A peer-reviewed version of this preprint was published in PeerJ on 28 August 2018.

View the peer-reviewed version (peerj.com/articles/5498), which is the preferred citable publication unless you specifically need to cite this preprint.

Rota J, Malm T, Chazot N, Peña C, Wahlberg N. 2018. A simple method for data partitioning based on relative evolutionary rates. PeerJ 6:e5498 https://doi.org/10.7717/peerj.5498

A simple method for data partitioning based on relative evolutionary rates

Jadranka Rota Corresp., [1] , **Tobias Malm** [2] , **Niklas Wahlberg** [1]

[1] Department of Biology, Lund University, Lund, Sweden

[2] Department of Zoology, Swedish Museum of Natural History, Stockholm, Sweden

Corresponding Author: Jadranka Rota
Email address: jadranka.rota@biol.lu.se

Background. Multiple studies have demonstrated that partitioning of molecular datasets is important in model-based phylogenetic analyses. Commonly, partitioning is done *a priori* based on some known properties of sequence evolution, e.g. differences in rate of evolution among codon positions of a protein-coding gene. Here we propose a new method for data partitioning based on relative evolutionary rates of the sites in the alignment of the dataset being analysed. The rates are inferred using the previously published Tree Independent Generation of Evolutionary Rates (TIGER), and the partitioning is conducted using our novel python script RatePartitions. We applied this method to eight published multi-locus phylogenetic datasets, representing different taxonomic ranks within the insect order Lepidoptera (butterflies and moths).

Methods. We used TIGER to generate relative evolutionary rates for all sites in the alignments. Then, using RatePartitions, we partitioned the data into bins based on their relative evolutionary rate. RatePartitions applies a simple formula that ensures a distribution of sites into partitions following the distribution of rates of the characters from the full dataset. This ensures that the invariable sites are placed in a partition with slowly evolving sites, avoiding the pitfalls of previously used methods, such as *k*-means. Different partitioning strategies were evaluated using BIC scores as calculated by PartitionFinder.

Results. In all eight datasets, partitioning using TIGER and RatePartitions was significantly better as measured by the BIC scores than other partitioning strategies, such as the commonly used partitioning by gene and codon position.

Discussion. We developed a new method of partitioning phylogenetic datasets without using any prior knowledge (e.g. DNA sequence evolution). This method is entirely based on the properties of the data being analysed and can be applied to DNA sequences (protein-coding, introns, ultra-conserved elements), protein sequences, as well as morphological characters. A likely explanation for why our method performs better than other tested partitioning strategies is that it accounts for the heterogeneity in the data to a much greater extent than when data are simply subdivided based on prior knowledge.

A simple method for data partitioning based on relative evolutionary rates

Jadranka Rota[a], Tobias Malm[b], and Niklas Wahlberg[a]

[a]*Department of Biology, Lund University, SE-223 62 Lund, Sweden*

[b]*Department of Zoology, Swedish Museum of Natural History, SE-10405 Stockholm, Sweden*

Corresponding author: *Jadranka Rota, Department of Biology, Lund University, SE-223 62 Lund, Sweden, jadranka.rota@biol.lu.se*

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3414v1 | CC BY 4.0 Open Access | rec: 15 Nov 2017, publ: 15 Nov 2017

Abstract

Background. Multiple studies have demonstrated that partitioning of molecular datasets is important in model-based phylogenetic analyses. Commonly, partitioning is done *a priori* based on some known properties of sequence evolution, e.g. differences in rate of evolution among codon positions of a protein-coding gene. Here we propose a new method for data partitioning based on relative evolutionary rates of the sites in the alignment of the dataset being analysed. The rates are inferred using the previously published Tree Independent Generation of Evolutionary Rates (TIGER), and the partitioning is conducted using our novel python script RatePartitions. We applied this method to eight published multi-locus phylogenetic datasets, representing different taxonomic ranks within the insect order Lepidoptera (butterflies and moths).

Methods. We used TIGER to generate relative evolutionary rates for all sites in the alignments. Then, using RatePartitions, we partitioned the data into bins based on their relative evolutionary rate. RatePartitions applies a simple formula that ensures a distribution of sites into partitions following the distribution of rates of the characters from the full dataset. This ensures that the invariable sites are placed in a partition with slowly evolving sites, avoiding the pitfalls of previously used methods, such as *k*-means. Different partitioning strategies were evaluated using BIC scores as calculated by PartitionFinder.

Results. In all eight datasets, partitioning using TIGER and RatePartitions was significantly better as measured by the BIC scores than other partitioning strategies, such as the commonly used partitioning by gene and codon position.

Discussion. We developed a new method of partitioning phylogenetic datasets without using any prior knowledge (e.g. DNA sequence evolution). This method is entirely based on the properties of the data being analysed and can be applied to DNA sequences (protein-coding, introns, ultra-conserved elements), protein sequences, as well as morphological characters. A likely explanation

for why our method performs better than other tested partitioning strategies is that it accounts for the heterogeneity in the data to a much greater extent than when data are simply subdivided based on prior knowledge.

Key words: BIC; intron; PartitionFinder; phylogenetics; phylogenomics; RatePartitions; UCEs; TIGER

Introduction

Phylogenetic analysis of DNA sequences is based on models of molecular evolution that estimate parameters such as base frequencies, substitution rates among nucleotides, as well as among-site rate variation. To reduce the heterogeneity in the data, datasets are often partitioned into subsets that are deemed to have undergone more similar molecular evolution. A number of studies have demonstrated that partitioning of data is important (Nylander et al., 2004; Brandley, Schmitz & Reeder, 2005; Brown & Lemmon, 2007; Rota, 2011; Rota & Wahlberg, 2012; Kainer & Lanfear, 2015), especially for model-based phylogenetic analyses, which are known to be more sensitive to underparameterization than overparameterization (Huelsenbeck & Rannala, 2004; Lemmon & Moriarty, 2004; Nylander et al., 2004).

Today, in most phylogenetic studies, partitions are defined *a priori* by the user, commonly by gene, gene and codon position, stems vs. loops in ribosomal RNA, or another feature of the sequence that the user believes to be important. In several studies, partitioning of protein-coding genes by gene and codon position was demonstrated to be a better option when compared to not partitioning or partitioning by gene (Nylander et al., 2004; Brandley, Schmitz & Reeder, 2005; Brown & Lemmon, 2007; Miller, Bergsten & Whiting, 2009; Rota, 2011). This approach is practical when a dataset consists of only a few genes. However, when data come from tens (or hundreds) of genes, this approach becomes unwieldy, although there are methods that allow one to combine many *a priori* established partitions into fewer, based on model testing with programs such as PartitionFinder (Lanfear et al., 2012).

