

# Family-Level Sampling of Mitochondrial Genomes in Coleoptera: Compositional Heterogeneity and Phylogenetics

Martijn J. T. N. Timmermans<sup>1,2,3,\*</sup>, Christopher Barton<sup>1</sup>, Julien Haran<sup>1,4</sup>, Dirk Ahrens<sup>1,5</sup>, C. Lorna Culverwell<sup>1,6</sup>, Alison Ollikainen<sup>1,7</sup>, Steven Dodsworth<sup>1,8</sup>, Peter G. Foster<sup>1</sup>, Ladislav Bocak<sup>1,9</sup>, and Alfried P. Vogler<sup>1,2</sup>

<sup>1</sup>Department of Life Sciences, Natural History Museum, London, United Kingdom

<sup>2</sup>Department of Life Sciences, Imperial College London - Silwood Park Campus, Ascot, United Kingdom

<sup>3</sup>Department of Natural Sciences, Middlesex University, Hendon Campus, London, United Kingdom

<sup>4</sup>Present address: INRA, UR633 Zoologie Forestière, Orléans, France

<sup>5</sup>Zoologisches Forschungsmuseum Alexander Koenig Bonn, Bonn, Germany

<sup>6</sup>Present address: Haartman Institute, Haartmaninkatu 3, University of Helsinki, Helsinki, Finland

<sup>7</sup>Present address: Department of Medical Genetics, Genome-Scale Biology Research Program, University of Helsinki, Helsinki, Finland

<sup>8</sup>Present address: Department of Comparative Plant and Fungal Biology, Royal Botanic Gardens, Kew, Richmond, Surrey, United Kingdom

<sup>9</sup>Department of Zoology, Faculty of Science UP, Olomouc, Czech Republic

\*Corresponding author: E-mail: m.timmermans@mdx.ac.uk.

Accepted: November 28, 2015

Data deposition: This project has been deposited in NCBI ([www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)) under accession numbers provided in [Supplementary Table S1](#).

## Abstract

Mitochondrial genomes are readily sequenced with recent technology and thus evolutionary lineages can be densely sampled. This permits better phylogenetic estimates and assessment of potential biases resulting from heterogeneity in nucleotide composition and rate of change. We gathered 245 mitochondrial sequences for the Coleoptera representing all 4 suborders, 15 superfamilies of Polyphaga, and altogether 97 families, including 159 newly sequenced full or partial mitogenomes. Compositional heterogeneity greatly affected 3rd codon positions, and to a lesser extent the 1st and 2nd positions, even after RY coding. Heterogeneity also affected the encoded protein sequence, in particular in the *nad2*, *nad4*, *nad5*, and *nad6* genes. Credible tree topologies were obtained with the nhPhyML (“nonhomogeneous”) algorithm implementing a model for branch-specific equilibrium frequencies. Likelihood searches using RAxML were improved by data partitioning by gene and codon position. Finally, the PhyloBayes software, which allows different substitution processes for amino acid replacement at various sites, produced a tree that best matched known higher level taxa and defined basal relationships in Coleoptera. After rooting with Neuropterida outgroups, suborder relationships were resolved as (Polyphaga (Myxophaga (Archostemata + Adephaga))). The infraorder relationships in Polyphaga were (Scirtiformia (Elateriformia ((Staphyliniformia + Scarabaeiformia) (Bostrichiformia (Cucujiformia)))). Polyphagan superfamilies were recovered as monophyla except Staphylinioidea (paraphyletic for Scarabaeiformia) and Cucujoidea, which can no longer be considered a valid taxon. The study shows that, although compositional heterogeneity is not universal, it cannot be eliminated for some mitochondrial genes, but dense taxon sampling and the use of appropriate Bayesian analyses can still produce robust phylogenetic trees.

**Key words:** mitogenomes, long-range PCR, rogue taxa, RY coding, mixture models, PhyloBayes.

## Introduction

Mitochondrial genomes have often been perceived as unreliable phylogenetic markers due to poor recovery of the expected relationships, in particular in early studies that were compromised by sparse taxon sampling (Bernt et al.

2013; Simon and Hadrys 2013). In insects, high rates of nucleotide change in mitochondrial genomes, together with high adenine-thymine (AT) content and constraints of protein function, limit the type of character variation and result in high levels of homoplasy (Talavera and Vila 2011). As rates of

change and nucleotide composition vary among lineages, mitogenome sequences are exposed to long-branch attraction, which confounds phylogenetic inferences. This phenomenon has received particular attention in studies of Coleoptera (beetles) showing that compositional heterogeneity is pervasive (Sheffield et al. 2009; Pons et al. 2010; Song et al. 2010; Bernt et al. 2013; Cameron 2014). However, although various likelihood models of DNA evolution assume stationarity, that is, an evolutionary process that keeps the character state distribution uniform across lineages, recent nonhomogeneous models accommodate changes in composition over the tree (Galtier and Gouy 1998; Foster 2004; Boussau and Gouy 2006; Foster et al. 2009).

An alternative approach for accommodating complex character variation is the site-heterogeneous CAT model implemented in PhyloBayes (Lartillot et al. 2009), which infers an infinite number of substitution processes (classes) from the empirical data, each of which are defined by different equilibrium frequencies of nucleotides or amino acids. This “heterogeneous mixture model” is widely used for the analysis of protein sequences, and was shown to reduce the susceptibility to long-branch attraction (Lartillot et al. 2007; Talavera and Vila 2011; Li et al. 2015). When applied to the Coleoptera, the use of PhyloBayes greatly improved the tree to match expected taxonomic groups over other models applied to the nucleotide sequences. For example, in the analysis of Timmermans et al. (2010) the single representative of the suborder Archostemata (genus *Tetraphalerus*) was placed incorrectly in a derived position within the suborder Polyphaga under various coding schemes and optimality criteria, as also observed in other studies (Pons et al. 2010; Song et al. 2010), but under the CAT model it was placed correctly outside of Polyphaga. Likewise, the CAT model was more successful than other approaches in recovering the major clades including the infraorders (“series”) within the Polyphaga (Timmermans et al. 2010). To some extent the effect of these mixture models can be achieved by partitioning the data according to a priori determined character sets and applying an independent GTR model, which can be implemented using the RAxML likelihood method (Stamatakis 2006).

The misleading signal from compositional heterogeneity is not produced by all nucleotides in equal measure, as rates are constrained in 1st and 2nd codon positions, which prevents rapid divergence in base composition (Song et al. 2010; Talavera and Vila 2011). Many previous studies therefore excluded 3rd codon positions from the analysis to reduce the effects of compositional heterogeneity. In addition, purine-pyrimidine (RY) coding can be used, which removes the AT versus GC compositional information in the assessment of character variation (Hassanin 2006). Finally, compositional heterogeneity has sometimes been shown to be concentrated in particular portions of the mitochondrial genome or in particular species or subclades, and hence data exclusion has been recommended, for example, omitting individual genes

that produce trees in conflict with the topology obtained from the full data (Talavera and Vila 2011). However, the link between topological incongruence among data partitions and compositional heterogeneity has not been widely explored. In Coleoptera, substitution rates are well known to differ among mitochondrial genes (Vogler et al. 2005; Pons et al. 2010), but the level of compositional heterogeneity has not been compared among genes.

With the application of high-throughput sequencing techniques, the number of mitochondrial genomes available for these analyses is increasing rapidly. The resulting denser taxon sampling may improve the estimation of molecular rates and variation in base composition, and thus result in improvements in estimates of tree topology, in particular through reduced long-branch attraction of convergent character variation. Here we generate a large set of mitochondrial genomes for the Coleoptera to test if the known problems for phylogenetic inference in this group previously ascribed to compositional heterogeneity can be overcome by denser taxon sampling. We also examine if high compositional heterogeneity affecting some terminals weakens the recovery of monophyletic groups and produce erroneous relationships. Not all such groups are expected to be strongly supported, but instead the effect of compositional heterogeneity may mainly reduce levels of support for otherwise well founded groups, and as their placement is ill-defined by the data they may appear as nuisance “rogue taxa” weakening an otherwise well supported topology (Wilkinson 1996). Their removal may reduce the compositional heterogeneity across the data and improve the overall tree topology.

We thus examine the evidence for compositional heterogeneity within and among genes, and test its impact on the topology. However, measuring compositional heterogeneity itself is challenging. A chi-square test (implemented in PAUP; Swofford 2002) has been widely used to assess if nucleotide composition in a data matrix is homogeneous, but this test suffers from a high probability of Type II error (the null hypothesis of homogeneity is false but fails to be rejected) because it does not assume phylogenetic relatedness (Kumar and Gadagkar 2001). As the effects of common ancestry are integral to the test quantity, they should be part of the null distribution as well. Such a null hypothesis can be generated by simulating data on the tree topology and model parameters of the empirical data, and the heterogeneity in the empirical data is then assessed against this distribution from simulations, again using the chi-square as a test quantity (Foster 2004). This approach is used here to address how compositional heterogeneity in different partitions of the mitogenome data matrix (e.g., various genes, codon positions, clades) affects the accuracy of the tree. We also examine whether these biases can be overcome by analyses of the translated protein sequences and by removal of certain data partitions or divergent lineages, including potential rogue taxa. We show that densely sampled mitogenomes can provide a well-supported

tree for the Coleoptera, even under moderate levels of compositional heterogeneity, and these relationships are best captured by the mixture models in PhyloBayes. The new tree consolidates the phylogenetic conclusions from previous studies and resolves several questionable nodes defining coleopteran superfamily and family-level relationships.

