BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Increased gene sampling strengthens support for higher-level groups within leaf-mining moths and relatives (Lepidoptera: Gracillariidae)

Akito Y Kawahara[1*], Issei Ohshima[2], Atsushi Kawakita[3], Jerome C Regier[4], Charles Mitter[1], Michael P Cummings[5], Donald R Davis[6], David L Wagner[7], Jurate De Prins[8] and Carlos Lopez-Vaamonde[9]

Abstract

Background: Researchers conducting molecular phylogenetic studies are frequently faced with the decision of what to do when weak branch support is obtained for key nodes of importance. As one solution, the researcher may choose to sequence additional orthologous genes of appropriate evolutionary rate for the taxa in the study. However, generating large, complete data matrices can become increasingly difficult as the number of characters increases. A few empirical studies have shown that augmenting genes even for a subset of taxa can improve branch support. However, because each study differs in the number of characters and taxa, there is still a need for additional studies that examine whether incomplete sampling designs are likely to aid at increasing deep node resolution. We target Gracillariidae, a Cretaceous-age (~100 Ma) group of leaf-mining moths to test whether the strategy of adding genes for a subset of taxa can improve branch support for deep nodes. We initially sequenced ten genes (8,418 bp) for 57 taxa that represent the major lineages of Gracillariidae plus outgroups. After finding that many deep divergences remained weakly supported, we sequenced eleven additional genes (6,375 bp) for a 27-taxon subset. We then compared results from different data sets to assess whether one sampling design can be favored over another. The concatenated data set comprising all genes and all taxa and three other data sets of different taxon and gene sub-sampling design were analyzed with maximum likelihood. Each data set was subject to five different models and partitioning schemes of non-synonymous and synonymous changes. Statistical significance of non-monophyly was examined with the Approximately Unbiased (AU) test.

Results: Partial augmentation of genes led to high support for deep divergences, especially when non-synonymous changes were analyzed alone. Increasing the number of taxa without an increase in number of characters led to lower bootstrap support; increasing the number of characters without increasing the number of taxa generally increased bootstrap support. More than three-quarters of nodes were supported with bootstrap values greater than 80% when all taxa and genes were combined. Gracillariidae, Lithocolletinae + *Leucanthiza*, and *Acrocercops* and *Parectopa* groups were strongly supported in nearly every analysis. *Gracillaria* group was well supported in some analyses, but less so in others. We find strong evidence for the exclusion of Douglasiidae from Gracillarioidea sensu Davis and Robinson (1998). Our results strongly support the monophyly of a G.B.R.Y. clade, a group comprised of Gracillariidae + Bucculatricidae + Roeslerstammiidae + Yponomeutidae, when analyzed with non-synonymous changes only, but this group was frequently split when synonymous and non-synonymous substitutions were analyzed together.

Conclusions: 1) Partially or fully augmenting a data set with more characters increased bootstrap support for particular deep nodes, and this increase was dramatic when non-synonymous changes were analyzed alone. Thus, the addition of sites that have low levels of saturation and compositional heterogeneity can greatly improve results. 2) Gracillarioidea, as defined by Davis and Robinson (1998), clearly do not include Douglasiidae, and

* Correspondence: kawahara@flmnh.ufl.edu
[1]Department of Entomology, University of Maryland, College Park, MD, USA
Full list of author information is available at the end of the article

changes to current classification will be required. 3) Gracillariidae were monophyletic in all analyses conducted, and nearly all species can be placed into one of six strongly supported clades though relationships among these remain unclear. 4) The difficulty in determining the phylogenetic placement of Bucculatricidae is probably attributable to compositional heterogeneity at the third codon position. From our tests for compositional heterogeneity and strong bootstrap values obtained when synonymous changes are excluded, we tentatively conclude that Bucculatricidae is closely related to Gracillariidae + Roeslerstammiidae + Yponomeutidae.

Background

Researchers conducting molecular phylogenetic studies are frequently faced with the decision of what to do when weak branch support is obtained for key nodes of importance. As one solution, the researcher may choose to sequence additional orthologous genes of appropriate evolutionary rate. Indeed, it is well known that increasing the number of characters can improve branch support (e.g. [1-5]). However, generating large, complete data matrices can become increasingly difficult as the number of characters increases. Two empirical studies [6,7] have concluded that augmenting genes even for a subset of taxa can improve branch support. However, because each study differs in the number of characters and taxa, there is still a need for additional studies that examine whether incomplete sampling designs are likely to aid at increasing deep node resolution.

In this paper, we target Gracillariidae, a Cretaceous-age (~100 Ma) group of leaf-mining moths [8] to test whether the strategy of adding genes for a subset of taxa can improve branch support for deep nodes. Gracillariidae, with 1,855 species [9,10], is one of the largest groups of leaf-mining Lepidoptera with numerous economically important species that cause agricultural damage [9,11-16]. Gracillariids show a diversity of life-history strategies, such as fruit mining, stem mining, leaf rolling, boring, and galling [11,17], and some species change strategies during development [17-20]. Despite the agricultural importance and diversity of life-history strategies, the systematics of Gracillariidae is poorly understood. Monophyly of the superfamily Gracillarioidea as currently defined by Davis and Robinson [11] remains uncertain. The phylogenetic position of Gracillarioidea in Lepidoptera is also relatively unclear, though recent molecular studies strongly support a close relationship to Yponomeutoidea [7,21,22].

Davis and Robinson's classification includes four families in Gracillarioidea, Bucculatricidae, Douglasiidae, Gracillariidae, and Roeslerstammiidae. Bucculatricidae and Douglasiidae were included in Gracillarioidea based on nine morphological features that they share with Gracillariidae and Roeslerstammiidae, including two from the larva, two from the pupa, and five from the adult [11]. Others have included Bucculatricidae, Gracillariidae, and Phyllocnistidae (the latter now in Gracillariidae [9,10,23,24]), Bucculatricidae, Gracillariidae, and Lyonetiidae [25], or Bucculatricidae, Gracillariidae and Roeslerstammiidae [26]. While some putative relationships have been postulated for higher-level relationships within Gracillarioidea based on morphology (e.g. [24,27], the trees presented in these studies were based on phenetic similarity rather than discrete character analysis. The only phylogenetic study that examined higher-level gracillariid relationships was a recent study aimed at resolving broader relationships of Lepidoptera that included 14 gracillarioid species [22]. The authors suggested that Gracillarioidea might not include Bucculatricidae or Douglasiidae. Most phylogenetic studies within Gracillarioidea have focused mainly at the genus level or below (e.g. host races of *Acrocercops transecta* [28,29], *Epicephala* and relatives [30-32], and *Phyllonorycter* [33,34]).

This study utilizes 21 nuclear protein-coding genes to evaluate the effect of augmenting sequence data for a subsample of taxa and to tackle the problem of the phylogeny of Gracillariidae and their relatives. Fifty-seven taxa, including exemplars representing the major lineages of Gracillarioidea plus outgroups, were initially sequenced for ten genes (8,418 bp). After discovering that many deep divergences within the superfamily could not be recovered with strong branch support, we sequenced 11 additional genes (6,375 bp) for 27 taxa representing the major lineages of Gracillarioidea (21 genes total, 14,793 bp). We compared results from four data sets differing in gene and taxon sampling design (Figure 1), to assess whether one design can be favored over another. We also examined the effect of excluding synonymous changes, which at deeper levels in our taxon sample are subject both to saturation and to divergence in base composition, possibly obscuring phylogenetic signal.

Methods

Taxon sampling

The present study included 45 species of Gracillarioidea, of which 39 were Gracillariidae (Additional file 1). Taxa were chosen to represent the major lineages as defined by the classification of Davis and Robinson [11]. Whenever possible, we included the type species or genus. Twelve outgroups were chosen based on the availability of sequence data and their phylogenetic proximity to Gracillarioidea in two recent molecular phylogenetic studies of

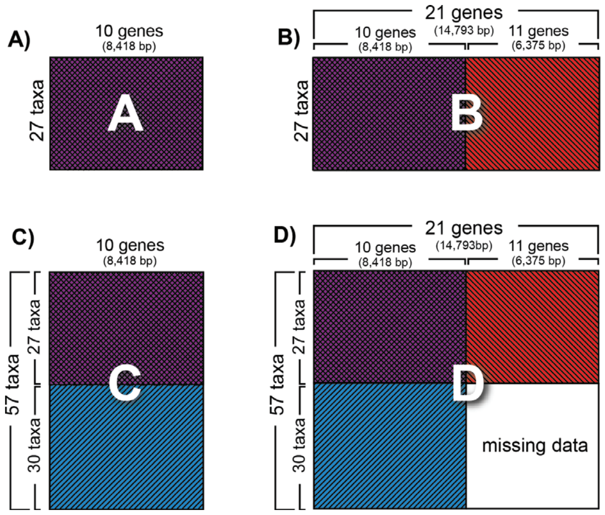

Figure 1 Four data sets representing different taxon/gene sampling strategies. A) 27 taxa × 10 genes, B) 27 taxa × 21 genes, C) 57 taxa × 10 genes, D) combination of B and C into a single data set with a block of missing data accounting for approximately a quarter of total data.

ditrysian Lepidoptera [7,21]. While we mention Oecophyllembiinae in our discussion, we follow Davis and Robinson [11] and Vári et al. [35] and do not formally recognize this subfamily.

