., 2007. Utility of mitochondrial genomes as phylogenetic markers for insect intraordinal relationships. A case study from flies (Diptera). Syst. Entomol. 32 (1), 40–59.

Castoe, T.A., Sasa, M.M., Parkinson, C.L., 2005. Modeling nucleotide evolution at the mesoscale,

The phylogeny of the Neotropical pitviper of the Porthidium group (Viperidae: Crotalidae). Mol. Phylogenet. Evol. 37 (3), 881–898.

Caterino, M.S., Shull, V.L., Hammond, P.M., Vogler, A.P., 2002. The basal phylogeny of the Coleoptera based on 18S rDNA sequences. Zool. Scripta 31 (1), 41–49.

Crowson, R.A., 1960. The phylogeny of Coleoptera. Annu. Rev. Entomol. 5, 111–134.

Crowson, R.A., 1981. The biology of Coleoptera. Academic Press, London. 802 p.

Dowton, M., Cameron, S.L., Austin, A.D., Whiting, M.F., 2009. Phylogenetic approaches for the analysis of mitochondrial genome sequences data in the Hymenoptera–A lineage with both rapidly and slowly evolving mitochondrial genomes. Mol. Phyl. Evol. 52 (2), 512–519.

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 (5), e88.

Farrell, B.D., 1998. "Inordinate Fondness" explained why are there so many beetles? Science 281 (5376), 555–559.

Felsenstein, J., 1981. Evolutionary trees from DNA sequences, a maximum likelihood approach. J. Mol. Evol. 17 (6), 368–376.

Felsenstein J., 2004. Inferring Phylogenies. Sunderland, Sinauer Associates Inc. 664 p.

Fenn, J.D., Song, H., Cameron, S.L., Whiting, M.F., 2008. A preliminary mitochondrial genome phylogeny of Orthoptera (Insecta) and approaches to maximizing phylogenetic signal found within mitochondrial genome data. Mol. Phylogenet. Evol. 49 (1), 59–68.

Forbes, W.T.M., 1926. The wing folding patterns of the Coleoptera. J. New York Entomol. Soc. 34, 42–139.

Friedrich, F., Farell, B.D., Beutel, R.G., 2009. The thoracic morphology of Archostemata and the relationships of the extant suborders of Coleoptera (Hexapoda). Cladistics 25 (1), 1–37.

Goldman, N., Yang, Z., 1994. A codon-based model of nucleotide substitution for protein coding

DNA sequences. Mol. Biol. Evol. 11 (5), 725–736.

Goldman, N., Anderson, J.P., Rodrigo, A.G., 2000. Likelihood–based tests of topologies in phylogenetics. Syst. Biol. 49 (4), 652–670.

Grimaldi, D., Engel, M.S., 2005. Evolution of the Insects. Cambridge University Press, Cambridge. 755 p.

Hammond, P.M., 1994. Practical approaches to the estimation of the extent of biodiversity in speciose groups. Phil. Trans. R. Soc. B 345 (1311), 119–136.

Hammond, P.M., 1979. Wing–folding mechanisms of beetles with special reference to investigations of Adephagan phylogeny (Coleoptera) In: Erwin, Ball, Whitehead (Eds). Carabid beetles their evolution natural history and classification. Junk Publishers, The Hague. pp. 113–180.

Hassanin, A., 2006. Phylogeny of Arthropoda inferred from mitochondrial sequences strategies for limiting the misleading effects of multiple changes in pattern and rates of substitution. Mol. Phylogenet. Evol. 38 (1), 100–116.

Hassanin, A., Léger, N., Deutch, J., 2005. Evidence for multiple reversals of assymmetric mutational constraints during the evolution of the mitochondrial genome of Metazoa and consequences for phylogenetic inferences. Syst. Biol. 54 (2), 277–298.

Hebert, P.D.N., Cywinska, A., Ball, S.L., DeWaard, J.R., 2003. Biological identifications through DNA barcodes. Proc. R. Soc. B 270 (1512), 313–321.

Huelsenbeck, J.P., Ronquist, F., 2001. MrBAYES Bayesian inference of phylogenetic trees. Bioinformatics 17 (8), 754–755.