## Materials and Methods

### Sampling and Laboratory Procedures

Mitogenome sequences were generated from long-range PCR amplicons using the Roche/454 sequencing platform. Specimens were selected for uniform coverage of major lineages of Coleoptera from existing DNA extractions of various age and quality of preservation (Hunt and Vogler 2008; Bocak et al. 2014), in addition to newly collected specimens, resulting in highly variable PCR success that limited the taxon choice (supplementary table S1, Supplementary Material online). Amplification primarily targeted a large *cob* to *cox1* fragment of ~10 kb. The remainder of the mitogenome was amplified using primer sites in the *cox1* and *cob* genes, to include the rRNA genes and the control region, but amplification success was lower (supplementary table S2, Supplementary Material online). Primers used are described in Timmermans et al. (2010).

Sequence reads were assembled using the MIRA or Newbler software as described previously (Timmermans et al. 2010; Haran et al. 2013) and the longest contig obtained with either assembler was retained. tRNA genes were annotated with COVE using beetle-specific covariance models (Timmermans and Vogler 2012). Protein-coding gene sequences were annotated using existing Coleoptera mitochondrial genomes as reference in Geneious (<http://www.geneious.com/>, last accessed December 17, 2015). For the rRNA genes, sequences were extracted from the newly generated and previously published mitogenome sequences, using BLAST searches on a *fasta* formatted database with methods described in Bocak et al. (2014). The taxonomic classification, voucher ID, GenBank accession numbers, and geographic origin for each specimen are given in supplementary table S1, Supplementary Material online.

### Phylogenetic Inference

The 13 protein-coding genes were aligned with ClustalW using the transAlign wrapper (Bininda-Emonds 2005). The *cox1* gene was split into the 5' "barcode" region (Hebert et al. 2003) and the 3' region widely used in Coleoptera systematics usually amplified with the Pat and Jerry primers (Simon et al. 1994). This was to account for the fact that the two PCR fragments with different amplification success are confined to the 5' or 3' ends for the short and long fragment, respectively. The two rRNA genes were aligned using MAFFT v. 7 (Katoh et al. 2009) under default parameters on the server <http://mafft.cbrc.jp/alignment/software/>, last accessed December 17, 2015. Protein-coding alignments were edited, trimmed, and

translated with Mesquite v. 2.75 (Maddison WP and Maddison DR 2014). The final concatenated matrix consisted of the 13 protein-coding genes (14 regions taking into account the split *cox1* gene) and 2 rRNA genes, with a minimum of 9 protein-coding genes represented in all taxa. All tree searches and analyses of evolutionary patterns were done without further outgroups, except for one case of a PhyloBayes analysis designed to test the basal branching order in the light of non-Coleoptera outgroups. Mutational saturation was assessed in Dambe5, using a simulation-based analysis of the critical substitution saturation beyond which the correct tree is unlikely to be recovered (Xia 2013).

Different partitioning strategies were compared for the nucleotide data matrix of protein-coding genes, by calculating likelihood scores on a fixed topology generated in RAXML (Stamatakis 2006). Twelve partitioning schemes for the protein-coding genes were compared, ranging from unpartitioned to a maximum of 42 partitions (by gene + codon position). Likelihood scores were compared with reference to the complexity of the partitioning schemes using the Akaike Information Criterion (AIC). Bayes Factors and Relative Bayes Factors (RBF) were calculated according to Castoe et al. (2005).

Phylogenetic trees were generated using ML and Bayesian methods for partitioned and unpartitioned data sets. All RAXML trees were generated at the CIPRES webserver, (Miller et al. 2010; <https://www.phylo.org/>, last accessed December 17, 2015) under the GTRCAT model of nucleotide substitution, which approximates a GTR +  $\Gamma$  model with a reduced computational cost (Stamatakis 2006). Where relevant, node support was assessed using a rapid bootstrap algorithm implemented in RAXML with 500 replicates.

PhyML (Galtier and Gouy 1998; Guindon and Gascuel 2003) was run on the ATGC webserver (Guindon et al. 2010; [www.atgc-montpellier.fr](http://www.atgc-montpellier.fr), last accessed December 17, 2015) and used a GTR substitution model using eight rate categories. The gamma shape parameter and the proportion of invariable sites were estimated from the data. To infer relationships under the nonhomogeneous model of Galtier and Gouy (1998), nhPhyML (Boussau and Gouy 2006) was used, again using eight rate categories. Topology, gamma shape parameter, and transition/transversion rates were evaluated, but no final optimization of parameters such as branch lengths was performed (setting: -quick = y). As starting tree for tree searches in PhyML and nhPhyML, we used the RAXML tree of the complete, partitioned data set rooted on the Archostemata. Both analyses used the Nearest Neighbor Interchange algorithm.

Finally, the translated data matrix was subjected to Bayesian analysis with PhyloBayes 3 under the CAT-Poisson model (Lartillot et al. 2009). Two Markov chain Monte Carlo chains were run after the removal of constant sites from the alignment. This Bayesian analysis was repeated with outgroups included. These outgroups were from three orders of Neuropterida, the presumed sister lineage of Coleoptera, and were obtained from GenBank (Accession

numbers: NC\_011277, NC\_011278, NC\_013257, NC\_015093, NC\_021415, NC\_023362, NC\_024825, NC\_024826). PhyloBayes tree searches were also conducted on the CIPRES webserver.

The R package “ape” (Paradis et al. 2004) was used to obtain root-to-tip branch lengths from the RY-coded ML and the Bayesian amino acid trees. Mean values and standard deviations of branch lengths were calculated for each suborder and each of the polyphagan subfamilies.

### Compositional Heterogeneity

Compositional heterogeneity in data matrices based on the protein-coding genes was assessed as described in Foster (2004), using the chi-square statistic. Significance was assessed using a null distribution generated by simulations on the ML tree with branch lengths and  $\alpha$  value ( $\alpha$  of the  $\Gamma$  distribution) optimized. If the procedure is performed on the entire matrix, this presumes that there is no among-partition rate variation and that branch lengths for all partitions are the same. Since we used a homogeneous model, these values form a valid null distribution by which to assess the chi-square of the original data. RY-coded partitions were analyzed as DNA with RAXML. For simulations of protein sequences, the null distribution for assessing chi-square was generated using simulations on the corresponding ML tree and the MtArt +  $\Gamma$  model (Abascal et al. 2007). Missing taxa will not contribute to the calculated chi-square value for the original data, and therefore the chi-square calculations were done without the taxa affected by missing data for a given locus. Assessment of significance was based on tail area probabilities  $P_t$ , and a value of 0.05 or less was taken to show compositional heterogeneity. We also used the conventional chi-square test of compositional heterogeneity for comparison. The analysis of heterogeneity was conducted on the ingroup sequences only.

### Identification of Rogue Taxa

The RogueNaRok algorithm (Aberer et al. 2013) was used to identify rogue taxa (Wilkinson 1996), that is, those taxa that, if excluded from the tree searches, yield a pruned consensus tree with increased support values. Using an RAXML tree on RY-coded data (see Results), two settings were tested, allowing either one taxon (run #1) or two taxa to be pruned simultaneously (run #2). The change of support values was assessed on the tree obtained from the ML tree. To handle the effect of interaction between long branches we ran an analysis with a maximum dropset size of 3.

## Results

### Mitochondrial Genomes of Coleoptera

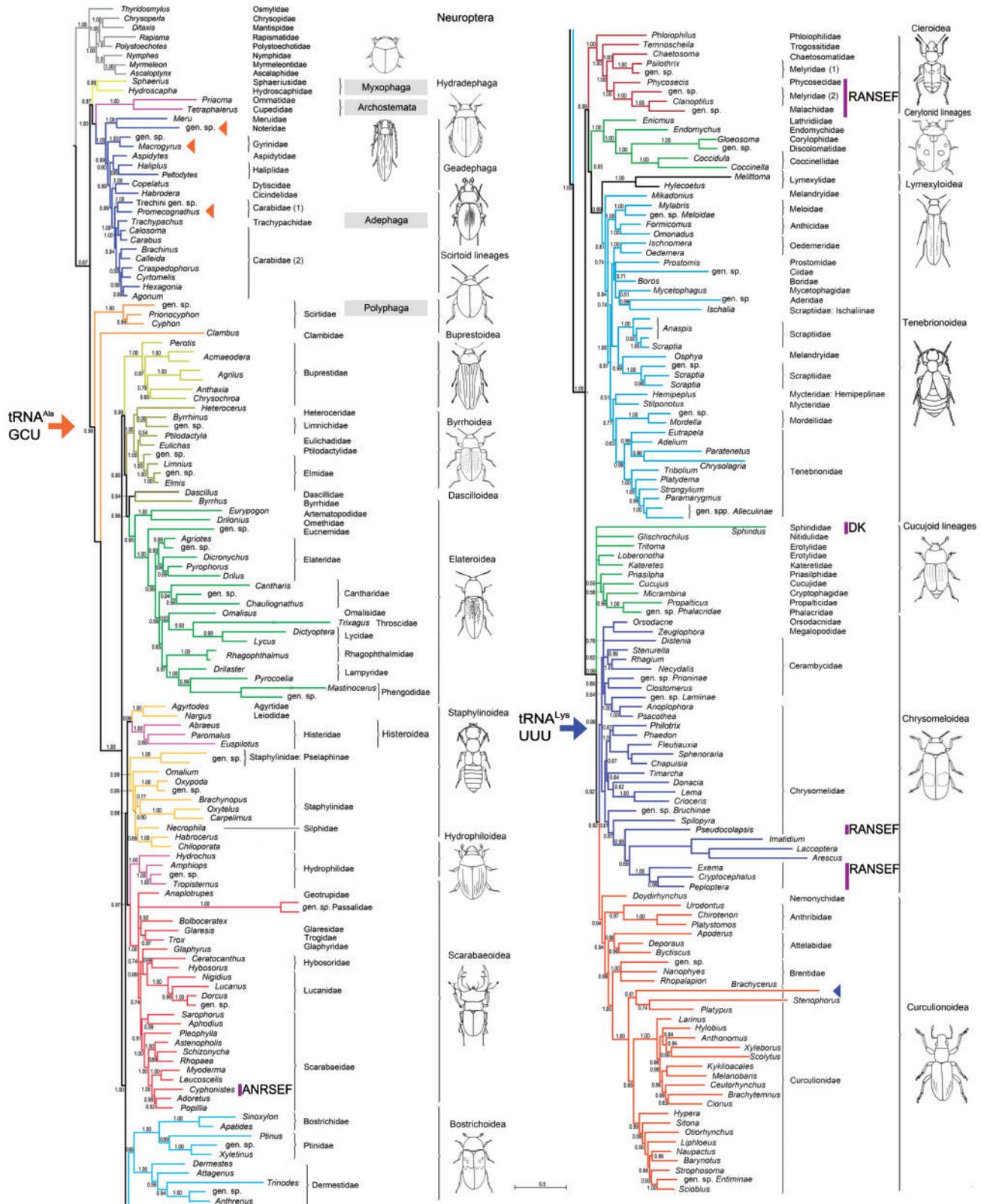
Full or partial mitochondrial genomes were newly generated for 159 taxa by sequencing LR-PCR fragments. In addition, 86 partial or full mitogenomes from previously published sources