Gene sampling

Ten nuclear protein-coding genes, totalling an alignment length of 8,418 bp, were initially chosen for this study (Table 1). These genes were included because they had some of the highest amplification success rates and had proven useful for estimating a "backbone" phylogeny of Lepidoptera (see http://www.leptree.net/) and are among the 68 gene regions originally developed for deep-level phylogenetics of Arthropoda [36]. We created two data sets for ten genes, one with 27 taxa (data set A, Figure 1A) and another with 57 taxa (data set C, Figure 1C). After discovering that ten genes did not adequately resolve phylogenetic relationships among subfamilies, we chose a resource-efficient approach to tackle the problem of weak branch support for deep nodes in the tree. We additionally sequenced eleven genes for 27 taxa and combined this additional sequence data with the original ten (data set B, 14,793 bp; Figure 1B). These additional eleven genes were chosen based on amplification success and examining the average rate of non-synonymous change from a previous study [36]. Throughout the paper, we refer to the "ten genes" and "eleven genes" as the original ten and additional eleven genes. Because our primary goal was to produce the best estimate for relationships of Gracillariidae and relatives, we combined data sets A through C to create data set D (57 taxa × 21 genes, 14,793 bp; Figure 1D). We assessed the differences in tree topology and branch support among these data sets and tested the effect of synonymous changes on each.

Specifically, data set A (27 taxa × 10 genes) was constructed to examine whether modest taxa and gene sampling can strongly resolve relationships of Gracillariidae and relatives. Data set B (27 taxa × 21 genes) was constructed to examine whether nearly doubling the number of characters would boost branch support for deep splits that were not well supported with ten genes. Data set C (57 taxa × 10 genes) was built to assess the effects of adding more taxa to data set A. Finally, data set D (57 taxa × 21 genes) was constructed to examine how trees generated from a data set that included all taxa and genes (but also approximately a quarter of the characters without data) compared to those from the other three data sets. The amount of missing data for each of the four data sets (A through D) was 13.6%, 14.3%, 31.2%, and 43.6% respectively.

For all genes except elongation factor-1 alpha (EF-1α [37]) and histone 3 (H3 [38]), nucleic acid sequences were generated from mRNAs amplified with RT-PCR following laboratory protocols, primer sequences, and amplification strategies of Regier et al. [39]. For EF-1α and H3, we followed methods outlined in Kawakita et al. [30], Kawakita and Kato [40], Ogden and Whiting [38], and amplified directly from genomic DNA. Each single-gene data set was individually translated and aligned with the "Translation Align" option in Geneious 5.1 [41] after making sure the data set began with the first codon position (nt1). The alignment was visually inspected, and checked twice for frame-shifts and the presence of termination codons. Difficult to align regions were assessed in GBlocks 0.91b [42] and removed as they can cause problems in phylogeny estimation [43,44]. Sequences were also assessed for contamination and sample-switching error, by generating pairwise distance tables for nt12, nt3, and nt123 in PAUP* 4b10 [45] and ML bootstrap trees in GARLI 1.0 [46] for each gene before all genes were concatenated. The 21 data matrices were concatenated with Geneious [41] and the entire edited sequence data set visually checked. GenBank accession numbers are listed in Additional file 1.

Phylogenetic analysis

Phylogenetic analyses were conducted with maximum likelihood (ML) as implemented in GARLI 1.0 [46] and GARLI-PART 0.97 [47]. All settings were kept as default except where indicated below. We used jModelTest [48] to determine the best substitution model for each data set, which in nearly all cases was the General-Time-Reversible (GTR) model [49,50], with among-site rate heterogeneity modeled according to a gamma (Γ) distribution [51] while allowing for a proportion of invariable sites (I) [52]. Two thousand ML and bootstrap tree searches were conducted

Table 1 Representation of genes and their amplicon names in each of the four data sets

			Data set			
			A	B	C	D
Gene	Amplicon name and reference	Length (bp)	10 g × 27 t	21 g × 27 t	10 g × 57 t	21 g × 27 t
40fin2_3	Phosphogluconate dehydrogenase [36]	750		X		X
42fin1_2	Putative GTP-binding protein [36]	840		X		X
109fin1_2	Gelsolin [36]	552	X	X	X	X
192fin1_2	Glutamyl- & prolyl-tRNA sybphetase [36]	402		X		X
197fin1_2	Triosephosphate isomerase [36]	444		X		X
262fin1_2	Proteasome subunit [36]	501		X		X
265fin2_3	Histidyl-tRNA sybphetase [36]	447	X	X	X	X
268fin1_2	AMP deaminase [36]	768	X	X	X	X
3007fin1_2	Glucose phosphosphate dehydrogenase [36]	621	X	X	X	X
3017fin1_2	Tetrahydrofolate sybphase [36]	594		X		X
3070fin4_5	Alanyl-tRNA sybphetase [36]	705		X		X
8028fin1_2	Nucleolar cysteine-rich protein [36]	324		X		X
8091fin1_2	Glucose phosphate isomerase [36]	666		X		X
acc2_4	Acetyl-coA carboxylase [36]	501	X	X	X	X
CAD	Pyrimidine biosynthesis [85]	2913	X	X	X	X
DDC	Dopa-decarboxylase [86]	708	X	X	X	X
EF-1alpha	Elongation factor-1 alpha [37]	519	X	X	X	X
enolase	Enolase [87]	1134	X	X	X	X
histone 3	Histone 3 [38]	273	X	X	X	X
period	Period [88]	747		X		X
wingless	Wingless [89]	402		X		X

A box with an “X” indicates a gene that was included in that particular data set.

for analyses that applied a nucleotide substitution model. We also ran codon model analyses [53] as implemented in GARLI. Due to computational limitations at the time of this study, each codon analysis was conducted with 100 ML tree searches and 100 bootstrap replicates. To expedite tree searches, we used Grid computing [54] through The Lattice Project [55]. For consistency in the characterization of results, we will refer to bootstrap support (BP) of 70-79% as “moderate,” support ≥ 80% (but < 90%) as “strong,” and ≥ 90% as “very strong.” We use the arbitrary cutoff of 80% BP as a measure to compare the number of nodes with strong support across individual genes.

Base compositional heterogeneity

Base compositional bias can cause unrelated lineages to incorrectly group together (e.g. [56-60]). While models for phylogenetic analysis assume compositional homogeneity, strong compositional bias is common at sites capable of undergoing synonymous substitution [21,36,61]. In order to examine the effect of compositional heterogeneity, we examined five different character partitions, with and without synonymous change: (a) “nt123": all nucleotides and all changes; (b) “codon": all nucleotides and changes, but implementing a codon model, which “down-weights” synonymous changes because of their relatively rapid evolution; (c) “degen1” [62,63]: all sequence sites with the potential to undergo synonymous changes fully degenerated, an extension of the RY coding scheme of Phillips et al. [64]; (d) “partitioned": all nucleotides partitioned into mostly synonymously evolving and mostly non-synonymously evolving sites, specifically, the partition, “noLRall1 + nt2” versus “LRall1 + nt3” of Regier et al. [62]; and (e) amino acids. As an alternative means to filter synonymous substitutions, in some cases we also analyzed the noLRall1 + nt2 data set alone (see Discussion).

To further investigate the potential influence of compositional heterogeneity, we conducted chi-square tests of among-taxon heterogeneity on data set B (27 taxa × 21 genes). We chose data set B because it includes the largest number of characters (14,793 bp) with a relatively low percentage of missing data (14.3%) out of the four data sets. Chi-square tests were conducted on a character set undergoing mostly synonymous change, nt3, and one undergoing mostly non-synonymous change, noLRall1 + nt2. We conducted the test for various groups in Gracillariidae and outgroups on both the entire character set, and after eliminating invariable sites in the degen1 data set. To gauge the possible effect of compositional heterogeneity on phylogeny inference, we compared neighbor-joining trees using two different distances: ML distances based on the GTR model, which can be influenced by compositional heterogeneity; and Euclidean distances calculated on the proportions of the four nucleotide states treated as independent characters, which will reflect only compositional heterogeneity. Compositional distances were generated using a Perl script that was written with modification of the MBE Toolbox [65] and calculated with PAUP* [45].

Testing alternative hypotheses

Morphology and larval mining patterns predict the monophyly of Gracillariidae + Bucculatricidae + Roeslerstammiidae [26], Gracillariinae + Lithocolletinae [24], and Oecophyllembiinae + Phyllocnistinae [11,35,66], but some of these proposed higher-level groups were not recovered. To test whether these differences between morphological (and behavioral) versus molecular inferences were "real," i.e. not attributable to sampling error in the molecular data, we used the Approximately Unbiased (AU) test of Shimodaira [67]. The AU test ranks trees and determines if trees under a topological constraint describe the data significantly worse than the best tree.