Huelsenbeck, J.P., Suchard, M.A., 2007. A nonparametric method for accommodating and testing across-site rate variation. Syst. Biol. 56 (6), 975–987.

Hughes, J., Longhorn, S.J., Papadopoulou, A., Theodorides, K., de Riva, A., Mejia–Chang, M., Foster, P.G., Vogler, A.P., 2006. Dense taxonomic EST sampling and its applications for molecular systematics of the Coleoptera (beetles). Mol. Biol. Evol. 23 (2), 268–278.

Hunt, T., Bergsten, J., Levkanicova, Z., Papadopoulou, A., John, O.S., 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 superradiation. Science 318 (5858), 1913–1916.

Hutchinson, G.E., 1959. Homage to Santa Rosalia or why are there so many kind of animals? Am. Nat. 93 (870), 145–159.

Jermiin, L.S., Ho, S.Y., Ababneh, F., Robinson, J., Larkum, A.W., 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. Syst. Biol. 53 (4), 638–644.

Kass, R.E., Raftery, A.E., 1995. Bayes Factors. J. Am. Stat. Assoc. 90 (430), 773–795.

Katoh, K., Kuma, K., Toh, H., Miyata, T., 2005. MAFFT version 5 improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33 (2), 511–518.

Klausnitzer, B., 1975. Probleme der Abgrenzung von Unterordnungen bei den Coleoptera. Entomol. Abh. Staatl. Mus. Tierk Dresden 40, 269–275.

Kukalová–Peck, J., Lawrence, J.F., 1993. Evolution of the hind wing in Coleoptera. Can. Entomol. 125 (2), 181–258.

Kukalová–Peck, J., Lawrence, J.F., 2004. Relationships among Coleoptera suborders and major Endomeopteran lineages evidence from hind wing characters. Eur. J. Entomol. 101 (1), 95–144.

Lartillot, N., Philippe, H., 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21 (6), 1095–1109.

Lartillot, N., Philippe, H., 2006. Computing bayes factors using thermodynamic integration. Syst. Biol. 55 (2), 195–207.

Lartillot, N., Lepage, T., Blanquart, S., 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25 (17), 2286–2288.

Lawrence, J.F., 2001. A new genus of Valdivian Scirtidae (Coleoptera) with comments on Scirtoidea

and the beetle suborders. Spec. Publ. Japan Coleopt. Soc. Osaka 1, 351–361.

Lemmon, A.R., Moriarty, E.C., 2004. The importance of proper model assumption in Bayesian phylogenetics. Syst. Biol. 53 (2), 265–277.

Leys, R., Watts, C.H., Cooper, S.J., Humphreys, W.F., 2003. Evolution of subterranean diving beetles (Coleoptera, Dytiscidae, Hydroporini Bidessini) in the arid zone of Australia. Evolution 57 (12), 2819–2834.

Lobry, J.R., 1995. Properties of a general model of DNA evolution under no-strand-bias conditions. J. Mol. Evol. 40 (3), 326–330.

Masta, S.E., Longhorn, S.J., Boore, J.L., 2009. Arachnid relationships based on mitochondrial genomes, Asymmetric nucleotide and amino acid bias affects phylogenetic analyses. Mol. Phylogenet. Evol. 50 (1), 117–128.

Meyer, A., Zardoya, R., 2003. Recent advances in the (molecular) phylogeny of vertebrates. Annu. Rev. Ecol. Syst. 34, 311–338.

Miller, K.B., Bergsten, J., Whiting, M.F., 2009. Phylogeny and classification of the tribe Hydaticini (Coleoptera, Dytiscidae), partition choice for Bayesian analysis with multiple nuclear and mitochondrial protein-coding genes. Zool. Scripta 38 (6), 591-615.

Mishmar, D., Ruiz–Pesini, E., Golik, P., Macaulay, V., Clark, A.G., Hosseini, S., Brandon, M., Easley, K., Chen, E., Brown, M.D. et al., 2003. Natural selection shaped regional mtDNA variation in humans. Proc. Natl. Acad. Sci. USA 100 (1), 171–176.

Mueller, R.L., 2006. Evolutionary rates, divergence dates, and the performance of mitochondrial genes in Bayesian phylogenetic analysis. Syst. Biol. 55(2), 289–300.