Using a method described by Cummins and McInerney (2011), it is possible to partition a dataset in a more objective way, based on the properties of the data. The method takes into account the relative evolutionary rates of characters by comparing the patterns in character-state distributions in homologous characters (i.e., nucleotides or amino acids in a molecular alignment or characters in a morphological matrix). Each character thus receives a value for its evolutionary

rate, which is based on comparisons to all other characters in the matrix. The rate values can then be used to group characters with similar rates by dividing the range of rates into bins, which can be user-defined so as to span equal ranges of rates. This usually leads to the first bin containing characters that are invariable, and the last bin consisting of characters with the highest relative rate of change (Cummins & McInerney, 2011). This method is implemented in the program TIGER – Tree Independent Generation of Evolutionary Rates (Cummins & McInerney, 2011).

Originally, the method was developed to identify and exclude the fastest-evolving characters in a dataset, but this approach has potential problems (see Simmons & Gatesy, 2016). We have extended the TIGER method to partitioning the data by sorting characters into data subsets with similar relative rates of evolution (Rota & Wahlberg, 2012; Rota & Miller, 2013; Wahlberg et al., 2014), where we arbitrarily combined neighbouring TIGER bins to form data partitions with enough characters for analysis. A similar approach has been used in a number of studies (Kaila et al., 2013; Rota & Miller, 2013; Heikkilä et al., 2014; Matos-Maravi et al., 2014; Wahlberg et al., 2014; Edger et al., 2015; Kristensen et al., 2015; Rajaei et al., 2015; Ounap, Viidalepp & Truuverk, 2016), and although this method works quite well, the downside is that it requires the user to make a subjective decision about the final partitioning strategy.

Recently, a different way of using TIGER together with *k*-means was described by Frandsen et al. (2015). They compared their new method to traditional *a priori* defined partitions, as well as to site rates calculated using a maximum likelihood function. In all test cases, partitioning by both TIGER calculated rates and likelihood calculated rates performed better than traditional methods, with likelihood rates doing much better (Frandsen et al., 2015). However, the *k*-means algorithm has been found to place all invariable characters into one partition (Baca et al., 2017), which leads to biased likelihood values. Indeed, the *k*-means algorithm has now been disabled for molecular data in PartitionFinder2

(https://github.com/brettc/partitionfinder/commit/19d7fe41d2e469c131a5b0cc30184a069867b7f2 accessed 13 November 2017).

Here, we describe a simple and objective method for partitioning using TIGER. TIGER is again used for sorting of sites based on their relative evolutionary rates, but now we introduce an algorithm – RatePartitions – for dividing the sites among partitions in an objective way. This method has already been used in several published studies (Heikkilä et al., 2015; Rota, Pena & Miller, 2016; Rota et al., 2016; Sahoo et al., 2016). We report our findings from further testing RatePartitions performance on eight published datasets, some of which were difficult to analyse using traditional partitioning strategies. We use the Bayesian Information Criterion (BIC) for comparison of partitioning strategies. We do not carry out phylogenetic analyses and compare resulting topologies because it has been previously established that partitioning does affect topology, branch support, and branch lengths (see Kainer & Lanfear, 2015 and references therein), and since true phylogenies in all of these cases are unknown, we can only select the best partitioning strategy using statistical model evaluation metrics, such as e.g. BIC.

Materials & Methods

RatePartitions

Although it is technically incorrect to use the word 'partition' when referring to a data subset, we use 'partition' in that sense since this is commonly done in phylogenetics. When partitioning is carried out using TIGER, one must take into account the general properties of the data. One of these properties is that with standard DNA sequence data of protein-coding genes, one to two thirds of the data consist of invariable characters. These tend to be binned together to the exclusion of other data when using the TIGER binning strategy or the *k*-means algorithm (Baca et al., 2017). A partition made of only such data contains no phylogenetic information and thus it is advisable to include a number of slowly evolving characters to create a data partition with low

variation. To deal with that problem we developed RatePartitions – an algorithm which works in the following way. The dataset is first run in TIGER to calculate the relative rate of evolution for each site (character). These values can range from 1 (invariable sites) to 0 (no common patterns, i.e. the fastest-evolving sites). The sites are then combined into partitions using RatePartitions, which applies a simple formula that ensures a distribution of sites into partitions following the distribution of rates of the characters from the full dataset. This leads to larger partitions for characters with slower rates and, conversely, smaller partitions for those with higher rates. Preliminary tests using MrModeltest v2.3 (Nylander, 2004) and PartitionFinder v.1.0.0 (Lanfear et al., 2012) suggested that this strategy led to models with uniform rate variation within partitions.

RatePartitions is a PYTHON script (Supplemental Script S1) that determines the rate-spans for a variable number of partitions based on a user-specified division factor and the original range of rates calculated by TIGER (with the "-rl" command), and subsequently defines character sets for each partition. The rate-spans are calculated for the first (and slowest) partition with the following function:

$z = x\text{-}((x\text{-}y)/d)$

and for the remaining partitions:

$z = x\text{-}((x\text{-}y)/(d+p*0.3))$

where z is the lower limit of the rate-span, x is the upper limit of the rate-span (determined iteratively for each partition, i.e. z becomes x in the following iteration), y is the minimum value of rates for the entire dataset, d is a user defined division factor (which must be greater than 1; a higher number gives a greater number of partitions) and p is the partition number (when >1), which is multiplied by a fixed value of 0.3. The latter reduces the rate-span exponentially as partition number grows, which we found leads to partitions with more uniform rate variation for model-based analyses. Thus, for a dataset with rates ranging from 1 to 0.2 and with d set to 1.5,

the first partition will consist of all characters with rates between 1 and 1-((1-0.2)/1.5) = 0.4667. For partition 2, x = 0.4667 and this partition will include characters with rates between 0.4667 and 0.4667-((0.4667-0.2)/(1.5+2*0.3) = 0.3397, and so on until less than 10% of all characters are remaining. At this point the iterations are stopped and the remaining characters are placed into their own partition (which becomes the last and fastest-evolving partition).