were incorporated for a combined data set of 245 terminals. The small PCR fragment was represented by fewer taxa, and thus *nad2*, *cox1-5'*, and the 12S and 16S rRNA (*rns* and *rnl*) genes were missing for 148, 142, 169, and 139, respectively, while the remaining set was nearly complete for all taxa (supplementary table S2, Supplementary Material online), and 51 taxa were represented by the complete set of genes. All terminals had a minimum of 9 protein-coding fragments (of 14 fragments in total, including 2 parts of *cox1*) and the average data completion was 13.1 fragments, with a total sequence length of 6,202–11,717 bp. The aligned supermatrix consisted of 11,141 characters for protein-coding genes, and 12,271 characters when the 2 rRNA genes were included. The two supermatrices contained 15.27% and 20.29% missing data, respectively. The sampling covered all 4 suborders of Coleoptera, 15 superfamilies of Polyphaga (only leaving out the Derodontoidea for which no sequences were available), and a total of 97 families.

We found several gene order rearrangements in addition to those already described by Timmermans and Vogler (2012), which mainly affected the ARNSEF (Ala, Arg, Asn, Ser, Glu, Phe) cluster between the *nad3* and *nad5* genes. Three species of Chrysomelidae (*Exema*, *Cryptocephalus*, and *Pseudocolapsis*) had the order of tRNA<sup>Ala</sup> and tRNA<sup>Arg</sup> reversed (RANSEF). This state had previously been observed in *Peplotera* (Timmermans and Vogler 2012), which was placed together with the other three suggesting a single origin of this gene order but the tree topology suggests this group to be paraphyletic for *Imatidium*, *Laccoptera*, and *Arescus* which apparently reverted to the ancestral state. In addition, the RANSEF gene order was also observed in a subclade of the distantly related meloid lineage (Cleroidea), represented by four species, while it was also previously reported from *Naupactus* (Curculionidae) (Song et al. 2010) and other weevil species (Haran et al. 2013; Gillett et al. 2014). A further rearrangement of this tRNA cluster was seen in *Cyphonistes* (Scarabaeidae: Dynastinae) (ANRSEF). This represents a new state not previously observed in Coleoptera. Finally, the order of the genes for tRNA<sup>Lys</sup> and tRNA<sup>Asp</sup> (KD) located between the *cox2* and *atp6* loci was reversed (DK) in *Sphindus* (Sphindidae). In addition to these various rearrangements, we observed two anticodon changes, including a GCG to GCU change in the tRNA<sup>Ala</sup> anticodon, present in all Polyphaga, and a change from CUU to UUU of the tRNA<sup>Lys</sup> anticodon, present in all Chrysomeloidea and also two species of Curculionoidea (only one of them represented in the tree) (fig. 1).

### Model Testing

Partitioning greatly improved the likelihood scores. The model testing under the AIC identified the most complex partitioning scheme (partitioning by gene and codon) as the most favorable, with highly significant Bayes Factors against all other



**FIG. 1.**—The tree of Coleoptera based on protein-coding genes obtained with PhyloBayes. Major groups at the level of superfamily and above are labeled, and each superfamily is illustrated with a representative line drawing. Numbers on the branches represent posterior probabilities. Changes in anticodons of tRNA<sup>Lys</sup> (in Chrysomeloidea and in taxon labeled with blue triangle) and tRNA<sup>Ala</sup> (in Polyphaga and taxa labeled with orange triangles) and several newly discovered gene order changes are mapped on the tree.

Downloaded from https://academic.oup.com/gbe/article/8/1/161/12574155 by guest on 13 January 2022

**Table 1**  
Likelihood and AIC Values under Various Partitioning Schemes

Partitioning	No. of Partitions	Parameters ( <i>k</i> )	ln( <i>L</i> )	AIC	ΔAIC	2 × ln ΔBF	RBF
None	1	9	−1,279,328.877	2,558,675.754	105,496.41	21.76	0.059
Forward/Reverse	2	18	−1,258,902.112	2,517,840.225	64,660.88	20.79	0.058
Homogeneous/Heterogeneous	2	18	−1,273,139.835	2,546,315.669	93,136.33	21.51	0.060
Gene	14	126	−1,256,482.92	2,513,217.84	60,038.51	20.64	0.082
Codon 1+2+3	3	27	−1,251,864.871	2,503,783.742	50,604.41	20.30	0.058
Codon 1+2+3+Forward/Reverse	6	54	−1,229,360.303	2,458,828.606	5,649.26	16.11	0.050
Gene × codon	42	378	−1,226,211.669	2,453,179.339	n/a	n/a	n/a

NOTE.—The likelihood of the data under each partitioning scheme was assessed on the fixed topology of a randomized parsimony tree under a GTR+G model, with the number of partitions, free parameters, and ln(*L*) scores used in the calculations given. ΔAIC refers to the decrease in likelihood relative to the most complex model (partitioning by gene and codon). Values for  $2 \times \ln \Delta BF_{10} > 10$  are usually considered to be highly significant. RBF was calculated according to Castoe et al. (2005) as  $2 \times \ln \Delta BF_{10} / \Delta$  parameters, to penalize greater model complexity.

**Table 2**  
Compositional Heterogeneity in Mitogenomes

	<i>n</i> missing	Conventional Chi-square				Foster (2004)					
		1st	2 <sup>nd</sup>	1st RY	1st RY	2nd	1st two-state	No rogue			Protein
								1st RY	2nd	1st two-state	
atp6	22	0.0999	1	1	1	1	0.2	1	1	0.21	1
atp8	1	1	1	1	1	0.36	0.53	1	0.24	0.51	0.02
cox1–5'	142	1	1	1	1	1	1	1	1	1	0.85
cox1–3'	43	1	1	1	1	1	0.95	1	1	0.99	1
cox2	1	1	1	1	1	1	1	1	1	1	1
cox3	2	1	1	1	1	1	0.82	1	1	0.85	0.98
Cytb	1	0.006	1	1	1	0.99	0.94	1	0.93	0.99	0.03
nad1	8	1	1	1	1	1	0.34	1	0.98	0.42	0.79
nad2	148	0	0.981	1	1	0	0	0.5	0	0	0
nad3	2	1	1	1	1	0.22	0.12	1	0.14	0.22	0.45
nad4	5	0	1	1	1	0.01	0.01	1	0	0	0
nad4L	5	1	1	1	1	0.96	0.95	1	0.96	0.99	0.73
nad5	5	0	1	1	1	0	0	1	0	0	0
nad6	5	0	1	1	1	0	0	1	0.01	0	0

NOTE.—Each gene was tested for the probability that the data are homogeneous and *P* values are provided in the table, separately for 1st and 2nd codon positions. Significance of the chi-square statistic was assessed either with the chi-square curve (“conventional chi-square”) or using a null distribution as described in Foster (2004). Note that four loci generally have low probability of homogeneity throughout. *n* missing, mitogenomes in the matrix not sequenced for a locus; no rogue, analysis conducted with rogue taxa omitted; protein, analysis based on amino acid sequence.

partitioning schemes (table 1). However, various partitioning schemes contributed in different ways. Based on the ΔAIC, partitioning by forward and reverse strand resulted in a major improvement over the unpartitioned model, and this could be improved only slightly by further partitioning by genes. Separating the genes according to those genes most strongly affected by compositional heterogeneity (see below) had little impact on the AIC score. In contrast, partitioning by codon positions had a strong effect, and this was further improved by partitioning according to coding on the forward and reverse strands, that is, using six partitions. The likelihood score for this partitioning scheme was closest to that from the full partitioning by gene and codon, and according to the RBF, it is the most efficient way of improving the likelihood scores per

parameter added to the model. However, based on the Bayes Factor the model distinguishing 42 partitions was still significantly better.