To compare the confidence between our results and prior morphology-based hypotheses, we conducted separate analyses in which groups believed to be monophyletic were constrained. ML trees were calculated with constraints enforced, and the ML tree from the constrained and the unconstrained analyses compared with the AU test. Each analysis applied the same number of ML runs determined to be appropriate for that character partition as described above. Site likelihoods were estimated with PAUP* [45]. For each data set, we combined the site likelihoods generated from all ML constraint analyses together into a single file with the unconstrained site likelihoods. In CONSEL 0.1j [68], the AU test statistic of Shimodaira [67] was used to determine the difference in fit to data of the constrained and unconstrained trees.

Results

Gene versus taxon sampling

Sampling design had the greatest influence on the recovery and bootstrap support of deep nodes in Gracillariidae and relatives, which was especially pronounced for the G. B.R.Y. clade (Gracillariidae + Bucculatricidae + Roeslerstammiidae + Yponomeutidae). Bootstrap support for this clade rose for all five analytical methods when the number of sampled characters was nearly doubled (data set A [27 taxa × 10 genes] versus B [27 taxa × 21 genes]; Table 2). Degen1 provided the strongest support for the G.B.R.Y. clade, rising from 74% (data set A) to 90% (data set B). An increase was also seen when we analyzed the complementary 11 gene, 27 taxa data set (data set B minus A), which had 84% BP for the G.B.R.Y. clade (data not shown). Conversely, doubling the number of taxa in data set A, yielding data set C, lowered support from 74% to < 50% BP for degen1 (data set A versus C; Table 2). Augmenting data set C with sequence data for 11 genes for just over half the number of taxa greatly improved branch support (all five analyses resulted in > 50% BP; data set C versus D, Table 2). There was little difference in bootstrap support for the G.B.R.Y. clade between the two data sets with the greatest amount of gene sampling (data set B versus D). Bootstrap support for other deep, non-G.B.R.Y. clades changed very little.

Gracillariinae was polyphyletic in all analyses conducted (data sets A through D), and the position of Phyllocnistinae remains unclear, as it was ancestral to the Lithocolletinae + *Leucanthiza* in data set D, but was sister to the *Parectopa* group in data sets A through C for degen1 and amino acids. For nt123 and codon analyses, the Phyllocnistinae was sister to the *Parectopa* group (nt123, data sets A, C), sister to *Dendrorcyter* + *Marmara* (nt123, data set D; codon, data sets C, D), or *Eumetriochroa* (nt123, data set C, codon, data sets A, B). Partitioned analyses ("noLRall1 + nt2" versus "LRall1 + nt3") gave the same placement for the Phyllocnistinae as did nt123 in all four data sets.

Agreement and conflict among individual genes

There were no strongly supported groups that conflicted with each other across genes, and few nodes above the subfamily level were moderately or strongly supported by any one gene alone. Instances of strong bootstrap support by only one gene were: 83% for Gracillariidae (CAD), 96% for the *Acrocercops* group (CAD), and 82% for *Eumetriochroa* + Phyllocnistinae (*Period*; Additional file 2).

Base compositional heterogeneity

Results of the chi-square tests for compositional heterogeneity are shown in Table 3. Homogeneity was not rejected for any group in the noLRall1 + nt2 character set. In contrast, nt3 showed highly significant heterogeneity across all taxa and the five taxon subsets. As a gauge of the possible misleading signal produced by compositional heterogeneity, we calculated neighbor-joining trees on distances reflecting only composition for nt123 and nt3. In these trees, Bucculatricidae clustered with five other gracillarioid and non-gracillarioid taxa that are together separated by long internal branches from the Tineidae and the remaining species in the tree (Additional file 3).

Relationships of Gracillariidae and Gracillarioidea

All analyses resulted in a polyphyletic Gracillarioidea *sensu* Davis and Robinson [11]; specifically, Douglasiidae was consistently separated from Bucculatricidae, Gracillariidae, and Roeslerstammiidae by two or more nodes.

Table 2 Bootstrap support values across data sets for selected clades.

Data set	Analysis	'G.B. R.Y.' clade	Gracillariidae + Bucculatricidae + Yponomeutidae ('G.B.Y.' clade)	Gracillariidae + Roeslerstammiidae + Yponomeutidae ('G.R.Y.' clade)	Gracillariidae + Yponomeutidae ('G.Y.' clade)	Gracillariidae	Lithocolletinae + *Leucanthiza*	*Acrocercops* group	*Gracillaria* group	*Parectopa* group	Phyllocnistinae + Oecophyllembiinae + *Dendrorycter* + *Marmara*
A	nt123	[< 50]	[< 50]	< 50	< 50	99	N/A	87	N/A	100	N/A
	codon	[< 50]	[< 50]	[< 50]	[< 50]	99	N/A	86	N/A	100	N/A
	degen	74	55	[< 50]	[< 50]	100	N/A	99	N/A	100	N/A
	partitioned	[< 50]	[< 50]	[< 50]	[< 50]	99	N/A	89	N/A	100	N/A
	aa	[< 50]	[< 50]	[< 50]	[< 50]	97	N/A	90	N/A	100	N/A
B	nt123	[53]	[< 50]	< 50	[< 50]	100	N/A	98	N/A	100	N/A
	codon	[54]	[< 50]	< 50	[< 50]	100	N/A	93	N/A	100	N/A
	degen	90	[< 50]	< 50	[< 50]	100	N/A	92	N/A	100	N/A
	partitioned	[62]	[< 50]	[< 50]	[< 50]	100	N/A	94	N/A	100	N/A
	aa	66	[< 50]	[< 50]	< 50	98	N/A	97	N/A	100	N/A
C	nt123	[< 50]	[< 50]	[< 50]	[< 50]	99	100	98	71	100	[< 50]
	codon	[< 50]	[< 50]	[< 50]	[< 50]	99	100	98	58	100	[< 50]
	degen	[< 50]	[< 50]	[< 50]	[< 50]	100	100	100	89	100	[< 50]
	partitioned	[< 50]	[< 50]	[< 50]	[< 50]	99	100	97	77	100	[< 50]
	aa	[< 50]	< 50	[< 50]	[< 50]	90	100	93	< 50	100	[< 50]
D	nt123	[55]	[< 50]	< 50	[< 50]	99	100	100	67	100	51
	codon	[61]	[< 50]	< 50	< 50	97	100	100	100	100	< 50
	degen	83	[< 50]	< 50	< 50	100	100	100	93	100	[< 50]
	partitioned	[59]	[< 50]	< 50	[< 50]	99	100	97	67	100	51
	aa	75	[< 50]	[< 50]	< 50	89	100	94	< 50	100	< 50

Square brackets indicate support values for clades that were not present in the ML tree.

Table 3 Results of Chi-square tests on nucleotide compositional homogeneity

	P value for character set	
Taxon (number of species)	**noLRall1 + nt2**	**nt3**
All (27)	> 0.999	< 0.001
Gracillariidae (11)	> 0.999	< 0.001
Oecophyllembiinae *sensu* Kumata + Phyllocnistinae (3)	0.969	< 0.001
Bucculatricidae + Tineidae (3)	0.953	< 0.001
Bucculatricidae + Outgroups + *Klimeschia* - Tineidae (10)	>0.999	< 0.001
Outgroups + *Klimeschia* - Tineidae (9)	> 0.999	< 0.001
Total number of characters	8701	4937

Monophyly of the superfamily was significantly rejected at $P \leq 0.015$ in six of eight AU tests (Table 4). In all analyses, branch support for the monophyly of Gracillariidae was robust (≥ 97% BP, Table 2). In general, nt123, codon and nt123 partitioned results were similar in topology and branch support, while degen1 and amino acids were similar to each other, but differed from the other results in topology. For data set D (57 taxa × 21 genes), degen1 resulted in a monophyletic 'G.B.R.Y.' clade with strong bootstrap support (83%, Figure 2). The G.B.R.Y. clade was never recovered in nt123, codon and nt123 partitioned ML trees, but the bootstrap consensus from these analyses supported the G.B.R.Y. clade with weak, but evident signal (up to 62% BP, Table 2). The latter three methods resulted in Bucculatricidae diverging before all taxa except the designated outgroup, Tineidae (e.g. Additional files 4, 5, 6).

While there was some support for the G.B.R.Y. clade, none of the analyses provided > 55% BP for any inter-family relationships within this clade. To assess the possible sister group of Gracillariidae, we examined whether the ML trees consistently recovered Gracillariidae plus any combination of Bucculatricidae, Roeslerstammiidae, or Yponomeutidae. ML analyses recovered a 'G.R.Y.' clade eight times, a 'G.Y.' clade five times, and a 'G.B.Y.' clade two times (Table 2). No 'G.B.', 'G.R.', or 'G.B.R.' clades were found in any of the best ML trees.

