

Genomic Outposts Serve the Phylogenomic Pioneers: Designing Novel Nuclear Markers for Genomic DNA Extractions of Lepidoptera

NIKLAS WAHLBERG^{1,3} AND CHRISTOPHER WEST WHEAT²

¹Department of Zoology, Stockholm University, S-106 91 Stockholm, Sweden

²Department of Biological and Environmental Sciences, PO Box 65 (Viikinkaari 1), FI-00014 University of Helsinki, Finland; and Department of Biology, 208 Mueller Lab, Pennsylvania State University, University Park, Pennsylvania 16802, USA

³Current Address: Laboratory of Genetics, Department of Biology, University of Turku, FI-20014 Turku, Finland; E-mail: niklas.wahlberg@utu.fi

Abstract.—Increasing the number of characters used in phylogenetic studies is the next crucial step towards generating robust and stable phylogenetic hypotheses—i.e., strongly supported and consistent across reconstruction method. Here we describe a genomic approach to finding new protein-coding genes for systematics in nonmodel taxa, which can be PCR amplified from standard, slightly degraded genomic DNA extracts. We test this approach on Lepidoptera, searching the draft genomic sequence of the silk moth *Bombyx mori*, for exons >500 bp in length, removing annotated gene families, and compared remaining exons with butterfly EST databases to identify conserved regions for primer design. These primers were tested on a set of 65 taxa primarily in the butterfly family Nymphalidae. We were able to identify and amplify six previously unused gene regions (Arginine Kinase, GAPDH, IDH, MDH, RpS2, and RpS5) and two rarely used gene regions (CAD and DDC) that when added to the three traditional gene regions (COI, EF-1 α and wingless) gave a data set of 8114 bp. Phylogenetic robustness and stability increased with increasing numbers of genes. Smaller taxonomic subsets were also robust when using the full gene data set. The full 11-gene data set was robust and stable across reconstruction methods, recovering the major lineages and strongly supporting relationships within them. Our methods and insights should be applicable to taxonomic groups having a single genomic reference species and several EST databases from taxa that diverged less than 100 million years ago. [Exons; Lepidoptera; Nymphalidae; Phylogenomics; PCR; primers.]

The field of molecular systematics is slowly but surely maturing. Researchers now generally try to avoid the mistakes of early molecular systematics, when dramatic rearrangements of phylogeny were often proposed on the basis of small and sparsely sampled data sets (e.g., Nardi et al., 2003). There is now general acceptance that phylogenetic inference based on a single gene using a single reconstruction method is rarely robust to the addition of new data and stable to changes in assumptions of analysis (e.g., parsimony versus modeling approaches), except perhaps at shallow nodes. Thus, the amount of data necessary for robust and stable phylogenetic inferences at the subfamily and family level is not clear. In the post-genomics era, we are in a unique position to start investigating the optimal amount of data necessary for robust and stable phylogenetic inferences. Recent studies have placed the recommended number of independent genes for robust inference at about 20 (Rokas et al., 2003). However, the Rokas et al. (2003) data set was not representative of most phylogenetic studies, as they focused on only eight taxa, of which five were closely related and three were very distantly related (Gatesy et al., 2007). Thus, more detailed analyses of more representative data sets are needed to begin addressing how analyses of data matrices at different taxonomic levels behave with increasing numbers of molecular characters (Simonsen et al., 2006; Gatesy et al., 2007).

One of the goals of phylogenomics, as we envision it, is to generate data sets that result in topologies that are robust to the addition of new data and stable to changes in assumptions of analyses. Such phylogenetic hypotheses may not be attainable for some taxa in which lateral gene transfer is common (e.g., in prokaryotes Ochman et al., 2000; Boucher et al., 2003), but in theory should be attainable for groups of organisms with sexual reproduction

and hybridization limited to closely related species. Stability of phylogenetic hypotheses to changing assumptions of analyses has not generally been considered to be an important confidence measure of an inferred hypothesis (Giribet, 2003). This has arisen mainly due to the differences in the philosophies of proponents of parsimony- and model-based methods (see, e.g., Kluge, 2001; Felsenstein, 2004).

Changing the assumptions of analysis can provide information about the strength of phylogenetic signal in a data set. Data sets with weak phylogenetic signals will be strongly influenced by assumptions of analysis, whereas data sets with strong phylogenetic signals will not be influenced as much. In a parsimony framework, assumptions about character weights can be changed, e.g., by down-weighting third codon positions or by weighting transversions more than transitions (Wheeler, 1995; Giribet, 2003; Wahlberg et al., 2005b). Another way of testing topological stability is to compare the results of unweighted parsimony analysis (which allows characters to evolve unrestricted by assumptions about their evolvability; Brower, 2000a) to the results of a model-based analysis, which have strict assumptions about how character evolution (Brooks et al., 2007). Although there is a long history of methodological debate in systematics (see, e.g., Kluge, 2001; Felsenstein, 2004), presumably as the amount of data increases, the different methods of analysis should converge on the same topology (Brooks et al., 2007). Phylogenomics thus may help resolve a central debate in systematics, namely one analysis method's primacy over another, through generating data sets where all methods of analysis yield identical topologies and thus render such debate moot through inclusion. Here we explore this potential in the butterfly family Nymphalidae (Lepidoptera).

Butterflies are model organisms for a diverse set of ecological and evolutionary questions (Boggs et al., 2003; Ehrlich and Hanski, 2004), and thus understanding their phylogenetic relationships is of significant importance. However, the phylogenetic relationships of various butterfly taxa have been contentious, although recent work is starting to resolve points of discrepancy (Caterino et al., 2001; Wahlberg et al., 2003, 2005a; Braby et al., 2006; Nazari et al., 2007). Common to all butterfly studies is the small number of molecular markers available and used (Sperling, 2003). Usually one to three genes are used with the only exceptions being two recent studies using seven genes (Mallarino et al., 2005; Nazari et al., 2007). The study by Mallarino et al. (2005) looked at the relationships of species in the genus *Ithomia* (Nymphalidae) using data from four nuclear genes and three mitochondrial genes and their results were robust and stable. The study by Nazari et al. (2007) looked at relationships of genera in the subfamily Parnassiinae (Papilionidae) using data from morphology, two nuclear and five mitochondrial genes. They found that the mitochondrial genes gave conflicting and weak results compared to the nuclear and morphological data, and that the nuclear genes were particularly good at resolving the deeper nodes in their phylogenetic hypothesis. Thus, increasing characters results in more robust phylogenetic inference at various taxonomic levels and is therefore a goal of many systematists. Although one avenue to acquiring more characters is to use morphological data, which has been shown to increase the robustness of phylogenetic hypotheses (Wahlberg and Nylin, 2003; Wahlberg et al., 2005a; Simonsen et al., 2006), morphological data are limited, difficult to code, and require extensive experience to identify character states correctly.

Phylogenomics, using many genes from across the genome, is likely to be the path to robust phylogenetic inference (Delsuc et al., 2005). Until recently, the only molecular markers easily sequenced in a broad range of taxa have been mitochondrial genes. However, the utility of using many mitochondrial genes is questionable, as there is a shared evolutionary history, and even entire mitochondrial genomes (15,000 to 20,000 bp in insects) fail to provide robust inferences at deep levels (Cameron et al., 2004). Ideally, many genes of independent evolutionary history should be used for phylogenomics.

Recently, new opportunities have arisen for identifying suitable nuclear genes for systematic work in Lepidoptera (and similarly in many other taxonomic groups). First, two initial drafts of the whole genome sequence (WGS) for the silkworm, *Bombyx mori*, have been generated (Biology Analysis Group, 2004; Mita et al., 2004), allowing one to identify nuclear genes that are single copy. Second, there are several groups generating Expressed Sequence Tag (EST) libraries for diverse butterfly taxa; e.g., *Pieris rapae* (<http://www.ice.mpg.de/tmo/research/InsectGenome.htm>), *Heliconius melpomene* (Jiggins et al., 2005), *Bicyclus anynana* (Beldade et al., 2006), and *Melitaea cinxia* (Vera et al., 2008). EST libraries provide DNA sequence of the mRNA of expressed genes (i.e., the joined exons, as the

introns have been removed during the mRNA processing). The mRNA to make these libraries are derived from tissues of interest particular to a given research group, providing thousands of individual genes expressed in that tissue (the gut of *Pieris rapae* larvae, the developing wing discs in larval *Heliconius melpomene* and *Bicyclus anynana*, and diverse tissues of *Melitaea cinxia*). The latter three species belong to the family Nymphalidae. This concentration of EST libraries in Nymphalidae presents an opportunity to find large numbers of nuclear protein coding genes potentially suitable for systematics.

