Reconstructing the phylogeny of aphids (Hemiptera: Aphididae) using DNA of the obligate symbiont Buchnera aphidicola

Eva Nováková a, b,*, Václav Hypša a, Joanne Klein b, Robert G. Footitt c, Carol D. von Dohlen d, Nancy A. Moran b

a Faculty of Science, University of South Bohemia, and Institute of Parasitology, Biology Centre, ASCR, v.v.i., Branisovka 31, 37005 Ceske Budejovice, Czech Republic
b Department of Ecology and Evolutionary Biology, Yale University, 300 Heffner Dr., West Haven, CT 06516-4150, USA
c Agriculture & Agri-Food Canada, Canadian National Collection of Insects, K.W. Neatby Bldg., 960 Carling Ave. Ottawa, Ontario, Canada K1A 0C6
d Department of Biology, Utah State University, UMC 5305, Logan, UT 84322-5305, USA

ABSTRACT

Reliable phylogenetic reconstruction, as a framework for evolutionary inference, may be difficult to achieve in some groups of organisms. Particularly for lineages that experienced rapid diversification, lack of sufficient information may lead to inconsistent and unstable results and a low degree of resolution. Coincidentally, such rapidly diversifying taxa are often among the biologically most interesting groups. Aphids provide such an example. Due to rapid adaptive diversification, they feature variability in many interesting biological traits, but consequently they are also a challenging group in which to resolve phylogeny. Particularly within the family Aphididae, many interesting evolutionary questions remain unanswered due to phylogenetic uncertainties. In this study, we show that molecular data derived from the symbiotic bacteria of the genus Buchnera can provide a more powerful tool than the aphid-derived sequences. We analyze 255 Buchnera gene sequences from 70 host aphid species and compare the resulting trees to the phylogenies previously retrieved from aphid sequences, only. We find that the host and symbiont data do not conflict for any major phylogenetic conclusions. Also, we demonstrate that the symbiont-derived phylogenies support some previously questionable relationships and provide new insights into aphid phylogeny and evolution.

© 2013 Elsevier Inc. All rights reserved.

1. Introduction

Aphids form a distinctive insect clade that features considerable variability in interesting biological traits, such as the presence of many distinct, yet genetically identical, forms of females during the life cycle (polyphenism), alternation of sexual and asexual reproduction, sterile soldier morphs, and seasonal alternation between unrelated groups of host plants. These traits vary among species, reflecting a long evolutionary history of biogeographical expansions and contractions and codiversification with plant hosts. Understanding the evolution of these traits thus requires a reliable phylogeny as a framework for particular evolutionary hypotheses. However, except for a few generally accepted aspects, studies on aphid phylogeny have not yet produced a clear picture of many relationships within this group.

Earlier attempts to reconstruct aphid phylogeny using morphology resulted in conflicting evolutionary scenarios (Heie, 1987; Wojciechowski, 1992). While an overall picture describing three distinct groups with either viviparous (Aphididae) or oviparous (Adelgidae and Phylloxeridae) parthenogenetic females was generally accepted, the phylogeny of the most diverse group, the Aphididae, remained unclear. More recent efforts have applied DNA sequence data to the reconstruction of Aphididae phylogeny. An emergent problem with such data is a lack of sufficient phylogenetic signal, mostly ascribed to a rapid adaptive radiation within Aphididae (von Dohlen and Moran, 2000; Martínez-Torres et al., 2001). Recently, this difficulty has been partly compensated by combining data from nuclear and mitochondrial genes (Ortiz-Rivas and Martínez-Torres, 2010); however, even this extended source of information leaves many relationships within Aphididae unresolved, and several evolutionary questions unanswered. Some questions concern the number of origins of particular aspects of biology. For example, three tribes with similar life cycles that include dwarf sexual forms lacking mouthparts (Eriosomatini, Fordini, Pemphigini) have been traditionally classified together in the Eriosomatinae (formerly Pemphiginae) (Heie, 1980; Remaudière and Remaudière, 1997). Furthermore, mainly on the basis of shared...
life cycle characteristics, Eriosomatinae, Hormaphidinae, and Anoecii- nae have been hypothesized to share a most recent common ancestor (Heie, 1987). Most molecular phylogenies fail to support these ideas, however, because they recover Eriosomatinae as paraphyletic, and position Eriosomatinae, Hormaphidinae, and Anoecii- nae as unrelated lineages. Other biological questions that could be illuminated by improved phylogenetic reconstruction concern whether the common ancestor of extant Aphididae fed on gymnosperms or angiosperms, and whether transitions between these major plant groups were accompanied by changes in diversification rates.