Within Gracillariidae, there was strong support for 28 of the 36 total nodes (78%) with data set D, and nearly every species was placed into one of six lineages that are either monotypic or strongly supported in our sampling. These lineages, labeled on Figure 2, include: 1) Lithocolletinae + *Leucanthiza*, 2) Phyllocnistinae (excluding *Eumetriochroa*), and four clades within Gracillariinae: 3) *Acrocercops* group, 4) *Dendrorycter* + *Marmara*; 5) *Gracillaria* group, and 6) *Parectopa* group. Kumata compared the morphology of these groups and recognized that each has unique features absent in others (e.g. [69-72], some noted in Figure 2).

Branch support within these six gracillariid clades was strong. There was > 90% BP for eight of the eleven nodes within Lithocolletinae + *Leucanthiza*, three of five nodes within the *Acrocercops* group, five of six nodes in the *Parectopa* group, and all nodes in the *Gracillaria* group. Monophyly of the subfamily Gracillariinae Stainton 1854, as previously defined, was not recovered, and statistically rejected in all eight AU tests ($P < 0.001$, Table 4). Rejection of gracillariine monophyly was also evident even when *Leucanthiza* was excluded from Gracillariinae in the AU tests (see Gracillariinae minus *Leucanthiza*, Table 4). Monophyly of Gracillariinae + Lithocolletinae, as previously proposed by Kuznetzov and Stekol'nikov [24] was rejected by nt123 and codon model analyses ($P < 0.05$). Monophyly cannot be rejected for the sister-group relationship of the Oecophyllembiinae (*sensu* Kumata) + Phyllocnistinae, however, as $P > 0.073$ under the AU test in all cases (Table 4). While we do not formally designate any new taxonomic names in this study,

Table 4 Results of Approximately Unbiased (AU) significance tests [67] for non-monophyly of predicted clades for data sets C and D

	P values: data sets C/D			
Predicted clade	**nt123**	**codon**	**degen**	**AA**
Gracillarioidea *sensu* Davis & Robinson	**0.015/0.002**	**0.007/0.006**	0.083/<**0.001**	0.081/**0.008**
Gracillariinae + Lithocolletinae	**0.011/0.002**	**0.013/0.011**	0.455/0.161	0.152/0.119
Gracillariinae	**< 0.001/< 0.001**	**< 0.001/< 0.001**	**< 0.001/< 0.001**	**< 0.001/< 0.001**
Gracillariinae minus *Leucanthiza*	**0.015/0.001**	**0.021/< 0.001**	0.107/0.084	0.104/0.202
Oecophyllembiinae *sensu* Kumata + Phyllocnistinae	0.467/0.165	0.385/0.739	0.339/0.352	0.472/0.073

nt123, all nucleotides; Codon, codon model; degen, degeneracy1; AA, amino acids. Groups that were significant at alpha = 0.05 are shown in bold.

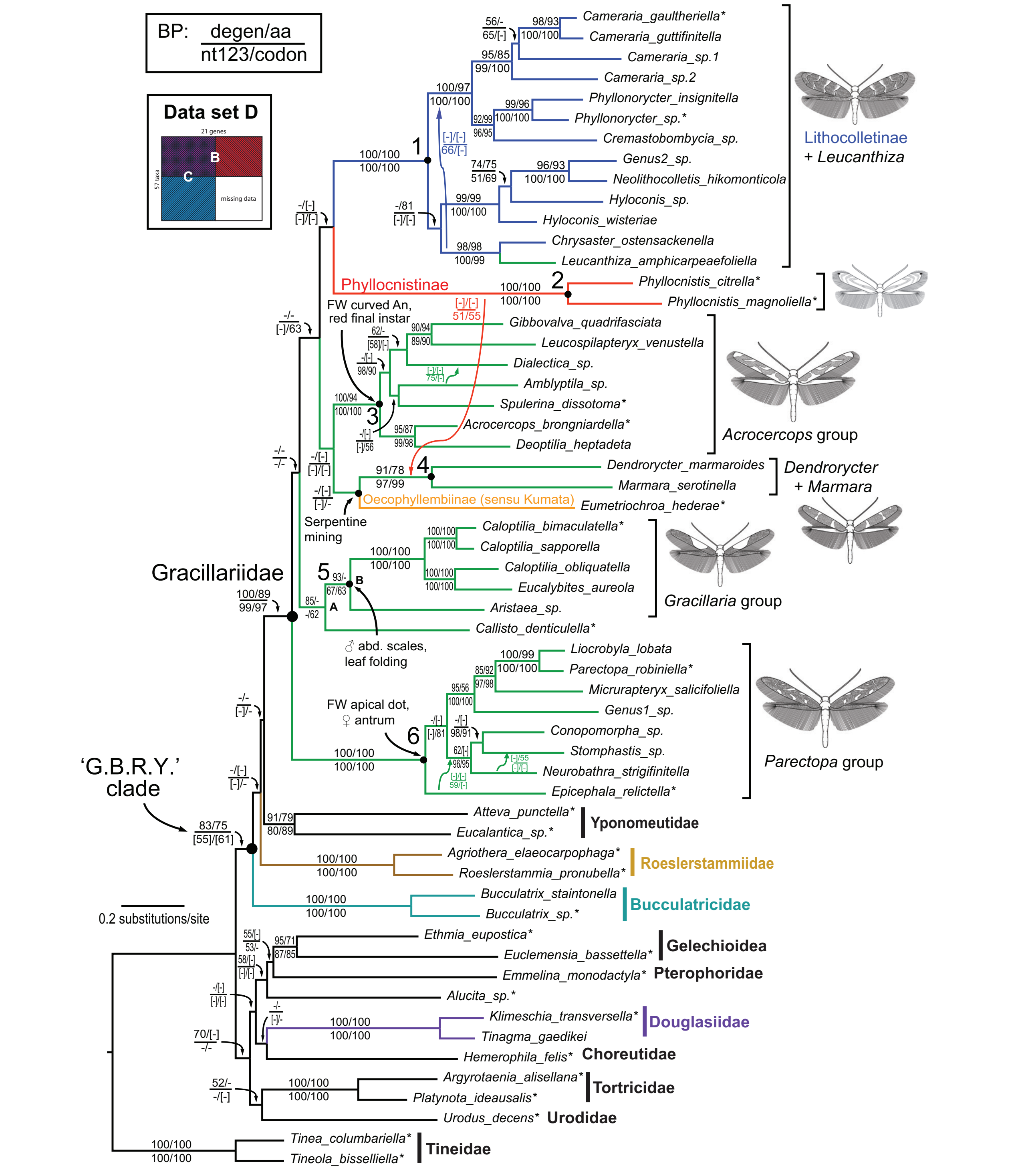

Figure 2 Maximum likelihood degen1 tree for data set D. Large numbers denote six major clades in Gracillariidae (see Results). Asterisks indicate taxa sequenced for 21 genes. Hyphens denote support values < 50%. Square brackets, shown only for nodes with support > 50% that conflict with the nt123 ML tree, denote groupings not present in the ML tree for that analysis. Green branches lead to taxa placed in Gracillariinae. Morphological and behavioral traits that are characteristic of each group are also noted.

we recognize these well-supported clades as the first step toward a phylogenetic reclassification of Gracillariidae.

Discussion

Augmentation of sequence data and its effect on branch support

Partial augmentation of gene sampling can improve estimates of deep relationships of Gracillariidae (comparison of data set C and D), as it increased support for the G.B.R.Y. clade for all five character treatments, most strikingly for degen1 (a BP increase from < 50% to 83%). While partial or full augmentation of gene sampling generally improved branch support for deep nodes, other nodes at the superfamily level are not robustly supported even with > 14 kb of sequence data (Table 2). Short internal branches for deep divergences and the difficulty of achieving strong support for some nodes even with 21 genes may reflect a rapid radiation, which likely characterizes many divergences among lepidopteran families [21,22,73] and other insect orders [74].

The amount of missing data in data set D, accounting for roughly a quarter of the total matrix, does not apparently induce phylogenetic artifacts of the kind postulated by Lemmon et al. [75]. For degen1, the problematic Bucculatricidae is left out of the G.B.R.Y. clade in the ML tree from non-augmented data set C (Figure 3C), but partially augmented data set D pulls it into the G.B.R.Y. clade. We favor the topology from the partially augmented data set D over the non-augmented data set C, as the close relationship of Bucculatricidae to Gracillariidae corroborates morphological [24,26] and molecular studies [22].