New gene regions have recently been successfully sequenced for Lepidoptera and these are a much-welcomed addition to a field hungry for novel molecular markers (Friedlander et al., 1996; Fang et al., 1997; Regier et al., 1998). However, protocol for sequencing these genes necessitates cDNA, which is the DNA form of mRNA and therefore has introns removed. Thus, primers designed from cDNA-generated sequences will often fail to amplify the same region when using genomic DNA, due to the presence of introns (Fig. 1). Introns are problematic for several reasons: (1) they can inflate the length between cDNA-based primers 10 to 10,000 fold (e.g., Deutsch and Long, 1999); (2) they may contain indel polymorphisms (i.e., insertion/deletion), which generates variable length PCR products that lead to unusable sequence data (due to overlap); and (3) they may result in an exon/intron breakpoint being within a designed cDNA primer.

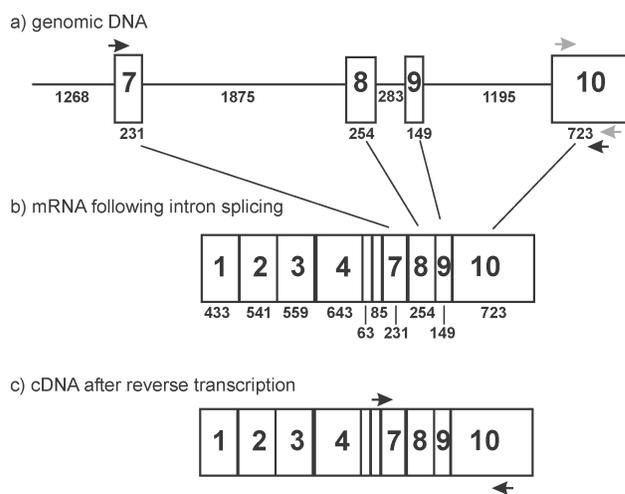


FIGURE 1. Comparison between PCR primer locations in genomic DNA and cDNA of the dopadecarboxylase gene (DDC) in Lepidoptera, based on *Bombyx mori* genomic structure. (a) The last four exons and introns of DDC are shown as boxes and lines, respectively, with their size shown below. Arrows represent PCR primers, with the black forward primer spanning both exon 6 (not shown) and exon 7 and the black reverse primer located fully in exon 10. Gray primers are fully located within the large and final exon 10, representing primers and their locations designed in this study. (b) mRNA showing all exons of DDC brought together after excision of introns. (c) cDNA showing the location of black PCR primers, the forward of which is now able to bind to continuous DNA as exons 6 and 7 are joined. The forward and reverse black primers are previously reported DDC amplification primers 1.7sF and 4sR, respectively (Fang et al., 1997).

Genomic DNA is by far the most common DNA used for molecular systematic studies, mainly due to the sensitivity of RNA to degradation and simplicity of genomic DNA preservation in the field. Genomic DNA extracts are also often degraded, as the individual samples the DNA was extracted from may not have been stored appropriately and the DNA may have been subjected to repeated freeze-thaw cycles during use. However, amplifying nuclear gene regions of about 500 bp is possible from such degraded DNA extracts and even dried leg material, as has been performed in the studies on Nymphalidae by the Niklas Wahlberg laboratory (Wahlberg and Nylin, 2003; Wahlberg et al., 2003, 2005a, 2005b; Peña et al., 2006; Wahlberg and Freitas, 2007).

Ideally then, newly designed genes for phylogenomics in the Lepidoptera would develop primers that can work specifically on genomic DNA. In this paper, we report a new method to search through genome databases for exons of suitable size (500 to 600 bp), comparing these exons to EST databases for related taxa of interest, and finally develop primers potentially universal across the taxa of interest. Using these primers, we amplify exemplar taxa from across Nymphalidae in an attempt to address the questions raised above, namely, what is the number of molecular characters needed to attain a robust and stable phylogenetic tree at the intrafamilial level of classification.

MATERIAL AND METHODS

Finding and Amplifying Exons

The exon-intron boundaries between studied species of Papilionoidea and *Bombyx mori* are completely conserved for studied metabolic enzymes and ribosomal proteins (Wheat, unpublished data). With this knowledge in hand, one can BLAST search the unique genes (unigenes) of *B. mori*, obtained from the extensive EST libraries for this species as well as gene prediction algorithms used on WGS, against its preliminary WGS contigs (contiguous stretches of DNA assembled from smaller, overlapping DNA sequence reads). This comparison between cDNA and genomic DNA reveals intron locations likely found across all of Papilionoidea and potentially higher Lepidoptera. We used the program Spidey, provided free from NCBI website, to compare the EST-derived *B. mori* unigene set against the released WGS contigs available on NCBI (Biology Analysis Group, 2004; Mita et al., 2004). Identified exons longer than 500 bp were then six-frame translated and BLAST searched against the protein reference database Swiss Prot for annotation. Annotation is crucial at this stage as we wished to use only single-copy, nuclear-coding exons that were not members of a gene family, as gene families have a high potential for concerted evolution or birth-death dynamics, which violate assumptions of orthology in phylogenetic reconstruction. Long exons that had multiple copies or were members of gene families were excluded from further analysis, except in the case of one gene region (HSP70), where we wanted to assess how such gene family dynamics might affect analysis.

The long exons were then used for cross species PCR primer design. In order to determine which regions of these long exon were most conserved and thus suitable for degenerate primer design, we searched for these long exons in publicly available Papilionoidea EST libraries of *Heliconius* sp. (Butterfly base, <http://heliconius.cap.ed.ac.uk/butterfly/db/>) and *Bicyclus anynana* (Beldade et al., 2006), as well as private EST collections of several species (*Pieris rapae*, *Colias eurytheme*, *Melitaea cinxia*) (Pierid EST data courtesy of H. Vogel). Identified orthologues were then aligned and conserved gene regions identified in light of their amino acid codon degeneracy. Primer design, using the defaults in the Web-based program Primer3 (<http://frodo.wi.mit.edu/cgi-bin/primer3/primer3-www.cgi>), attempted to maximize amplicon length while minimizing degeneracy. In order to facilitate high-throughput PCR and sequencing, either a universal forward or reverse primer was attached to each degenerate primer (F or R, respectively). These nondegenerate, nonhomologous 5' tails were then used to sequence all PCR products regardless of exon origin, sidestepping traditional gene region-specific sequencing primer design and/or cloning of PCR products followed by sequencing. A similar hybrid primer design has been also implemented by other laboratories (Regier and Shi, 2005) to increase yield and facilitate sequencing. We used the universal primer pair T7/T3 (T7PromoterF 5' TAA TAC GAC TCA CTA TAG GG 3', T3R 5' ATT AAC CCT CAC TAA AG 3').

All test primers were run on a standard set of test taxa (table in online Appendix; available at www.systematicbiology.org), including *B. mori* as a positive control. Our goal was to increase the number of gene regions available for the molecular systematics of the butterfly family Nymphalidae (Brower, 2000b; Wahlberg et al., 2003, 2005b; Mallarino et al., 2005; Whinnett et al., 2005; Brower et al., 2006), but we included 11 species from a range of Lepidoptera families as well as 1 species of Trichoptera to test the universality of our primers. The 54 species of Nymphalidae were chosen to represent all major lineages of the family based on a number of recent publications (Brower, 2000b; Wahlberg et al., 2003, 2005a, 2005b; Freitas and Brown, 2004; Brower et al., 2006). All subfamilies are represented by at least two species for Nymphalidae (table in online Appendix). Almost all specimens have been used in previous publications, which describe the methods of DNA extraction (Wahlberg et al., 2003, 2005a, 2005b; Peña et al., 2006).

We performed all PCRs in a 20- μ L reaction volume using 1 μ L of DNA extract (with varying unmeasured concentrations of DNA). In initial trials, the PCR reaction had a primer concentration of 1 μ M, dNTP concentration of 200 μ M, 2 units of AmpliTaq Gold polymerase, and a MgCl₂ concentration of 1.5 mM. For successful primer pairs, the concentration of primers was decreased to a standard 0.5 μ M. The PCR protocols are given in the online Appendix. The initial PCR cycling profile was 95°C for 7 min, 40 cycles of 95°C for 30 s, 50°C for 30 s, 72°C for 2 min, and a final extension period of 72°C for 10 min. Successful PCRs were sequenced in

both directions using the nondegenerate, nonhomologous 5' tails (universal primers). We also tested several primer pairs used successfully in butterflies (Caterino et al., 2001; Wahlberg et al., 2003, 2005a; Braby et al., 2006; Nazari et al., 2007) on a wider range of Lepidoptera. Sequencing was performed either with a Beckman-Coulter CEQ8000 capillary sequencer (Stockholm) or an ABI PRISM 3130xl capillary sequencer (Turku) using dye terminator sequencing kits according to the recommendations of manufacturers.

