Systematic Entomology

Systematic Entomology (2011), **36**, 31–43 DOI: 10.1111/j.1365-3113.2010.00543.x

Increased gene sampling yields robust support for higher-level clades within Bombycoidea (Lepidoptera)

ANDREAS ZWICK[1,2], JEROME C. REGIER[1,3,4], CHARLES MITTER[3] and MICHAEL P. CUMMINGS[5]

[1]Center for Biosystems Research, University of Maryland Biotechnology Institute, College Park, MD, U.S.A., [2]State Museum of Natural History Stuttgart, Entomology, Stuttgart, Germany, [3]Department of Entomology, University of Maryland, College Park, MD, U.S.A., [4]Institute for Biosciences and Biotechnology Research, University of Maryland, College Park, MD, U.S.A. and [5]Laboratory of Molecular Evolution, Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD, U.S.A.

Abstract. This study has as its primary aim the robust resolution of higher-level relationships within the lepidopteran superfamily Bombycoidea. Our study builds on an earlier analysis of five genes (~6.6 kbp) sequenced for 50 taxa from Bombycoidea and its sister group Lasiocampidae, plus representatives of other macrolepidoteran superfamilies. The earlier study failed to yield strong support for the monophyly of and basal splits within Bombycoidea, among others. Therefore, in an effort to increase support specifically for higher-level nodes, we generated 11.7 kbp of additional data from 20 genes for 24 of 50 bombycoid and lasiocampid taxa. The data from the genes are all derived from protein-coding nuclear genes previously used to resolve other lepidopteran relationships. With these additional data, all but a few higher-level nodes are strongly supported. Given our decision to minimize project costs by augmenting genes for only 24 of the 50 taxa, we explored whether the resulting pattern of missing data in the combined-gene matrix introduced a nonphylogenetic bias, a possibility reported by others. This was achieved by comparing node support values (i.e. nonparametric bootstrap values) based on likelihood and parsimony analyses of three datasets that differ in their number of taxa and level of missing data: 50 taxa/5 genes (dataset A), 50 taxa/25 genes (dataset B) and 24 taxa/25 genes (dataset C). Whereas datasets B and C provided similar results for common nodes, both frequently yielded higher node support relative to dataset A, arguing that: (i) more data yield increased node support and (ii) partial gene augmentation does not introduce an obvious nonphylogenetic bias. A comparison of single-gene bootstrap analyses identified four nodes for which one or two of the 25 genes provided modest to strong support for a grouping not recovered by the combined-gene result. As a summary proposal, two of these four groupings (one each within Bombycoidea and Lasiocampidae) were deemed sufficiently problematic to regard them as unresolved trichotomies. Since the alternative groupings were always highly localized on the tree, we did not judge a combined-gene analysis to present a problem outside those regions. Based on our robustly resolved results, we have revised the classification of Bombycoidea: the family Bombycidae is restricted to its nominate subfamily, and its tribe Epiini is elevated to subfamily rank (Epiinae **stat.rev.**), whereas the bombycid subfamily Phiditiinae is reinstated as a separate family (Phiditiidae **stat.rev.**). The bombycid subfamilies Oberthueriinae

Correspondence: Andreas Zwick, State Museum of Natural History Stuttgart, Entomology, Rosenstein 1, D-70191 Stuttgart, Germany. E-mail: andreas.zwick@smns-bw.de

Kuznetzov & Stekolnikov, 1985, **syn.nov.** and Prismostictinae Forbes, 1955, **syn.nov.**, and the family Mirinidae Kozlov, 1985, **syn.nov.** are established as subjective junior synonyms of Endromidae Boisduval, 1828. The family Anthelidae (Lasiocampoidea) is reincluded in the superfamily Bombycoidea.

Introduction

Among Lepidoptera, the superfamily Bombycoidea sensu stricto (hereafter referred to as 'Bombycoidea' and 'bombycoids') has garnered disproportionate interest from experimentalists (see Goldsmith & Wilkins, 1995; Goldsmith & Marec, 2010), and includes numerous model organisms and the first complete lepidopteran genome sequence (Xia *et al.*, 2004, 2009). A robust higher-level phylogeny of bombycoids would provide a valuable comparative framework for the interpretation of previous and ongoing studies. This report focuses on our continued efforts to robustly resolve higher-level relationships within the Bombycoidea by greatly expanding the dataset of Regier *et al.* (2008a).

A recent molecular phylogenetic study (Regier *et al.*, 2008a) of 38 bombycoid species representing most subfamilies and tribes plus 28 affiliated Macrolepidoptera, all sequenced for five protein-coding nuclear genes (∼6.75 kb/taxon), yielded strong support for numerous clades within Bombycoidea (summarized in Fig. 1, left side), and resulted in substantial differences from an earlier morphology-based phylogenetic proposal (Minet, 1994; Lemaire & Minet, 1998). In particular, the polyphyly of the nominate family Bombycidae, sensu Minet (1994) and Lemaire & Minet (1998), found to comprise five distantly related groups, was strongly supported. None of the remaining bombycoid families was polyphyletic within the limits of taxon sampling, but the family Anthelidae, previously placed within Lasiocampoidea, was strongly supported as deeply nested within Bombycoidea. Additionally, two suprafamily-level groups (informally labelled as 'CAPOPEM' and 'BALE') and some relationships therein were strongly supported in the all-nucleotide analysis. However, other higher-level groupings (e.g. relationships among the well-supported Saturniidae, Bombycinae, Sphingidae, 'CAPOPEM' and 'BALE' groups) received low bootstrap support (<50%), and showed sensitivity to the method of analysis. Even the superfamily itself received low bootstrap support in the five-gene study, although a subsequent study that included three divergent bombycoids sampled for 26 genes, plus ten more bombycoids sampled for five genes, yielded somewhat stronger support for Bombycoidea (62% bootstrap) and strong support (85% bootstrap) for its sister-group relationship with Lasiocampidae (seven taxa sampled), when analysed together with 104 other species of diverse ditrysian Lepidoptera (Cho *et al.*, 2010).

The increased node support within Ditrysia resulting from increased gene sampling (Cho *et al.*, 2010, as compared with Regier *et al.*, 2009) is encouraging, and suggests that expanded gene sampling specifically within Bombycoidea and its sister group Lasiocampidae might prove similarly useful. However, various designs of such an expanded gene-sampling approach are possible, with three obvious ones being: (i) expanded gene sampling of all bombycoid and lasiocampid taxa previously

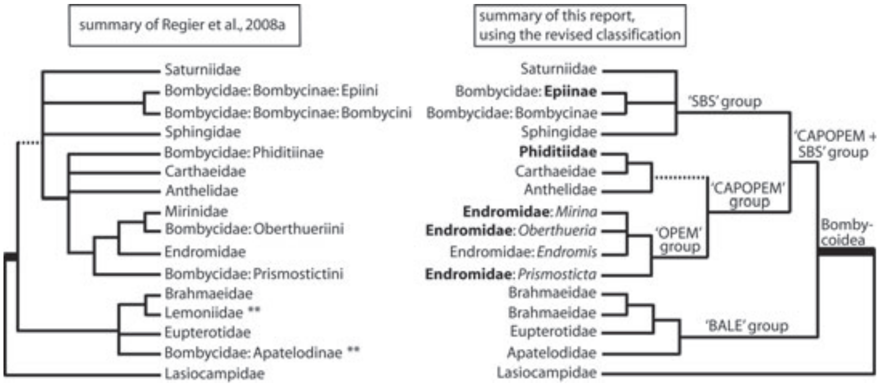

Fig. 1. Summaries of higher-level phylogenetic relationships within Bombycoidea. The tree on the left is a partial redrawing of fig. 1 from Regier *et al.* (2008a), in which taxa were sampled for five genes. The tree on the right is based on our best maximum likelihood topology generated in the current 25-gene study (dataset B), using either the codon or degen1 datasets, except that the Oberthueriini + Endromidae + Mirinidae group and the 'SBS' group are left unresolved (see text for explanation). Dashed lines identify groups that are favoured but that do not receive strong support. The classificatory names on the left correspond to those of Minet (1994) and Lemaire & Minet (1998), as used in Regier *et al.* (2008a). **Formally altered by Zwick (2008): Lemoniidae was synonymized with Brahmaeidae, and Apatelodinae was elevated to Apatelodidae. Throughout the remainder of the current report we use these two revised names rather than those in Regier *et al.* (2008a). The classificatory names on the right include six changes (five shown in bold plus the restriction of Bombycidae to Bombycinae) that are based on results of the current study.