Data partitioning and analyses

We analysed eight previously published lepidopteran datasets (Kodandaramaiah et al., 2010; Sihvonen et al., 2011; Penz, Devries & Wahlberg, 2012; Rota & Wahlberg, 2012; Zahiri et al., 2013; Matos-Maravi et al., 2014; Wahlberg et al., 2014; Rönkä et al., 2016) (Table 1). From the published datasets we excluded sites from the alignment that had more than 80% of missing data unless they had 1% or fewer of such sites (Table 2). These were the following datasets: Arctiina, Geometridae, *Morpho*, and Pieridae. All datasets are provided as Supplemental Information (Data S1). The datasets varied in base pair length from 4435 to 6372 and in number of taxa from 31 to 164 (Table 1). All datasets included one mitochondrial gene (COI) and four to seven nuclear genes that are commonly used in lepidopteran phylogenetics (CAD, EF-1α, GAPDH, IDH, MDH, RpS5, wingless) (Wahlberg & Wheat, 2008). We compared 14 partitioning strategies (Table 2), including user-defined ones such as partitioning by gene and by gene and codon position, and a number of different strategies devised based on the relative evolutionary rates assigned by TIGER and division of sites into partitions using the RatePartitions algorithm. We varied the parameter d in the RatePartitions algorithm between 1.5 and 4.5 in increments of 0.5. For comparison of the partitioning strategies we used the BIC score as calculated by PartitionFinder 1.1 (Lanfear et al., 2012). We did two types of searches with PartitionFinder. The first was a user-defined search for direct evaluation of the partitioning strategy obtained with

TIGER and RatePartitions. The second was a greedy search, which searches for partitions with similar parameter estimates and combines them so as to reduce the final number of partitions. For example, for a dataset with eight genes that are *a priori* partitioned by gene and codon position (24 partitions), a greedy search may result in a total of nine partitions because some of the original partitions were combined into a larger subset of data with similar parameter values. BIC was chosen as a statistical model evaluation metric because it has been shown to perform well in model selection for phylogenetic analysis (Abdo et al., 2005). We refer to analyses with different values of *d* as TIG1.5, TIG2.0, etc. The greedy search was not performed on TIG1.5, TIG2.0, TIG2.5, and TIG3.0 partitioning strategies because these were shown to have inferior BIC values in preliminary analyses.

Results

The eight datasets analysed covered a range of taxonomic ranks within Lepidoptera, from genus level (*Morpho* and *Calisto*), subtribes (Arctiina and Coenonymphina), two small to medium-sized families (Choreutidae and Pieridae, with about 400 and 1100 species, respectively), to two very large families (Geometridae and Noctuidae, with over 23,000 and 11,000 species, respectively) (van Nieukerken et al., 2011). They varied in sequence length from 4423 to 6716 base pairs (Table 1). The amount of missing data was quite variable. The most complete dataset, Coenonymphina, had more than 90% of sites with less than 20% of missing data, while the least complete dataset, Arctiina, had only 21% of sites with less than 20% of missing data (Table 3).

TIGER partitioning resulted in a different number of partitions for each dataset, with Geometridae and Pieridae being split into many more partitions than the other datasets (Table 4). For example, at *d* equalling 4.5, *Morpho*, the dataset with fewest taxa was split into only seven partitions, Pieridae into 20, Geometridae into 24, while all the other datasets ranged 10–14 in their number of partitions.

In all cases partitioning by gene region was clearly the worst way to subdivide the data, as determined by BIC scores, and applying the greedy search made little improvement (Fig. 1, Table S2). In all datasets, partitioning using TIGER and RatePartitions was the best strategy. However, in two datasets (Geometridae and Pieridae), partitioning by gene and codon position with a greedy search came close to the best TIGER strategy, although the BIC scores were still significantly higher for the TIGER strategy (Table S1). In all datasets, the improvement in the BIC score from TIG1.5 to TIG3.0 was quite steep, but further differences between TIG3.5, TIG4.0, and TIG4.5, with and without greedy search were relatively small, although the analyses with the greedy search always received a significantly better BIC score. TIG4.5Gr was the best strategy in *Calisto*, Choreutidae, Noctuidae, and Pieridae, whereas TIG4.0Gr was the best strategy in Arctiina, Coenonymphina, Geometridae, and *Morpho* (Fig. 1, Table S1).

An examination of the plots of the relative evolutionary rates estimated by TIGER for each gene fragment and codon position reveal differences among gene fragments, as well as sites belonging to the same codon position in the same gene fragment (Figs 2, S1). As expected, in general, first and second codon positions receive a much higher rate (i.e. implying slower change) than third codon positions, but in some genes there is a large proportion of third codon positions that also receive a rate of one, e.g. in the *Morpho* dataset for CAD, EF-1α, and RpS5 (Fig. S1). Conversely, there are genes that tend to have some fast-changing first and second codon positions, which then receive a relatively low rate. This is usually the case in COI, the mitochondrial gene, but also in several nuclear genes (wingless in all datasets, but also CAD, MDH, and RpS5 in some of the datasets; Figs 2, S1).

Discussion

Many studies have shown that partitioning of DNA sequence data for phylogenetic analysis is important because it affects the resulting tree topology, branch support, as well as branch lengths (see Kainer & Lanfear, 2015 and references therein). A common approach is to define partitions *a priori* based on some feature(s) of the DNA sequences such as genes, codon positions, stems, loops, introns, exons, etc., but this can be problematic because the properties of the sequence data are not fully known to the user to begin with. To avoid *a priori* partitioning, we developed a method of partitioning based on relative evolutionary rates of sites in an alignment. In our analyses, we demonstrated that this method outperforms other commonly used partitioning strategies, such as partitioning by gene and codon position, in all datasets that we tested. This method is entirely based on the alignment – not on trees or some features of the data deemed important by the user. It can be applied to any kind of categorical data (nucleotides, amino acids, morphological characters), to protein-coding genes, RNA, introns, exons, as well as ultra-conserved elements (UCEs). It can be especially useful for sequences derived from introns or UCEs, where *a priori* partitioning is difficult, as one does not need to provide user-defined partitions.

A possible explanation for why TIGER partitioning performed better than partitioning by codon position is that there are significant differences among sites belonging to the same codon position of the same gene in their relative evolutionary rate (Figs 2, S1), and this leads to high heterogeneity in the data when they are simply grouped by codon position. Since our method groups sites based on the pattern present in the alignment, the models of molecular evolution have to account for less variation within each partition.

In all of our analyses, partitioning by gene was much worse than the other strategies. A protein-coding gene, with its first, second, and third codon positions, each of which evolve differently, is highly heterogeneous, and applying the same model to such a sequence most likely leads to an underparameterized model. It has been demonstrated that underpartitioning can result

229 in in a more severe error in most datasets than overpartitioning (Brown & Lemmon, 2007; Ward
230 et al., 2010; Kainer & Lanfear, 2015), and our recommendation is to take this into account when
231 devising a partitioning strategy.