Phylogenetic Analyses

Amplified and sequenced gene regions were aligned by eye using the program BioEdit (Hall, 1999) and analyzed separately and combined. Both parsimony analyses and model-based methods (maximum likelihood and Bayesian inference) were used to analyze the data. Initially all Lepidoptera sequences were included, but preliminary analyses showed that the long branches of nonpapilionoid species confounded results by attaching to internal branches of Nymphalidae. Thus we analyzed only papilionoid sequences, rooting our trees on *Papilio glaucus*. The combined data matrix is available from TREEBASE (www.treebase.org, accession number SN3759).

For parsimony analyses, the data were subjected to 100 random addition rounds of successive Sectorial, Ratchet, Drift, and Tree Fusing searches (Goloboff, 1999; Moilanen, 1999; Nixon, 1999) with the program TNT (Goloboff et al., 2004). We evaluated the character support for the clades in the resulting cladograms using Bremer support (Bremer, 1988, 1994) and bootstrap (Felsenstein, 1985). The scripting feature of TNT was used to calculate BS values (see Peña et al., 2006). We assessed the contribution of each data partition to the BS values of the combined analyses using partitioned Bremer support (Baker and DeSalle, 1997; Gatesy et al., 1999) using another script in TNT (scripts available at <http://www.zmuc.dk/public/phylogeny/>). The degree of congruence between the three separate data sets was summarized using the partition congruence index (PCI; Brower, 2006). This index is equal to the Bremer support value when there is no conflict between data sets and has negative values when there is strong conflict between data sets (Brower, 2006). Bootstrap values were calculated with 1000 pseudoreplicates of 100 random addition rounds of successive Sectorial, Ratchet, Drift, and Tree Fusing searches (Goloboff, 1999; Moilanen, 1999; Nixon, 1999) with the program TNT (Goloboff et al., 2004).

Bayesian inference was implemented with a DNA substitution model selected based on AIC values obtained using the program MrModelTest (Nylander, 2002). The best-fit model for each gene was the most complex model available (GTR+ Γ +I). However, it has been noted that the Γ shape parameter and the I parameter are highly correlated and are considered to be "pathological" when estimated together (Ren et al., 2005); thus, we also analyzed our data with the reduced model GTR+ Γ . We used

Bayesian methods to estimate parameter values using the program MrBayes 3.1 (Ronquist and Huelsenbeck, 2003). The single-gene data sets were then subjected to two independent simultaneous runs (one cold and three heated chains per run) of 5 million generations each, with every 500th generation sampled and the first 1000 sampled generations discarded as burn-in (cut-off point confirmed after the analysis using the *sump* command in MrBayes). The convergence of topology for the two runs was monitored by following the standard deviation of split frequencies. Single-gene data sets were not partitioned in any way.

For the combined analysis, the Bayesian analysis was performed with parameter values estimated separately for each gene region using the "unlink" command and the rate prior (ratepr) set to "variable." For each gene, we used the same model chosen in the above analysis, although parameter values were estimated again. The combined analyses were partitioned in two ways, either with one partition per gene (11 partition analysis) or with the genes partitioned further by codon positions (33 partition analysis). Four independent analyses were run simultaneously for 5 million generations, with every 500th generation sampled and the first 5000 sampled generations discarded as burn-in (cut-off point determined after the analysis using the *sump* command in MrBayes). The convergence of topology of the four runs was verified by monitoring the standard deviation of split frequencies during the run.

Maximum likelihood analyses were implemented using the online version of RAxML (<http://phylobench.vital-it.ch/raxml-bb/index.php>) (Stamatakis, 2006). The default GTR+ Γ was used on the concatenated data set with 11 genes and the data set was analyzed both unpartitioned and partitioned into 11 gene regions. Node support was assessed through bootstrapped data sets with 500 pseudoreplicates. ML analyses were also initially run using Garli (Zwickl, 2006), but these gave identical results to RAxML, although being considerably slower.

In order to investigate the amount of data needed to arrive at a stable phylogenetic hypothesis, we tested the ability of three subsets of the full data set to recover the deeper nodes found in the analyses of the all 11 genes. Subsets were structured as (1) single genes (i.e., each gene on its own), (2) 11 sets each of three genes, and (3) 11 sets of five genes, with each set randomly sampled without replacement. The composition of the data sets is in the online Appendix. Each data set was analyzed in RAxML with 100 bootstrapped pseudoreplicates, and the proportion of time that each node from the full data set was recovered was calculated based on the 100 bootstrapped trees. The bootstrap values were then averaged over all replicates in the single-gene, three-gene, and five-gene analyses. The data were also analyzed in a parsimony framework with 1000 bootstrap pseudoreplicates analyzed as described above.

We also investigated the effects of taxon sampling on our results, as preliminary analyses suggested that this might have an effect on tree topology. We created two

subsampled data sets of 51 taxa each, one that was unbalanced as possible, while retaining all major lineages (taxa deleted were *Libytheana*, *Anetia*, *Methona*, *Pseudergolis*, *Marpesia*, *Chitoria*, and *Hypanartia*), and one that had the same number of taxa deleted from larger inclusive clades (taxa deleted were *Euxanthe*, *Pararge*, *Brintesia*, *Vagrans*, *Biblis*, *Hypanartia*, and *Anartia*). The data were analyzed using parsimony, as described above, and with maximum likelihood in RAxML with 100 bootstrap pseudoreplicates.

RESULTS

Finding, Amplifying, and Assessing Long Exons

NCBI *B. mori* unigene build of December 2005 consisted of 6534 unigenes. Of these we identified 461 exons longer than 500 bp. Only one such exon was taken per gene and the longest open reading frame determined. Of the remaining 366 coding exons, the mean coding region was 415 bp long (SE = 21.48, min. = 103, max. = 2959) and only 88 had coding regions longer than 500 bp. Thus, assuming a random starting sample, our survey returned 88 "long" coding exons from a potential of 6534 unigenes, or 1.3%. Of these 88 in-frame long exons, a protein BLAST search against Swiss Prot Unigene set found good hits for 82. Known ribosomal protein cDNA sequences were also long exon screened as per the unigene set, adding an additional four long exon candidate genes (see online Appendix).

Several of these annotated genes turned out to be members of large gene families (e.g., tubulin, histone, collagen, lipase, HSP90, HSP70, etc.). One of these genes, HSP70, was assessed for its phylogenetic performance, whereas the rest of these gene family genes were discarded. Of the remaining long exons, we were able to find a sufficient number of hits in Lepidopteran databases for degenerate primer design in 15 candidate genes matching our selection criteria (see online Appendix). Degenerate primers with universal tails amplified 9 of the 15 gene regions with good success (see online Appendix). The six gene regions that failed, either failed completely (no visible bands in the gel), gave very weak bands that did not sequence, or amplified for only certain samples (such as *Bombyx*, *Heliconius*, and/or *Bicyclus*). For those that amplified poorly, it is likely that MgCl₂ concentrations were not optimal, although this was not tested, as we were specifically interested in gene regions that were successful under standard conditions for high-throughput processing. For those that amplified in specific samples, these were always the same species used when designing primers from EST libraries, possibly meaning that primer areas were not conserved enough in other species of Nymphalidae.

The utility of each gene region for phylogenetic reconstruction was assessed by investigating their ability to recover the five major nymphalid lineages identified by Wahlberg et al. (2003) when analyzed singly. These lineages were recovered with strong support in the combined analysis of all gene regions (see below) and are considered to be bona fide clades of Nymphalidae in this

study. Of the 11 gene regions tested, two, CAD and IDH, recovered all five lineages as monophyletic (figures in online Appendix). Six other gene regions (COI, DDC, EF-1 α , RpS5, MDH, GAPDH) recovered most of the major lineages as monophyletic entities, although there appeared to be some rooting problems due to the long branches of outgroup taxa (figures in online Appendix). Three gene regions (wingless, RpS2, and ArgKin) showed low resolution and recovered only some clades with strong support (figures in online Appendix). These clades are composed of taxa that belong to the same subfamilies. One gene, HSP70, showed strongly supported polyphyly of the major lineages (figure in online Appendix). This is likely to be due to paralogy as the heat shock proteins form a large gene family evolving by gene conversion and birth death dynamics (Rensing and Maier, 1994; Bettencourt and Feder, 2002). Indeed, initial trials with this gene gave multiple bands after amplification. Based on these results, we omitted HSP70 from further analyses.