Murataa ,Y., Nikaidoa, M., Sasakia, T., Caob, Y., Fukumotoc, Y., Hasegawab, M., Okada, N., 2003. Afrotherian phylogeny as inferred from complete mitochondrial genomes. Mol. Phylogenet. Evol. 28 (2), 253–260.

Nardi F., Spinsanti, G., Boore, J.L., Carapelli, A., Dallai, R., Frati, F., 2003. Hexapod origins monophyletic or paraphyletic? Science 299 (5614), 1887–1889.

Ødegaard, F., 2000. How many species of arthropods? Erwin's estimate revised. Biol. J. Linn. Soc. 71 (4), 583–597.

Pagel, M., Meade, A., 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character state data. Syst. Biol. 53 (4), 571–581.

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. Mol. Biol. Evol. 23 (9), 1731–1740.

Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N., Delsuc, F., 2005. Heterotachy and long-branch attraction in phylogenetics. BMC Evol. Biol. 5 (1), 50.

Phillips, M.J., 2009. Branch-length estimation bias misleads molecular dating for a vertebrate mitochondrial phylogeny. Gene 441 (1-2), 132–140.

Ponomarenko, A.G., 1969. Historical development of Archostemata beetles. Trudy Paleontologicheskogo Instituta AN SSSR 125, 70–115.

Ponomarenko, A.G., 1973. The nomenclature of wing venation in beetles (Coleoptera). Entomol. Rev. 51, 454–458.

Ponomarenko, A.G., 1995. The geological history of beetles. In: Pakaluk , Slipinski (Eds). Biology Phylogeny and classification of Coleoptera, Papers celebrating the 80th birthday of Roy A Crowson. Museum I Instytut Zoologii PAN, Warszawa. pp. 155–171.

Pons, J., Vogler, A.P., 2005. Complex pattern of coalescence and fast evolution of a mitochondrial rRNA pseudogene in a recent radiation of tiger beetles. Mol. Biol. Evol. 22 (4), 991–1000.

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. Syst. Biol. 55 (4), 595–609.

Posada, D., 2008. jModelTest Phylogenetic Model Averaging. Mol. Biol. Evol. 25 (7), 1253–1256.

Posada, D., Buckley, T.R., 2004. Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood

Ratio Tests. Syst. Biol. 53(5), 793–808.

Rambaut, A., Drummond, A.J., 2007. Tracer v1.4 Available from http, //beastbioedacuk/Tracer

Ribera, I., Hogan, J.E., Vogler, A.P., 2002. Phylogeny of hydradephagan water beetles inferred from 18S rRNA sequences. Mol. Phylogenet. Evol. 23 (1), 43–62.

Ribera, I., Fresneda, J., Bucur, R., Izquierdo, A., Vogler, A.P., Salgado, J.M., Cieslak, A. 2010. Ancient origin of a Western Mediterranean radiation of subterranean beetles. BMC Evol. Biol. 10, 29.

Roger, J.A., Hug, L.A., 2006. The origin and diversification of eukaryotes, problems with molecular phylogenetics and molecular clock estimation. Phil. Trans. R. Soc. B 361 (1470), 1039–1054.

Rota-Stabellia, O., Telford, M.J., 2008.A multi criterion approach for the selection of optimal outgroups in phylogeny, Recovering some support for Mandibulata over Myriochelata using mitogenomics. Mol. Phylogenet. Evol. 48 (1), 103–111.

Russo, C.A.M., Takezaki, N., Nei, M., 1996. Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. Mol. Biol. Evol. 13 (3), 525–536.

Sanderson, MJ., 2002. Estimating absolute rates of molecular evolution and divergence times a penalized likelihood approach. Mol. Biol. Evol. 19 (1), 101–109.

Shapiro, B., Rambaut, A., Drummond, A.J., 2006. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Mol. Biol. Evol. 23 (1), 7-9.

Sheffield, N.C., Song, H., Cameron, S.L., Whiting, M.F., 2008. A comparative analysis of mitochondrial genomes in Coleoptera (Arthropoda, Insecta) and genome descriptions of six new beetles. Mol. Biol. Evol. 25 (11), 2499–2509.

Sheffield, N., Song, H., Cameron, S.L., Whiting, M.F. 2009. Nonstationary evolution and compositional heterogeneity in beetle mitochondrial phylogenomics. Syst. Biol. 58 (4), 381–394.