**Tests of Compositional Heterogeneity**

The conventional chi-square test showed that the data are heterogeneous (*P*=0 that the data are homogeneous). We then asked if heterogeneity is uniform across the data partitions by performing the test separately on each gene partition and codon position. The 3rd codon positions were heterogeneous for all genes (table 2) and also showed significant levels of saturation for about half of the gene partitions (supplementary table S3, Supplementary Material online). Therefore they were not considered further for tests of heterogeneity. In

Downloaded from https://academic.oup.com/gbe/article/8/1/161/2574155 by guest on 13 January 2022

contrast, all 2nd codon position partitions appeared homogeneous by this test. The 1st codon positions failed for some genes, notably *cytb*, *nad2*, *nad4*, *nad5*, and *nad6*, but showed compositional homogeneity in the others. When the 1st codon positions were RY recoded, the data set as a whole was still heterogeneous ( $P=0$ ), but heterogeneity was no longer apparent in the 1st codon positions when tested for each gene individually (table 2).

The data were also assessed against data simulated under a homogeneous model (Foster 2004), which revealed heterogeneity ( $P < 0.05$ ) in 2nd codon positions in genes *nad2*, *nad4*, *nad5*, and *nad6*, despite appearing homogenous in the conventional chi-square test. The RY-recoded 1st positions remained compositionally homogeneous. However, it could be argued that using a two-state model would be more valid for analysis of RY-coded matrices, rather than calculations with DNA models. We found that this approach detected highly significant levels of heterogeneity in the *nad2*, *nad4*, *nad5*, and *nad6* genes that were already implicated in 2nd position heterogeneity above (table 2). Finally, we conducted the test of heterogeneity on the translated protein sequence. This showed that out of 14 gene partitions, six were heterogeneous ( $P < 0.05$ ), and eight were not. The highest level of significance was again observed for *nad2*, *nad4*, *nad5*, and *nad6* (table 2).

The RogueNaRok algorithm identified 14 (run #1) and 30 (run #2) taxa as being inconsistently placed when investigating the placement of a single terminal or a set of two terminals, respectively, for a total of 33 rogue taxa (supplementary table S4, Supplementary Material online). Compositional heterogeneity was investigated for a reduced data set that had these 33 taxa excluded. The results were very similar to those obtained with the full matrix, with heterogeneity in 2nd positions and in the two-state model of RY-recoded 1st position sites limited to *nad2*, *nad4*, *nad5*, and *nad6* partitions (table 2). Rogue taxa instead seemed to be affected by slightly lower data completion, specifically the sequences for the short amplicon coding for *nad2* and *cox1-5'*, which was missing from 22 or 23 respectively of the 33 rogue taxa. Yet, the average completion of the data set for rogue taxa was similar to the complete data set (12.24 vs. 12.40 protein-coding loci per taxon; supplementary table S4, Supplementary Material online), and >120 other taxa in the matrix were also lacking the short fragment (supplementary table S2, Supplementary Material online).

### Phylogenetic Analysis

A series of phylogenetic analyses was conducted to assess the effects of nonhomogeneity on tree topology. We used three different approaches for tree searches to make use of the available phylogenetic methods, and scored these trees for about 30 nodes defining deep relationships that were expected based on previous work or appeared noteworthy

because they differed among the tree searches here (table 3 and supplementary table S4, Supplementary Material online). We used PhyML and nhPhyML for assessing the sensitivity of the topology to the introduction of branch-specific parameters in the nonhomogeneous model. The tree generated with PhyML was unsatisfactory in many regards due to the failure of recovering several key groups, including the large suborders Adephaga and Polyphaga, four of the five infraorders, and the superfamilies in the species-rich Cucujiformia. We then compared the topology from the nhPhyML model, which adds a separate parameter for the nucleotide composition for each branch. The nhPhyML tree (supplementary fig. S1, Supplementary Material online) was greatly improved, including the monophyly of the suborders and all infraorders. However, in the Cucujiformia only the (reciprocal) monophyly of Tenebrionoidea and Lymexyloidea was recovered, whereas paraphyly remained surrounding Chrysomeloidea, Curculionoidea, and Cucujoidea.

The RAxML software was used to assess nonhomogeneity across the data (not across the tree, as in nhPhyML) implementing independent GTR models for different partitions of the matrix (although without allowing among-partition rate variation that is not implemented in this software). A tree from the unpartitioned data had many of the same undesirable features as the PhyML tree, including the nonmonophyly of Adephaga and Polyphaga, although with a better outcome overall including the recovery of three of five infraorders. Partitioning the data according to the 42 codon and gene partitions improved the topology by recovering all 4 suborders, the 5 infraorders, and most superfamilies, but problems with the recovery of the cucujiform superfamilies were not fully solved. The impact of including and excluding the two rRNA genes was limited (table 3 and supplementary table S4, Supplementary Material online). We further used the RAxML algorithm to explore the effects of removing the most compositionally heterogeneous data, first by removal of 3rd codon positions and RY coding of 1st positions, and in an additional search we also removed the four loci showing the greatest level of heterogeneity. Finally, we used the amino acid translation (on all protein-coding genes) (table 3 and supplementary table S4, Supplementary Material online). Although most of the correctly recovered higher groupings were robust to the specific data treatment, there was a general decrease in power with the removal of data, and none of these analyses performed better than the partitioned analysis of all nucleotides. Notably, the removal of the rate-heterogeneous genes (*nad2*, *nad4*, *nad5*, *nad6*) resulted in the loss of monophyly of both small suborders, Myxophaga and Archostemata (supplementary fig. S2, Supplementary Material online, for a tree from a matrix RY recoded for 1st positions and 3rd positions removed). Equally, the amino acid coding resulted in the failure to recover several key groups, including the suborder Polyphaga that was paraphyletic due to the misplaced *Tetraphalerus* and *Priacma* (Archostemata). Hence, the

**Table 3**

Recovery of Key Nodes and Other Features in Trees Obtained from Different Analyses with PhyML, RAxML or PhyloBayes (PB), Before and After (excl. heterogeneous) Removing the Composition Heterogeneous Markers *nad2*, *nad4*, *nad5*, and *nad6*

Taxon	PhyML										RAxML										PhyloBayes																										
	1	2	3	4	4	Partitioned no 125/16S	RY code	RY no 125/16S	Excluding heterogeneous	Amino acid	Plus outgroups (fig. 1)	No outgroups (supplementary fig. S2, Supplementary Material online)	Excluding rogue	Excluding heterogeneous	1	2	3	4	4	Partitioned no 125/16S	RY code	RY no 125/16S	Excluding heterogeneous	Amino acid	Plus outgroups (fig. 1)	No outgroups (supplementary fig. S2, Supplementary Material online)	Excluding rogue	Excluding heterogeneous	1	2	3	4	4	Partitioned no 125/16S	RY code	RY no 125/16S	Excluding heterogeneous	Amino acid	Plus outgroups (fig. 1)	No outgroups (supplementary fig. S2, Supplementary Material online)	Excluding rogue	Excluding heterogeneous					
Position in figure 3	1	2	3	4	4		X	5	6	7	x	8	9	1	2	3	4	4		X	5	6	7	x	8	9	1	2	3	4	4		X	5	6	7	x	8	9								
All suborders monophyletic	N	Y	N	Y	Y		Y	N	N	Y	Y	Y	N	N	Y	N	Y	Y		Y	N	(P+Ar)	(P+Ar)	Y	Y	N	N	Y	N	Y	Y		Y	N	(P+Ar)	(P+Ar)	Y	Y	N								
Suborders relationships	n/a	P (Ar)	(P+Ar)	P (Ar)	P (Ar)		P (Ar)	n/a	(M+Ad)	(M+Ad)	(M+Ad)	(Ar+Ad)	(M+)	n/a	P (Ar)	(P+Ar)	P (Ar)	P (Ar)		P (Ar)	(M+Ad)	(M+Ad)	(M+Ad)	(Ar+Ad)	(M+)	(Ar+Ad)	(Ar+Ad)	(Ar+Ad)	(Ar+Ad)	(Ar+Ad)	(Ar+Ad)	(Ar+Ad)	(Ar+Ad)	(Ar+Ad)	(Ar+Ad)	(Ar+Ad)	(Ar+Ad)	(Ar+Ad)									
Geodephaga	M*	M*	M*	M*	M*		M	M*	M	M*	M*	M*	M*	M*	M*	M*	M*	M*		M	(M+Ad)	(M+Ad)	(M+Ad)	(M+Ad)	(M+Ad)	(M+Ad)	(M+Ad)	(M+Ad)	(M+Ad)	(M+Ad)	(M+Ad)	(M+Ad)	(M+Ad)	(M+Ad)	(M+Ad)	(M+Ad)	(M+Ad)	(M+Ad)									
Elateriformia	P	M	P	M	M		M	M	M	M	M	M	M	M	M	P	M	M		M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M				
Staphyliniformia + Scarabaeiformia	P	M	M	M	M		M	M	M	M	M	M	M	M	M	M	M	M		M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M			
Scarabaeiformia	P	M	M	M	M		M	M	M	M	M	M	M	M	M	M	M	M		M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M		
Bostrichiformia	P	M	M	M	M		M	M	M	M	M	M	M	M	M	M	M	M		M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	
Bostrichiformia sister	n/a	Elat	Elat	Elat	Elat		Cuc	Cuc	Cuc	Cuc	Cuc	Cuc	Cuc	Cuc	Elat	Elat	Elat	Elat		Cuc	Cuc	Cuc	Cuc	Cuc	Cuc	Cuc	Cuc	Cuc	Cuc	Cuc	Cuc	Cuc	Cuc	Cuc	Cuc	Cuc	Cuc	Cuc	Cuc	Cuc	Cuc	Cuc	Cuc	Cuc			
Cucujiformia	M	M	M	M	M		M	M	M	M	M	M	M	M	M	M	M	M		M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	
Clerolea	M	M	M	M	M		M	M	M	M	M	M	M	M	M	M	M	M		M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	
Cerylonid Series	M	M	M	M	M		M	M	M	M	M	M	M	M	M	M	M	M		M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	
Nitidulid Series	M	P	M	M	M		M	M	M	M	M	M	M	M	P	M	M	M		M	M	M	M	M	M	M	P	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M
Cucujoid Series	M	M	M	M	M		M	M	M	M	M	M	M	M	M	M	M	M		M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M
Nitidulid + Cucujoid	M	M	M	M	M		M	M	M	M	M	M	M	M	M	M	M	M		M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M
Tenebrionoidea + Lymexyloidea	M	M	M	M	M		M	M	M	M	M	M	M	M	M	M	M	M		M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M
Ten. + Lym. recipr.monophyly	N	Y	Y	N	N		N	N	N	N	N	N	N	N	Y	Y	N	N		N	N	N	N	N	N	N	Y	Y	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
Chrysomeloidea	P	P	P	P	P		P	P	P	P	P	P	P	P	P	P	P	P		P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	
Curculionoidea	P	P	P	P	P		P	P	P	P	P	P	P	P	P	P	P	P		P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	
Chrys. + Curc. recipr. monophyly	N	N	N	N	N		N	N	N	N	N	N	N	N	N	N	N	N		N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N