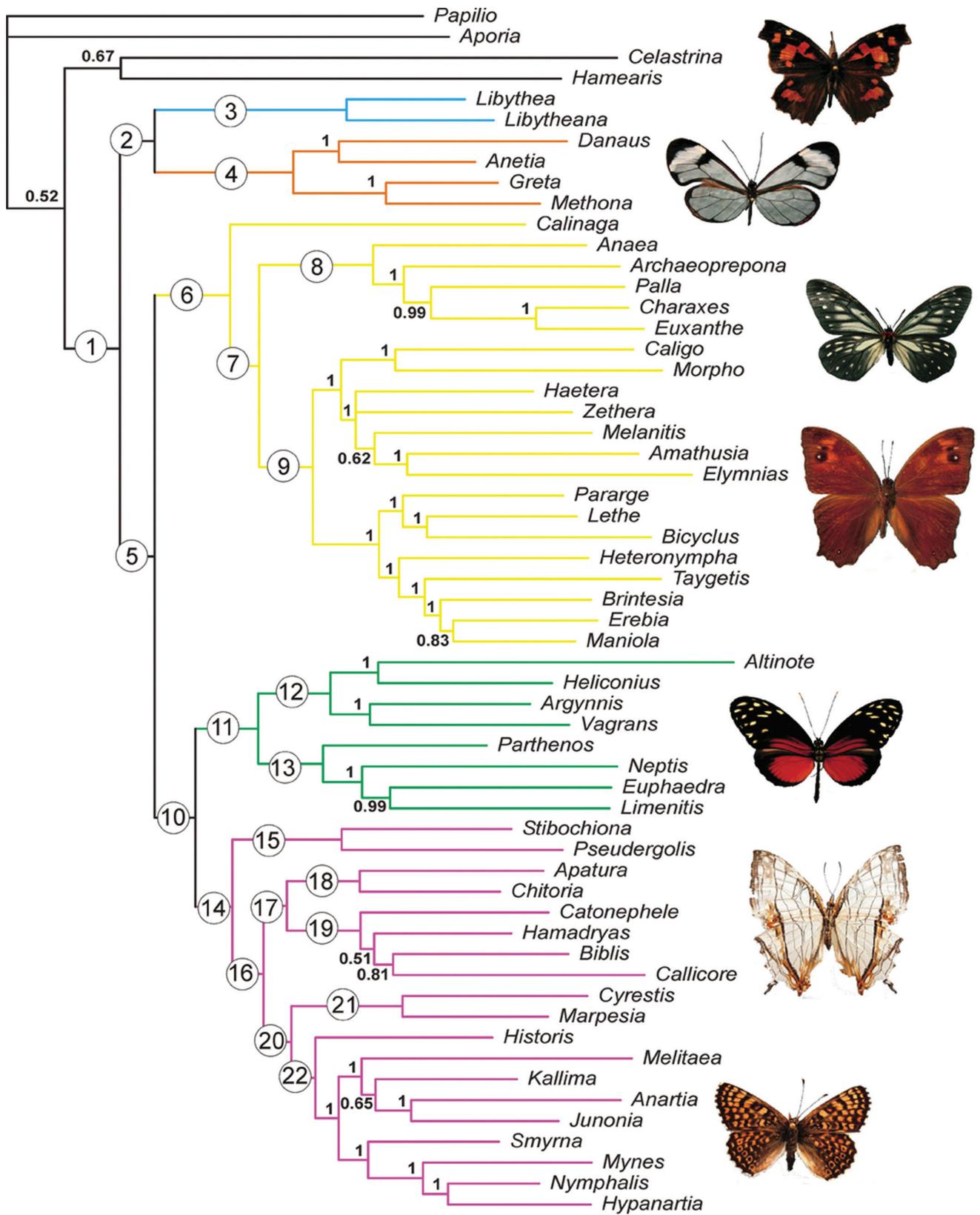
Phylogenetic Analyses

The newly amplified gene regions showed similar levels of variation to the traditional three genes (COI, EF-1 α , and wingless), with parsimony-informative sites varying between 30% and 50% of all sequenced sites (Table 1). The combined 11 gene regions gave a data set comprising 8114 bp. This data set recovers the five major nymphalid lineages identified by Wahlberg et al. (2003) as monophyletic entities with very strong support (Bayesian posterior probabilities of 1.0, bootstraps >0.95, and Bremer support >20 with no or little conflict among the 11 data partitions), regardless of method used for analysis (Fig. 2, Table 2, PBS values in online Appendix). The sister-group relation of the heliconiine (green clade in Fig. 2) and nymphaline (red) clades is also strongly supported, with little conflict between the partitions. The monophyly of Nymphalidae is strongly supported by the Bayesian (posterior probability of 1.0) and maximum likelihood (bootstrap of 0.93) analyses, but only weakly supported by the parsimony analysis (bootstrap 0.5, Bremer support 5).

Within the major lineages, most subfamilies identified by Wahlberg et al. (2003) are strongly supported

TABLE 1. Summary information of 11 gene regions sequenced for exemplars of Lepidoptera. Summary statistics calculated only for papilionoid taxa.

Gene	Taxa amplified	Variable	Parsimony informative	Overall mean K2P distance
ArgKin	40	221 (37%)	186 (31%)	0.117 \pm 0.015
CAD	57	458 (54%)	401 (47%)	0.228 \pm 0.013
COI	57	712 (48%)	571 (38%)	0.143 \pm 0.006
DDC	41	201 (54%)	163 (44%)	0.196 \pm 0.019
EF-1 α	57	482 (39%)	390 (31%)	0.125 \pm 0.006
GAPDH	40	272 (39%)	248 (36%)	0.178 \pm 0.011
IDH	54	337 (47%)	284 (40%)	0.224 \pm 0.016
MDH	53	370 (50%)	319 (44%)	0.208 \pm 0.013
RpS2	47	165 (40%)	150 (36%)	0.202 \pm 0.017
RpS5	56	270 (44%)	238 (39%)	0.191 \pm 0.013
Wgl	56	259 (64%)	195 (48%)	0.236 \pm 0.016



0.1

FIGURE 2.

TABLE 2. Support values for nodes of interest numbered in Figure 2. PCI = Partition congruence index (see text for details).

Node number	Node name	Bayesian PP (11 partitions)	Bayesian PP (33 partitions)	ML bootstrap	Parsimony bootstrap	Bremer support	PCI
1	Nymphalidae	1	1	0.93	0.5	5	-3.4
2	Libytheinae+Danainae	0.86	0.93	0.8	0.3	5	-3.4
3	Libytheinae	1	1	1	1	81	80.5
4	Danainae	1	1	1	1	86	85.8
5	Satyrine+heliconiine+nymphaline	0.93	0.74	0.69	0.87	20	18.7
6	Satyrine clade	1	1	1	0.99	40	39.3
7	Charaxinae+Satyrinae	1	1	0.77	0.59	11	8.7
8	Charaxinae	1	1	1	1	64	63.7
9	Satyrinae	1	1	1	0.75	14	12.3
10	Heliconiine+nymphaline	1	1	0.95	0.92	25	24.7
11	Heliconiine clade	1	1	1	0.97	32	31.3
12	Heliconiinae	1	1	1	0.99	33	33.1
13	Limenitidinae	1	1	1	0.99	47	46.9
14	Nymphaline clade	1	1	1	0.96	21	19.0
15	Pseudergolinae	1	1	1	1	97	96.9
16	Cyrestinae+Nymphalinae+Apaturinae+Biblidinae	1	1	0.99	0.64	10	7.9
17	Biblidinae+Apaturinae	1	1	0.85	0.61	7	0.1
18	Apaturinae	1	1	1	1	63	62.2
19	Biblidinae	1	1	1	1	50	49.8
20	Cyrestinae+Nymphalinae	1	1	0.99	0.37	0	0.0
21	Cyrestinae	1	1	1	1	90	90.0
22	Nymphalinae	1	1	0.99	0.86	13	11.0

monophyletic entities (Table 2). The satyrine lineage comprises the subfamilies Calinaginae, Charaxinae, and Satyrinae, with the latter two being sister subfamilies, with strong support in the Bayesian analyses. Satyrinae is found to contain the tribes Morphini, Brassolini, and Amathusiini (formerly considered a subfamily of their own, Morphinae; e.g., by Wahlberg et al., 2003) with strong support, thus corroborating the findings of Peña et al. (2006), which was based on the three traditional genes. The heliconiine clade comprises two subfamilies, Heliconiinae and Limenitidinae, which are very strongly supported monophyletic subfamilies and each others' sister clades. This unconventional finding, initially reported by Brower (2000b) based on a single gene and by Wahlberg et al. (2003) based on the three traditional genes, is well supported by the full data set analysis.