sampled for five genes in Regier *et al.* (2008a); (ii) expanded gene sampling for at least two, but not necessarily all, of the terminal taxa that represent each of the higher-level bombycoid and lasiocampid groups of interest, while still including in the analysis the remaining taxa sequenced only for five genes; and (iii) expanded gene sampling of a subset of such taxa as in design no. 2, but excluding taxa sequenced only for five genes. Although the first, most complete implementation would seem the obvious preference, its downside is that it requires the maximal outlay of resources, which are frequently limiting, as in this study, without any certainty that all of the extra effort would even be needed to achieve strong support. The second implementation has the advantage that higher-level groups are still multiply sampled, and that no previous data are excluded, but the resulting data matrix will thereby have blocks of missing data, which may (Lemmon *et al.*, 2009) or may not (Wiens, 2003, 2006) compromise phylogenetic accuracy by introducing a nonphylogenetic bias. The third implementation minimizes missing data but at the expense of discarding sequence data and taxa. The current report compares the benefits of the latter two designs relative to each other, and to an initial 50-taxon/five-gene dataset of Bombycoidea and Lasiocampidae very similar to that in Regier *et al.* (2008a). This is accomplished by augmenting 24 of these 50 taxa already sequenced for five genes with data from 20 additional genes, representing a nearly three-fold increase in the overall size of the data matrix.

A similar, but higher-level, comparison of experimental designs incorporating incomplete gene sampling, this time across 123 Ditrysia (a group that comprises 98% of all lepidopteran species), but for gene sets nearly identical to this study (five genes for all taxa versus 26 genes for a subset of 41 taxa), has recently been published (Cho *et al.*, 2010). The results show that incomplete gene augmentation (analogous to design no. 2 above) and complete gene augmentation (analogous to design no. 3) both yielded consistently, sometimes dramatically, higher bootstrap support than the original 123-species/five-gene data matrix for groups represented by at least two species in all data matrices, while introducing no strongly supported conflicts between them. However, further empirical tests, particularly at higher (e.g. across Arthropoda: see Regier *et al.*, 2008b) and lower levels (e.g. this study) remain necessary to explore the empirical circumstances under which the potential advantages of biased gene sampling designs might hold in practice.

Another feature of data matrix design that has received recent consideration is based on the distinction between synonymous and nonsynonymous character change. On average, synonymous change occurs more rapidly, leading to multiple substitutions per site and nonhomogeneous base composition, which in turn can degrade the phylogenetic signal. For example, the major taxonomic finding in Cho *et al.* (2010), namely, the identification of Gracillarioidea + Yponomeutoidea as sister group to all other Ditrysia, only received strong support from nonsynonymous change. By contrast, the current study explores more recent divergences, within Bombycoidea and Lasiocampidae, and might benefit more from synonymous change, given that fewer nonsynonymous changes would have accumulated. Whether greater reliance on synonymous change presents analytical problems at this level or not is an empirical question that this study addresses.

A separate issue that was noted in the earlier five-gene study (Regier *et al.*, 2008a) was the strongly conflicting support for two unrelated nodes – one within Bombycoidea and one within Lasiocampidae – from independent analyses of single genes (see also Regier *et al.*, 1998). Among the possible explanations, such intergene conflicts could be the result of incomplete allele sorting, species hybridization or an analytical artifact, for example, arising from the inadequacy of the substitution model. Regardless, they could pose a serious challenge for phylogenetic estimation of the species tree (McCormack *et al.*, 2009), and especially for resolving rapid species radiations, such as may be widespread within Lepidoptera and other insect orders (e.g. Regier *et al.*, 2009). The current study provides further documentation of the occurrence of conflicting signals within our relatively large gene sample.

Materials and methods

Taxon sampling and classification

Our sampling within Bombycoidea covers all recognized families and nearly all subfamilies, except for some subfamilies of Eupterotidae (no Janinae, Striphnopteryginae and Panacelinae were sampled; sensu Nässig & Oberprieler, 2008) and of Anthelidae (no Munychryiinae were sampled); see Table S1 for details. This taxon sampling within the in-group (Bombycoidea) is almost identical to the sampling in Regier *et al.* (2008a), merely differing in the substitution of three saturniid species (Ceratocampinae, *Eacles imperialis*; Saturniinae, *Saturnia mendocino* and *Antheraea polyphemus*), with two equivalent saturniid species for which we have more complete datasets (Ceratocampinae, *Citheronia sepulcralis*; Saturniinae, *Saturnia naessigi*), and an additional subfamily (Agliinae, *Aglia tau*) not sampled in Regier *et al.* (2008a). Unlike Regier *et al.* (2008a), we restrict outgroup sampling in our current analyses to the family Lasiocampidae (18 species sampled, see Table S1), which is strongly supported as the sister group of Bombycoidea (85% bootstrap based on a likelihood analysis that included 13 bombycoids, seven lasiocampids and 103 other diverse species of ditrysian Lepidoptera; Cho *et al.*, 2010; see also Regier *et al.*, 2009 for a complete list of taxa). In that same study, the monophyly of Bombycoidea was supported with 62% bootstrap.

Specimens used for this study and obtained from numerous collectors (see Acknowledgements) are stored at −85°C in 100% ethanol as part of the ATOLep collection at the University of Maryland (details at http://www.leptree.net). DNA 'bar codes' for all specimens, confirming their identities, have been kindly generated by the All-Leps Barcode of Life project (http://www.lepbarcoding.org).

One outcome of the current report is a revised classification of Bombycoidea. Because multiple classification schemes are necessarily discussed, we summarize our usage to avoid confusion. The left side of Fig. 1 summarizes the phylogenetic results of Regier *et al.* (2008a), and it uses the classificatory terms found in Minet (1994) and Lemaire & Minet (1998). This classification is used throughout this paper, but with the subsequent modifications introduced by Zwick (2008), namely, with Lemoniidae being a subjective junior synonym of Brahmaeidae, and the bombycid subfamily Apatelodinae sensu Minet (1994) being regarded as a separate family. Finally, the revised classification based on our current results is discussed in the section 'Revised classification', and is summarized in the right side of Fig. 1 (changes are set in bold) and in Table S1.

Gene sampling

In addition to the portions of five genes (*CAD*, *DDC*, *enolase*, *period* and *wingless*; 6633 bp) used by Regier *et al.* (2008a), and sequenced here for all 50 taxa (with some amplification and sequencing failures noted), we sequenced portions of 20 additional protein-coding nuclear genes (11 688 bp combined) for 24 of the 50 taxa, representing all families, but not all subfamilies, sampled in this study. Therefore, the matrix contains a mixture of up to 18 321 bp (25 genes) for about half of all taxa, representing all major lineages of Bombycoidea, and up to 6633 bp (five genes) for the remaining taxa, representing lineage diversity within larger families, mainly Saturniidae, Sphingidae, Bombycidae and the out-group Lasiocampidae. The proportion of sequence completeness of each individual gene for each species is given in Table S2. GenBank accession numbers for new sequences are listed in Table S3. Gene names can be found in Table S4. For more information about these genes (i.e. putative protein function, amplicon length, rate of nonsynonymous change), see Table 2 of Regier *et al.*, 2008b.

Amplification of nucleic acid sequences and sequence editing

A detailed protocol of all laboratory procedures has been published (Regier *et al.*, 2008b; downloadable under 'Appendices and data' at http://www.systematicbiology.org). PCR primer sequences can be found in Regier *et al.* (2008a, b) (for *CAD*, *DDC*, *enolase*, *period* and *wingless* primers, see Table S2 in Regier *et al.* (2008a); for all others, see Regier *et al.*, 2008b under 'Appendices and data' at http://systematicbiology.org). In summary, templates for DNA sequencing were generated by reverse transcription-polymerase chain reaction of extracts of total nucleic acids. After gel isolation, templates were either sequenced directly or reamplified using one original primer and one new, internal primer, again followed by gel isolation. Sequences were generated on a 3730 DNA Analyzer (Applied Biosystems, Carlsbad, CA). Sequences were edited and assembled using the Pregap4 and Gap4 programs in the Staden package (Staden *et al.*, 2001). Multiple sequence alignments were performed manually on the conceptually translated sequences using the sequence editor Genetic Data Environment 2.2 (Smith *et al.*, 1994). Alignments were generally straightforward, given the overall conservation of the protein-coding sequences. A data-exclusion mask of 387 nt out of 18 708 nt total aligned sequences (i.e. 2.1% of the total) for all 50 species was applied.