232 Our partitioning method has been applied in analyses of several other lepidopteran
233 datasets: 1) the subfamily Acronictinae (Noctuidae) (Rota et al., 2016) analyzed in MrBayes
234 (Ronquist et al., 2012) and RAxML (Stamatakis, 2014); 2) an expanded dataset for the family
235 Choreutidae (Rota, Pena & Miller, 2016) in MrBayes, RAxML, and BEAST (Drummond et al.,
236 2012); 3) the family Hesperiidae (skippers) (Sahoo et al., 2017) in BEAST; and 4) for inferring
237 relationships among Ditrysian superfamilies and families using molecular and morphological
238 characters (Heikkilä et al., 2015) in RAxML. In the Heikkilä et al. study (2015), in addition to
239 applying our partitioning method, the authors also explored the effect of exclusion of fastest
240 evolving characters from the analyses. They found that phylogenetic signal was lost especially
241 when the fastest evolving morphological characters were excluded, and that branch support was
242 lowered with the exclusion of fastest evolving molecular characters, which also resulted in a
243 spurious placement of some groups, and therefore is not at all recommended (see Simmons &
244 Gatesy, 2016 for a detailed exploration of this topic).

245 An issue we would like to stress with our approach, however, is that it should only be
246 applied to studies where concatenation of data is justified, i.e. where gene tree/species tree
247 problems are minimized. This is because our approach of partitioning by specific properties of
248 each character removes any connections between characters belonging to the same gene region.
249 This reshuffling of characters based on relative rates of evolution does have a biological basis to
250 it (sites evolving at a similar rate are modelled together), but at the risk of losing other
251 biologically relevant information (such as differential evolutionary histories of gene regions). We
252 do feel that for studies looking at deeper relationships, such as among genera, tribes, families, and
253 orders, our approach is very useful and overcomes problems of overpartitioning for large

multigene datasets that might be partitioned by codon position, as well as underpartitioning when users might be inclined to analyse their data unpartitioned because they are uncertain of how to partition *a priori*.

Conclusions

Here we present a way of partitioning data based on relative rates of evolution as calculated by TIGER (Cummins & McInerney, 2011). We find that this approach works better than the traditional approaches to partitioning in all of our test cases. Further utility of TIGER calculated rates and RatePartitions needs to be ascertained on other datasets. The program could certainly be used on amino acid (or any other categorical) data in the same way as done here for nucleotides. However, to establish how useful partitioning based on TIGER calculated rates is for phylogenomic data containing sequences from hundreds or thousands of genes, additional testing needs to be conducted.

Funding

This work was supported by the Kone Foundation (JR and TM), Academy of Finland (NW) and the Swedish Research Council (NW).

Author contributions

Jadranka Rota conceived and designed the experiments, performed the experiments, analysed the data, wrote the paper, prepared figures and/or tables, and reviewed drafts of the paper.

Tobias Malm conceived the RatePartition algorithm, analysed the data and reviewed drafts of the manuscript.

Niklas Wahlberg conceived and designed the experiments, performed the experiments, analysed the data, wrote the paper, prepared figures, and reviewed drafts of the paper.

References

Abdo Z, Minin VN, Joyce P, and Sullivan J. 2005. Accounting for uncertainty in the tree topology has little effect on the decision-theoretic approach to model selection in phylogeny estimation. *Molecular Biology and Evolution* 22:691-703. 10.1093/molbev/msi050

Baca SM, Toussaint EFA, Miller KB, and Short AEZ. 2017. Molecular phylogeny of the aquatic beetle family Noteridae (Coleoptera: Adephaga) with an emphasis on data partitioning strategies. *Molecular Phylogenetics and Evolution* 107:282-292. 10.1016/j.ympev.2016.10.016

Brandley MC, Schmitz A, and Reeder TW. 2005. Partitioned Bayesian analyses, partition choice, and the phylogenetic relationships of scincid lizards. *Systematic Biology* 54:373–390.

Brown JM, and Lemmon AR. 2007. The importance of data partitioning and the utility of Bayes factors in Bayesian phylogenetics. *Systematic Biology* 56:643–655.

Cummins CA, and McInerney JO. 2011. A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. *Systematic Biology* 60:833-844. 10.1093/sysbio/syr064

Drummond AJ, Suchard MA, Xie D, and Rambaut A. 2012. Bayesian phylogenetics with BEAUTi and BEAST 1.7. *Molecular Biology and Evolution* 29:1969-1973.

Edger PP, Heidel-Fischer HM, Bekaert M, Rota J, Glöckner G, Platts AE, Heckel DG, Der JP, Wafula EK, Tang M, Hofberger JA, Smithson A, Hall JC, Blanchette M, Bureau TE, Wright SI, dePamphilis CW, Schranz ME, Barker MS, Conant GC, Wahlberg N, Vogel H, Pires JC, and Wheat CW. 2015. The butterfly plant arms-race escalated by gene and genome duplications. *Proceedings of the National Academy of Sciences, USA* 112:8362–8366. 10.1073/pnas.1503926112

Frandsen PB, Calcott B, Mayer C, and Lanfear R. 2015. Automatic selection of partitioning schemes for phylogenetic analyses using iterative k-means clustering of site rates. *BMC Evolutionary Biology* 15. 10.1186/s12862-015-0283-7

Heikkilä M, Mutanen M, Kekkonen M, and Kaila L. 2014. Morphology reinforces proposed molecular phylogenetic affinities: a revised classification for Gelechioidea (Lepidoptera). *Cladistics* 30:563-589. 10.1111/cla.12064

Heikkilä M, Mutanen M, Wahlberg N, Sihvonen P, and Kaila L. 2015. Elusive ditrysian phylogeny: an account of combining systematized morphology with molecular data (Lepidoptera). *BMC Evolutionary Biology* 15:27. 10.1186/s12862-015-0520-0

Huelsenbeck JP, and Rannala B. 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Systematic Biology* 53:904–913.

Kaila L, Epstein ME, Heikkilä M, and Mutanen M. 2013. The assignment of Prodidactidae to Hyblaeoidea, with remarks on Thyridoidea (Lepidoptera). *Zootaxa* 3682:485–494.

Kainer D, and Lanfear R. 2015. The Effects of Partitioning on Phylogenetic Inference. *Molecular Biology and Evolution* 32:1611-1627. 10.1093/molbev/msv026

Kodandaramaiah U, Peña C, Braby MF, Grund R, Müller CJ, Nylin S, and Wahlberg N. 2010. Phylogenetics of Coenonymphina (Nymphalidae: Satyrinae) and the problem of rooting rapid radiations. *Molecular Phylogenetics and Evolution* 54:386-394. 10.1016/j.ympev.2009.08.012

Kristensen NP, Hilton DJ, Kallies A, Milla L, Rota J, Wahlberg N, Wilcox SA, Glatz RV, Young DA, Cocking G, Edwards T, Gibbs GW, and Halsey M. 2015. A new extant family of primitive moths from Kangaroo Island, Australia, and its significance for understanding early Lepidoptera evolution. *Systematic Entomology* 40:5-16.