The relationships of the subfamilies in the nymphaline clade differ from previous studies (Wahlberg et al., 2003, 2005b). There are five well-supported clades that represent the subfamilies Nymphalinae, Biblidinae, and Apaturinae, and the tribes Cyrestini and Pseudergolini (Table 2). The latter two tribes were found to be sister groups by Wahlberg et al. (2003) and were placed in the same subfamily Cyrestinae. However, Wahlberg et al. (2005b) found with greater taxon sampling that the two tribes did not group with each other, and results here show that they are independent lineages deserving

subfamilial rank, with Pseudergolinae being sister to the rest of the nymphaline clade, and Cyrestinae being sister to Nymphalinae. The tribe Coeini was problematic in the study by Wahlberg et al. (2005b), but current results (the tribe represented by the genus *Historis* here) suggest that it is sister to the rest of the subfamily Nymphalinae (Fig. 2). Relationships of the five subfamilies in the nymphaline clade are stable only with 11 gene regions and over 8000 bp sampled, although the sister relationships of Cyrestinae + Nymphalinae and Apaturinae + Biblidinae have strong support only in the Bayesian and maximum likelihood analyses (Table 2).

The Bayesian analysis of the combined data set allows comparison of the estimated rate multiplier parameter (m) across the 11 gene partitions. This parameter describes the relative differences in rates between the partitions and is reported in the output of the program MrBayes (Nylander et al., 2004). Genes vary considerably in their rates of change. The traditional three genes show the full range of variation, with rates for COI \gg wingless $>$ EF-1 α (Fig. 3). There appeared to be a bimodal distribution of rates, with five of the newly designed genes having substitution rates between that of EF-1 α and wingless, and three (CAD, DDC, and IDH) having a much greater rate similar to COI. Single-gene performance (resolution of the five major clades) did not correlate with variation in rate parameters (Fig. 3). In sum, the performance of the traditional three genes may

FIGURE 2. Results of Bayesian analysis of the combined data set partitioned according to gene region (11 partitions). Circled numbers to the left of a node refer to nodes of interest. These were found in all analyses of the combined data set. Support values for these nodes are found in Table 2. Uncircled numbers to the left of a node are Bayesian posterior probabilities from the 11-partition Bayesian analysis. Clades are colored according to the five major lineages identified in this and previous studies (e.g., Wahlberg et al., 2003): Blue = Libytheinae; orange = Danainae; yellow = satyrine clade; green = heliconiine clade; red = nymphaline clade. Butterflies shown are voucher specimens used in this study, from top to bottom: *Libythea*, *Greta*, *Calinaga*, *Melanitis*, *Heliconius*, *Cyrestis*, and *Melitaea*.

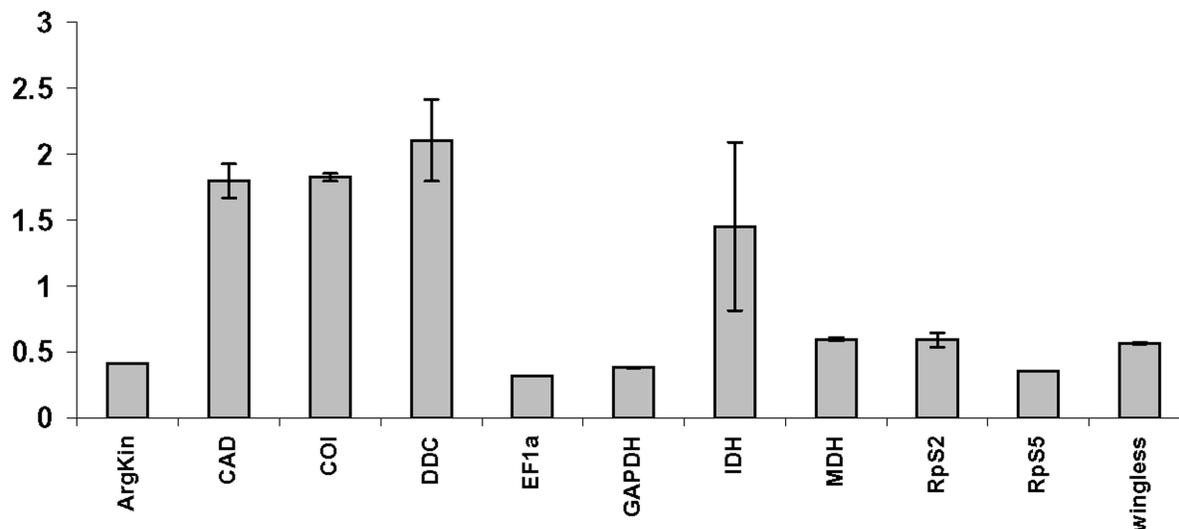


FIGURE 3. Mean estimates of the rate multiplier parameter m by codon position by gene (variance shown in error bars). Means calculated as the average across four independent Bayesian runs, each 5 million generations.

stem from their significant differences in rates of molecular evolutionary change. Our additional genes appear to be adding increased signal at rates roughly similar to these traditional nuclear phylogenetic markers.

Our analysis of data subsets shows a steep increase in bootstrap support values with increasing data set size (Fig. 4), across nodes of interest (Fig. 2). Analyses based on single genes rarely recovered the deeper phylogenetic relationships with any confidence, although several subfamilies were recovered as well-supported monophyletic clades. In addition, there was a lot of variation in bootstrap support among nodes. Three gene data sets also showed a lot of variation among nodes, although almost all subfamily clades are well-supported clades. In the five genes subsets, the bootstrap support variation among nodes is lower and all subfamilies, as well as the five major lineages, are well supported. However, the relationships of subfamilies within the major lineages and the relationships of the major lineages themselves are not well supported. With each of the 11 gene regions included, all nodes of interest are well supported except the sister relationship of the satyrine clade and the heliconiine+nymphaline clade.

Analyses of the data sets with taxa deleted in a balanced or unbalanced manner did not change results for the nodes of interest in any way for the maximum likelihood approach (figures in online Appendix). However, parsimony results were affected by changes in taxa sampled, although the changes were weakly supported and were mainly within the five major clades, which were recovered as monophyletic (figures in online Appendix).

DISCUSSION

A robust and stable phylogenetic understanding of the evolutionary relationships among subfamily to family level taxa is critical for inference of evolutionary processes. In the postgenomic era, attaining a robust or at

least a “more” robust and stable phylogenetic understanding is theoretically possible through the use of increased genomic data. Although phylogenomics holds the promise of providing such phylogenetic inferences (Philippe et al., 2004; Brinkmann et al., 2005; Delsuc et al., 2005), developing new molecular markers in nonmodel species is still a significant roadblock for many systems. Here we present an approach that utilizes EST libraries in conjunction with the nearest genomic reference species for the development of a new suite of molecular markers. This approach can be applied to any taxonomic group with these minimal genomic resources. Our results triple the number of genes and base pairs available for phylogenetic studies in the butterfly family Nymphalidae and possibly in all Lepidoptera. Importantly, these new molecular makers are backward compatible, in that they can be used with the large specimen samples already collected around the world.

Phylogenetic reconstruction assumes that orthologous genes are being analyzed across species. Determining orthology across species is greatly complicated by gene birth and death dynamics inherent in gene families. The use of gene family members would therefore likely result in serious violations of orthologous assumptions. Several of the annotated genes that passed our initial screening for long exons turned out to be members of well-known gene families. Three examples are the tubulin, histone, and HSP70 genes. Tubulin is a member of a highly conserved gene family, with the β form showing no amino acid variation across 60 million years of drosophilid evolution (Nielsen et al., 2006). Nielsen et al. (2006) found other tubulin gene copies that did have faster rates of molecular evolution, in specific structural subsections, but these minor copies gave conflicting phylogenetic information likely due to a birth and death process of gene evolution common to gene families (Nei and Rooney, 2005). Histones are highly duplicated across genomes,