Dataset construction

Three combined-gene datasets were constructed in order to explore the effects of incomplete gene augmentation on the robustness of higher-level group recovery (Fig. 2). Dataset A consists of all taxa sequenced for the five genes previously sequenced and analysed in Regier *et al.* (2008a; 50 taxa/five genes; 6633 bp/taxon and 331 650 bp in total; 18.8% missing data), except for the three taxon substitutions within Saturniidae mentioned above. Dataset B consists of dataset A plus new sequences from 20 additional genes generated for a 24-taxon subset (50 taxa/25 genes; 18 321 bp/taxon and 916 050 bp in total; 44.9% missing data). Dataset C consists only of the 24-taxon subset for 25 genes (24 taxa/25 genes; 18 321 bp and 439 705 bp in total; 14.6% missing data). With identical numbers of taxa, datasets A and B are used to assess the effect of increased sequence data. Dataset C provides a check on possible artifacts resulting from the much larger incompleteness of dataset B relative to A, particularly where their results differ. In principle, dataset C may also be used to uncover deeper groupings that are less well supported when many taxa are included. However, because the overall matrix sizes of A and C differ (∼332 kbp versus 440 kbp, respectively), and because bootstrap support is inversely correlated with taxon number, other things being equal (Zharkikh & Li, 1995; Susko, 2009), caution is required in comparing results from datasets C with the others. All three datasets can be found in Appendices S1–3.

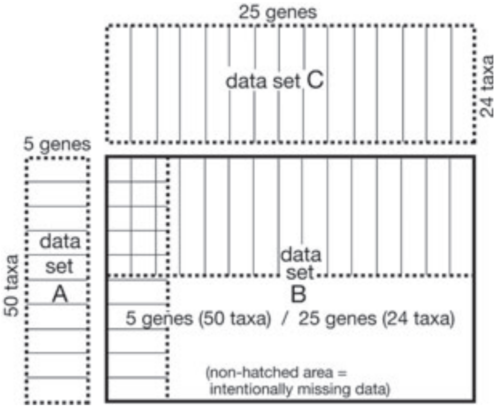
25 genes
data set C
24 taxa
5 genes
data set A
50 taxa
data set B
5 genes (50 taxa) / 25 genes (24 taxa)
(non-hatched area = intentionally missing data)

Fig. 2. Composition of the three datasets used for this study. When combined, dataset A (50 taxa/5 genes) and dataset C (24 taxa/25 genes) yield dataset B (50 taxa/25 genes, some intentionally missing data).

Single-gene datasets were constructed for each of the 25 genes, but only for the taxa for which sequence data were available.

Phylogenetic analyses

For each of the combined-gene datasets (A, B and C), up to six different analyses were performed. Within a likelihood framework, the 'codon analysis' consisted of analysing an entire dataset (minus the masked, unalignable portion, as for all analyses) under a model of codon substitution (Goldman & Yang, 1994), with an underlying general time-reversible model of nucleotide substitution (Lanave *et al.*, 1994; GTR) for individual nucleotide substitutions, observed codon frequencies and four estimated nonsynonymous-to-synonymous rate categories. Codon analyses were performed on datasets A and B, but not on dataset C because of computational limitations. The other five analyses were all based on the GTR model of nucleotide substitution with discrete gamma distributed rate heterogeneity (Yang, 1994) and invariant sites (GTR + G + I). The 'nt123 analysis' consisted of analysing the entire dataset. The 'nt12 analysis' consisted of analysing only the first two codon positions (nt1 and nt2), while excluding the third (nt3). In an attempt to restrict the analysis largely to nonsynonymous change, the 'noLRall1 + nt2 analysis' consisted of analysing only nt2 characters plus the nt1 characters that encode no leucine or arginine residues (Regier *et al.*, 2008b). Only nt1 characters that encode leucine and arginine (LRall1) have the potential to undergo synonymous change. In a separate attempt to analyse nonsynonymous change but without excluding entire characters, a 'degen1 analysis' was performed on characters in which all nt1 and nt3 characters that could possibly undergo synonymous change based on the 'universal' genetic code were fully degenerated, using standard IUPAC codenames, such that four-fold degenerate sites were recoded as 'N', three-fold as 'H' (methionine only) and two-fold as 'Y' or 'R' (Regier *et al.*, 2010). Finally, an nt123 analysis of dataset B with separate models for two character subsets ('partitioned' analysis) was performed: noLRall1 + nt2 and its complementary LRall1 + nt3 for dataset B (Note that 'noLRall1 + nt2' + 'LRall1 + nt3' = nt123.). Partitioning in this manner placed most nonsynonymous change in one subset (noLRall1 + nt2) and almost all inferred synonymous change in the other subset (LRall1 + nt3). Scripts written in PERL to generate the noLRall1 and LRall1 character sets and the degen1 data matrix are freely available at http://www.phylotools.com.

Maximum-likelihood analyses of datasets were implemented in GARLI (Genetic Algorithm for Rapid Likelihood Inference; v0.961, v1.0 and 'partition 0.97'; Zwickl, 2006) using grid computing (Cummings & Huskamp, 2005) on computational resources provided through 'The Lattice Project' (Bazinet & Cummings, 2008). For each nucleotide-model analysis, 500 searches were carried out and the best tree was chosen, whereas bootstrap analyses consisted of 1000 bootstrap pseudoreplicates with ten search replicates each, except for single-gene bootstrap analyses, for which approximately 300 pseudoreplicates each were performed. Codon model analyses consisted of 108 searches and 804 bootstrap pseudoreplicates of a single search each.

Although dataset B has a large block of intentionally missing data, there are smaller sections of missing data from all datasets caused by failures in amplification or sequencing (see 'Data set construction', Fig. 2). Two genes (*42fin* and *69fin*) and one taxon (*Oxytenis*) are particularly depauperate (Table S2), and so we independently tested the effect of their exclusion in dataset B without removing the large block of intentionally missing data. Differences in likelihood bootstrap percentages of greater than 50% were always within 10%, and were usually less than 3% (data not shown).

Maximum parsimony analyses were implemented in PAUP* 4.0b10 (Swofford, 2003) for dataset B only. For each of three analyses (nt123, nt12 and degen1), 250 search replicates were carried out and the best tree was chosen, whereas bootstrap analyses consisted of 1000 bootstrap pseudoreplicates with ten heuristic search replicates each.

For practical ease of reference, we will refer to nodes as 'strongly supported' and 'modestly supported' when they have bootstrap values of 80–100% and 70–79%, respectively, in one or more analyses.

Results

Combined gene analyses of the complete 'initial' dataset with fewer genes: 50 taxa/5 genes (~331 kbp total, dataset A)

Five likelihood analyses were performed on dataset A: a codon analysis and four nucleotide-model analyses (i.e. nt123, nt12, noLRall1 + nt2 and degen1; as described in Materials and methods). The results are illustrated in Fig. 3, in which bootstrap values are mapped onto the maximum likelihood topology from the codon analysis of dataset B (discussed below). It is noteworthy that the codon analyses from datasets A and B differ in their maximum likelihood topologies at only one weakly supported node, the basal split within the 'SBS' group, with bootstrap support of less than 50% (Fig. 3).