323 Lanfear R, Calcott B, Ho SYW, and Guindon S. 2012. PartitionFinder: Combined Selection of
324 Partitioning Schemes and Substitution Models for Phylogenetic Analyses. *Molecular Biology*
325 *and Evolution* 29:1695-1701. 10.1093/molbev/mss020
326 Lemmon AR, and Moriarty EC. 2004. The importance of proper model assumption in Bayesian
327 phylogenetics. *Systematic Biology* 53:265–277.
328 Matos-Maravi P, Aguila RN, Pena C, Miller JY, Sourakov A, and Wahlberg N. 2014. Causes of
329 endemic radiation in the Caribbean: evidence from the historical biogeography and
330 diversification of the butterfly genus Calisto (Nymphalidae: Satyrinae: Satyrini). *BMC*
331 *Evolutionary Biology* 14. 199
332 10.1186/s12862-014-0199-7
333 Miller KB, Bergsten J, and Whiting MF. 2009. Phylogeny and classification of the tribe
334 Hydaticini (Coleoptera: Dytiscidae): partition choice for Bayesian analysis with multiple
335 nuclear and mitochondrial protein-coding genes. *Zoologica Scripta* 38:591-615.
336 10.1111/j.1463-6409.2009.00393.x
337 Nylander JAA. 2004. MrModeltest v2. Evolutionary Biology Centre, Uppsala University:
338 Program distributed by the author.
339 Nylander JAA, Ronquist F, Huelsenbeck JP, and Nieves-Aldrey JL. 2004. Bayesian phylogenetic
340 analysis of combined data. *Systematic Biology* 53:47–67.
341 Ounap E, Viidalepp J, and Truuverk A. 2016. Phylogeny of the subfamily Larentiinae
342 (Lepidoptera: Geometridae): integrating molecular data and traditional classifications.
343 *Systematic Entomology* 41:824-843. 10.1111/syen.12195
344 Penz CM, Devries PJ, and Wahlberg N. 2012. Diversification of Morpho butterflies (Lepidoptera,
345 Nymphalidae): a re-evaluation of morphological characters and new insight from DNA
346 sequence data. *Systematic Entomology* 37:670-685. 10.1111/j.1365-3113.2012.00636.x
347 Rajaei H, Greve C, Letsch H, Stuning D, Wahlberg N, Minet J, and Misof B. 2015. Advances in
348 Geometroidea phylogeny, with characterization of a new family based on Pseudobiston
349 pinratanai (Lepidoptera, Glossata). *Zoologica Scripta* 44:418-436. 10.1111/zsc.12108
350 Rönkä K, Mappes J, Kaila L, and Wahlberg N. 2016. Putting Parasemia in its phylogenetic place:
351 a molecular analysis of the subtribe Arctiina (Lepidoptera). *Systematic Entomology* 41:844-
352 853. 10.1111/syen.12194
353 Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L,
354 Suchard MA, and Huelsenbeck JP. 2012. MrBayes 3.2: Efficient Bayesian phylogenetic
355 inference and model choice across a large model space. *Systematic Biology* 61:539-542.
356 10.1093/sysbio/sys029
357 Rota J. 2011. Data partitioning in Bayesian analysis: molecular phylogenetics of metalmark
358 moths (Lepidoptera: Choreutidae). *Systematic Entomology* 36:317-329. 10.1111/j.1365-
359 3113.2010.00563.x
360 Rota J, and Miller SE. 2013. New genus of metalmark moths (Lepidoptera: Choreutidae) with
361 Afrotropical and Australasian distribution. *ZooKeys* 355:29-47. 10.3897/zookeys.355.6158
362 Rota J, Pena C, and Miller SE. 2016. The importance of long-distance dispersal in small insects:
363 historical biogeography of metalmark moths (Lepidoptera, Choreutidae). *Journal of*
364 *Biogeography* 43:1254-1265. 10.1111/jbi.12721
365 Rota J, and Wahlberg N. 2012. Exploration of data partitioning in an eight-gene data set:
366 phylogeny of metalmark moths (Lepidoptera, Choreutidae). *Zoologica Scripta* 41:536-546.
367 10.1111/j.1463-6409.2012.00551.x
368 Rota J, Zacharczenko BV, Wahlberg N, Zahiri R, Schmidt BC, and Wagner DL. 2016.
369 Phylogenetic relationships of Acronictinae with discussion of the abdominal courtship brush
370 in Noctuidae (Lepidoptera). *Systematic Entomology* 41:416–429. 10.1111/syen.12162

371 Sahoo RK, Warren AD, Collins SC, and Kodandaramaiah U. 2017. Hostplant change and
372 paleoclimatic events explain diversification shifts in skipper butterflies (Family: Hesperiidae).
373 *BMC Evolutionary Biology* 17:174. 10.1186/s12862-017-1016-x
374 Sahoo RK, Warren AD, Wahlberg N, Brower AVZ, Lukhtanov VA, and Kodandaramaiah U.
375 2016. Ten genes and two topologies: an exploration of higher relationships in skipper
376 butterflies (Hesperiidae). *Peerj* 4. 10.7717/peerj.2653
377 Sihvonen P, Mutanen M, Kaila L, Brehm G, Hausmann A, and Staude HS. 2011. Comprehensive
378 molecular sampling yields a robust phylogeny for geometrid moths (Lepidoptera:
379 Geometridae). *PlosOne* 6:e20356. 10.1371/journal.pone.0020356
380 Simmons MP, and Gatesy J. 2016. Biases of tree-independent-character-subsampling methods.
381 *Molecular Phylogenetics and Evolution* 100:424-443. 10.1016/j.ympev.2016.04.022
382 Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
383 phylogenies. *Bioinformatics* 30:1312-1313. 10.1093/bioinformatics/btu033
384 van Nieukerken EJ, Kaila L, Kitching IJ, Kristensen NP, Lees DC, Minet J, Mitter C, Mutanen
385 M, Regier JC, Simonsen TJ, Wahlberg N, Yen S-H, Zahiri R, Adamski D, Baixeras J, Bartsch
386 D, Bengtsson BÅ, Brown JW, Bucheli SR, Davis DR, De Prins J, De Prins W, Epstein ME,
387 Gentili-Poole P, Gielis C, Hättenschwiler P, Hausmann A, Holloway JD, Kallies A, Karsholt
388 O, Kawahara A, Koster JC, Kozlov M, Lafontaine JD, Lamas G, Landry J-F, Lee S, Nuss M,
389 Park K-T, Penz C, Rota J, Schmidt BC, Schintlmeister A, Sohn JC, Solis MA, Tarmann GM,
390 Warren AD, Weller S, Yakovlev RV, Zolotuhin VV, and Zwick A. 2011. Order Lepidoptera.
391 In: Zhang Z-Q, ed. *Animal biodiversity: An outline of higher-level classification and survey*
392 *of taxonomic richness*: Zootaxa, 212–221.
393 Wahlberg N, Rota J, Braby MF, Pierce NP, and Wheat CW. 2014. Revised systematics and higher
394 classification of pierid butterflies (Lepidoptera: Pieridae) based on molecular data. *Zoologica*
395 *Scripta* 43:641-650. 10.1111/zsc.12075
396 Wahlberg N, and Wheat CW. 2008. Genomic outposts serve the phylogenomic pioneers:
397 designing novel nuclear markers for genomic DNA extractions of Lepidoptera. *Systematic*
398 *Biology* 57:231-242. 10.1080/10635150802033006
399 Ward PS, Brady SG, Fisher BL, and Schultz TR. 2010. Phylogeny and Biogeography of
400 Dolichoderine Ants: Effects of Data Partitioning and Relict Taxa on Historical Inference.
401 *Systematic Biology* 59:342-362. 10.1093/sysbio/syq012
402 Zahiri R, Lafontaine D, Schmidt C, Holloway JD, Kitching IJ, Mutanen M, and Wahlberg N.
403 2013. Relationships among the basal lineages of Noctuidae (Lepidoptera, Noctuoidea) based
404 on eight gene regions. *Zoologica Scripta* 42:488-507. 10.1111/zsc.12022