NOTE.—RAxML trees were obtained with the RY-coded 1st positions and 3rd positions removed, or on all data (including the rRNA genes). All PhyloBayes trees were conducted on the amino acid coded matrix. M, monophyletic; P, paraphyletic or polyphyletic; U, unresolved, consistent with monophyly; Y, yes, a feature is present; N, no, a feature is not present. In some cases, the groups were recovered but with certain member taxa absent (–) or other taxa included (+), as indicated. Note that *Sphindus* (Sphindidae) was disregarded when scoring Nitidulid and Cucujid series. The asterisks mark the trees that are monophyletic for Geodephaga only if *Habrodera* (Cicindelidae) is disregarded.

RAXML analysis was not greatly distorted by compositional heterogeneity and instead suffered more from the loss of data when the most heterogeneous positions were removed.

Finally, the CAT model in PhyloBayes also partitions the data, but unlike the RAXML analysis these partitions are not determined a priori but are estimated from the data themselves. The resulting tree (fig. 1) showed most of the features of the trees from the partitioned RAXML analysis, but also recovered the two large superfamilies Curculionoidea and Chrysomeloidea that were otherwise polyphyletic with respect to each other and included portions of Cucujoidea in all other analyses (supplementary table S4, Supplementary Material online). This tree also recovered a different relationship of the four suborders, linking Adephaga with Archostemata and not Myxophaga, and when rooted with the neuropteroid outgroups, the relationships were (Neuropteroid (Polyphaga (Myxophaga (Archostemata + Adephaga))), consistent with the findings of transcriptome analyses (Misof et al. 2014). Removal of the four heterogeneous *nad* genes did not greatly change the tree topology, although the resolution was reduced, indicating the loss of phylogenetic signal (supplementary table S4, Supplementary Material online). Finally, the Bayesian analysis was run again after removal of rogue taxa, which produced a tree very similar to that based on the complete data set, with the main improvement simply due to the absence of the inconsistently placed rogue taxa themselves. For example, only after removing several rogue taxa, in particular the divergent sequence for *Sphindus* (Sphindidae), the Nitidulid and Cucujid series of Cucujoidea each resolved as monophyletic and combined they were the sister group to Curculionoidea + Chrysomeloidea (supplementary table S4, Supplementary Material online).

### The Branch Length across Superfamilies

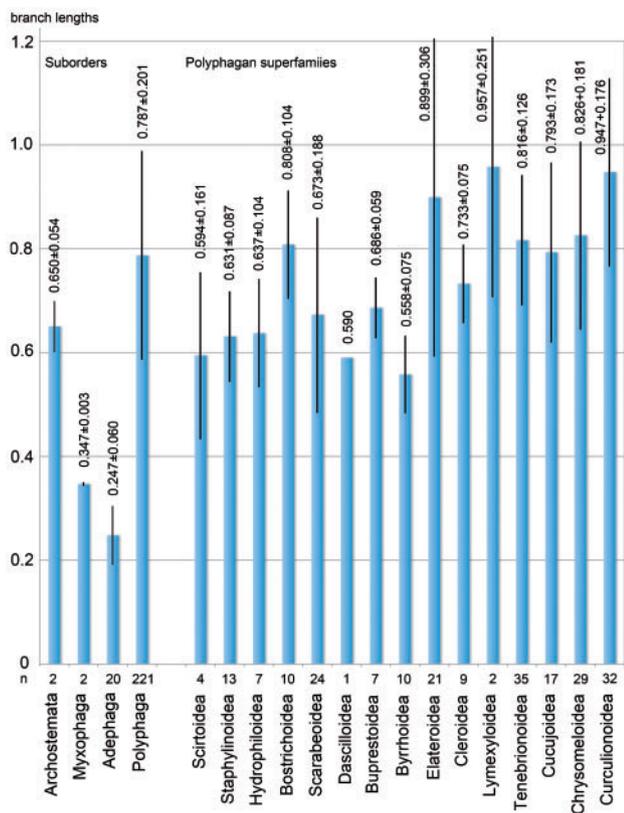
Root-to-tip branch lengths were investigated on the RY-coded ML tree (supplementary fig. S2, Supplementary Material online) and the Bayesian amino acid tree (supplementary fig. S3, Supplementary Material online) for each suborder and polyphagan superfamily (fig. 2 and supplementary fig. S4, Supplementary Material online). Variation among these groups was very similar for each data set. The Adephaga and Myxophaga showed substantially shorter branches than the two other suborders Archostemata and Polyphaga. Shorter branches were found in several polyphagan superfamilies, compared with Bostrichiformia and all superfamilies of Cucujiformia, which are sister groups in most analyses and occupy a derived position in the tree. Within some superfamilies branch lengths were highly variable, for example, the two sequences of Passalidae with extremely long branches, which were responsible for shifting the average branch length in Scarabaeoidea beyond the rate of other staphyliniform lineages. Similarly high variation in branch lengths was found in Elateroidea due to extremely long branches in

*Trixagus* and *Mastinocerus*. Extremely long branches compared with their sister taxa were found additionally in *Melittomma* (Lymexylidae), *Sphindus* (Sphindidae), Cassidinae (Chrysomelidae), and others. In addition, the rogue taxa had a tendency to exhibit faster rates of nucleotide change, with an average branch length higher than for the complete set of taxa (0.86997 vs. 0.73820) and many terminals in the top part of the range of root-to-tip distances, and a generally higher proportion of rogue taxa was found in superfamilies with higher branch length variability (supplementary table S5, Supplementary Material online).

### Discussion

This study generated a large number of new mitogenome sequences for the Coleoptera that more than doubles the available sequences and now permits an analysis of molecular evolution at the resolution of the family level. Early studies of Coleoptera using mitochondrial genomes noted the great heterogeneity in nucleotide composition and molecular rate that apparently misled the trees (Pons et al. 2010; Song et al. 2010). The sparse taxon sampling of studies conducted with conventional Sanger sequencing may have exacerbated these problems. If nucleotide heterogeneity is high and localized in the tree, and if similar composition arises convergently, there will be a tendency to create biases that overwhelm the phylogenetic signal. Already denser taxon sampling, the removal of synonymous codon positions, and the use of protein sequences were shown to partly overcome these problems (Timmermans et al. 2010). This is confirmed here for a much greater set of mitogenomes. However, it was not clear if the improved phylogenetic inference is correlated with reduced compositional heterogeneity, and to what degree heterogeneity can be reduced by removal of the most affected bases and by translation to protein sequences that might reduce the compositional bias from different codon usage.

Previous studies have established the distribution of compositional heterogeneity using the disparity index  $I_D$  (Song et al. 2010) that is based on the differences in substitution pattern for pairs of sequences deviating from expectations under a process of uniform nucleotide change. This analysis produced a measure of compositional heterogeneity for each terminal relative to other taxa in the data set and found that the more densely sampled Polyphaga exhibit the lowest cumulative disparity across all pairwise comparisons, whereas *Tetraphalerus* as the single representative of Archostemata had the highest disparity when summing the  $I_D$  values from comparisons with all other taxa (Song et al. 2010). These findings suggest that compositional heterogeneity is increased between distantly related taxa and therefore greater sampling density, as available in the Polyphaga, ameliorates the problem, although residual heterogeneity remains even in very



**FIG. 2.**—Mean branch length for major groups at suborder and superfamily levels. The corresponding numbers for the amino acid tree are provided in [supplementary figure 4, Supplementary Material](#) online.

densely sampled mitogenome trees, for example, in a tree of ~100 taxa in the family Curculionoidea (Gillett et al. 2014).