A comparison of data sets A and B reveals that increasing the number of genes from 10 to 21 can improve bootstrap support values, a result consistent with other studies (e.g. [1-5]). In contrast, increasing the number of taxa (comparison of data set A and C) depressed branch support for higher groups (e.g. G.B.R.Y. clade). Therefore, if the goal is to achieve strong branch support for deep divergences, it may be favorable to focus on sequencing many genes for fewer taxa. However, sequencing many characters for few taxa can lead to problems such as long-branch attraction [76,77] and produce relationships that are misleading (e.g. [78-81]). Thus, if the researcher is faced with limited resources and seeks to improve support for deep divergences without introducing misleading artifacts, a solution may be to partially augment more characters for a selected but broad taxonomic sample. In our study, a deep divergence that was initially weak in support (the G.B.R.Y. clade) now has stronger branch support and the relationships are largely consistent with prior morphological hypotheses. While we have chosen a broad taxonomic sample to augment additional characters,

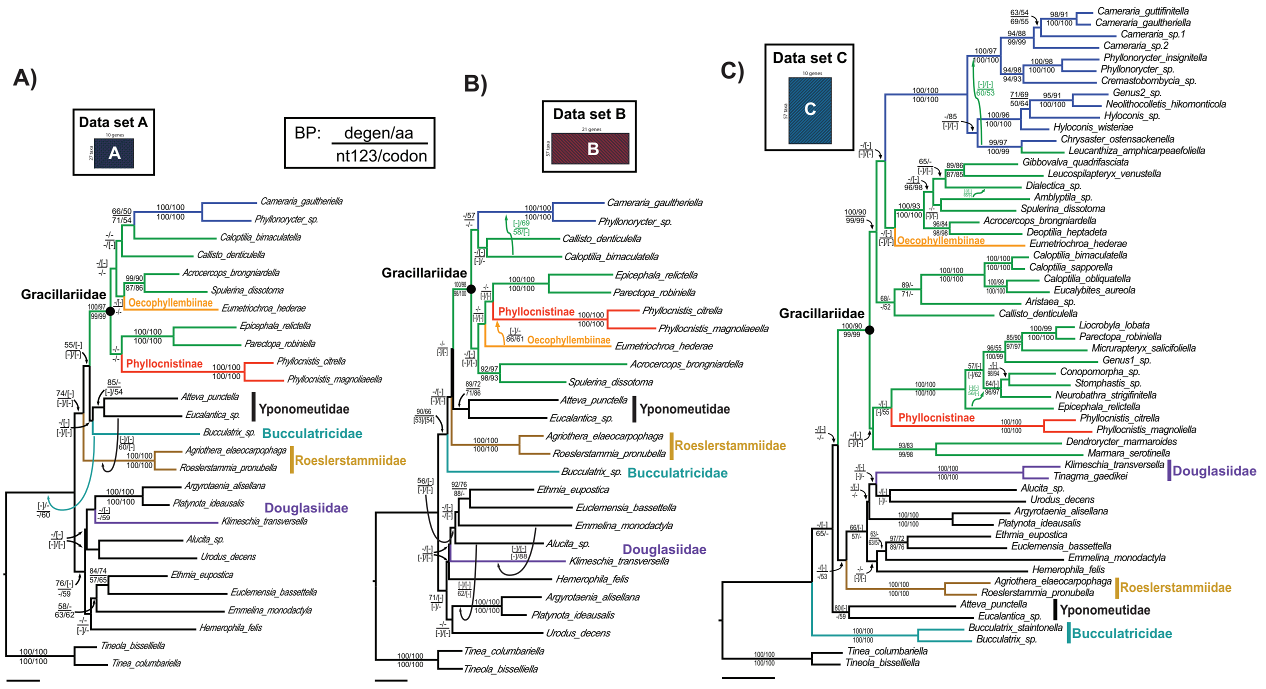

Figure 3 ML trees based on non-synonymous differences only (degen1) of data sets A through C. Bucculatricidae + Gracillariidae + Roeslerstammiidae + Yponomeutidae form a monophyletic group for data sets A and B. Scale bar = 0.02 substitutions/site.

future studies should examine how the choice of taxa for augmentation can influence the support for the correct tree.

Synonymous changes and base compositional heterogeneity

Overall, two non-synonymous characters sets, degen1 and amino acids, provided the highest support for deep-level relationships in Gracillarioidea. These analyses resulted in a monophyletic G.B.R.Y. clade, for which support in some cases was very robust. In contrast, nt123, the codon model and nt123 partitioned analyses provided little or no support for deep relationships among gracillarioid families. When synonymous sites are added, only weak signal (≤ 62% BP) remains for the G.B.R.Y. clade. We speculate that analyses that include synonymous changes, even when that signal is down-weighted or modeled separately, do not effectively correct for the strong compositional heterogeneity found at nt3, leaving that non-phylogenetic signal to conflict with and obscure the true signal from non-synonymous change.

Strong compositional heterogeneity can incorrectly group unrelated taxa together [56,57], or equivalently, widely separate a taxon with strongly divergent base composition from its true relatives. Under the three character treatments, nt123, codon model, and nt123 partitioned, *Bucculatrix* (Bucculatricidae) was placed between the Tineidae and the remaining taxa (Additional files 4, 5, 6), a position difficult to reconcile with morphology. Compositional heterogeneity may account for the strikingly different placement of Bucculatricidae. Because non-synonymous changes strongly support the monophyly of the G.B.R.Y. clade, synonymous changes, mostly at nt3, must account for the less decisive placement of Bucculatricidae in nearly all nt123 trees. These results support our previous findings (e.g. [21]) that filtering synonymous substitutions (and thereby compositional heterogeneity) can result in more robust phylogenetic inference at deep levels.

A comparison of the ML topology with the neighbor-joining GTR ML distance and with Euclidean compositional distance trees for nt123 and nt3 suggests that the uncertain placement of Bucculatricidae in the nt123 data set is largely due to nt3 (Additional file 3). In the compositional distance trees, six taxa (*Bucculatrix sp., Atteva punctella*, *Eumetriochroa hederae*, *Hemerophila felis*, *Phyllocnistis citrella*, and *P. magnoliella*) fall between the Tineidae and the remaining taxa in a long internal branch. In the nt123 ML tree, in contrast, all taxa but *Bucculatrix* move to parts of the tree that are generally well supported and expected based on morphology and existing classifications (e.g. [11,35,82], *Eumetriochroa* with *Phyllocnistis*, and *Atteva* with *Eucalantica*).

Results of the ML nt3 analysis are very different, providing further evidence that compositional heterogeneity can affect trees based on nt3 alone. Despite providing about 90% of the total character change, the nt3 character set alone yields > 50% BP for only 7 nodes as compared to the full data set (nt123; 15 supported nodes), fewer even than the degen1 character set (13 supported nodes). Some unexpected relationships are found, such as *Bucculatrix* + *Eumetriochroa*, which break up well-supported groups, in this case the monophyletic Gracillariidae (Additional file 3, F).

Phylogenetic relationships of Gracillariidae and Gracillarioidea

Our results provide some of the first molecular evidence on the composition of and relationships within Gracillariidae and Gracillarioidea *sensu* Davis and Robinson [11]. Some previous hypotheses about those relationships were confirmed, and several novel ones proposed. Because rate variation between synonymous and non-synonymous sites was dramatic in the present study (see Table 3), we focus our discussion on the degen1 analyses unless otherwise noted. The discussion focuses on the degen1 tree from data set D (Figure 2) because it includes the most number of taxa and characters.

Gracillarioidea *sensu* Davis and Robinson [11], i.e. including Douglasiidae, was polyphyletic in all analyses, and monophyly of the superfamily was rejected significantly in six of eight AU tests (Table 4). Recently, Mutanen et al. [22] came to a similar conclusion based on fewer gracillarioid taxa and genes. In their analyses, Gracillarioidea were never monophyletic, and Douglasiidae was consistently placed in Apoditrysia. Mutanen et al. [22] also had difficulty in placing Bucculatricidae, which was paraphyletic with respect to *Tritymba* (Plutellidae), and Bucculatricidae + *Tritymba* was sister to Gracillariidae with weak (< 50%) ML bootstrap support. Based on our study and Mutanen et al. [22], Gracillarioidea as currently defined, will need to be reevaluated. Future studies should also sequence the genes included in the present study for *Ogmograptis* (Bucculatricidae), a genus that could not be obtained.

To resolve a possible sister group of Gracillariidae remains a challenging problem. Comparing ML trees from all analyses, Gracillariidae was most often closely related to Roeslerstammiidae + Yponomeutidae (the latter relationship which is congruent with morphology [83,84]). The close relationship of Yponomeutidae to Gracillarioidea (excluding Douglasiidae) is also consistent with other molecular studies [7,21,22]. These reports suggest, at least tentatively, that the putative morphological apomorphies proposed for Gracillarioidea by Davis and Robinson [11] may be homoplasies. In order to restore monophyly of the superfamily, we would need to exclude Douglasiidae from Gracillarioidea or include Yponomeutidae. However, more convincing resolution of inter-family relationships is desirable before any formal taxonomic changes are made.