Tables

Table 1. Datasets analysed. List of analysed datasets providing the reference, the number of sampled taxa and gene regions in the dataset, and the length of the dataset in base pairs (bp).

Table 2. The amount of missing data in each of the eight datasets analysed. All alignment columns were pulled into one of the ten categories based on the range of missing data being 0–10%, 10–20%, etc. to more than 90% missing. The Cumulative missing data refers to summing percentage of missing data from one range category to the next. All datasets had 1% or less of columns in the alignment with missing more than 80% of data, and overall all datasets had 50% or more columns with less than 40% of missing data.

Table 3. List of partitioning strategies evaluated for each of the analysed datasets. TIGER refers to the program that assigns each site in the alignment a relative evolutionary rate, and d is the division factor in the RatePartitions script used to group sites into subsets based on their relative evolutionary rates. See text for more details.

Table 4. The number of partitions for each dataset and partitioning strategy. *Gene* refers to partitioning by gene fragment, *Codon* to partitioning by codon position, and *TIG* to partitioning by relative evolutionary rate as estimated with the program TIGER with different values for the *d*, division factor in the RatePartitions script. See Table 3 and text for more details.

Figures

Figure 1. A comparison of BIC values for the 14 partitioning strategies tested in all eight datasets. The partitioning strategies are plotted on the horizontal axis, and the BIC values are plotted on the vertical axis. The lower the BIC value, the better the partitioning strategy. *Gene*

refers to partitioning by gene fragment, *Codon* to partitioning by codon position, and *TIG* to partitioning by relative evolutionary rate as estimated with the program TIGER with different values for the *d*, division factor in the RatePartitions script. See text and Table 3 for more details.

Figure 2. Relative evolutionary rate estimates for codon positions in the Noctuidae dataset. Plots are showing the assigned TIGER relative evolutionary rates for codon positions of each of the eight genes in the Noctuidae dataset. TIGER rates are shown on the horizontal axis, and the number of codon positions that were assigned the rate between 0.0–0.1, 0.1–0.2, etc. is shown on the vertical axis. The lower the number, the higher the rate of evolution, with rate of 1 being assigned to invariable sites in the alignment. As expected, most of the first and second codon positions received the rate of 1, but there are exceptions, with some first and/or second codon positions receiving a relatively low rate (especially in e.g. COI, and wgl). Likewise, most of the third codon positions received lower rates, but in some genes (e.g. EF-1α), the number of third positions that received the TIGER rate of 1 is relatively high. Such plots for the other seven datasets are in supplemental information files Fig. S1.

Supplemental Information

Supplemental Data File S1. Datasets analysed in this study. The datasets are provided in the PHYLIP format, together with the RAxML style partition definitions for the best partitioning strategy.

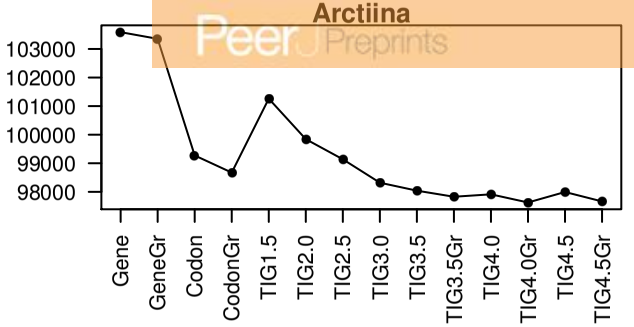
Supplemental Script S1. RatePartitions script. Python script for grouping sites in the alignment based on the relative evolutionary rate assigned by the program TIGER.

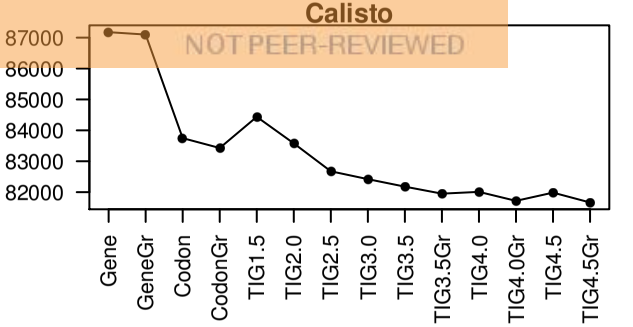
Supplemental Table S1. Comparison of BIC values. A BIC value is provided for each partitioning strategy for each of the eight datasets analysed.

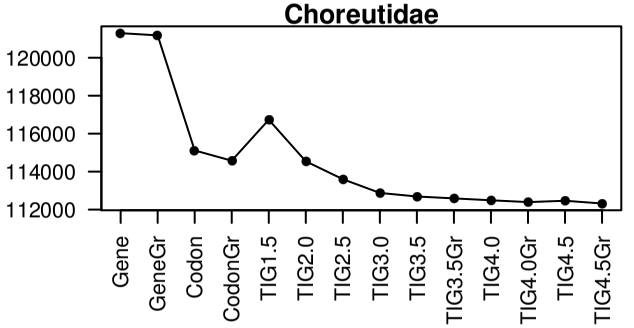
Supplemental Figure S1. Relative evolutionary rate estimates for codon positions of all gene fragments. Plots showing the assigned TIGER relative evolutionary rates for codon positions of each of the gene fragments analysed: Arctiina (a), *Calisto* (b), Choreutidae (c), Coenonymphina (d), Geometridae (e), *Morpho* (f), and Pieridae (g).