Combined gene analyses of the incompletely augmented dataset: 50 taxa/25 genes (~916 kbp total, dataset B)

The same five likelihood analyses were performed on dataset B (Fig. 3). Forty of 47 nodes are strongly supported (i.e. ≥80% bootstrap) by at least one of the five analyses. Thirty of these 40 nodes are supported by noLRall1 + nt2 and/or degen1, which largely derive their signal from nonsynonymous change. Support for above-family relationships within Bombycoidea has increased relative to results from dataset A (see also below), such that now only two nodes, namely, the basal splits within the 'SBS' and 'CAPOPEM' groups, fail to receive strong support by at least one analysis. For these two groups the same topologies are favoured by either three or four out of five analyses (Fig. 3, left side),

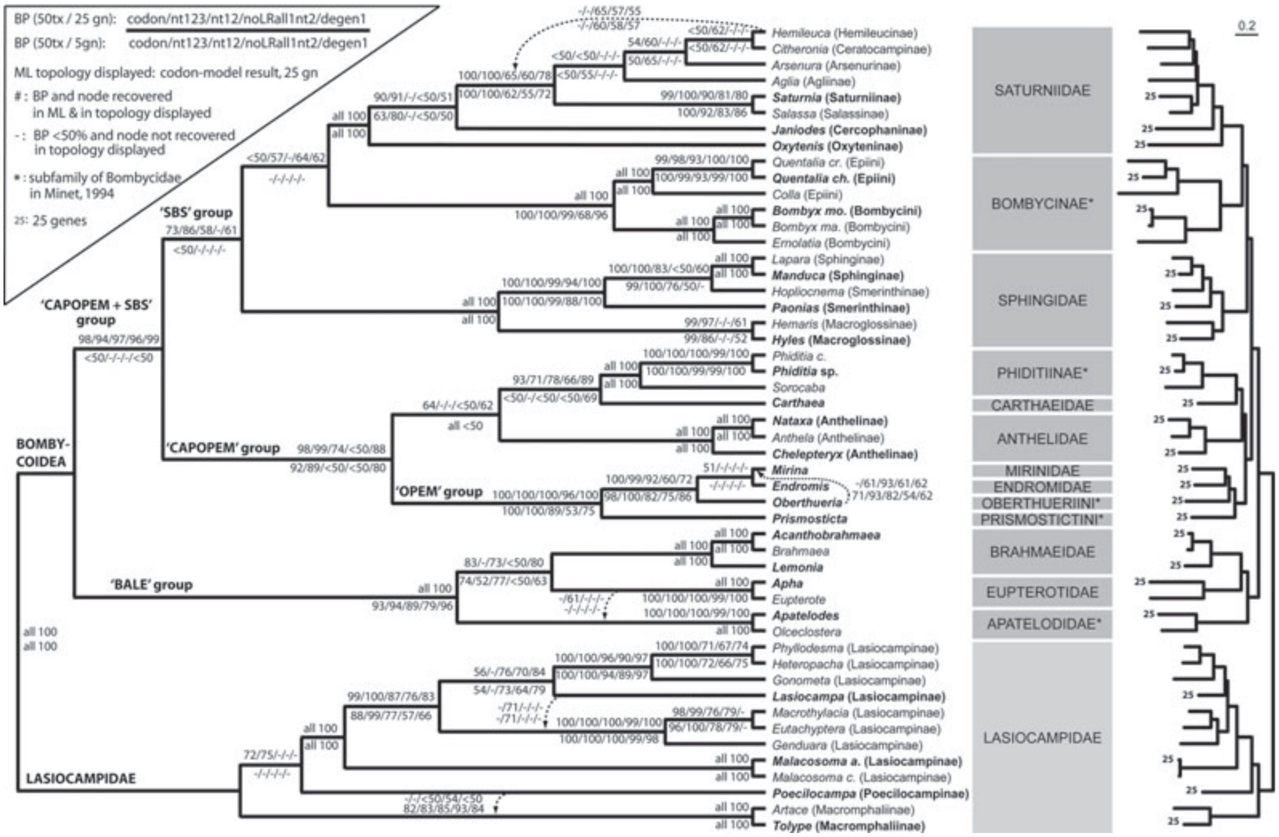

Fig. 3. Maximum likelihood topology (cladogram, left side; phylogram, right side) found under a codon model for 50 taxa sequenced for 25 genes (dataset B), with bootstrap percentages. Bootstrap percentages (BPs) above branches are separately calculated for dataset B (50 taxa/25 genes) using five analyses (in order: codon, nt123, nt12, noLRall1 + nt2, degen1). Bootstrap percentages below branches are separately calculated for dataset A (50 taxa/5 genes) using the same five analysis types, and are displayed in the same relative order. Dashes denote bootstrap support of less than 50% *and* failure to recover that particular node in the topology shown. Dashed arrows (five in total) identify alternative topologies (relative to the topology shown) that receive at least 60% bootstrap support by one or more of the approaches. Terminal taxa are identified by their genus names. The 24 taxa sampled for 25 genes (versus five genes) are highlighted in bold and labelled '25' on the terminal branches of the phylogram. Higher-level taxonomic names are in the shaded column; those marked with an asterisk were included in Bombycidae by Minet (1994) and Lemaire & Minet (1998) (see also Fig. 1, left side). Informal higher-level group names, here used for descriptive purposes only, are in quotes on selected branches of the cladogram. Branch lengths of the phylogram (on the right) are proportional to total nucleotide change per character as calculated under the codon model.

but both nodes have short subtending internodes (Fig. 3, right side). The family Anthelidae remains strongly anchored within the 'CAPOPEM' group, well removed from its former placement in Lasiocampoidea (Minet, 1994). However, there are also five nodes in which one or more analyses favour a different topology from the codon model with bootstrap ≥60% (see the dashed, curved arrows in Fig. 3). Yet, in none of these instances are both the codon-model topology and its incongruent alternative strongly supported.

Parsimony analysis of the nt12 character set recovers a nearly identical topology to that shown in Fig. 3, although the 'SBS' group is no longer recovered (but not contradicted either, as bootstrap support for the alternative is less than 50%), and the 'CAPOPEM' group is recovered but not strongly supported (Figure S1). The two other character sets analysed by parsimony yield similar results (Figure S1).

Combined gene analyses of the complete expanded dataset with fewer taxa: 24 taxa/25 genes (~439 kbp total, dataset C)

Four of the five nucleotide-model analyses mentioned above (i.e. nt123, nt12, noLRall1 + nt2 and degen1; no codon analysis) were performed for dataset C, which is restricted to the 24 taxa sequenced for all 25 genes (Fig. 4). Taking into account taxa missing relative to datasets A and B, there are only weakly supported (i.e. less than 50% bootstrap) differences relative to the topology recovered by datasets A and B (cf. Figs 3, 4).

Partitioned analyses of nt123 for datasets A, B and C

The aforementioned analytical approaches either utilise the total dataset (nt123 and codon), exclude a portion (nt12

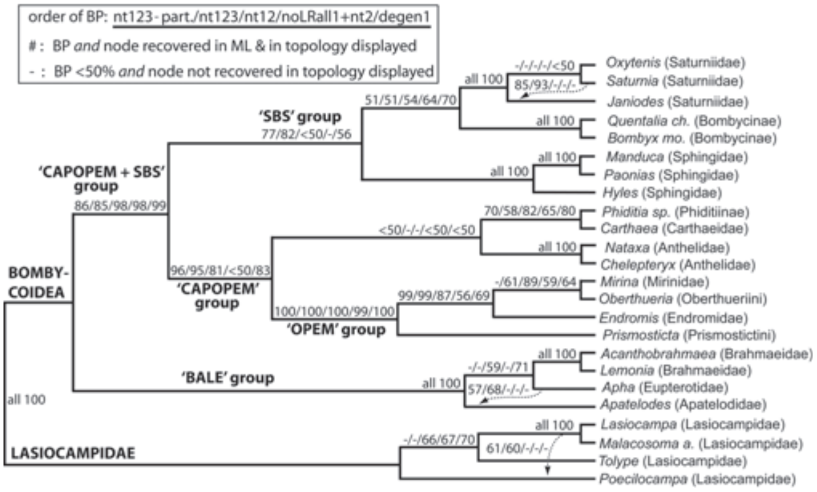

Fig. 4. Maximum likelihood topology found in a degen1 analysis of 24 taxa sequenced for 25 genes (dataset C), with bootstrap percentages. Bootstrap percentages above branches are separately calculated using five analyses [in order: nt123-partitioned, nt123-unpartitioned (nt123), nt12, noLRall1 + nt2, degen1]. Dashes denote bootstrap support of less than 50% *and* failure to recover that particular node in the topology shown. Dashed arrows (three in total) identify alternative topologies (relative to the topology shown) that receive at least 60% bootstrap support by one or more of the approaches. Terminal taxa are identified by their genus names followed in parentheses by their higher classification. Informal higher-level group names, here used for descriptive purposes only, are given in quotes on selected branches.

and noLRall1 + nt2), or degenerate potential synonymous change (degen1). A final approach is to partition the total dataset without character exclusion (Figures S2, 4). A node-by-node comparison of bootstrap percentages reveals that the maximum difference between partitioned and unpartitioned analyses is 12%, with most being much less, and that neither analysis yields consistently higher bootstrap percentages.