Monophyly of Gracillariidae was strongly supported in nearly all analyses, a relationship that is corroborated by at least two morphological synapomorphies [26]. The grouping of Oecophyllembiinae (*sensu* Kumata) + Phyllocnistinae, which share unique serpentine mine morphology [11,66] and a highly specialized spinning instar [17], was supported weakly or not at all in our analyses. However, this sister-group relationship could not be rejected by any AU test, and this grouping was well supported by *Period* (82% BP, Additional file 2). The sister-group relationship of Gracillariinae to Lithocolletinae proposed by Kuznetzov and Stekol'nikov [24] was rejected by four of eight AU tests (Table 4). Our results strongly support the inclusion of *Leucanthiza* in Lithocolletinae, suggesting that this genus should be transferred from Gracillariinae, a conclusion that is also supported by larval morphology (Wagner and Davis unpubl. data). Monophyly of Gracillariinae (both with and without *Leucanthiza*) was rejected by the AU test in more than half of the data sets, suggesting that this subfamily needs to be revised. However, we did identify four genus-level groups with strong support within Gracillariinae: *Acrocercops*, *Gracillaria*, *Parectopa* groups and *Dendrorycter* + *Marmara*, all which were previously postulated based on morphology and life-history comparisons [18,69-72].

Conclusions

Several main conclusions can be drawn from this study. First, branch support was maximized when gene sampling was increased, especially when we analyzed only the non-synonymous changes. Second, augmenting a taxon-rich data set (data set C; 57 taxa × 10 genes) with additional sequence data for approximately half the taxa substantially increased deep node support in a resource-efficient manner, apparently without inducing phylogenetic artifacts due to large blocks of missing data. While these two conclusions were drawn specifically from the data sets in this study, they are congruent with the results of Cho et al. [7]. Third, Gracillariidae were monophyletic in all our analyses, and nearly all species can be placed into one of six strongly supported clades, though relationships among them remain largely unclear. Fourth, Gracillarioidea, as defined by Davis and Robinson [11], clearly do not include Douglasiidae, and changes to the classification will be required. Fifth, the difficulty in placing Bucculatricidae is probably attributable to base compositional heterogeneity at nt3. From our tests for compositional heterogeneity and strong bootstrap values obtained when synonymous changes are excluded, we tentatively conclude that Bucculatricidae is closely related to Gracillariidae + Roeslerstammiidae + Yponomeutidae. Finally, the different levels of branch support seen under our different character treatments reinforce the importance of assessing confidence in groups under multiple phylogenetic approaches. Factors such as compositional heterogeneity, which can influence phylogenetic accuracy, are most easily assessed when data are partitioned into largely non-synonymous and mostly synonymous character sets. Branch support overall is strongest when all changes are included, but for several deep divergences, strong support is obtained only when synonymous changes were excluded. Because branch support for many deep splits was weak, we are exploring whether greater branch support for gracillariids and relatives can be achieved by means of genomic (next-generation) sequencing – the focus of a future project.

Additional material

Additional file 1: Exemplar species included, their classification, and GenBank accession numbers. For Gracillariidae the number of taxa in each subfamily and genus is listed in parentheses (number of taxa sampled/number of taxa known). "x" denotes a sequence that could not be amplified.

Additional file 2: Single gene bootstrap values for all nodes in the nt123 tree of data set B. Shaded boxes are those with > 80% bootstrap support. "ALL" refers to dataset B (all genes included). See Additional file 1 for taxon code names.

Additional file 3: Comparison of Euclidean compositional distance (NJ), GTR ML distance (NJ), and ML trees for nt123 and nt3. Arrows indicate a long internal branch in the Euclidean compositional distance trees.

Additional file 4: Maximum likelihood nt123 trees for data sets A through D. Scale bar = 0.07 substitutions/site.

Additional file 5: Maximum likelihood trees based on a codon model. Scale bar = 0.03 substitutions/site.

Additional file 6: Maximum likelihood trees based on a partitioned model. Scale bar = 0.2 substitutions/site.

Additional file 7: Maximum likelihood trees based on inferred amino acids. Scale bar = 0.03 substitutions/site.

Acknowledgements
We thank Andreas Zwick (State Museum of Natural History, Stuttgart) for continuous support, critical comments and advice, and Tosio Kumata (Hokkaido University) for many insights into the systematics of Gracillariidae. Zaile Du and Anamika Verma assisted with sequencing. Adam Bazinet (University of Maryland) was instrumental in helping access and utilize The Lattice Grid. Joaquin Baixeras (University of Valencia), Soowon Cho (Chungbuk National University), Terry Harrison (University of Illinois), Kenji Nishida (University of Costa Rica), Jae-Cheon Sohn (University of Maryland), and Kazuhiro Sugisima (Okayama City, Japan) provided some important taxa used in this study. We also thank Rodolphe Rougerie (Canadian Centre for DNA Barcoding) for sending valuable extracts, and Kim Mitter (University of Maryland) for assisting with the preparation of samples. Jean-François Landry (Canadian National Collection) helped clarify taxonomic inconsistencies related to some taxa in the study. Julie Byrd (University of Maryland) and two anonymous reviewers provided comments that improved the quality of the manuscript. The U.S. National Science Foundation awards 0531626, 0531769, and 0755048 provided funding for this project. Financial support also came from the Belgian Science Policy Research Programmes for the

Federal Scientific Institutions, project number MO/37/022; The Research Fellowships of the Japan Society for the Promotion of Science (JSPS) for Young Scientists (award 20-5555); and a Grant-in Aid from JSPS (20870021).

Author details

[1]Department of Entomology, University of Maryland, College Park, MD, USA. [2]Division of Evolutionary Biology, National Institute for Basic Biology, Okazaki, Japan. [3]Center for Ecological Research, Kyoto University, Kyoto, Japan. [4]Institute for Bioscience and Biotechnology Research, University of Maryland, College Park, MD, USA. [5]Laboratory of Molecular Evolution, Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD, USA. [6]Department of Entomology, Smithsonian Institution, Washington, D.C., USA. [7]Department of Ecology & Evolutionary Biology, University of Connecticut, Storrs, CT, USA. [8]Royal Museum for Central Africa, Tervuren, Belgium. [9]INRA, UR0633 Zoologie Forestière, F-45000, Orléans, France.

Authors' contributions

AYK carried out the RT-PCR experiments, sequence alignment, phylogenetic analyses, and drafted the manuscript. IO and AK conducted PCR work on two genes, H3 and EF-1α. JCR and CM provided funds for conducting the molecular work and helped design the study. MPC provided phylogenetic programs and hardware through the Lattice Project. AK, DRD, DLW, IO, JDP, and CLV collected valuable specimens necessary for the project. All authors read and approved the final manuscript.