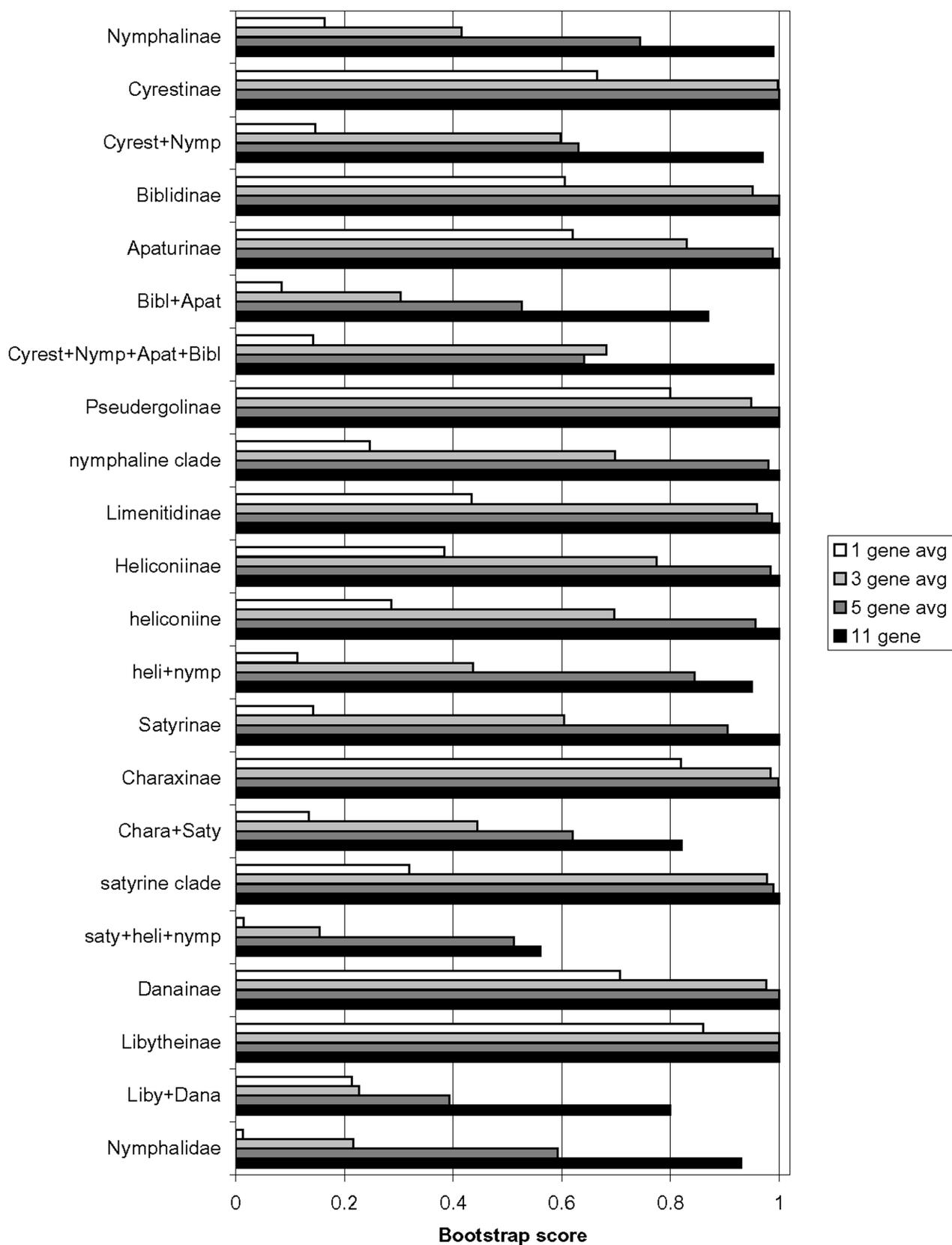


FIGURE 4. Average bootstrap values for the nodes of interest in Figure 2 and Table 2 for subsets of the full data set. Values were taken from maximum likelihood analysis of 100 bootstrapped data sets.

with over 100 copies in *Drosophila melanogaster* alone (Nei and Rooney, 2005). To highlight the possible violations of orthology such genes can present, we amplified and sequenced a long exon from HSP70, finding that it gave strong polyphyly of the major lineages. Future development of novel molecular markers must ensure, as best as possible, that single copy genes are used.

Of the 11 gene regions we have identified in this study, several have been used in previous studies. In addition to the three "traditional" genes, CAD has been used successfully in studies of Diptera and Lepidoptera (e.g., Wiegmann et al., 2000; Regier et al., 2008; Zwick, 2008) and DDC in studies of Lepidoptera (Fang et al., 1997, 2000). CAD appears to be an exceptional gene in that it has a large exon of 3008 bp and is thus potentially useful for getting a large number of base pair with the same evolutionary history. We have, however, chosen a region of only 850 bp, which can easily be amplified in one PCR reaction and sequenced in both directions to allow efficient use of resources. DDC has previously been amplified using reverse transcriptase-PCR and published primers (e.g., Fang et al., 1997) have not worked on genomic extracts.

Our approach allows efficient searching of new gene regions for molecular systematics that can be used with standard genomic extracts of DNA. We used a genomic reference species for gene structure insight (i.e., the intron/exon structure of protein-coding genes). This approach to novel molecular marker design is not limited to Lepidoptera (see, e.g., Li et al., 2007) and should function in any taxonomic group with several EST libraries and a genomic reference species that diverged roughly less than 100 million years ago. Currently we are limited by our search criteria (i.e., using only single-copy nuclear genes), EST library coverage (both across and within taxa), and *B. mori* WGS contig assembly for designing more exon specific primers. However, more ESTs are in the pipeline for several species across Lepidoptera and a new assembly of the combined sequences from both the Japanese and Chinese *B. mori* sequencing consortia is due to be released soon.

By comparing the genome of *Bombyx mori* with EST libraries of nymphalid butterflies, we were able to design primers that appear to be universal across Papilionoidea and perhaps Lepidoptera as well. The exact relationship of Bombycidae (to which *Bombyx* belongs to) and Nymphalidae is not known at the moment, although both belong to the higher Ditrysia crown group of Lepidoptera (Kristensen, 1999). Our primers were successful in two species that are distantly related to *Bombyx* and Nymphalidae, the monotrysian *Hepialus* and the basal ditryisian *Depressaria*. However, the universality of the primers did not extend beyond Lepidoptera, as most primers failed with the exemplar of Trichoptera, which is the sister order of Lepidoptera (Kristensen, 1999). Lepidoptera is one of the megadiverse orders, along with Coleoptera, Hymenoptera, and Diptera, with some 170,000 species described to date (Kristensen, 1999). Thus our universal primers are potentially useful in studies of evolutionary processes leading to high diversity.

Our focus on long exons of protein-coding genes allows the new gene regions to be amplified from standard genomic extracts of DNA, even from specimens with degraded DNA. In studies of Lepidoptera, DNA is often extracted from dried specimens that may have been stored for several years in collections in suboptimal conditions. In our study, such specimens are *Lethe*, *Elymnias*, *Erebia*, and *Smyrna* (table in online appendix), all of which gave good PCR amplifications of the new gene regions.

Increasing the number of gene regions that can be easily sequenced is of interest for several groups where taxon sampling is no longer an issue. The debate over "more taxa versus more data" (Graybeal, 1998; Mitchell et al., 2000) is moving on to the question of "How much data are enough?" and how to analyze them (Rokas et al., 2003; Delsuc et al., 2005; Gatesy et al., 2007). Our results suggest that for a group of insects that has experienced about 100 million years of evolution (Wahlberg, 2006), sequences from 1 gene region are not enough to resolve the relationships of major lineages with confidence, sequences from 3 and 5 gene regions are enough to find strongly supported clades, and sequences from 11 gene regions are enough to solidify most clades that are ambiguous with less data. Interestingly, the method of analysis did not affect the results of the combined analyses and each method contributed significantly to the understanding of the behavior of the data. Additionally, we attempted to address taxon sampling issues by creating two smaller, equal-sized subset of taxa, one "balanced" and the other "unbalanced" phylogenetically according to our final consensus tree. There was little, if any, topological difference between these two taxa subsets, regardless of assumptions of analysis, which agreed with the full taxa set consensus. Thus, using a reasonable sampling of taxa, considering both number and likely phylogenetic relationships, coupled with a large gene data set of diverse substitution rates, can result in very robust phylogenetic inference. A more detailed analysis of the relationship between substitution rates (i.e., functional constraint) and phylogenetic signal is beyond the scope of this study as we lack sufficient genes among functional constraint types. However, such a data set is of interest for future research into the number and types of molecular markers needed for specific phylogenetic questions.

Our current results suggest that the five well-supported major lineages in the family Nymphalidae identified by Wahlberg et al. (2003) are robust to the addition of new data. In addition to this, the sister relationship of the heliconiine and nymphaline clades is robust and stable in all analyses. We also identify 12 clades that correspond to subfamilies of Nymphalidae. Most of the subfamilies are the same as those suggested by Wahlberg et al. (2003), but with Morphinae being within Satyrinae (as found by Peña et al. 2006) and Cyrestinae being split into Cyrestinae and Pseudergolinae. The 12 subfamilies identified in this study are strongly supported in all the combined analyses and are likely to remain robust to the addition of new data.

In sum, our newly designed primers (see online Appendix) will facilitate lepidopteran phylogenomics.

These primers appear to be universal in Lepidoptera (particularly those for CAD, IDH, MDH, and Rp55; see online Appendix) and can be used with standard genomic extracts from dried specimens. We expect that a further 10 new gene regions should be easily discovered using our methods when the next version of the *Bombyx mori* genome emerges and more EST libraries become public.