Inclusion of additional markers with appropriate information capacity, as well as additional taxa to bisect long terminal branches, are useful approaches to resolving ambiguous nodes of a rapid radiation (Whitfield and Lockhart, 2007). Nuclear gene markers would be the obvious, potentially informative additions for reconstructing the late Cretaceous radiation of Aphididae. However, due to the possibility of paralogy, designation of orthologous nuclear sequences across a broad selection of species could require extensive experimental work. This is particularly relevant for aphids, because a vast amount of gene duplication affecting more than 2000 gene families was confirmed within the *Acrithosiphon pismum* genome (The International Aphid Genomics Consortium, 2010). This extensive duplication of nuclear genes has been ongoing during aphid evolution, resulting in complex gene trees in which orthologs and paralogs are difficult to distinguish and the species phylogeny is obscured (e.g., Rispe et al., 2008; Nováková and Moran, 2012).

Aphids fortunately possess an additional source of inherited genetic material apart from their own genomes, in the form of obligate, maternally transferred, and highly derived bacterial mutualists. The association between aphids and bacteria of the genus *Buchnera* is one of the most extensively studied symbiotic systems since Buchner (1933) suggested their mutualistic association. As in many other mutualistic relationships, the bacteria supply essential nutrients to their hosts, which in turn provide a stable environment for their bacterial partners. Genomic studies have revealed that particular amino acids, vitamins and sterols are supplied by *Buchnera* (Akman Gunduz and Douglas, 2009; Moran et al., 2005; Moran and Degnan, 2006; Moya et al., 2008; Hansen and Moran, 2011). *Buchnera* is one of numerous groups of insect-associated symbionts for which evidence supports a long-term, strict cosegregation with the host (Moran et al., 1993, 1995; Clark et al., 2000). The acquisition of *Buchnera* symbionts by aphids is thought to have been a single event that took place by 150–200 MYA and was followed by cosegregation (Moran et al., 1993; Martínez-Torres et al., 2001). Due to this close evolutionary association, the symbiont and host phylogenies mirror each other for deeper evolutionary divergences. Thus, symbiont-derived data in principle can be used to reconstruct the evolutionary history of hosts.

Such an approach is appealing, because the small and simple genomes of symbionts may be an easier source of suitable sequences for phylogenetic analysis. This idea has been considered for some other insect-symbiont associations. For example, Kölsch and Pedersen (2009) suggested using endosymbionts for elucidating unresolved questions of reed beetle phylogeny. In addition to resolving host relationships, endosymbiont genes have been used for calculating divergence times in their hosts (e.g., *Cryptocercus* woodroaches (Maekawa et al., 2005)). *Buchnera* genomes consist of single-copy genes, most of which are shared across different *Buchnera* genomes (Moran et al., 2009); thus, they lack the complications presented by gene duplication and paralogy. Bacterial genomes possess several additional advantages, such as haplidity and absence of introns. Attempts to infer aphid and *Buchnera* phylogenies in a common framework have been undertaken several times (e.g., Munson et al., 1991; Moran et al., 1993; Martínez-Torres et al., 2001), mainly with the aim of testing phylogenetic correspondence of host and symbiont. Only a few studies were focused explicitly on reconstructing aphid phylogeny using *Buchnera* markers; these included analyses performed on *Uroleucon*, *Brachycoccus*, and *Mollitrichosiphum* species (Clark et al., 2000; Jousselin et al., 2009; Liu et al., 2013). All three studies ruled out the possibility of occasional horizontal transfers, even between closely related aphids sharing the same host plant. Even within species, phylogenies based on *Buchnera* genes are congruent with those based on mitochondrial sequences, confirming that this symbiont is strictly maternally inherited (Pecoud et al., 2009; Funk et al., 2000). Within an individual aphid, *Buchnera* genes also have higher copy number than aphid nuclear genes, so that *Buchnera* genes are relatively easy to amplify. DNA of *Buchnera* symbionts thus serves as a useful source of information for inferring aphid phylogeny. However, phylogenetic studies using *Buchnera* genes so far have been limited by sparse sampling across aphid taxa, and especially by taxonomic bias towards the subfamily Aphidinae.