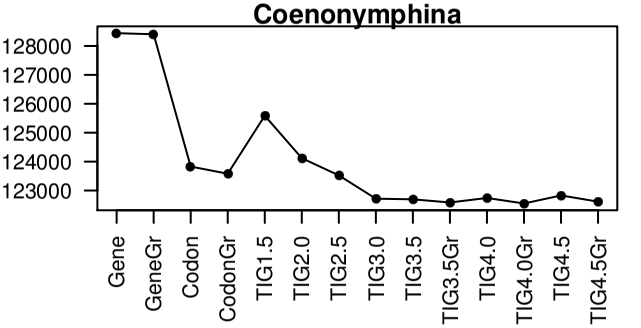
Figure 1(on next page)

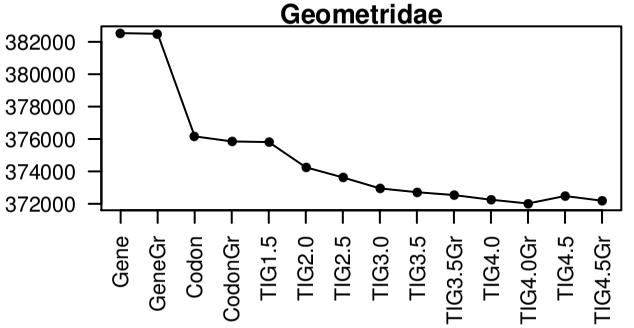
A comparison of BIC values for the 14 partitioning strategies tested in all eight datasets

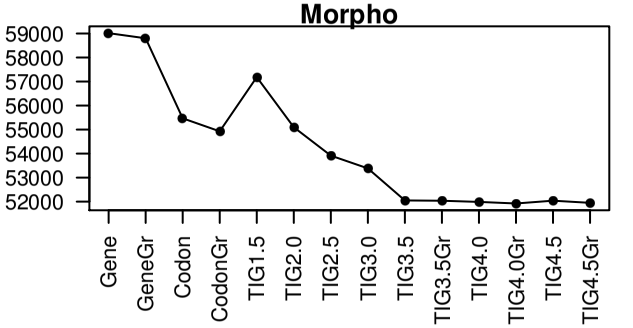


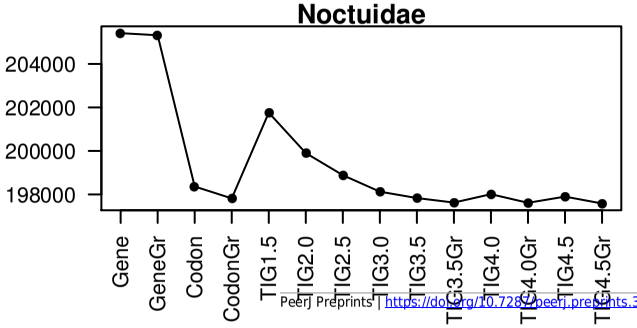












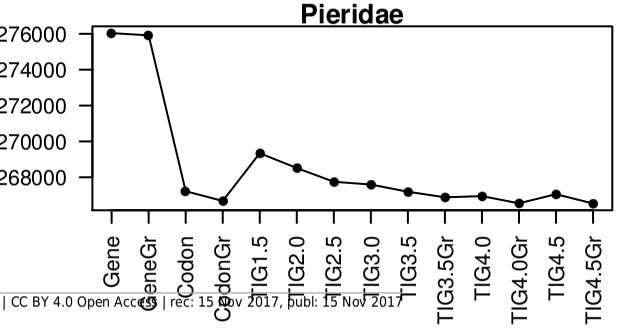
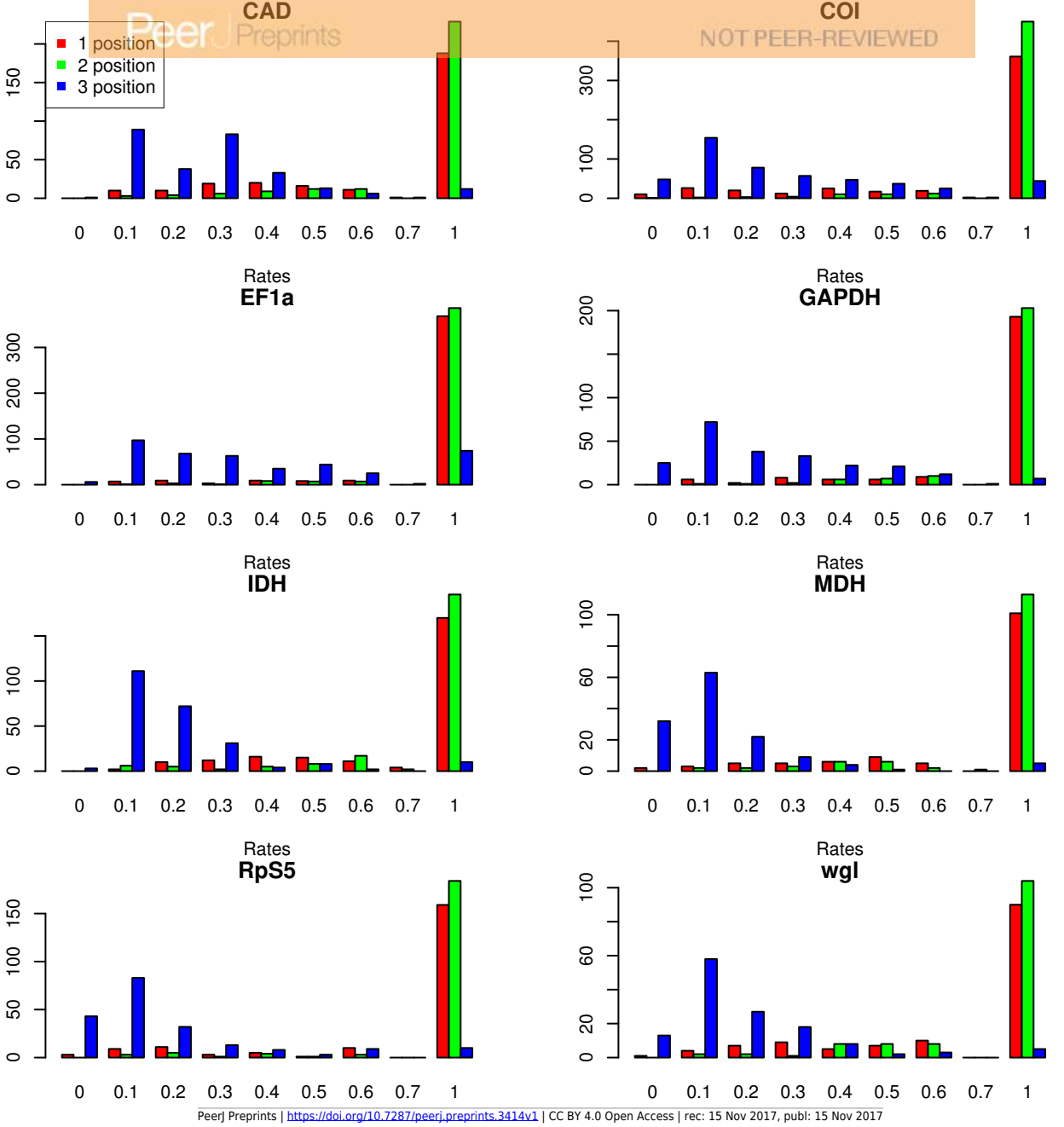