Shimodaira, H., Hasegawa, M., 1999. Multiple comparisons of log–likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16 (8), 1114–1116.

Shull, V.L., Vogler, A.P., Baker, M.D., Maddison, D.R., Hammond, P.M., 2001. Sequence alignment of 18S ribosomal RNA and the basal relationships of Adephagan beetles evidence for monophyly of aquatic families and the placement of Trachypachidae. Syst. Biol. 50 (6), 945–969.

Soares, P., Ermini, L.,Thomson, N., Mormina, M., Rito, T., Röhl, A., Salas, A., Oppenheimer, S., Macaulay, V., Richards, M.B., 2009. Correcting for purifying selection an improved human mitochondrial molecular clock. Am. J. Hum. Genet. 84 (6), 740–759.

Stamatakis, A., Ludwig, T., Meier, H., 2005. Raxml–iii A fast program for maximum likelihood–based inference of large phylogenetic trees. Bioinformatics 21 (4), 456–463.

Strimmer, K., Rambaut, A., 2002. Inferring confidence sets of possibly misspecified gene trees. Proc. Roy. Soc. B 269 (1487), 137–142.

Suchard, M.A., Weiss, R.E., Sinsheimer, J.S., 2001. Bayesian selection of continuous-time markov chain evolutionary models. Mol. Biol. Evol. 18 (6), 1001–1013.

Sullivan, J., Joyce, P., 2005. Model selection in phylogenetics. Annu. Rev. Ecol. Evol. Syst. 36, 445–466.

Supek, F., Vlahovicek, K., 2004. INCA synonymous codon usage analysis and clustering by means of self–organizing map. Bioinformatics 20 (14), 2329–2330.

Tamura, K., Dudley, J., Nei, M., Kumar, S., 2007. MEGA4, Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24 (8), 1596–1599.

Vogler, A.P., 2005. Molecular systematics of Coleoptera. What has been achieved so far? In: Beutel, Leschen (Eds). Handbook of Zoology Vol IV. Arthropoda, Insecta. Part 38, Coleoptera, Volume Vol 1. Morphology and Systematics (Archostemata Adephaga Myxophaga Polyphaga (partim). De Gruyter, Berlin and New York. pp. 17–22.

Wiegmann, B.M., Trautwein, M.D., Kim, J.W., Cassel, B.K., Bertone, M.A., Winterton, S.L.,

Yeates, D.K., 2009. Single-copy nuclear genes resolve the phylogeny of the holometabolous insects. BMC Biol. 7, 34.

Wilgenbusch, J.C., Warren, D.L., Swofford, D.L., 2004. AWTY: A system for graphical exploration of MCMC convergence in Bayesian phylogenetic inference. http, //ceb.csit.fsu.edu/awty.

Wilson, A.C., Cann, R.L., Carr, S.M., George, M., Gyllensten, U.B., Helm-Bychowski, K.M., Higuchi, R.G., Palumbi, S.R., Prager, E.M., Sage, R.D., Stoneking, M., 1985. Mitochondrial DNA and two perspectives on evolutionary genetics. Biol. J. Linn. Soc. 26 (4), 375–400.

Zardoya, R., Meyer, A., 1996. Phylogenetic performance of mitochondrial protein–coding genes in resolving relationships among vertebrates. Mol. Biol. Evol. 13 (7), 933–942.

Legends

Figure 1. Plots of AT skew (x axis) *versus* GC skew (y axis) in mitochondrial protein coding genes (MPCGs) among the Coleoptera, estimated across species (triangles) and genes (dots). Color indicates that skew was estimated using all codon sites (black) and first (dark gray), second (light gray) or third (white) codon positions only. Note that MPCGs coded on the minus mtDNA strand are all enclosed in the top left panel. Relative nucleotide skews between intrastrand complementary nucleotides were calculated as follows: AT skew = (A − T)/(A + T) and GC skew = (G − C)/(G + C). Positive values indicate a skew towards purines (G or A) and negative numbers shown a skew towards pyrimidines (C or T).