Received: 8 November 2010 Accepted: 24 June 2011
Published: 24 June 2011

References

1. Rokas A, Williams BL, King N, Carroll SB: **Genome-scale approaches to resolving incongruence in molecular phylogenies.** *Nature* 2003, **425**:798-804.
2. Bapteste E, Brinkmann H, Lee JA, Moore DV, Sensen CW, Gordon P, Duruflé L, Gaasterland T, Lopez P, Müller M, *et al*: **The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*.** *Proceedings of the National Academy of Sciences, USA* 2002, **99**:1414-1419.
3. Wiens JJ, Kuczynski CA, Smith SA, Mulcahy DG, Sites JW Jr, Townsend TM, Reeder TW: **Branch lengths, support, and congruence: testing the phylogenomic approach with 20 nuclear loci in snakes.** *Systematic Biology* 2008, **57(3)**:420-431.
4. Zwick A, Regier JC, Mitter C, Cummings MP: **Increased gene sampling yields robust support for higher-level clades within Bombycoidea (Lepidoptera).** *Systematic Entomology* 2011, **36**:31-43.
5. Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, *et al*: **Broad phylogenomic sampling improves resolution of the animal tree of life.** *Nature* 2008, **452**:745-749.
6. Burleigh JG, Hilu KW, Soltis DE: **Inferring phylogenies with incomplete data sets: a 5-gene, 567-taxon analysis of angiosperms.** *BMC Evolutionary Biology* 2009, **9**:61.
7. Cho S, Zwick A, Regier JC, Mitter C, Cummings MP, Yao J, Du Z, Zhao H, Kawahara AY, Weller SJ, *et al*: **Deliberately unequal gene sampling: boon or bane for phylogenetics of Lepidoptera (Hexapoda)?** *Systematic Biology* 2011.
8. Labandeira CC, Dilcher DL, Davis DR, Wagner DL: **Ninety-seven million years of angiosperm-insect association: paleobiological insights into the meaning of coevolution.** *Proceedings of the National Academy of Sciences, USA* 1994, **91**:12278-12282.
9. De Prins J, De Prins W: **Global Taxonomic Database of Gracillariidae (Lepidoptera).** World Wide Web electronic publication;[http://www.gracillariidae.net], accessed 1 Feb 2011.
10. Nieukerken EJ, Kaila L, Kitching IJ, Kristensen NP, Lees DC, Minet J, Mitter C, Mutanen M, Regier JC, Simonsen TJ, *et al*: **Order Lepidoptera Linnaeus, 1758.** In *Zootaxa.* Edited by: Zhang ZQ. Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness; 2011:.
11. Davis DR, Robinson GS: **The Tineoidea and Gracillarioidea.** In *Lepidoptera, moths and butterflies 1: Evolution, Systematics and Biogeography..* 35 edition. Edited by: Kristensen NP. Berlin, New York: De Gruyter; 1998:91-117, 4. Arthropoda: Insecta, Part.
12. Gilbert M, Guichard S, Freise J, Grégoire J-C, Heitland W, Straw N, Tilbury C, Augustin S: **Forecasting *Cameraria ohridella* invasion dynamics in recently invaded countries: from validation to prediction.** *Journal of Applied Ecology* 2005, **45**:805-813.
13. Heppner JB: **Citrus leafminer, *Phyllocnistis citrella* Stainton (Lepidoptera: Gracillariidae: Phyllocnistinae).** *Entomology Circular No 359, Florida Department of Agriculture & Consumer Services* 1993, 1-2.
14. Shapiro LH, Scheffer SJ, Maisin N, Lambert S, Purung HB, Sulistyowati E, Vega FE, Gende P, Laup S, Rosmana A, *et al*: ***Conopomorpha cramerella* (Lepidoptera: Gracillariidae) in the Malay Archipelago: genetic signature of a bottlenecked population?** *Annals of the Entomological Society of America* 2008, **101**:930-938.
15. Valade R, Kenis M, Hernandez-Lopez A, Augustin S, Mari Mena N, Magnoux E, Rougerie R, Lakatos F, Roques A, Lopez-Vaamonde C: **Mitochondrial and microsatellite DNA markers reveal a Balkan origin for the highly invasive horse-chestnut leaf miner *Cameraria ohridella* (Lepidoptera, Gracillariidae).** *Molecular Ecology* 2009, **18**:3458-3470.
16. Davis DR, Wagner DL: **Biology and systematics of the New World *Phyllocnistis* leafminers of the avocado genus *Persea* (Lepidoptera: Gracillariidae).** *ZooKeys* 2011, **97**:39-73.
17. Davis DR: **Gracillariidae.** In *Immature insects. Volume 1.* Edited by: Stehr FW. Dubuque: Kendall/Hunt; 1987:372-374.
18. Wagner DL, Loose JL, Fitzgerald TD, DeBenedictis JA, Davis DR: **A hidden past: The hypermetamorphic development of *Marmara arbutiella* (Lepidoptera: Gracillariidae).** *Annals of the Entomological Society of America* 2000, **93**:59-64.
19. Kumata T: **Japanese species of the subfamily Oecophyllembiinae Réal and Balachowsky (Lepidoptera: Gracillariidae), with descriptions of a new genus and eight new species.** *Insecta Matsumurana, New Series* 1998, **54**:77-131.
20. Fitzgerald TD, Simeone JB: **Serpentine miner *Marmara fraxinicola* (Lepidoptera: Gracillariidae) in stems of white ash.** *Annals of the Entomological Society of America* 1971, **64**:770-773.
21. Regier JC, Zwick A, Cummings MP, Kawahara AY, Cho S, Weller S, Roe A, Baixeras J, Brown JW, Parr C, *et al*: **Toward reconstructing the evolution of advanced moths and butterflies (Lepidoptera: Ditrysia): an initial molecular study.** *BMC Evolutionary Biology* 2009, **9**:280.
22. Mutanen M, Wahlberg N, Kaila L: **Comprehensive gene and taxon coverage elucidates radiation patterns in moths and butterflies.** *Proceedings of the Royal Society of London, Series B* 2010, **277**:2839-2848.
23. Gerasimov AM: **Lepidoptera-the butterflies.** *Opredelitel Nasekomykh Evropeyskoy Chasti SSSR* 1948, 920-1094.
24. Kuznetzov VI, Stekol'nikov AA: **Functional morphology of the male genitalia and notes on the classification and phylogenetic relationships of mining moths of superfamily Gracillarioidea (Lepidoptera).** *Entomological Review* 1987, **66**:16-30.
25. Zimmerman EC: **Insects of Hawaii 9, Microlepidoptera, pt. 1.** Honolulu: University of Hawaii Press; 1978.
26. Robinson GS: **A phylogeny for the Tineoidea (Lepidoptera).** *Entomologica Scandinavica* 1988, **19**:117-129.
27. Kuznetzov VI, Kozlov MV, Seksyaeva SV: **To the systematics and phylogeny of mining moths Gracillariidae, Bucculatricidae and Lyonetiidae (Lepidoptera) with consideration of functional and comparative morphology of male genitalia.** *Trudy Zoologicheskogo Instituta, Akademija Nauk SSSR* 1988, **176**:52-71.
28. Ohshima I: **Host race formation in the leaf-mining moth *Acrocercops transecta* (Lepidoptera: Gracillariidae).** *Biological Journal of the Linnaean Society* 2008, **93**:135-145.
29. Ohshima I: **Differential introgression causes genealogical discordance in host races of *Acrocercops transecta* (Insecta: Lepidoptera).** *Molecular Ecology* 2010, **19**:2106-2119.
30. Kawakita A, Takimura A, Terachi T, Sota T, Kato M: **Cospeciation analysis of an obligate pollination mutualism: have *Glochidion* trees (Euphorbiaceae) and pollinating *Epicephala* moths (Gracillariidae) diversified in parallel?** *Evolution* 2004, **10**:2201-2214.
31. Kawakita A, Kato M: **Repeated independent evolution of obligate pollination mutualism in the Phyllantheae-*Epicephala* association.** *Proceedings of the Royal Society of London, Series B* 2009, **276**:417-426.
32. Kawakita A, Okamoto T, Goto R, Kato M: **Mutualism favours higher host specificity than does antagonism in plant-herbivore interaction.**

Proceedings of the Royal Society of London, Series B 2010, **277(1695)**:2765-2774.