In this study, rather than using pairwise comparisons, heterogeneity was assessed for the matrix as a whole, but only after the data were partitioned by gene. This analysis showed that compositional heterogeneity is concentrated in four genes, all of them NADH dehydrogenases. Two of these (*nad4* and *nad5*) are on the reverse strand, while *nad6*, but not *nad2*, is in proximity to these genes, encoded by the forward strand. It is intriguing that these genes did not deviate in their impact on model fit in the partitioning, as splitting them and all others did not greatly improve the likelihood of the model (table 1). This was in contrast to data partitioning by forward and reverse strand that accounted for a large improvement in statistical fit in GTR models (i.e., under compositional homogeneity assumed by the GTR, and hence indicating different evolutionary patterns on either strand unrelated to compositional heterogeneity). Although RY recoding reduced the problem of compositional heterogeneity, it remains strong if applying a two-state model. Equally, the problem of compositional heterogeneity was not removed by using the amino acid sequences. Nucleotide bias has been shown to feed through the amino acid level; for example,

there is a correlation of AT or GC-rich mitogenomes with a prevalence of particular amino acids, which was established mainly in inter-phyla and inter-order comparisons of mitogenomes greatly differing in base composition (Foster et al. 1997; Bert et al. 2013; Li et al. 2015), and this seems to be confirmed here at a lower hierarchical level. The finding that predominantly the *nad* genes were affected by heterogeneity, which are functionally linked, might suggest that variation in the protein level and possible covariation in the NAD protein complex drives compositional heterogeneity, rather than some unspecified genomic process driven by strand bias. Evolutionary shifts in mitochondrial genes have been associated with positive selection, for example, with changes to respiratory function (Tomasco and Lessa 2011), although because compositional heterogeneity in the four affected genes is encountered in all codon positions, other explanations due to gene-wide effects may also apply.

We also tested if exclusion of the so-called rogue taxa improves the tree topology for the remaining taxa. There are different reasons for a taxon to be rogue, and here we speculated that compositional heterogeneity is a contributing factor, but their removal had virtually no impact on the degree of compositional heterogeneity in the data. It is not clear what causes their inconsistent placement instead, but multiple factors probably contribute. Rogue taxa have a slightly lower representation of the *nad2* and *cox1-5'* markers located on the shorter PCR fragment than the matrix as a whole. Rogue taxa also have a tendency to show higher rates of nucleotide variation, which appears to interfere with stability of their placement on the tree. These factors may affect the strength of the signal through limited data or weak long-branch attraction.

Taken together, the compositional heterogeneity in Coleoptera mitogenomes is moderate and it is spread over the tree somewhat evenly, and therefore heterogeneity per se might not have a great impact on the difficulties to recover the correct tree, in particular for those lineages where the true phylogenetic signal is strong. We can see the effect of nucleotide composition alone if we construct a tree based on the composition of each taxon. Therefore we constructed distance matrices based on nucleotide compositions and made neighbor joining trees based on the distance matrices with the BIONJ algorithm (Gascuel 1997). Using 100 bootstrap replicates, a consensus tree showed hardly any strong (>50%) support for any lineage, and most support was weak at <20% (data not shown). This confirms the idea that the effect of compositional biases on the tree topology is moderate and not localized.

### Heterogeneity and Tree Topology

The three major approaches using the PhyML, RAxML, and PhyloBayes algorithms are implementations of very different likelihood models and search strategies, whose performance

was assessed in the light of information about the level of compositional heterogeneity. As the tree of Coleoptera remains insufficiently known, the quality of different models cannot be tested against a “true tree,” but the knowledge on coleopteran phylogeny is now sufficiently good to rely on the recovery of numerous well-established monophyletic groups to assess the quality and thus provides guidance on how to select the most defensible topology. In turn, the assessment against those “known” nodes also provides information on the less well-known parts of the tree to establish basal relationships.

Only the nhPhyML analysis provides a means for testing the effect of nonhomogeneity explicitly, as it accommodates changing the GC/AT ratio at every node in the tree (Galtier and Gouy 1998), although perhaps at the risk of overparameterization. The algorithm is implemented only for DNA data. Other tree-heterogeneous models are also implemented for protein sequences, such as the nonhomogeneous nhPhyloBayes, and the NDCH and the NDRH models (node-discrete composition heterogeneity and node-discrete rate heterogeneity, respectively) which allow different compositions and different rate matrices on different branches, implemented in P4 (Foster et al. 2009). However, neither of these can be applied on the scale required here. The improvement obtained from nhPhyML over the homogeneous PhyML model was considerable, indicating the importance of taking into account the nonhomogeneity of nucleotide composition across the tree. This approach clearly increases the number of higher taxa recovered, although the tree remains unsatisfactory in some parts. We also conducted a RAxML analysis that only implements the standard GTR model (i.e., it does not parameterize tree heterogeneity), but permits partitioning of the data according to genes and codon positions. Data partitioning clearly improved the tree topology, to a similar degree as the use of the nonhomogeneous model in nhPhyML. However, there was no improvement after RY coding and removal of 3rd positions, while the removal of the heterogeneous *nad* genes or the recoding as amino acids caused a deterioration. The only obvious improvement from omitting the 3rd position was the avoidance of long-branch attraction for two lineages in Elateriformia, *Trixagus* and *Mastinocerus*, which are members of distantly related families, yet display very long terminal branches that group them together in the RAxML tree based on all data including 3rd positions, but not in the other analyses. Interestingly, at least with the search strategy applied here, the nhPhyML analysis does not overcome this problem, suggesting that the cause of the long-branch attraction is not primarily due to nucleotide heterogeneity of branches. The great rate acceleration in a few isolated taxa is a curious feature of mitogenome evolution of Coleoptera and affects nucleotide and amino acid variation alike (fig. 1 and supplementary figs. S2 and S3, Supplementary Material online). There is concern that taxa affected by this increased rate are misplaced in the tree, in particular if multiple

such sequences attract each other, but for the most striking cases the removal of 3rd position suffices to avoid this type of long-branch attraction.

Finally, the CAT model generated the most defensible trees, and although the method does not address tree heterogeneity explicitly, apparently it is best equipped to deal with the complex sequence variation in mitogenomes, as it provides greater flexibility for modeling different classes of sites with independent substitution processes (Lartillot and Philippe 2004). Due to the size of the data set we used the simpler CAT-Poisson model whose estimate of global exchange rates (obtained empirically from the data) is shared by all sites. Yet, the CAT and CAT-GTR models are efficient in dealing with long-branch attraction due to their ability to account more accurately for saturation and thus the greater power for estimating the evolutionary process (Lartillot et al. 2007). These analyses were conducted only at the level of protein sequences, but the improvement over other analyses is not due to the use of protein data per se. These data are also affected by compositional heterogeneity, and amino acid coding performed with RAxML did not result in any improvement over the analysis of the nucleotide variation (table 3 and fig. 3). These findings further support the power of the CAT model, at least at the level of divergence within the Coleoptera, where saturation may still be limited. An additional conclusion from this analysis was that the removal of rogue taxa does not greatly improve the tree topologies, while the level of heterogeneity also is not reduced. Rogue taxa were, however, affected by longer average branches and hence were more prone to long-branch attraction, and their removal facilitated the recognition of higher taxa whose limits were blurred otherwise. For example, the sequence for *Sphindus* consistently interfered with the recognition of other lineages in Cucujoidea, and the extremely long-branched *Trixagus* interfered with relationships in Elateroidea. Both were recognized as rogue taxa.

### Implications for the Phylogenetic Tree of Coleoptera

The tree topology obtained from mitochondrial genomes adds to the growing confidence in the principal lineages of Coleoptera attained in the last two decades (Lawrence and Newton 1995; Hunt et al. 2007; McKenna and Farrell 2009; Bouchard et al. 2011; Lawrence et al. 2011; McKenna, Farrell, et al. 2015). A schematic summary of basal relationships from various analyses is given in figure 3. The results confirm the monophyly of the four beetle suborders; the monophyly of the infraorders within Polyphaga; the monophyly of most of Crowson's superfamilies (Crowson 1970); and the monophyly of most families (where multiple representatives were used). The study also paints an increasingly clearer picture of the relationships of these groups to each other, in particular in the species-rich Polyphaga.

Specifically, the PhyloBayes analysis is the first to favor the sister relationship of Polyphaga to the three other suborders based on mitochondrial genes, which is supported by the transcriptome study of Misof et al. (2014). Rooting was critical for this inference; data from ESTs (Hughes et al. 2006) and a smaller set of mitogenomes (Timmermans et al. 2010) included coleopteran ingroup taxa only and were rooted on Archostemata, which was supported by morphological studies (Beutel and Haas 2000; Friedrich et al. 2009) and by the abundance of fossils of this presumed earliest radiation of Coleoptera (Crowson 1960). However, rerooting these trees with Polyphaga produces the same ingroup topology as found here after inclusion of Neuropterida outgroups. All other molecular studies based on mitogenomic analyses to date favored Myxophaga + Adephaga (Pons et al. 2010; Song et al. 2010; Timmermans et al. 2010), which was also supported by the RAxML and PhyML analyses conducted here, and which could easily be explained by the convergent low evolutionary rates in both suborders (fig. 2 and [supplementary fig. S4, Supplementary Material](#) online). Previous studies combining mitochondrial data with nuclear rRNA genes generally support a yet different topology of Polyphaga + Adephaga (Caterino et al. 2002; Hunt et al. 2007; Bocak et al. 2014). If indeed Polyphaga is the sister to the other suborders, this would reduce the imbalance of species diversity at the basal node of the tree, given that in previous work Archostemata and Myxophaga with less than 100 species each were thought to be the sister of all other Coleoptera and the Polyphaga, respectively.