Figure 2. Phylogram showing relationships among the four Coleoptera suborders estimated in MrBayes using the nucleotide sequences of the 13 MPCGs. Numbers on nodes indicate Bayesian posterior probabilities (right) and bootstrap support estimated after 1000 replicates in RAxML (left). Asterisks indicate that a node was not retrieved in the RAxML analysis. Nucleotide sequences were partitioned by codon sites (first, second and third) and analyzed using an independent GTR+I+G model and nucleotide frequencies.

Supplementary material:

Supplementary text. Annotation of the mitochondrial genomes of the *Hydroscapha granulum* Motschulsky, 1855 (Myxophaga, Hydroscaphidae) and *Aspidytes niobe* Ribera et al., 2002 (Adephaga, Aspidytidae).

Supplementary Table 1. Primers list. Universal primers used to amplify short mitochondrial chunks, which allowed the design of more specific primers to amplify longer fragments.

Supplementary Figure 1. Gene order of the mitochondrial genome of *Hydroscapha granulum* and *Aspydites niobe*. Genes highlighted on gray are coded on the minus strand, and those without color on the plus strand. Note that the secondary structure of the putative origin of replication within the control region is shown for *H. granulum*. Numbers indicate the nucleotide position in the complete mitochondrial DNA. The region highlighted in black in *A. niobe* could be not obtained. Single letters indicates each particular tRNA gene.

Supplementary Figure 2. Secondary structure of the 22 tRNAs found in the mitochondrial genome of *H. granulum*.

Supplementary Figure 3. Secondary structure of the 19 tRNAs found in the mitochondrial genome of *A. niobe*.

Figure 1

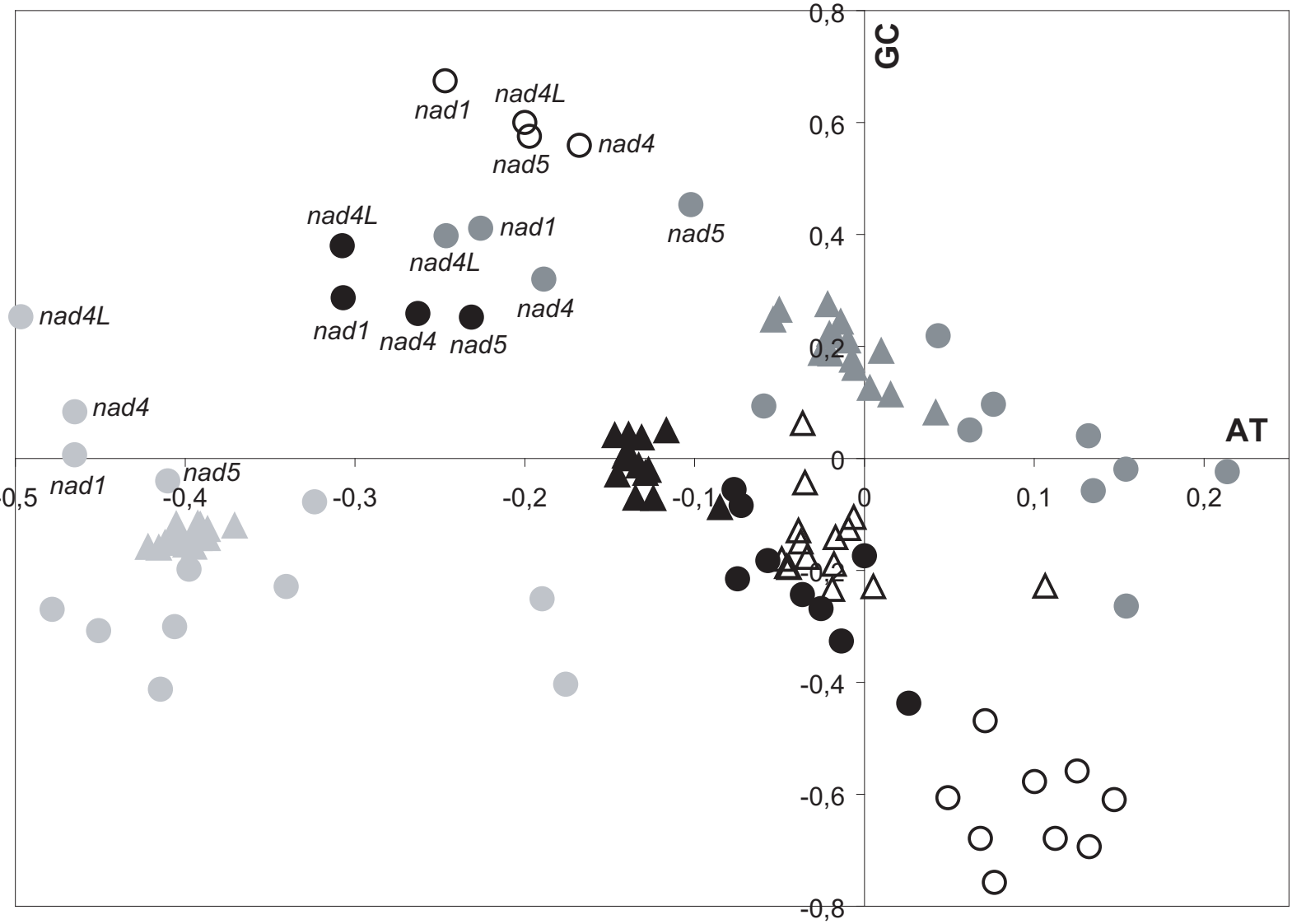
GC
AT
0,8
0,6
0,4
0,2
0
-0,2
-0,4
-0,6
-0,8
-0,5
-0,4
-0,3
-0,2
-0,1
0
0,1
0,2
nad1
nad4L
nad5
nad4
nad4L
nad1
nad5
nad4L
nad4
nad1
nad4
nad5
nad4L
nad4
nad1
nad5