33. Lopez-Vaamonde C, Godfray C, Cook JM: **Evolutionary dynamics of host-plant use in a genus of leaf mining moth.** *Evolution* 2003, **57**:1804-1821.
34. Lopez-Vaamonde C, Wikström N, Labandeira C, Godfrey HCJ, Goodman SJ, Cook JM: **Fossil-calibrated molecular phylogenies reveal that leaf-mining moths radiated millions of years after their host plants.** *Journal of Evolutionary Biology* 2006, **19**:1314-1326.
35. Vári L, Kroon DM, Krüger M: **Classification and checklist of the species of Lepidoptera recorded in southern Africa.** Chatswood: Simple Solutions; 2002.
36. Regier JC, Shultz JW, Ganley ARD, Hussey A, Shi D, Ball B, Zwick A, Stajich JE, Cummings MP, Martin JW, *et al*: **Resolving arthropod phylogeny: exploring phylogenetic signal within 41 kb of protein-coding nuclear gene sequence.** *Systematic Biology* 2008, **57**:920-938.
37. Cho S, Mitchell A, Regier JC, Mitter C, Poole RW, Friedlander TP, Zhao S: **A highly conserved nuclear gene for low-level phylogenetics: elongation factor-1α recovers morphology-based tree for heliothine moth.** *Molecular Biology and Evolution* 1995, **12**:650-656.
38. Ogden TH, Whiting MF: **The problem with "the Paleoptera problem:" sense and sensitivity.** *Cladistics* 2003, **19**:432-442.
39. Regier JC: **Protocols, Concepts, and Reagents for preparing DNA sequencing templates.**[http://www.umbi.umd.edu/users/jcrlab/PCR_primers.pdf], Version 12/4/08.
40. Kawakita A, Kato M: **Assessment of the diversity and species specificity of the mutualistic association between *Epicephala* moths and *Glochidion* trees.** *Molecular Ecology* 2006, **15**:3567-3581.
41. Drummond AJ, Ashton B, Cheung M, Heled J, Kearse M, Moir R, Stones-Havas S, Thierer T, Wilson A: **Geneious ver. 5.1.** 2010 [http://www.geneious.com].
42. Castresana J: **GBlocks, ver. 0.91b.** 2002 [http://molevol.cmima.csic.es/castresana/].
43. Talavera G, Castresana J: **Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments.** *Systematic Biology* 2007, **56**:564-577.
44. Castresana J: **Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis.** *Molecular Biology and Evolution* 2000, **17**:540-552.
45. Swofford DL: **PAUP*: Phylogenetic analysis using parsimony (*and other methods). Version 4.0b10.** Sunderland: Sinauer Associates; 2002.
46. Zwickl DJ: **Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion.** Ph.D. dissertation: The University of Texas at Austin; 2006.
47. Zwickl DJ: **GARLI-PART Version 0.97. Genetic Algorithm for Rapid Likelihood Inference.**[https://www.nescent.org/wg_garli/Partitioned_version], accessed 12 Aug. 2010.
48. Posada D: **jModelTest: Phylogenetic Model Averaging.** *Molecular Biology and Evolution* 2008, **25**:1253-1256.
49. Lanave C, Preparata G, Saccone C, Serio G: **A new method for calculating evolutionary substitution rates.** *Journal of Molecular Evolution* 1984, **20**:86-93.
50. Tavaré S: **Some probablistic and statistical problems on the analysis of DNA sequences.** *Lectures on Mathematics in the Life Sciences* 1986, **17**:57-86.
51. Yang Z: **Maximum-likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods.** *Journal of Molecular Evolution* 1994, **39**:306-314.
52. Gu X, Fu YX, Li WH: **Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites.** *Molecular Biology and Evolution* 1995, **12**:546-557.
53. Goldman N, Yang Z: **A codon-based model of nucleotide substitution for protein-coding DNA sequences.** *Molecular Biology and Evolution* 1994, **11**:725-736.
54. Cummings MP, Huskamp JC: **Grid computing.** *EDUCAUSE Review* 2005, **40**:116-117.
55. Bazinet AL, Cummings MP: **The Lattice Project: a Grid research and production environment combining multiple Grid computing models.** In *Distributed & Grid Computing - Science Made Transparent for Everyone Principles, Applications and Supporting Communities Tectum.* Edited by: Weber WHW. Marburg; 2009:2-13.
56. Lockhart PJ, Stell MA, Hendy MD, Penny D: **Recovering evolutionary trees under a more realistic model of sequence evolution.** *Molecular Biology and Evolution* 1994, **11**:605-612.
57. Foster PG, Hickey DA: **Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions.** *Journal of Molecular Evolution* 1999, **48**:284-290.
58. Nesnidal MP, Helmkampf M, Bruchhaus I, Hausdorf B: **Compositional heterogeneity and phylogenomic inference of metazoan relationships.** *Molecular Biology and Evolution* 2010, **27**:2095-2104.
59. Song H, Sheffield NC, Cameron SL, Miller KB, Whiting MF: **When phylogenetic assumptions are violated: base compositional heterogeneity and among-site rate variation in beetle mitochondrial phylogenomics.** *Systematic Entomology* 2010, **35**:429-448.
60. Gibson A, Gowri-Shankar V, Higgs PG, Rattray M: **A comprehensive analysis of mammalian mitochondrial genome base composition and improved phylogenetic methods.** *Molecular Biology and Evolution* 2005, **22**:251-264.
61. Regier JC, Cook CP, Mitter C, Hussey A: **A phylogenetic study of the 'bombycoid complex' (Lepidoptera) using five protein-coding nuclear genes, with comments on the problem of macrolepidopteran phylogeny.** *Systematic Entomology* 2008, **33**:175-189.
62. Regier JC, Shultz JW, Zwick A, Hussey A, Ball B, Wetzer R, Martin JW, Cunningham CW: **Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences.** *Nature* 2010, **463**:1079-1083.
63. Zwick A: **Degeneracy Coding Web Service. PhyloTools.** 2010 [http://www.phylotools.com/ptdegen1webservice.htm], Web, accessed 25 Feb. 2010.
64. Phillips MJ, Delsuc F, Penny D: **Genome-scale phylogeny and the detection of systematic biases.** *Molecular Biology and Evolution* 2004, **21**:1455-1458.
65. Cai JJ, Smith DK, Xia X, Yuen KY: **MBE Toolbox: a Matlab toolbox for sequence data analysis in molecular biology and evolution.** *BMC Bioinformatics* 2005, **6**:64.
66. Davis DR: **New leaf-mining moths from Chile, with remarks on the history and composition of Phyllocnistinae (Lepidoptera: Gracillariidae).** *Tropical Lepidoptera* 1994, **5**:65-75.
67. Shimodaira H: **An approximately unbiased test of phylogenetic tree selection.** *Systematic Biology* 2002, **51**:492-508.
68. Shimodaira H, Hasegawa M: **CONSEL: for assessing the confidence of phylogenetic tree selection.** *Bioinformatics* 2001, **17**:1246-1247.
69. Kumata T: **A taxonomic revision of the *Gracillaria* group occurring in Japan (Lepidoptera: Gracillariidae).** *Insecta Matsumurana, New Series* 1982, **26**:1-186.
70. Kumata T, Kuroko H, Ermolaev VP: **Japanese species of the *Acrocercops*-group (Lepidoptera: Gracillariidae).** *Insecta Matsumurana, New Series* 1988, **38**:1-111, 40:1-133.
71. Kumata T: **A new stem-miner of alder in Japan, with a review of the larval transformation in the Gracillariidae (Lepidoptera).** *Insecta Matsumurana, New Series* 1978, **13**:1-27.
72. Kumata T: **On the Japanese species of the genera *Macarostola*, *Aristaea* and *Systoloneura*, with descriptions of three new species (Lepidoptera: Gracillariidae).** *Insecta Matsumurana, New Series* 1977, **9**:1-51.
73. Kristensen NP, Scoble M, Karsholt O: **Lepidoptera phylogeny and systematics: the state of inventorying moth and butterfly diversity.** Linnaeus Tercentenary: Progress in Invertebrate Taxonomy. *Zootaxa* 2007, **1668**:699-747.
74. Whitfield JB, Kjer KM: **Ancient rapid radiations of insects: challenges for phylogenetic analysis.** *Annual Review of Entomology* 2008, **53**:449-472.
75. Lemmon AR, Brown JM, Stanger-Hall K, Lemmon EM: **The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference.** *Systematic Biology* 2009, **58**:130-145.
76. Felsenstein J: **Cases in which parsimony or compatibility methods will be positively misleading.** *Systematic Zoology* 1978, **27**:401-410.
77. Hendy MD, Penny D: **A framework for the quantitative study of evolutionary trees.** *Systematic Zoology* 1989, **38**:297-309.
78. Soltis DE, Soltis PS: ***Amborella* not a "basal angiosperm"? Not so fast.** *American Journal of Botany* 2004, **2004**:997-1001.
79. Philippe H, Germot A: **Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution.** *Molecular Biology and Evolution* 2000, **17**:830-834.

80. Hedtke SM, Townsend TM, Hillis DM: **Resolution of phylogenetic conflict in large data sets by increased taxon sampling.** *Systematic Biology* 2006, **55(3)**:522-529.
81. Kawahara AY, Mignault AA, Regier JC, Kitching IJ, Mitter C: **Phylogeny and biogeography of hawkmoths (Lepidoptera: Sphingidae): evidence from five nuclear genes.** *PLoS ONE* 2009, **4(5)**:e5719.
82. Kyrki J: **The Yponomeutoidea: a reassessment of the superfamily and its suprageneric groups (Lepidoptera).** *Entomologica Scandinavica* 1984, **15**:71-84.
83. Kuznetzov VI, Stekol'nikov AA: **Phylogenetic relationship between the superfamilies Psychoidea, Tineoidea, and Yponomeutoidea (Lepidoptera), taking into account the functional morphology of the male genital apparatus. Part 1. Functional morphology of the male genitalia.** *Entomological Review* 1976, **55**:533-548.
84. Kuznetzov VI, Stekol'nikov AA: **Phylogenetic relationship between the superfamilies Psychoidea, Tineoidea, and Yponomeutoidea (Lepidoptera), taking into account the functional morphology of the male genital apparatus. Part 2. Phylogenetic relationships of the families and subfamilies.** *Entomological Review* 1977, **56**:14-21.
85. Moulton JK, Wiegmann B: **Evolution and phylogenetic utility of CAD (rudimentary) among Mesozoic-aged eremoneuran Diptera (Insecta).** *Molecular Phylogenetics and Evolution* 2003, **31**:363-378.
86. Fang QQ, Cho S, Regier JC, Mitter C, Matthews M, Poole RW, Friedlander TP, Zhao SW: **A new nuclear gene for insect phylogenetics: dopa decarboxylase is informative of relationships within Heliothinae (Lepidoptera: Noctuidae).** *Systematic Biology* 1997, **46**:269-283.
87. Farrell BD: **Evolutionary assembly of the milkweed fauna: Cytochrome oxidase I and the age of *Tetraopes* beetles.** *Molecular Phylogenetics and Evolution* 2001, **18**:467-478.
88. Regier JC, Fang QQ, Mitter C, Peigler RS, Friedlander TP, Solis MA: **Evolution and phylogenetic utility of the *period* gene in Lepidoptera.** *Molecular Biology and Evolution* 1998, **15**:1172-1182.
89. Brower AVZ, DeSalle R: **Patterns of mitochondrial versus nuclear DNA sequence divergence among nymphalid butterflies: the utility of *wingless* as a source of characters for phylogenetic inference.** *Insect Molecular Biology* 1998, **7**:73-82.

doi:10.1186/1471-2148-11-182
Cite this article as: Kawahara *et al.*: Increased gene sampling strengthens support for higher-level groups within leaf-mining moths and relatives (Lepidoptera: Gracillariidae). *BMC Evolutionary Biology* 2011 **11**:182.