Comparison of combined gene analyses for higher-level groupings (datasets A, B and C)

Bootstrap values for 11 higher-level groupings have been compared for the six maximum likelihood analyses across the three datasets (Table 1). Nine groups (listed as the top nine groups in Table 1) receive strong support in one or more analyses; whereas, Saturniidae + Bombycidae is modestly supported (i.e. 70% bootstrap) only in the degen1 analysis of dataset C, and Phiditiinae + Carthaeidae + Anthelidae receives no bootstrap support ≥65%.

Data sets A and B can be most straightforwardly compared because they have identical taxon samples. Of the 34 instances (not to be confused with distinct nodes!) in which bootstrap values change by 10% or more, 29 have higher values with dataset B, consistent with an overall beneficial effect of additional characters. Degen1 and nt12 analyses show the greatest relative increases in bootstrap scores (seven groups increase ≥10%), whereas nt123 and nt123-partitioned show the least (three groups each).

There are also five instances in which bootstrap values decrease by more than 10% in dataset B relative to dataset A. Four instances involve only the three species that constitute the group Oberthueriinae + Mirinidae + Endromidae, and we suggest that this reflects conflicting gene signals (see the next section and the Discussion).

A comparison of datasets C and A reveals that of the 24 instances in which bootstrap values change by 10% or more, 19 have higher values with dataset C, again consistent with an overall beneficial effect of additional characters. Three of the five instances in which bootstrap values decrease by more than 10% again involve the group Oberthueriinae + Mirinidae + Endromidae. A fourth instance involves the group Brahmaeidae + Eupterotidae, but only in the nt12 analysis (also see the next section on conflicting gene signals).

A comparison of datasets C and B reveals only four instances in which bootstrap values differ by 10% or more, and all four correspond to relative increases in dataset B, consistent with a contributing signal from the 'additional' taxa sequenced for only five genes.

Single-gene analyses: agreement and conflict

Single-gene nt123 bootstrap analyses were performed, and all groups that are recovered with at least 50% bootstrap are shown (Table S4). Few individual genes support many groups, except the genes of greatest length, e.g. *CAD* and *DDC*. Considering the nine higher-level groups that were strongly supported by one or more all-gene analyses (listed in Table 1), the Bombycoidea/Lasiocampidae split receives at least modest support from 14 genes, the 'OPEM' group from four genes, the 'BALE' group from between two and four genes (key taxa are missing for two genes), and Mirinidae + Oberthueriinae from one gene. However, no single gene recovers the 'SBS' group,

Table 1. Comparison of bootstrap values of selected higher-level nodes (see Fig. 2) based on analysis of datasets B (50 taxa/25 genes), A (50 taxa/5 genes) and C (24 taxa/25 genes) by up to six analytical approaches[a].

Taxonomic analysis:	Codon		nt123			degen1		
Group dataset:	B	A	B	A	C	B	A	C
'SBS' group	73	<50	86	—	82	61	—	56
Phiditiinae + Carthaeidae	93	<50	71	—	58	89	69	80
Oberthueriinae + Mirinidae + Endromidae	100	98	99	100	99	72	86	69
'OPEM' group	100	100	100	100	100	100	75	100
'CAPOPEM' group	98	92	99	89	95	88	80	83
'CAPOPEM + SBS' group	98	<50	94	—	85	99	<50	99
Brahmaeidae + Eupterotidae	83	74	—	52	—	80	63	71
'BALE' group	100	93	100	94	100	100	96	100
Oberthueriinae + Mirinidae	—	71	61	93	61	62	62	64
Saturniidae + Bombycidae	<50	—	57	—	51	62	—	70
Phiditiinae + Carthaeidae + Anthelidae	64	<50	—	—	—	62	<50	<50
≥10% Δ for A → B comparison:	4↑, 1↓		3↑, 1↓			7↑, 1↓		
≥10% Δ for B → C comparison:					0↑, 1↓			0↑, 1↓
≥10% Δ for A → C comparison:					2↑, 1↓			4↑, 1↓

Taxonomic analysis:	nt12			noLRall1 + nt2			nt123: partitioned		
Group dataset:	B	A	C	B	A	C	B	A	C
'SBS' group	58	—	<50	—	<50	—	84	—	77
Phiditiinae + Carthaeidae	78	<50	82	66	<50	65	82	—	70
Oberthueriinae + Mirinidae + Endromidae	92	82	87	60	75	56	100	100	99
'OPEM' group	100	89	100	96	53	99	100	100	100
CAPOPEM group	74	<50	81	<50	<50	<50	98	91	96
'CAPOPEM + SBS' group	97	—	98	96	—	98	94	—	86
Brahmaeidae + Eupterotidae	73	77	59	<50	<50	—	53	57	—
'BALE' group	100	89	100	100	79	100	100	93	100
Oberthueriinae + Mirinidae	93	82	89	61	54	59	53	90	—
Saturniidae + Bombycidae	—	—	54	64	<50	64	56	—	51
Phiditiinae + Carthaeidae + Anthelidae	—	<50	—	<50	<50	<50	—	—	<50
≥10% Δ for A → B comparison:	7↑, 0↓			5↑, 1↓			3↑, 1↓		
≥10% Δ for B → C comparison:			0↑, 1↓			0↑, 0↓			0↑, 1↓
≥10% Δ for A → C comparison:			5↑, 1↓			5↑, 1↓			3↑, 1↓

[a] Bootstraps values are displayed for up to six analyses (codon, nt123, degen1, nt12, noLRall1 + nt2 and nt123-partitioned) applied to each of three datasets (A, B and C). Dashes denote bootstrap support of less than 50% *and* the failure to recover that node in its own maximum likelihood topology; otherwise, the designated nodes were recovered. Results for across-dataset changes (A → B, B → C and A → C) in bootstrap values of 10% or more are tabulated at the bottom, with '↑' indicating a relative rise in bootstrap (i.e. bootstrap percentages are higher for B relative to A, for C relative to B and for C relative to A), and '↓' indicating a relative decline (i.e. bootstraps percentages are lower for B relative to A, for C relative to B and for C relative to A). For the purposes of these calculations, all bootstrap values of less than 50% are assigned a value of 49%.

'CAPOPEM' group, 'CAPOPEM + SBS' group, Phiditiinae + Carthaeidae or Brahmaeidae + Eupterotidae with 50% or more bootstrap.

In three instances (see the coloured boxes in Table S4), conflicting groupings are each strongly supported (i.e. ≥80% bootstrap) by one individual gene, plus additional genes of lesser support. In a fourth instance (see yellow boxes in Table S4), only one of the two conflicting groups is strongly supported, but this group strongly conflicts with the combined gene result. The first instance of strong conflict is at the base of Saturniidae, which are represented by only three species in the 24-taxa set (dataset C), but which form a monophyletic group that is strongly supported by six individual genes without conflict. Given that combined-gene evidence quite strongly places Oxyteninae as sister group to other saturniids (90% bootstrap by codon model), and that only one gene strongly (*40fin*, 82% bootstrap) and another modestly (*CAD*, 77%) support an alternative to the combined-gene result, this instance does not argue against a combined-gene analysis, nor does it raise serious doubts about species relationships at the base of Saturniidae (also see the Discussion).

The second instance of individual gene conflict is the aforementioned problem within Oberthueriinae + Mirinidae + Endromidae, a group that is itself strongly supported by three individual genes, and without even modest conflict from the others. In particular, *113fin* strongly (90% bootstrap) supports Mirinidae + Endromidae, whereas *period* strongly (88% bootstrap) supports Mirinidae + Oberthueriinae, as do four of five analyses with the combined-gene dataset B (93% bootstrap for nt12 but significantly lower for the other analyses). Both topologies receive weaker support from two or three other single genes. Given these findings, a combined gene analysis is still warranted, although we suggest caution in interpreting the basal split within Oberthueriinae + Mirinidae + Endromidae.