Figure 2

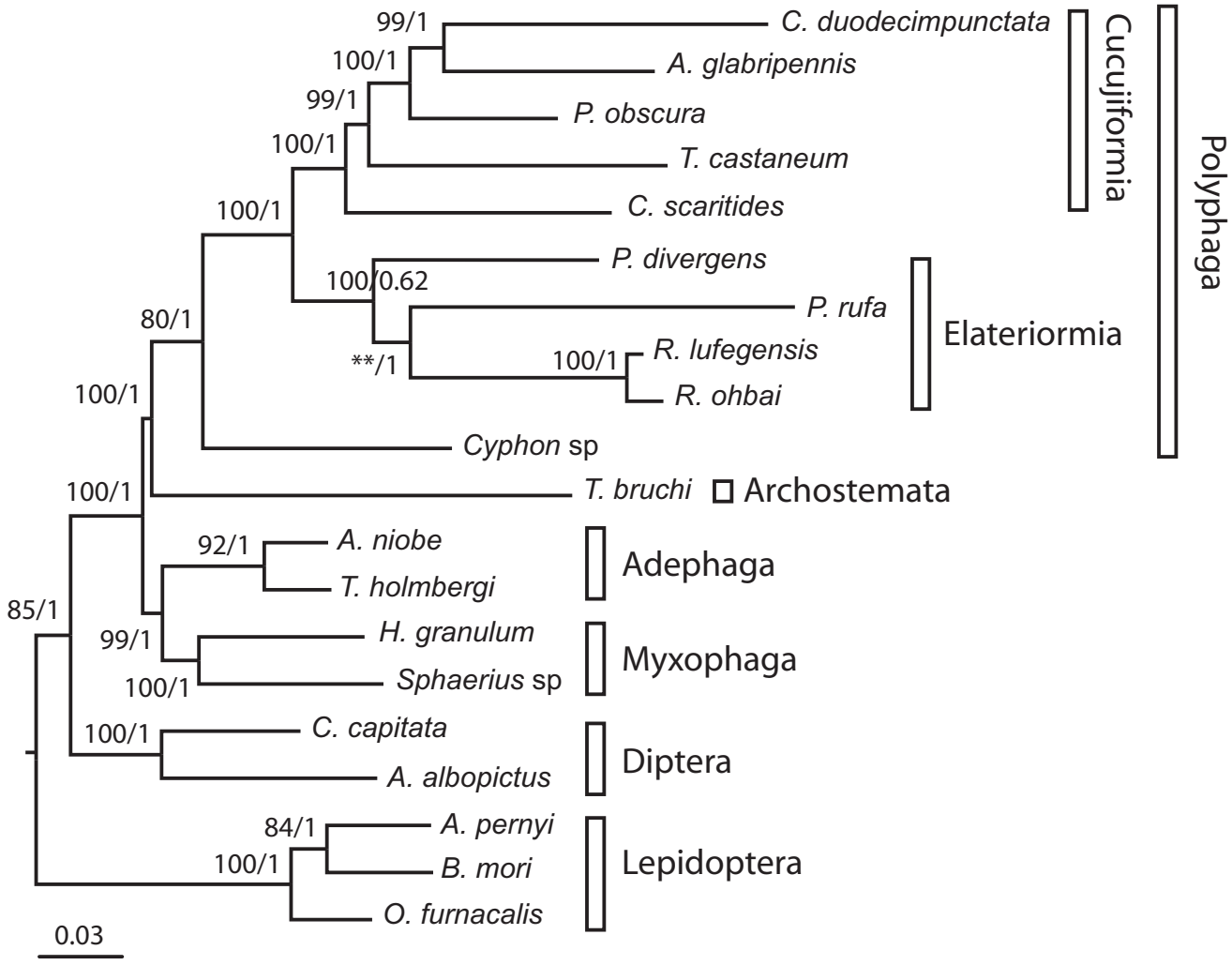
99/1
C. duodecimpunctata
100/1
A. glabripennis
99/1
P. obscura
100/1
T. castaneum
100/1
C. scaritides
P. divergens
100/0.62
P. rufa
80/1
**/1
R. lufegensis
100/1
R. ohbai
100/1
Cyphon sp
100/1
T. bruchi
92/1
A. niobe
T. holmbergi
85/1
99/1
H. granulum
100/1
Sphaerius sp
100/1
C. capitata
A. albopictus
84/1
A. pernyi
100/1
B. mori
O. furnacalis
0.03
Cucujiformia
Elateriformia
Polyphaga
Archostemata
Adephaga
Myxophaga
Diptera
Lepidoptera