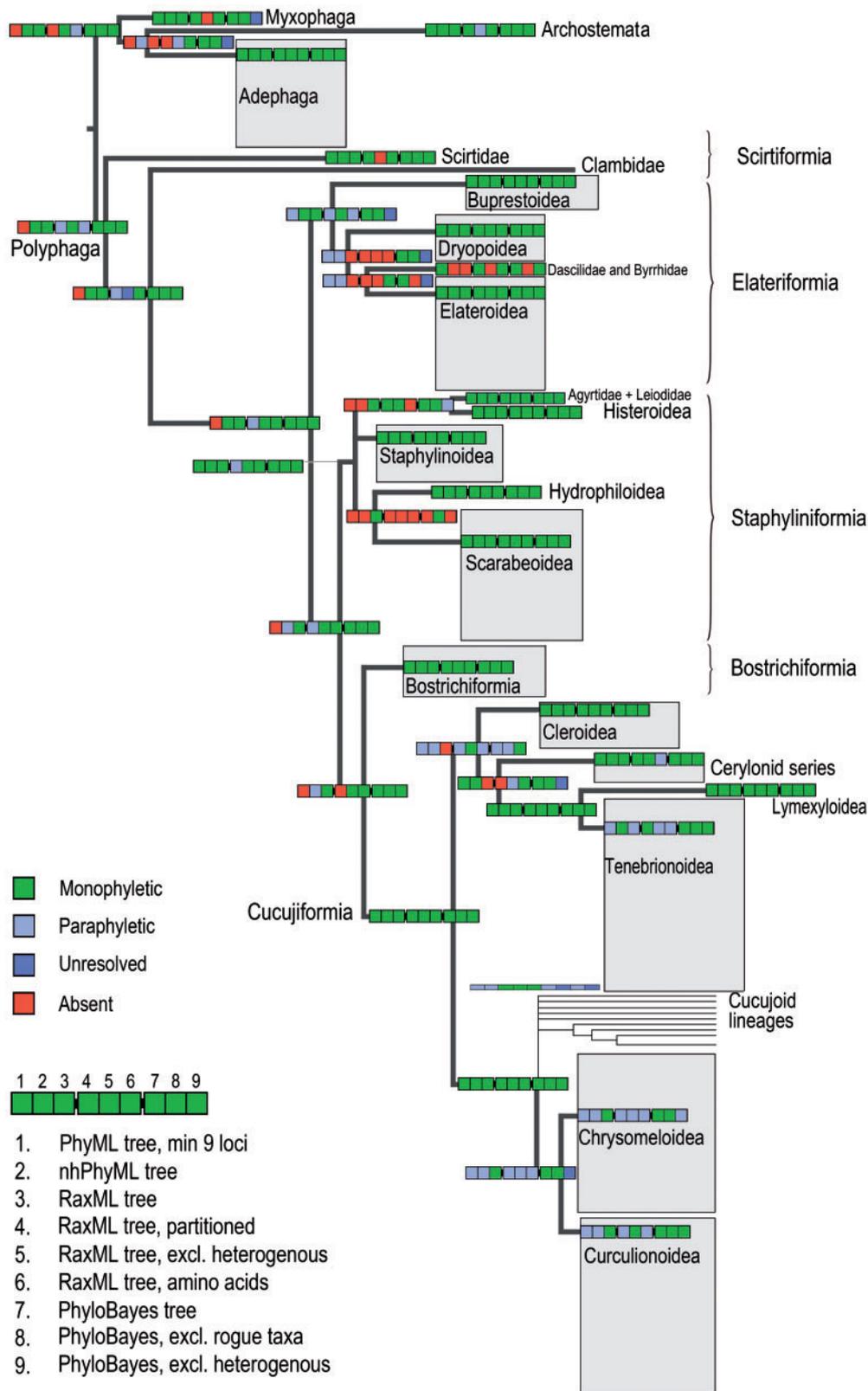
Within Polyphaga, we confirm the Scirtidae/Clambidae grade as the earliest branching lineages in Polyphaga, as proposed by Hunt et al. (2007) and Lawrence (2001), to form the new series Scirtiformia. The Elateriformia is the sister to all remaining Polyphaga, again in agreement with studies from ESTs (Hughes et al. 2006), although the RAxML (all nucleotides) and nhPhyML analyses group them as sister to Bostrichiformia. Internal relationships of Elateriformia recover the three large groups Buprestoidea, Elateroidea, and Dryopoidea (=Byrrhoidea minus Byrrhidae). The latter is defined by a unique rearrangement of tRNA gene order (Timmermans and Vogler 2012), which is confirmed here for all members of this clade, but the position of Byrrhidae (Byrrhoidea) and Dascilloidea remains ambiguous ([supplementary table S4, Supplementary Material](#) online). The Staphyliniformia occupying the next node is composed of three major groups (Histeroidea, Hydrophiloidea, Staphylinioidea) and also includes the Scarabaeiformia (Scarabaeoidea), which should no longer be considered at the rank of an infraorder. The staphylinoid families Leiodidae + Agyrtidae were repeatedly recovered as sister to Histeroidea, which interfered with the expected sister relationship of Histeroidea and Hydrophiloidea (McKenna, Wild, et al. 2015) recovered only in the PhyML analyses or when excluding the heterogeneous loci in PhyloBayes.

Bostrichiformia were split into two clades composed of Anobiidae (Anobiinae) + Ptiniidae and Dermestidae, and were the sister of Cucujiformia (except in some RAxML and nhPhyML).

Cucujiformia, the infraorder encompassing about half of all species of beetles, was always monophyletic and consists of sequential nodes defining major lineages including Cleroidea, Cerylonid series (Cucujoidea), Lymexyloidea + Tenebrionoidea, remaining Cucujoidea, Chrysomeloidea, and Curculionoidea. The Tenebrionoidea were found as sister to Lymexyloidea (Timmermans et al. 2010; Bocak et al. 2014; Gunter et al. 2014). The Cucujoidea can no longer be considered a valid taxonomic group (Hunt et al. 2007; Marvaldi et al. 2009). The mitogenomes now confirm that the Cerylonid series (Robertson et al. 2008) is only distantly related to the other cucujoid lineages, which include sets of families referred to as Nitidulid, Erotylid, and Cucujid series by Hunt et al. (2007). These groups cluster closely in the tree, either as an unresolved grade at the base of, or as sister to, the Curculionoidea + Chrysomeloidea. Only the PhyloBayes analysis recovers the reciprocal monophyly of Curculionoidea + Chrysomeloidea, which was partly interdigitated in all other analyses, but the monophyly of Chrysomeloidea is supported by the unique GCU tRNA<sup>Lys</sup> anticodon (fig. 1).

## Conclusion

The possibilities for rapid sequencing of mitochondrial genomes have brought a new perspective to the phylogenetics of Coleoptera. Although compositional heterogeneity is pervasive in these data sets, the study joins others (Talavera and Vila 2011; Li et al. 2015) in suggesting the power of the CAT model that produced highly satisfactory trees. Partitioned likelihood models with the RAxML software were not much worse, but missed a few critical relationships apparently affected by different rates of molecular change. The problem of compositional heterogeneity has been considered to be a major driver of long-branch attraction, and is frequently thought to be reduced by RY coding and removal of 3rd codon positions, or by using the translated protein sequence. Here we show that these strategies cannot remove compositional heterogeneity completely, and that heterogeneity is not uniformly distributed among the various mitochondrial genes. Although removing and recoding of codon or gene partitions may reduce heterogeneity, tree resolution and support are diminished. As it has become possible to sequence mitochondrial genomes very rapidly (Gillett et al. 2014; Tang et al. 2015), the challenge is to have implementations of the Bayesian mixture models that can be used at the much larger scale required for future studies.



**Fig. 3.**—Schematic representation of the basal relationships from mitogenome sequences. The tree is based on the PhyloBayes analysis of figure 1, with outgroups removed. Key nodes were scored for nine trees obtained in various analyses described in table 3.

Downloaded from https://academic.oup.com/gbe/article/8/1/161/2574155 by guest on 13 January 2022

## Supplementary Material

Supplementary tables S1–S5 and figures S1–S4 are available at *Genome Biology and Evolution* online (<http://www.gbe.oxfordjournals.org/>).

## Acknowledgments

This work was supported by a grant from the Leverhulme Trust to A.P.V., C.B., D.A., and L.B. (grant F/00969/H). M.T.J.N. was supported by an NERC Postdoctoral Fellowship (NE/I021578/1). We thank Junying Lim and Benjamin Linard for assistance with the editing and annotation of the mitochondrial sequences. We are obliged to M. Balke, M. Barclay, M. Bednarik, S. Fabrizi, J. Gomez-Zurita, P. Hammond, R. A. B. Leschen, and C. Murria for donating specimens, and to two anonymous reviewers for valuable comments.