The third instance of individual gene conflict is the basal split in the 'BALE' group, which is itself strongly and moderately supported by two genes each, and in the analyses of dataset B (100% bootstrap with all analyses). In particular, *40fin* (86%) and *CAD* (70%) support Eupterotidae + Apatelodidae, whereas four of the five combined gene analyses (but not nt123) recover Brahmaeidae + Eupterotidae with up to 83% bootstrap (codon model). No single gene even moderately supports Brahmaeidae + Eupterotidae. Given these findings, this instance does not argue against a combined-gene analysis, nor does it raise serious doubts about species relationships at the base of the 'BALE' group.

The fourth instance of individual gene conflict is at the base of the family Lasiocampidae (see also Regier *et al.*, 2008a), which is itself strongly supported by four individual genes, and is not contradicted by others at the level of 50% or more bootstrap. The strong single-gene conflict hinges on whether Poecilocampinae groups with Lasiocampinae (85 and 83% bootstrap from *109fin* and *period*, respectively) or with Macromphaliinae (83% bootstrap from CAD). Given that combined-gene results are never strongly supported, these substantial single-gene conflicts argue for caution when interpreting the basal split within Lasiocampidae, but do not argue in general against a combined gene analysis.

Discussion

The effect of additional data on node support

A previous five-gene analysis of relationships within Bombycoidea (Regier *et al.*, 2008a) yielded strong resolution of many relationships, but left several higher-level nodes weakly supported. The principal aim of the current study was to determine whether a nearly three-fold increase in sequence data, while keeping the number of taxa constant, would provide further support for higher-level relationships. Indeed, it did. A comparison of the likelihood analysis results for datasets A (50 taxa/5 genes) and B (50 taxa/25 genes) shows that whereas dataset A provides strong support for only four of 11 higher-level nodes within Bombycoidea ('CAPOPEM' group, 'OPEM' group, 'BALE' group and Oberthueriini + Endromidae + Mirinidae), dataset B strongly supports the same four, plus four additional relationships (Fig. 3; Table 1; support for the 'SBS' group increases from less than 50 to 86%, for the 'CAPOPEM + SBS' group from less than 50 to 98%, for Phiditiinae + Carthaeidae from less than 50 to 93% and for Brahmaeidae + Eupterotidae from 74 to 83%). Interpretation of support levels for Mirinidae + Oberthueriinae, although strong in the nt12 combined-gene analysis of dataset B, is more complicated because of single-gene conflict, and this issue is considered separately below. In summary, eight of the 11 nodes in Table 1 become strongly supported with increased data size (Fig. 1, right tree).

There remain two higher-level groups in Table 1 (i.e. Phiditiinae + Carthaeidae + Anthelidae and Saturniidae + Bombycinae) that are generally recovered in the combined-gene analyses, but for which the support is consistently low. Neither group displays even modest levels of conflict in the single-gene analyses (Table S4), and their lack of robust resolution may therefore be caused by the very short branches and insufficient data, or even by hard polytomies (Fig. 3, right side). We do note, however, that both the codon and degen1 analyses result in major increases in bootstrap support for Phiditiinae + Carthaeidae + Anthelidae (from less than 50 to 64 and 62%, respectively), consistent with an increasing phylogenetic signal. The situation with Saturniidae + Bombycinae is more murky, however, especially in light of support (albeit weak) for Bombycinae + Sphingidae, with a much reduced taxon sample and a slightly different gene sampling scheme (Cho *et al.*, 2010). The status of this trichotomy remains uncertain.

Within Lasiocampidae, analyses of datasets A and B yield similarly robust support for eight out of ten nodes, and weak to modest support for identical alternative placements of *Lasiocampa*. *Lasiocampa* and its sister group are subtended by a very short internode, suggesting either insufficient data or a hard polytomy (Fig. 3, right side). Data sets A and B provide somewhat differing signals at the base of Lasiocampidae, and this may result from conflict across the signals of single genes, introduced with the expanded gene sampling (discussed below).

Intentionally incomplete data matrices: an efficient strategy or an analytical quagmire?

As just discussed, our results provide clear support for the effectiveness of partial gene augmentation in order to improve the node support of higher-level groupings within Bombycoidea (dataset B versus dataset A). However, it is possible that the major blocks of nonrandomly missing data in dataset B, amounting to ∼45% of the total possible sequence for a complete matrix of these dimensions, could induce phylogenetic artifacts (systematic errors) that result in inflated support for incorrect groupings, as was recently demonstrated through the intentional manipulation of real datasets (Lemmon *et al.*, 2009). In our case, this is not a likely explanation. In the codon-model analyses of datasets A and B, the favoured topologies agree at 35 of 37 nodes within Bombycoidea, including all eight strongly supported, higher-level groups identified in the previous section, and nine out of ten nodes within Lasiocampidae (Fig. 3; Table 1). This striking degree of similarity, despite some major differences in bootstrap support, indicates that datasets A and B have qualitatively very similar signals, contra the scenario described in Lemmon *et al.* (2009). Instead, the increase in node support derived from dataset B appears to be the direct consequence of the additional data.

Additional key evidence against systematic errors in the inferences based on dataset B comes from the observation that the complete matrix of dataset C (24 taxa × 25 genes) widely supports the groupings of dataset B. Indeed, there are only three instances out of 44 comparisons in Table 1 in which bootstrap values from datasets C and B differ by 10% or more. Only one of these is strongly supported (Phidiitinae + Carthaeidae), although both datasets recover the same three nodes, and dataset B values are always greater than dataset C values, consistent with the hypothesis that incomplete data for some taxa can contribute to bootstrap values (Wiens, 2003, 2006).

The utility of synonymous and nonsynonymous changes in the phylogenetic analyses of Bombycoidea

In two studies across ditrysian Lepidoptera (Regier *et al.*, 2009; Cho *et al.*, 2010) we demonstrated that, relative to nonsynonymous substitutions, the signal from synonymous substitutions, which constitute more than 90% of total nucleotide change, can present analytical challenges at particular nodes because of faster divergence in nucleotide composition and accumulation of multiple substitutions per site (Regier *et al.*, 2008b). Indeed, this issue is implicitly acknowledged quite widely in studies of higher-level phylogeny, in that synonymous substitutions are frequently and intentionally minimized, e.g. through the removal of third codon position characters or by analysing amino acids, whereas similar attempts to selectively reduce nonsynonymous substitutions, e.g. through the removal of second codon position characters, are nonexistent. However, the current study is at a lower taxonomic level (i.e. within a single lepidopteran superfamily), and therefore the generally faster synonymous change might be particularly informative. In practical terms the issue of interest is not whether nonsynonymous changes outperform synonymous change, but whether the so-called total evidence (i.e. synonymous + nonsynonymous changes, e.g. derived from nt123 and codon analyses) is more or less informative than evidence derived mostly from nonsynonymous change, e.g. degen1 analysis.

We addressed this question by comparing the levels of bootstrap support for the eight bombycoid groups listed in Table 1 that are generally strongly supported by dataset B, and do not provide any evidence of single-gene conflict (Table S4). As an initial result, we see from Table 2 that more groups are recovered under likelihood than parsimony for both nt123 and degen1 (eight versus seven for degen1, seven versus five for nt123), so our further discussion will focus exclusively on the likelihood results. Whereas degen1 does recover all eight groups in its favored maximum likelihood topology (versus seven for nt123), the codon and nt123-partitioned analyses do this as well, but in addition yield strong support for a higher number of groups (seven versus six for degen1 and nt123). Based on these observations, our suggestion is that within Bombycoidea, total character change, including synonymous change, becomes more useful with parameter-rich models, particularly within codon and partitioned frameworks, and that a reliance on nonsynonymous change alone is unnecessary (contra Regier *et al.*, 2008b), but is still a useful comparison (Regier *et al.*, 2008a; Cho *et al.*, 2010).

Table 2. Comparison of bootstrap values for ‘nonsynonymous-only’and ‘all-nt’ approaches under parsimony and likelihood[a].