Table 1. The five partitioning strategies used in the analyses, with the number of partitions (n), likelihood score of harmonic mean, and Bayes factors (upper diagonal) and PM factor (lower diagonal, Miller et al. 2009) in pairwise comparisons. Bayes factors (BFs) were calculated as (LnL H1 – LnL H2) and PM factor as ΔLnL/Δp (where p = the number of free parameters). BF > 150 and PM factor > 10 were considered to give very strong support in favoring the tree with the higher likelihood score. Last two columns indicate the values for Akaike Information Criterion (AIC) and increment of AIC (ΔAIC), see Material and methods.

	n	Free parameters	Harmonic mean (Ln)	H1	H2	H3	H4	H5	AIC	ΔAIC
H1 single	1	13	-144931.68	-	-3434.90	**-4520.47**	684.79	-2404.47	289841.32	9035.62
H2 (1^{st} +2^{nd}) : 3^{rd} codon	2	27	-141496.78	245.35	-	**-1085.56**	4119.69	-1030.43	282955.80	2150.20
H3 1^{st} : 2^{nd} : 3^{rd} codon	3	40	**-140411.21**	**167.42**	**83.51**	**-**	**5205.26**	**2116.00**	**280805.70**	**0**
H4 codon (omega equal)	1	70	-145616.47	-12.01	-95.81	**-173.51**		-3089.26	291234.40	10428.70
H5 13 genes	13	169	-142527.21	15.41	-7.26	**-16.40**	31.20	-	285279.88	4447.18

Table 2. Rates of nucleotide substitution per site per million years per lineage ($\times 10^{-2}$ subs/s/my/l) estimated on the 13 mitochondrial protein coding genes of 15 Coleopteran species. Rates were estimated in BEAST using a relaxed clock with log normal distribution with a fixed topology (see results). The age of the Coleoptera was set to 250 million years (MY), allowing a standard error of 25 MY. The table also includes the mean and the standard deviation of the mean of the branch rates (ucld). The last column shows the mean rates calculated using penalized likelihood in the r8s software.