ACKNOWLEDGMENTS

We are grateful to Kelvin Guererra, Dan Janzen, and Torben B. Larsen for help with getting specimens for this study. We thank Julien Leneveu for help in the laboratory. Comments by Jack Sullivan, Lacey Knowles, Jerome Regier, Felix Sperling, and two anonymous referees substantially improved the manuscript. N.W. wishes to acknowledge funding from the Swedish Research Council (grant no. 621-2004-2853) and the Academy of Finland (grant no. 118369). C.W.W. wishes to acknowledge funding from the US National Science Foundation (grant IBN-0412651 awarded to J. H. Marden and I. Hanski) and support from the Academy of Finland (grant nos. 38604 and 44887, Finnish Centre of Excellence Programme, 2000–2005, awarded to I. Hanski).

REFERENCES

- Baker, R. H., and R. DeSalle. 1997. Multiple sources of character information and the phylogeny of Hawaiian drosophilids. *Syst. Biol.* 46:654–673.
- Beldade, P., S. Rudd, J. D. Gruber, and A. D. Long. 2006. A wing expressed sequence tag resource for *Bicyclus anynana* butterflies, an evo-devo model. *BMC Genomics* 7:130.
- Bettencourt, B. R., and M. E. Feder. 2002. Rapid concerted evolution via gene conversion at the *Drosophila* hsp70 genes. *J. Mol. Evol.* 54:569–586.
- Biology Analysis Group. 2004. A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science* 306:1937–1940.
- Boggs, C. L., W. B. Watt, and P. R. Ehrlich, eds. 2003. *Butterflies: Evolution and ecology taking flight*. University of Chicago Press, Chicago.
- Boucher, Y., C. J. Douady, R. T. Papke, D. A. Walsh, M. E. Boudreau, C. L. Nesbo, R. J. Case, and W. F. Doolittle. 2003. Lateral gene transfer and the origins of prokaryotic groups. *Ann. Rev. Genet.* 37:283–328.
- Braby, M. F., R. Vila, and N. E. Pierce. 2006. Molecular phylogeny and systematics of the Pieridae (Lepidoptera: Papilionoidea): Higher classification and biogeography. *Zool. J. Linn. Soc.* 147:239–275.
- Bremer, K. 1988. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. *Evolution* 42:795–803.
- Bremer, K. 1994. Branch support and tree stability. *Cladistics* 10:295–304.
- Brinkmann, H., M. Van der Giezen, Y. Zhou, G. P. De Raucourt, and H. Philippe. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Syst. Biol.* 54:743–757.
- Brooks, D. R., J. Bilewicz, C. Condy, D. C. Evans, K. E. Folsinbee, J. Fröbisch, D. Halas, S. Hill, D. A. McLennan, M. Mattern, L. A. Tsuji, J. L. Ward, N. Wahlberg, D. Zamparo, and D. Zanatta. 2007. Quantitative phylogenetic analysis in the 21st century. *Rev. Mex. Biodivers.* 78:225–252.
- Brower, A. V. Z. 2000a. Evolution is not a necessary assumption of cladistics. *Cladistics* 16:143–154.
- Brower, A. V. Z. 2000b. Phylogenetic relationships among the Nymphalidae (Lepidoptera), inferred from partial sequences of the *wingless* gene. *Proc. R. Soc. Lond. Ser. B Biol. Sci.* 267:1201–1211.
- Brower, A. V. Z. 2006. The how and why of branch support and partitioned branch support, with a new index to assess partition incongruence. *Cladistics* 22:378–386.
- Brower, A. V. Z., and R. DeSalle. 1998. Patterns of mitochondrial versus nuclear DNA sequence divergence among nymphalid butterflies: The utility of *wingless* as a source of characters for phylogenetic inference. *Ins. Mol. Biol.* 7:73–82.
- Brower, A. V. Z., A. V. L. Freitas, M.-M. Lee, K. L. Silva Brandão, A. Whinnett, and K. R. Willmott. 2006. Phylogenetic relationships among the Ithomiini (Lepidoptera: Nymphalidae) inferred from one mitochondrial and two nuclear gene regions. *Syst. Entomol.* 31:288–301.
- Cameron, S. L., K. B. Miller, C. D'Haese, M. F. Whiting, and S. C. Barker. 2004. Mitochondrial genome data alone are not enough to unambiguously resolve the relationships of Entognatha, Insecta and Crustacea *sensu lato* (Arthropoda). *Cladistics* 20:534–557.
- Chen, M. S., R. D. Reed, M. M. Kuo, and F. A. H. Sperling. 2001. A partitioned likelihood analysis of swallowtail butterfly phylogeny (Lepidoptera: Papilionidae). *Syst. Biol.* 50:106–127.
- Cho, S. W., A. Mitchell, J. C. Regier, C. Mitter, R. W. Poole, T. P. Friedlander, and S. W. Zhao. 1995. A highly conserved nuclear gene for low-level phylogenetics—Elongation factor-1-alpha recovers morphology-based tree for heliothine moths. *Mol. Biol. Evol.* 12:650–656.
- Delsuc, F., H. Brinkmann, and H. Philippe. 2005. Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.* 6:361–375.
- Deutsch, M., and M. Long. 1999. Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res.* 27:3219–3228.
- Ehrlich, P. R., and I. Hanski, eds. 2004. *On the wings of checkerspot: A model system for population biology*. Oxford University Press, New York.
- Fang, Q. Q., S. Cho, J. C. Regier, C. Mitter, M. Matthews, R. W. Poole, T. P. Friedlander, and S. Zhao. 1997. A new nuclear gene for insect phylogenetics: Dopa decarboxylase is informative of relationships within Heliothinae (Lepidoptera: Noctuidae). *Syst. Biol.* 46:269–283.
- Fang, Q. Q., A. Mitchell, J. C. Regier, C. Mitter, T. P. Friedlander, and R. W. Poole. 2000. Phylogenetic utility of the nuclear gene dopa decarboxylase in noctuid moths (Insecta: Lepidoptera: Noctuoidea). *Mol. Phylogenet. Evol.* 15:473–486.
- Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39:783–791.
- Felsenstein, J. 2004. *Inferring phylogenies*. Sinauer Associates, Sunderland, Massachusetts.
- Folmer, O., M. B. Black, W. Hoch, R. A. Lutz, and R. C. Vrijehock. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol. Mar. Biol. Biotechnol.* 3:294–299.
- Freitas, A. V. L., and K. S. J. Brown. 2004. Phylogeny of the Nymphalidae (Lepidoptera: Papilionoidea). *Syst. Biol.* 53:363–383.
- Friedlander, T. P., J. C. Regier, C. Mitter, and D. L. Wagner. 1996. A nuclear gene for higher level phylogenetics: Phosphoenolpyruvate carboxykinase tracks Mesozoic-age divergences within Lepidoptera (Insecta). *Mol. Biol. Evol.* 13:594–604.
- Gatesy, J., R. deSalle, and N. Wahlberg. 2007. How many genes should a systematist sample? Conflicting insights from a phylogenomic matrix characterized by replicated incongruence. *Syst. Biol.* 56:355–363.
- Gatesy, J., P. O'Grady, and R. H. Baker. 1999. Corroboration among data sets in simultaneous analysis: Hidden support for phylogenetic relationships among higher level artiodactyl taxa. *Cladistics* 15:271–313.
- Giribet, G. 2003. Stability in phylogenetic formulations and its relation to nodal support. *Syst. Biol.* 52:554–564.
- Goloboff, P. A. 1999. Analyzing large data sets in reasonable times: Solutions for composite optima. *Cladistics* 15:415–428.
- Goloboff, P. A., J. S. Farris, and K. C. Nixon. 2004. T. N. T. (Tree Analysis using New Technology), version 1.0. Published by the authors.
- Graybeal, A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst. Biol.* 47:9–17.
- Hall, T. A. 1999. BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids. Symp. Ser.* 41:95–98.
- Jiggins, C. D., J. Mavarez, M. Beltrán, W. O. McMillan, J. S. Johnston, and E. Bermingham. 2005. A genetic linkage map of the mimetic butterfly, *Heliconius melpomene*. *Genetics* 171:557–570.
- Kluge, A. G. 2001. Philosophy and phylogenetic inference: A comparison of likelihood and parsimony methods in the context of Karl Popper's writings on corroboration. *Cladistics* 17:395–399.
- Kristensen, N. P., ed. 1999. *Lepidoptera, moths and butterflies*. 1. Evolution, systematics and biogeography. *Handbook of zoology* 4 (35), Lepidoptera. de Gruyter, Berlin.
- Li, C. H., G. Orti, G. Zhang, and G. Q. Lu. 2007. A practical approach to phylogenomics: The phylogeny of ray-finned fish (Actinopterygii) as a case study. *BMC Evol. Biol.* 7:44.