	Parsimony		Likelihood		
Taxonomic group:	degen1	nt123	degen1	nt123	codon
‘SBS’ group	—	—	61	86	73
Phiditiinae + Carthaeidae	70	—	89	71	93
Oberthueriinae + Mirinidae + Endromidae	64	100	72	99	100
‘OPEM’ group	98	100	100	100	100
‘CAPOPEM’ group	<50	69	88	99	98
‘CAPOPEM + SBS’ group	75	53	99	94	98
Brahmaeidae + Eupterotidae	61	—	80	—	83
‘BALE’ group	100	100	100	100	100
Number recovered in topology	7	5	8	7	8
Number with BP≥80%	2	3	6	6	7

[a]Bootstrap values for eight higher-level groups (column 1; see also Figs 2, S1) are displayed for degen1 and nt123 analyses under the parsimony criterion (columns 2, 3) and for degen1, nt123 and codon analyses under the likelihood criterion (columns 4–6). All analyses were performed on dataset B (50 taxa/25 genes). Dashes denote bootstrap support of less than 50% *and* failure to recover that node in its own favoured topology; otherwise, the designated nodes were recovered. The bottom two rows summarise for each analysis the number of groups recovered in the favoured topology (either most-parsimonious or highest-likelihood), and the number of strongly supported groups.

Conflicting gene signals: how serious a problem for this study?

In each of four nodes on the combined-gene tree (Fig. 3), a single gene contributes a strongly conflicting signal, plus there are one or two additional genes that contribute lesser (but still more than 50% bootstrap) support to the conflicting alternatives in each region (Table S4). Whether these conflicts result from systematic errors (e.g. inadequate taxon sampling), stochastic errors (e.g. insufficient characters to accurately infer phylogenetic relationships), gene-tree/species-tree conflicts (e.g. incomplete sorting of alleles at speciation) or something else is beyond the scope of this study. However, the phylogenetic consequences of conflicting signals from individual genes should not be ignored simply because they are relatively few in number. In fact, most single genes strongly support very few groups, so the absence of data should lead us to be neutral about how widespread true gene conflict is, rather than to conclude that it is minor. Having said that, two of the four regions where the single-gene conflict is localized receive strong support from combined-gene analyses (Saturniidae excluding Oxyteninae, and Brahmaeidae + Eupterotidae), and we think it is reasonable to accept these as the current best estimates of the species phylogeny. On the other hand, the conflict at the bases of Oberthueriinae + Mirinidae + Endromidae and of Lasiocampidae is such that the combined-gene support never rises to strong levels, so we suggest that these regions of the tree remain outstanding problems. Our summary of our best estimate of bombycoid phylogeny is displayed in Fig. 1, right side.

The placement of Carthaea saturnioides and its implications for the biogeography of Bombycoidea

Common (1966) erected for the single species *C. saturnioides* the monotypic family Carthaeidae within the Bombycoidea, noting that this taxon was 'possibly the most primitive family of the Bombycoidea' because of its retention of numerous symplesiomorphies that are variously lost in most other bombycoid families. Similarly, Minet (1994) postulated synapomorphies that support an early divergence of Carthaeidae from other bombycoid families, i.e. as sister to Sphingidae + (Brahmaeidae + Lemoniidae), with this group of four families in turn sister to all other Bombycoidea. Whereas *C. saturnioides* is undoubtedly unusual among Bombycoidea in retaining numerous symplesiomorphies, our analysis strongly supports a very different placement of this taxon, as sister to the Phiditiinae (93% bootstrap in codon-model analysis), and rather deeply nested within the Bombycoidea. This sister-group relationship is also of interest from a biogeographic perspective. Whereas *C. saturnioides* is restricted to the south-western corner of Australia (southern Western Australia), the Phiditiinae are exclusively neotropical. The most parsimonious hypothesis explaining this current distribution of extant taxa is to postulate a common ancestor that occurred prior to the complete isolation of Australia from South America + Antarctica in the late Eocene (35 Mya), but not necessarily prior to the initial separation in the late Cretaceous (90 Mya; McLoughlin, 2001; Sanmartín & Ronquist, 2004). A split of such an age would presumably be in line with the strong support from the relatively slowly evolving nonsynonymous data alone (89% bootstrap in degen1 analysis), as well as from the total evidence (93% bootstrap in codon analysis). Obviously, a remnant Gondwanan origin of the Carthaeidae + Phiditiinae clade would mark a minimum age for the origin of the entire superfamily Bombycoidea.

Revised classification

In this section, we address the formal classification of Bombycoidea and Lasiocampoidea, which at present is largely based on morphological studies by Minet (1994), and is partly in conflict with phylogenetic hypotheses that are now very strongly supported by molecular data. In particular, our current and previous analyses (Regier *et al.,* 2008a; Zwick, 2008) demonstrate the polyphyly of 'Bombycidae' sensu Minet (1994). Therefore, the family Bombycidae is here restricted to its nominate subfamily Bombycinae, comprising only the Old World tribe Bombycini and the New World tribe Epiini sensu Minet (1994), with the latter reinstated here as the subfamily Epiinae **stat.rev.** of the Bombycidae sensu auct. Apatelodinae sensu Minet (1994) were previously elevated to family rank (Zwick, 2008), and the New World subfamily Phiditiinae is again elevated here to family level (Phiditiidae **stat.rev.**). The 'OPEM' group of this study, namely, Oberthueriinae + Prismostictinae + Endromidae + Mirinidae (each represented by their respective nominate genus), is very strongly supported (100% bootstrap), and its members show less sequence divergence among each other than do the sequenced members of other bombycoid families (Fig. 3, right side). Therefore, as relationships among these rather closely related taxa are either not strongly resolved or else show single-gene conflicts that are confined to the 'OPEM' group, these taxa are best retained in a single family, rather than being split into four small families. Endromidae Boisduval, 1828 is the oldest name, and the other family group names are here placed in synonymy with it (Mirinidae Kozlov, 1985, **syn.nov.**; Oberthueriinae Kuznetzov & Stekolnikov, 1985, **syn.nov.**; Prismostictinae Forbes, 1955, **syn.nov.**).

Minet (1994) hypothesized that the Australo-New Guinean family Anthelidae is the sister group of the cosmopolitan Lasiocampidae, separating both families from Bombycoidea by placing them in the superfamily Lasiocampoidea. Current and past (Regier *et al.*, 2008a) molecular results very strongly support the inclusion of the Anthelidae deeply nested within Bombycoidea. Therefore, we reinstate Anthelidae within the Bombycoidea.

In summary, changes in classification are shown in Fig. 1, right side, and in Table S1.

Concluding statement

With only a few exceptions, higher-level groupings within Bombycoidea are now robustly supported in our analysis of dataset B. It is clear that this increase in support relative to our earlier study (Regier *et al.*, 2008a), and to analysis of dataset A, has occurred because of an increase in data (for 24 of the 50 taxa), as our taxon sample was identical between datasets A and B. Although we have not addressed how many of the new data were actually needed to achieve our improved result, or what the effect of generating additional data for all 50 taxa would have been, we note that no single gene provides even modest support for the new, robustly resolved clades in the combined-gene result (see Table S4), suggesting that multiple additional genes were indeed required. This might suggest that it will be a daunting task to similarly resolve relationships within the other 32 lepidopteran superfamilies. However, pessimism should be mitigated on two accounts. First, not all superfamilies are likely to be as challenging as Bombycoidea. For example, five-gene analyses of substantially fewer than 24 representatives of Geometroidea, Noctuoidea, Pyraloidea and Zyganoidea each showed robust support for many of their respective higher-level relationships (Regier *et al.*, 2009). Secondly, next-generation-sequencing approaches are already being applied to phylogenetic problems (e.g. Hittinger *et al.*, 2010), and these promise to make available for analysis, at an affordable cost, a much larger fraction of the genome.

Supporting Information

Additional Supporting Information may be found in the online version of this article under the DOI reference: 10.1111/j.1365-3113.2010.00543.x

Appendix S1. Dataset A.

Appendix S2. Dataset B.

Appendix S3. Dataset C.

Figure S1. Maximum parsimony topology found in an nt12 analysis of 50 taxa sequenced for 25 genes (data set B), with bootstrap percentages (BP) from nt123, nt12 and degen1 analyses.