## Literature Cited

- Abascal F, Posada D, Zardoya R. 2007. MtArt: a new model of amino acid replacement for Arthropoda. *Mol Biol Evol.* 24:1–5.
- Aberer AJ, Krompass D, Stamatakis A. 2013. Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. *Syst Biol.* 62:162–166.
- Bernt M, et al. 2013. A comprehensive analysis of bilaterian mitochondrial genomes and phylogeny. *Mol Phylogenet Evol.* 69:352–364.
- Beutel RG, Haas F. 2000. Phylogenetic relationships of the suborders of Coleoptera (Insecta). *Cladistics* 16:103–141.
- Bininda-Emonds ORP. 2005. transAlign: using amino acids to facilitate the multiple alignment of protein-coding DNA sequences. *BMC Bioinformatics* 6:156.
- Bocak L, et al. 2014. Building the Coleoptera tree-of-life for >8000 species: composition of public DNA data and fit with Linnaean classification. *Syst Entomol.* 39:97–110.
- Bouchard P, et al. 2011. Family-group names in Coleoptera (Insecta). *ZooKeys* 88:1–972.
- Boussau B, Gouy M. 2006. Efficient likelihood computations with nonreversible models of evolution. *Syst Biol.* 55:756–768.
- Cameron SL. 2014. Insect mitochondrial genomics: implications for evolution and phylogeny. *Ann Rev Entomol.* 59:95–117.
- Castoe TA, Sasa MM, Parkinson CL. 2005. Modeling nucleotide evolution at the mesoscale: the phylogeny of the Neotropical pitvipers of the *Porthidium* group (Viperidae: Crotalinae). *Mol Phylogenet Evol.* 37:881–898.
- Caterino MS, Shull VL, Hammond PM, Vogler AP. 2002. The basal phylogeny of the Coleoptera inferred from 18S rDNA sequences. *Zool Scripta.* 31:41–49.
- Crowson RA. 1960. The phylogeny of Coleoptera. *Ann Rev Entomol.* 5:111–134.
- Crowson RA. 1970. Classification and biology. London: Heinemann Educational Books Ltd.
- Foster PG. 2004. Modeling compositional heterogeneity. *Syst Biol.* 53:485–495.
- Foster PG, Cox CJ, Embley TM. 2009. The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. *Phil Trans R Soc B.* 364:2197–2207.
- Foster PG, Jermini LS, Hickey DA. 1997. Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. *J Mol Evol.* 44:282–288.
- Friedrich F, Farrell BD, Beutel RG. 2009. The thoracic morphology of Archostemata and the relationships of the extant suborders of Coleoptera (Hexapoda). *Cladistics* 25:1–37.
- Galtier N, Gouy M. 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol Biol Evol.* 15: 871–879.
- Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol.* 14:685–695.
- Gillett CPDT, et al. 2014. Bulk *de novo* mitogenome assembly from pooled total DNA elucidates the phylogeny of weevils (Coleoptera: Curculionoidea). *Mol Biol Evol.* 31:2223–2237.
- Guindon S, et al. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst Biol.* 59:307–321.
- Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52:696–704.
- Gunter NL, et al. 2014. Towards a phylogeny of the Tenebrionoidea (Coleoptera). *Mol Phylogenet Evol.* 79:305–312.
- Haran J, Timmermans MJTN, Vogler AP. 2013. Mitogenome sequences stabilize the phylogenetics of weevils (Curculionoidea) and establish the monophyly of larval ectophagy. *Mol Phylogenet Evol.* 67:156–166.
- 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.* 38:100–116.
- Hebert PDN, Cywinska A, Ball SL, DeWaard JR. 2003. Biological identifications through DNA barcodes. *Proc Biol Sci.* 270:313–321.
- Hughes J, et al. 2006. Dense taxonomic EST sampling and its applications for molecular systematics of the Coleoptera (beetles). *Mol Biol Evol.* 23:268–278.
- Hunt T, Vogler AP. 2008. A protocol for large-scale rRNA sequence analysis: towards a detailed phylogeny of Coleoptera. *Mol Phylogenet Evol.* 47:289–301.
- Hunt T, et al. 2007. A comprehensive phylogeny of beetles reveals the evolutionary origins of a superradiation. *Science* 318:1913–1916.
- Katoh K, Asimenos G, Toh H. 2009. Multiple alignment of DNA sequences with MAFFT. *Methods Mol Biol.* 537:39–64.
- Kumar S, Gadagkar SR. 2001. Disparity index: a simple statistic to measure and test the homogeneity of substitution patterns between molecular sequences. *Genetics* 158:1321–1327.
- Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol Biol.* 7(Suppl 1):S4.
- Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25:2286–2288.
- Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol.* 21:1095–1109.
- Lawrence JF. 2001. A new genus of Valdivian Scirtidae (Coleoptera) with comments on Scirtoidea and the beetle suborders. In: Morimoto K, et al., editors. *Sukunahikona*. Osaka (Japan): The Japan Coleopterological Society (Special Publication No. 1). p. 351–361.
- Lawrence JF, Newton AF. 1995. Families and subfamilies of Coleoptera (with selected genera, notes, references and data on family-group names). In: Pakaluk J, Slipinski SA, editors. *Biology, phylogeny, and classification of Coleoptera*. Warszawa (Poland): Museum i Instytut Zoologii PAN. p. 779–1066.
- Lawrence JF, et al. 2011. Phylogeny of the Coleoptera based on morphological characters of adults and larvae. *Ann Zool.* 61:1–217.
- Li H, et al. 2015. Higher-level phylogeny of paraneopteran insects inferred from mitochondrial genome sequences. *Sci Rep.* 2015 Feb 23;5:8527. doi: 10.1038/srep08527.
- Maddison WP, Maddison DR. 2014. Mesquite: a modular system for evolutionary analyses. Version 3.0. Available from: <http://mesquiteproject.org>.

- Marvaldi AE, Duckett CN, Kjer KM, Gillespie JJ. 2009. Structural alignment of 18S and 28S rDNA sequences provides insights into phylogeny of Phytophaga (Coleoptera: Curculionoidea and Chrysomeloidea). *Zool Scripta*. 38:63–77.
- McKenna D, Farrell B. 2009. Coleoptera. In: Hedges S, Kumar S, editors. *The Timetree of Life*. Oxford : Oxford University Press. p. 278–289.
- McKenna DD, Farrell BD, et al. 2015. Phylogeny and evolution of Staphyliniformia and Scarabaeiformia: forest litter as a stepping stone for diversification of nonphytophagous beetles. *Syst Entomol*. 40:35–60.
- McKenna DD, Wild AL, et al. 2015. The beetle tree of life reveals that Coleoptera survived end-Permian mass extinction to diversify during the Cretaceous terrestrial revolution. *Syst Entomol*. 40:835–880.
- Miller MA, Pfeiffer W, Schwartz T. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *Proceedings of the Gateway Computing Environments Workshop (GCE)*, 14 Nov. 2010, New Orleans, LA pp 1–8.
- Misof B, et al. 2014. Phylogenomics resolves the timing and pattern of insect evolution. *Science* 346:763–767.
- Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20:289–290.
- Pons J, Ribera I, Bertranpetit J, Balke M. 2010. Nucleotide substitution rates for the full set of mitochondrial protein-coding genes in Coleoptera. *Mol Phylogenet Evol*. 56:796–807.
- Robertson JA, Whiting MF, McHugh JV. 2008. Searching for natural lineages within the Cerylonid Series (Coleoptera: Cucujoidea). *Mol Phylogenet Evol*. 46:193–205.
- Sheffield NC, Song HJ, Cameron SL, Whiting MF. 2009. Nonstationary evolution and compositional heterogeneity in beetle mitochondrial phylogenomics. *Syst Biol*. 58:381–394.
- Simon C, et al. 1994. 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.
- Simon S, Hadrys H. 2013. A comparative analysis of complete mitochondrial genomes among Hexapoda. *Mol Phylogenet Evol*. 69:393–403.
- Song HJ, Sheffield NC, Cameron SL, Miller KB, Whiting MF. 2010. When phylogenetic assumptions are violated: base compositional heterogeneity and among-site rate variation in beetle mitochondrial phylogenomics. *Syst Entomol*. 35:429–448.
- Stamatakis A. 2006. RAXML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.
- Swofford DL. 2002. PAUP\*: Phylogenetic Analysis using Parsimony. Version 4.0b. Sunderland (MA): Sinauer Associates.
- 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 Evol Biol*. 11:15.
- Tang M, et al. 2015. High-throughput monitoring of wild bee diversity and abundance via mitogenomics. *Meth Ecol Evol*. 6:1034–1043.
- Timmermans MJTN, Vogler AP. 2012. Phylogenetically informative rearrangements in mitochondrial genomes of Coleoptera, and monophyly of aquatic elateriform beetles (Dryopoidea). *Mol Phylogenet Evol*. 63:299–304.
- Timmermans MJTN, et al. 2010. Why barcode? High-throughput multiplex sequencing of mitochondrial genomes for molecular systematics. *Nucleic Acids Res*. 38:e197.
- Tomasco IH, Lessa EP. 2011. The evolution of mitochondrial genomes in subterranean caviomorph rodents: adaptation against a background of purifying selection. *Mol Phylogenet Evol*. 61:64–70.
- Vogler AP, Cardoso A, Barraclough TG. 2005. Exploring rate variation among and within sites in a densely sampled tree: species level phylogenetics of North American tiger beetles (genus *Cicindela*). *Syst Biol*. 54:4–20.
- Wilkinson M. 1996. Majority-rule reduced consensus trees and their use in bootstrapping. *Mol Biol Evol*. 13:437–444.
- Xia X. 2013. DAMBE5: a comprehensive software package for data analysis in molecular biology and evolution. *Mol Biol Evol*. 30:1720–1728.

Associate editor: Dennis Lavrov