Figure 2(on next page)

Relative evolutionary rate estimates for codon positions in the Noctuidae dataset



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3414v1 | CC BY 4.0 Open Access | rec: 15 Nov 2017, publ: 15 Nov 2017

Table 1(on next page)

Datasets analysed.

List of analysed datasets providing the reference, the number of sampled taxa and gene regions in the dataset, and the length of the dataset in base pairs (bp).

Table 1.

Taxon	Study	No. taxa	No. genes	base pairs
Arctiina	Rönkä et al. 2016	113	8	5809
Calisto	Matos-Maravi et al. 2014	90	6	5297
Choreutidae	Rota & Wahlberg 2012	41	8	6293
Coenonymphina	Kodandaramaiah et al. 2010	69	5	4435
Geometridae	Sihvonen et al. 2011	164	8	5998
Morpho	Penz et al. 2012	31	8	6372
Noctuidae	Zahiri et al. 2013	78	8	6365
Pieridae	Wahlberg et al. 2014	110	8	6247

Table 2 (on next page)

The amount of missing data in each of the eight datasets analysed.

All alignment columns were pulled into one of the ten categories based on the range of missing data being 0–10%, 10–20%, etc. to more than 90% missing. The Cumulative missing data refers to summing percentage of missing data from one range category to the next. All datasets had 1% or less of columns in the alignment with missing more than 80% of data, and overall all datasets had 50% or more columns with less than 40% of missing data.

Table 2.

Missing data range	Arctiina	Calisto	Choreutidae	Coenonymphina	Geometridae	Morpho	Noctuidae	Pieridae
0-10%	11%	10%	37%	47%	10%	16%	45%	28%
10-20%	10%	1%	26%	46%	26%	26%	22%	16%
20-30%	14%	5%	1%	5%	29%	20%	12%	35%
30-40%	22%	33%	27%	1%	15%	8%	4%	8%
40-50%	12%	24%	2%	0%	5%	15%	9%	3%
50-60%	17%	22%	3%	0%	3%	3%	5%	5%
60-70%	2%	2%	4%	0%	13%	6%	1%	2%
70-80%	13%	2%	0%	0%	1%	6%	1%	2%
80-90%	0%	0%	0%	1%	0%	0%	0%	0%
90-100%	0%	0%	0%	0%	0%	0%	1%	0%
Cumulative missing data								
0-10%	11%	10%	37%	47%	10%	16%	45%	28%
0-20%	21%	11%	62%	93%	36%	42%	67%	44%
0-30%	34%	16%	63%	97%	64%	62%	79%	79%
0-40%	56%	50%	90%	99%	79%	70%	84%	88%
0-50%	68%	74%	93%	99%	84%	85%	92%	91%
0-60%	85%	96%	96%	99%	87%	88%	97%	96%
0-70%	87%	98%	100%	99%	99%	94%	98%	98%
0-80%	100%	100%	100%	99%	100%	100%	99%	100%
0-90%	100%	100%	100%	100%	100%	100%	99%	100%
100%	100%	100%	100%	100%	100%	100%	100%	100%

Table 3(on next page)

List of partitioning strategies evaluated for each of the analysed datasets.

TIGER refers to the program that assigns each site in the alignment a relative evolutionary rate, and d is the division factor in the RatePartitions script used to group sites into subsets based on their relative evolutionary rates. See text for more details.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3414v1 | CC BY 4.0 Open Access | rec: 15 Nov 2017, publ: 15 Nov 2017

Table 3.

Partitioning strategy	Description
Gene	each gene fragment as separate subset
GeneGr	as above but with PF greedy algorithm combined into similar subsets
Codon	each codon position of each gene as separate subset
CodonGr	as above but with PF greedy algorithm combined into similar subsets
TIG1.5	TIGER partitioning strategy with d=1.5
TIG2.0	TIGER partitioning strategy with d=2.0
TIG2.5	TIGER partitioning strategy with d=2.5
TIG3.0	TIGER partitioning strategy with d=3.0
TIG3.5	TIGER partitioning strategy with d=3.5
TIG3.5Gr	as above but with PF greedy algorithm combined into similar subsets
TIG4.0	TIGER partitioning strategy with d=4.0
TIG4.0Gr	as above but with PF greedy algorithm combined into similar subsets
TIG4.5	TIGER partitioning strategy with d=4.5
TIG4.5Gr	as above but with PF greedy algorithm combined into similar subsets

Table 4(on next page)

The number of partitions for each dataset and partitioning strategy.

Gene refers to partitioning by gene fragment, *Codon* to partitioning by codon position, and *TIG* to partitioning by relative evolutionary rate as estimated with the program TIGER with different values for the *d*, division factor in the RatePartitions script. See Table 3 and text for more details.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3414v1 | CC BY 4.0 Open Access | rec: 15 Nov 2017, publ: 15 Nov 2017

Table 4.

Partitioning strategy	Arctiina	Calisto	Choreutidae	Coenonymphina	Geometridae	Morpho	Noctuidae	Pieridae
Gene	8	6	8	5	8	8	8	8
GeneGr	3	4	4	4	6	4	6	5
Codon	24	18	24	15	24	24	24	24
CodonGr	9	7	10	9	15	7	12	12
TIG1.5	4	4	3	4	7	2	4	6
TIG2.0	5	6	5	5	10	3	6	8
TIG2.5	7	7	6	7	13	4	8	11
TIG3.0	8	9	7	8	15	5	9	13
TIG3.5	10	10	8	9	18	5	11	16
TIG3.5Gr	6	5	6	6	12	4	7	10
TIG4.0	11	12	9	11	21	6	13	18
TIG4.0Gr	5	6	6	7	12	4	6	9
TIG4.5	12	13	10	12	24	7	14	20
TIG4.5Gr	5	7	6	7	16	4	6	8

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3414v1 | CC BY 4.0 Open Access | rec: 15 Nov 2017, publ: 15 Nov 2017