	mean rate	standard deviation	rate median	lower rate	upper rate	-Ln likelihood	ucld	ucld stdev	mean rate r8s
by codon (1,2,3)	1.342	0.017	1.300	0.881	1.890	105120	1.562	0.493	0.899
1st	0.175	0.001	0.171	0.124	0.233	33960	0.202	0.521	0.183
2nd	0.085	0.001	0.083	0.058	0.117	22400	0.108	0.607	0.095
3rd	2.420	0.007	2.370	1.748	3.224	48630	2.517	0.204	1.799
by codon (1+2,3)	1.115	0.014	1.089	0.747	1.523	105965	1.261	0.465	0.879
single	0.222	0.003	0.216	0.158	0.301	108979	0.245	0.455	0.210

Table 3. Rates of nucleotide substitution ($\times 10^{-2}$ subs/s/my/l) estimated on the 13 mitochondrial protein coding genes at the suborder and superfamily level. Rates were estimated in BEAST using a relaxed clock with log normal distribution with a fixed topology (see results). The age of the Coleoptera was set to 250 MY, allowing a standard error of 25 MY. The last column shows the mean rates calculated using different local clocks in r8s.

Taxa	Mean rate in BEAST	Number of Local Clocks and rates (r8s)
		Two clock
Polyphaga + Archostemata	1.51	1.03
Adephaga+Myxophaga	0.99	0.53
		Five clock
Archostemata	1.01	0.90
Adephaga	0.97	0.45
Myxophaga	1.01	0.53
Cucujiformia	1.70	1.17
Elateriformia	1.78	1.26

Table 4

Table 4. Rates of nucleotide substitution ($\times 10^{-2}$ subs/s/my/l) for each mitochondrial protein coding gene, estimated across 15 Coleoptera species. Rates were estimated in BEAST using a relaxed clock with log normal distribution with a fixed topology (see results). The age of the Coleoptera was set to 250 MY, allowing a standard error of 25 MY. For the analyses of the individual genes, first and second codon sites were merged in a single partition. The last column shows the mean rates calculated using penalized likelihood in the r8s software. Mu, mutation rate; ucld, mean of the branch rates.

gene	strand	length (bp)	mean rate	standard deviation	median rate	lower rate	upper rate	-Ln likelihood	ucld	ucld stdev	mu 1st+2nd	mean rate 1st+2nd	mu 3rd	mean rate 3rd	rate r8s
atp6	+	687	2.552	0.168	1.953	0.550	6.290	6087	2.763	0.471	0.076	0.193	2.848	7.268	1.003
atp8	+	168	4.179	0.174	2.821	0.578	11.200	1769	7.497	1.109	0.384	1.605	2.234	9.336	2.845
cob	+	1152	1.715	0.042	1.558	0.707	3.111	9772	2.150	0.576	0.060	0.105	2.880	5.040	1.837
cox1	+	1545	8.606	0.338	7.578	2.509	17.600	1156	10.600	0.556	0.009	0.079	2.982	25.663	6.311
cox2	+	687	2.610	0.083	2.253	0.684	5.325	6070	3.285	0.613	0.053	0.138	2.894	7.553	3.147
cox3	+	792	5.499	0.316	4.433	1.290	12.800	6782	6.587	0.565	0.023	0.125	2.954	16.244	3.383
nad1	-	1044	1.281	0.025	1.177	0.581	2.234	8442	1.380	0.405	0.096	0.123	2.808	3.597	1.392
nad2	+	1053	1.248	0.031	1.141	0.579	2.188	11730	1.522	0.588	0.202	0.252	2.596	3.240	0.876
nad3	+	372	3.079	0.087	2.503	0.624	7.092	3464	3.970	0.685	0.080	0.245	2.840	8.719	2.252
nad4	-	1377	0.879	0.008	0.831	0.436	1.422	12800	1.024	0.629	0.190	0.167	2.620	2.303	1.118
nad4L	-	297	0.555	0.016	0.476	0.218	1.055	2785	0.611	0.389	0.314	0.174	2.371	1.316	n/a
nad5	-	1767	1.678	0.036	1.583	0.857	2.790	16120	1.923	0.564	0.100	0.167	2.800	4.698	1.449
nad6	+	537	0.918	0.012	0.833	0.401	1.650	5858	1.121	0.697	0.360	0.330	2.280	2.093	0.561
all		11478	1.115	0.014	1.089	0.747	1.523	105965	1.261	0.465	0.107	0.119	2.786	3.106	0.879