Figure S2. Maximum likelihood topology from a partitioned analysis of 50 taxa sequenced for 25 genes (data set B), with bootstrap percentages from nt123-partitioned and nt123-unpartitioned data from 25 genes and 5 genes (data sets B and A, respectively).

Table S1. New higher classification and exemplar species sampled in this study, with taxonomic notes and indicating group diversity and geographic distributions.

Table S2. Percentage of sequence completeness displayed by gene and by taxon, plus total length of each gene segment.

Table S3. GenBank accession numbers.

Table S4. Bootstrap values in likelihood, single-gene, nt123 analyses for all groups recovered with ≥50% bootstrap.

Please note: Neither the Editors nor Wiley-Blackwell are responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Acknowledgements

We are indebted to many generous colleagues for supplying fresh specimens for this and our previous study (Regier *et al.*, 2008a), including James Adams, Vitor Becker, Tom Burbidge, Michael Collins, Andrea Coccuci, Ted Edwards, Tim Friedlander, Kurt Horst, Roger Hutchings, Nils Hydén, Daniel Janzen, Akito Kawahara, William Kelly, Ian Kitching, Marcus Matthews, Andrew Mitchell, Marcela Moré, Stefan Naumann, Judd Nelson, Richard Peigler, Rodolphe Rougerie, Pierre Tripotin, Adrienne Venables, Kirby Wolfe and S. Wolfe. Suwei Zhao and Kongyi Jiang provided valuable technical assistance at the Center for Biosystems Research's core DNA Sequencing Facility, and Adam Bazinet provided valuable assistance with grid computing through The Lattice Project. Derrick Zwickl kindly made available, through The Lattice Project, a version of GARLI that could analyse partitioned datasets. Kim Mitter made essential contributions as specimen collection and database manager. Financial support was provided by the US National Science Foundation's Assembling the Tree of Life programme grants 0531626 and 0531769, and by NSF grant DEB 0212910.

References

Bazinet, A.L. & Cummings, M.P. (2008) The Lattice Project: a Grid research production environment combining multiple Grid computing models. *Distributed & Grid Computing – Science Made Transparent for Everyone. Principles, Applications and Supporting Communities*, pp. 2–13. Rechenkraft.net, Marburg.

Cho, S., Zwick, A., Regier, J.C. *et al.* (2010) Deliberately unequal gene sampling: boon or bane for phylogenetics of Lepidoptera (Hexapoda)?, submitted.

Common, I.F.B. (1966) A new family of Bombycoidea (Lepidoptera) based on *Carthaea saturnioides* Walker from Western Australia. *Journal of the Entomological Society of Queensland*, **5**, 29–36.

Cummings, M.P. & Huskamp, J.C. (2005) Grid computing. *EDUCAUSE Review*, **40**, 116–117.

Goldman, N. & Yang, Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology & Evolution*, **11**, 725–736.

Goldsmith, M.R. & Marec, F. (2010) *Molecular Biology and Genetics of the Lepidoptera*. CRC Press, New York, New York.

Goldsmith, M.R. & Wilkins, A.S. (1995) *Molecular Model Systems in the Lepidoptera*. Cambridge University Press, New York, New York.

Hittinger, C.T., Johnston, M., Tossberg, J.T. & Rokas, A. (2010) Leveraging skewed transcript abundance by RNA-Seq to increase the genomic depth of the tree of life. *Proceedings of the National Academy of Science United States of America*, **107**, 1476–1481.

Lanave, C., Preparata, G., Saccone, C. & Serio, G. (1994) A new method for calculating evolutionary substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution*, **20**, 86–93.

Lemaire, C. & Minet, J. (1998) The Bombycoidea and their relatives. *Lepidoptera, Moths and Butterflies, Volume 1: Evolution, Systematics, and Biogeography*, Chapter 18 (ed. by N.P. Kristensen), pp. 321–353. Walter de Gruyter, Inc., Hawthorne, New York.

Lemmon, A.R., Brown, M.M., Stanger-Hall, K. & Lemmon, E.M. (2009) The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. *Systematic Biology*, **58**, 130–145.

McCormack, J.E., Huang, H. & Knowles, L.L. (2009) Maximum likelihood estimates of species trees: how accuracy of phylogenetic inference depends upon the divergence history and sampling design. *Systematic Biology*, **58**, 501–508.

McLoughlin, S. (2001) The breakup history of Gondwana and its impact on pre-Cenozoic floristic provincialism. *Australian Systematic Botany*, **49**, 271–300.

Minet, J. (1994) The Bombycoidea: phylogeny and higher classification (Lepidoptera: Glossata). *Entomologica Scandinavica*, **25**, 63–88.

Nässig, W.A. & Oberprieler, R.G. (2008) An annotated catalogue of the genera of Eupterotidae (Insecta, Lepidoptera, Bombycoidea). *Senckenbergiana Biologica*, **88**, 53–80.

Regier, J.C., Fang, Q.Q., Mitter, C., Peigler, R.S., Friedlander, T.P. & Solis, M.A. (1998) Evolution and phylogenetic utility of the *period* gene in Lepidoptera. *Molecular Biology and Evolution*, **15**, 1172–1182.

Regier, J.C., Cook, C.P., Mitter, C. & Hussey, A. (2008a) A phylogenetic study of the 'bombycoid complex' (Lepidoptera) using five protein-coding nuclear genes, with comments on the problem of macrolepidoteran phylogeny. *Systematic Entomology*, **33**, 175–189.

Regier, J.C., Shultz, J.W., Ganley, A.R.D. *et al.* (2008b) Resolving arthropod phylogeny: Exploring phylogenetic signal within 41 kb of protein-coding nuclear gene sequence. *Systematic Biology*, **57**, 920–938.

Regier, J.C., Zwick, A., Cummings, M.P. *et al.* (2009) Toward reconstructing the evolution of advanced moths and butterflies (Lepidotera: Ditrysia): an initial molecular study. *BMC Evolutionary Biology*, **9**, 280. doi: 10.1186/1471-2148-9-28.

Regier, J.C., Shultz, J.W., Zwick, A. *et al.* (2010) Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature*, **463**, 1079–1083.

Sanmartín, I. & Ronquist, F. (2004) Southern Hemisphere biogeography inferred by event-based models: plant versus animal patterns. *Systematic Biology*, **53**, 216–243.

Smith, S.W., Overbeck, R., Woese, C.R., Gilbert, W. & Gillevet, P.M. (1994) The genetic data environment and expandable GUI for multiple sequence analysis. *Computer Applications in the Biosciences*, **10**, 671–675.

Staden, R., Judge, D. & Bonfield, J. (2001) Sequence assembly and finishing methods. *Methods of Biochemical Analysis*, **43**, 303–322.

Susko, E. (2009) Bootstrap support is not first-order correct. *Systematic Biology*, **58**, 211–223.

Swofford, D.L. (2003) *PAUP*: Phylogenetic Analysis Using Parsimony (* and other methods)*, Version 4.0b10. Sinauer Associates, Sunderland, Massachusetts.

Wiens, J.J. (2003) Missing data, incomplete taxa, and phylogenetic accuracy. *Systematic Biology*, **52**, 528–538.

Wiens, J.J. (2006) Missing data and the design of phylogenetic analyses. *Journal of Biomedical Informatics*, **39**, 34–42.

Xia, Q., Zhou, Z., Lu, C. *et al.* (2004) A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science*, **306**, 1937–1940.

Xia, Q., Guo, Y., Zhang, Z. *et al.* (2009) Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (*Bombyx*). *Science*, **326**, 433–436.

Yang, Z. (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution*, **39**, 306–314.

Zharkikh, A. & Li, W.-H. (1995) Estimation of confidence in phylogeny: the complete-and-partial bootstrap technique. *Molecular Phylogenetics and Evolution*, **4**, 44–63.

Zwickl, D.J. (2006) *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion*. PhD dissertation, The University of Texas, Austin, Texas.

Zwick, A. (2008) Molecular phylogeny of Anthelidae and other bombycoid taxa (Lepidoptera: Bombycoidea). *Systematic Entomology*, **33**, 190–209.

Accepted 1 July 2010
First published online 30 September 2010