- Mallarino, R., E. Bermingham, K. R. Willmott, A. Whinnett, and C. D. Jiggins. 2005. Molecular systematics of the butterfly genus *Ithomia* (Lepidoptera: Ithomiinae): A composite phylogenetic hypothesis based on seven genes. *Mol. Phylogenet. Evol.* 34:625–644.
- Mita, K., M. Kasahara, S. Sasaki, Y. Nagayasu, T. Yamada, H. Kanamori, N. Namiki, M. Kitagawa, H. Yamashita, Y. Yasukochi, K. Kadono-Okuda, K. Yamamoto, M. Ajimura, G. Ravikumar, M. Shimomura, Y. Nagamura, T. Shin-I, H. Abe, T. Shimada, S. Morishita, and T. Sasaki. 2004. The genome sequence of silkworm, *Bombyx mori*. *DNA Res.* 11:27–35.
- Mitchell, A., C. Mitter, and J. C. Regier. 2000. More taxa or more characters revisited: Combining data from nuclear protein-encoding genes for phylogenetic analyses of Noctuoidea (Insecta: Lepidoptera). *Syst. Biol.* 49:202–224.
- Moilanen, A. 1999. Searching for most parsimonious trees with simulated evolutionary optimization. *Cladistics* 15:39–50.
- Nardi, F., G. Spinsanti, J. L. Boore, A. Carapelli, R. Dallai, and F. Frati. 2003. Hexapod origins: Monophyletic or paraphyletic. *Science* 299:1887–1889.
- Nazari, V., E. V. Zakharov, and F. A. H. Sperling. 2007. Phylogeny, historical biogeography, and taxonomic ranking of Parnassiinae (Lepidoptera, Papilionidae) based on morphology and seven genes. *Mol. Phylogenet. Evol.* 42:131–156.
- Nei, M., and A. P. Rooney. 2005. Concerted and birth-and-death evolution of multigene families. *Ann. Rev. Genet.* 39:121–152.
- Nielsen, M. G., J. M. Caserta, S. J. Kidd, and C. M. Phillips. 2006. Functional constraint underlies 60 million year stasis of Dipteran testis-specific beta-tubulin. *Evol. Dev.* 8:23–29.
- Nixon, K. C. 1999. The parsimony ratchet, a new method for rapid parsimony analysis. *Cladistics* 15:407–414.
- Nylander, J. A. A. 2002. MrModelTest v2.1. Available from author.
- Nylander, J. A. A., F. Ronquist, J. P. Huelsenbeck, and J. L. Nieves-Aldrey. 2004. Bayesian phylogenetic analysis of combined data. *Syst. Biol.* 53:47–67.
- Ochman, H., J. G. Lawrence, and E. A. Groisman. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* 405:299–304.
- Peña, C., N. Wahlberg, E. Weingartner, U. Kodandaramaiah, S. Nylin, A. V. L. Freitas, and A. V. Z. Brower. 2006. Higher level phylogeny of Satyrinae butterflies (Lepidoptera: Nymphalidae) based on DNA sequence data. *Mol. Phylogenet. Evol.* 40:29–49.
- Philippe, H., E. A. Snell, E. Baptiste, P. Lopez, P. W. H. Holland, and D. Casane. 2004. Phylogenomics of eukaryotes: Impact of missing data on large alignments. *Mol. Biol. Evol.* 21:1740–1752.
- Reed, R. D., and F. A. H. Sperling. 1999. Interaction of process partitions in phylogenetic analysis: An example from the swallowtail butterfly genus *Papilio*. *Mol. Biol. Evol.* 16:286–297.
- Regier, J. C., C. P. Cook, C. Mitter, and A. Hussey. 2008. A phylogenetic study of the “bombycoid complex” (Lepidoptera) using five protein-coding nuclear genes, with comments on the problem of macrolepidopteran phylogeny. *Syst. Entomol.* 33:175–189.
- Regier, J. C., Q. Q. Fang, C. Mitter, R. S. Peigler, T. P. Friedlander, and M. A. Solis. 1998. Evolution and phylogenetic utility of the *period* gene in Lepidoptera. *Mol. Biol. Evol.* 15:1172–1182.
- Regier, J. C., and D. Shi. 2005. Increased yield of PCR product from degenerate primers with nondegenerate, nonhomologous 5' tails. *BioTechnology* 38:34–38.
- Ren, F., H. Tanaka, and Z. Yang. 2005. An empirical examination of the utility of codon-substitution models in phylogeny reconstruction. *Syst. Biol.* 54:808–818.
- Rensing, S. A., and U. G. Maier. 1994. Phylogenetic analysis of the stress-70 protein family. *J. Mol. Evol.* 39:80–86.
- Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425:798–804.
- Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574.
- Simon, C., F. Frati, A. Beckenbach, B. Crespi, H. Liu, and P. Flook. 1994. Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. *Ann. Entomol. Soc. Am.* 87:651–701.
- Simonsen, T. J., N. Wahlberg, A. V. Z. Brower, and R. de Jong. 2006. Morphology, molecules and Fritillaries: Approaching a stable phylogeny for Argynnini (Lepidoptera: Nymphalidae). *Ins. Syst. Evol.* 37:405–418.
- Sperling, F. A. H. 2003. Butterfly species and molecular phylogenies. Pages 431–458 in *Butterflies: Evolution and ecology taking flight* (C. L. Boggs, W. B. Watt, and P. R. Ehrlich, eds.). University of Chicago Press, Chicago.
- Stamatakis, A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.
- Vera, J. C., C. W. Wheat, H. W. Fescemyer, M. J. Frilander, D. L. Crawford, I. Hanski, and J. H. Marden. 2008. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Mol. Ecol.* 17:1636–1647.
- Wahlberg, N. 2006. That awkward age for butterflies: Insights from the age of the butterfly subfamily Nymphalinae. *Syst. Biol.* 55:703–714.
- Wahlberg, N., M. F. Braby, A. V. Z. Brower, R. de Jong, M.-M. Lee, S. Nylin, N. Pierce, F. A. Sperling, R. Vila, A. D. Warren, and E. Zakharov. 2005a. Synergistic effects of combining morphological and molecular data in resolving the phylogeny of butterflies and skippers. *Proc. R. Soc. Lond. Ser. B Biol. Sci.* 272:1577–1586.
- Wahlberg, N., A. V. Z. Brower, and S. Nylin. 2005b. Phylogenetic relationships and historical biogeography of tribes and genera in the subfamily Nymphalinae (Lepidoptera: Nymphalidae). *Biol. J. Linn. Soc.* 86:227–251.
- Wahlberg, N., and A. V. L. Freitas. 2007. Colonization of and radiation in South America by butterflies in the subtribe Phyciodina (Lepidoptera: Nymphalidae). *Mol. Phylogenet. Evol.* 44:1257–1272.
- Wahlberg, N., and S. Nylin. 2003. Morphology versus molecules: Resolution of the positions of *Nymphalis*, *Polygonia* and related genera (Lepidoptera: Nymphalidae). *Cladistics* 19:213–223.
- Wahlberg, N., E. Weingartner, and S. Nylin. 2003. Towards a better understanding of the higher systematics of Nymphalidae (Lepidoptera: Papilionoidea). *Mol. Phylogenet. Evol.* 28:473–484.
- Wheeler, W. C. 1995. Sequence alignment, parameter sensitivity, and the phylogenetic analysis of molecular data. *Syst. Biol.* 44:321–331.
- Whinnett, A., A. V. Z. Brower, M.-M. Lee, K. R. Willmott, and J. Mallet. 2005. Phylogenetic utility of *Tektin*, a novel region for inferring systematic relationships among Lepidoptera. *Ann. Entomol. Soc. Am.* 98:873–886.
- Wiegmann, B. M., C. Mitter, J. C. Regier, T. P. Friedlander, D. M. Wagner, and E. S. Nielsen. 2000. Nuclear genes resolve Mesozoic-aged divergences in the insect order Lepidoptera. *Mol. Phylogenet. Evol.* 15:242–259.
- Zwick, A. 2008. Molecular phylogeny of Anthelidae and other bombycoid taxa (Lepidoptera: Bombycoidea). *Syst. Entomol.* 33:190–209.
- Zwickl, D. J. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. PhD dissertation, Department of Biology The University of Texas at Austin, Austin, Texas.

First submitted 7 June 2007; reviews returned 14 August 2007; final acceptance 23 December 2007

Associate Editor: L. Lacey Knowles