In this study, we use *Buchnera*-derived sequences of five genes (*groEL*, *trpB*, *dnaB*, *ilvD* and 16S rDNA) from an extended taxonomic set to reconstruct phylogenetic relationships within the Aphididae. We then compare the resulting topologies to those reported from recent multilocus analyses of aphid-derived relationships. We focus mainly on the most problematic questions in aphid evolution, such as monophyly of individual subfamilies/tribes and their relationships, variation in DNA substitution rates, and rooting of the Aphididae tree. We show that symbiont genes yield informative phylogenetic signal and have several methodological advantages.

2. Materials and methods

2.1. Sample collection

Our study was designed to reconstruct evolutionary relationships among a wide diversity of aphids, represented by a broad sample of most major Aphididae taxa. The collection includes 70 species from 15 of 25 subfamilies and 25 of 36 tribes recognized in the most recent, comprehensive classifications of aphids (Remaudière and Remaudière, 1997; Nieto Nafria et al., 1997) (Table S1).

2.2. DNA extraction, primer design and sequencing

For all species, several individuals were pooled and homogenized, and genomic DNA was extracted using QiAamp DNA Micro Kit (Qiagen). In efforts to compile sufficient data for a reliable phylogenetic reconstruction, we aimed to amplify five *Buchnera*-derived genes, namely 16S rDNA, *groEL*, *trpB*, *dnaB*, and *ilvD*. To confirm the aphid host species identification, the sequence of the aphid mitochondrial gene COI was obtained for each sample.

Three primer pairs allowing for nine combinations were designed for each of the targeted *Buchnera* genes, except for 16S rDNA. Pairs of highly degenerate primers, corresponding to the coding region of *groEL*, *trpB*, and *dnaB* genes, were designed according to alignments of partial sequences available in GenBank using Primaclade software (Gadberry et al., 2005). Primers for *ilvD* were designed in CLC Genomic Workbench (CLC bio A/S) based on an alignment of *Buchnera* genome sequences, with the reverse primers corresponding to a highly conserved region of the flanking tRNA. Primers for amplifying the 16S rDNA gene and aphid mitochondrial gene COI were from previous studies (Folmer et al., 1994; Håjibabaei et al., 2006; Munson et al., 1991; Mateos et al., 2006). More detailed information on primer pairs, including the sequence and the length of amplified regions, is summarized in Table S2.

To amplify PCR products from diverse aphid lineages, we used the two best-performing primer pairs for each gene (Table S2). In
the case of the 16S rRNA gene, primers 10F and 35R were used to detect the presence of secondary symbionts commonly associated with aphids (Fukatsu and Ishikawa, 1993; Sandström et al., 2001; Russell et al., 2003). This primer pair spans roughly a 1600 bp region of the 16S–23S rRNA operon in most bacterial genomes. However, in the reduced genomes of Buchnera, genes encoding 16S rRNA and 23S rRNA are widely separated on the chromosome (Munson et al., 1993). Thus, in contrast to almost all other bacteria, the 10F/35R primer pair does not produce an amplicon in Buchnera. According to results of the first amplification, the product of the reaction using the universal 16S rRNA primer pair (10F and 1507R, amplifying most of the 16S rRNA gene) was directly sequenced or was subcloned to enable separation of the Buchnera and secondary symbiont sequences.

To recover Buchnera 16S rDNA sequences in samples with positive 10F/35R primed products indicating presence of other bacteria, amplicons of successful 10F/1507R PCRs were subcloned into Promega pGEM-T Easy vectors. On average, three transformant colonies from each species were picked and their inserts amplified by colony PCR using T7 and SP6 primers. Products of colony PCR as well as PCR from each species were picked and their inserts amplified by colony megap GEM-T Easy vectors. On average, three transformant colonies of successful 10F/1507R PCRs were subcloned into Pro- tective 10F/35R primed products indicating presence of other bacteria, as well as PCR products of success ful 10F/35R PCRs were subcloned into Promega pGEM-T Easy vectors. On average, three transformant colonies from each species were picked and their inserts amplified by colony PCR using T7 and SP6 primers. Products of colony PCR as well as PCR products for groEL, trpB, dnaB, ilvD, and the mitochondrial COI gene were Sanger-sequenced in both directions on an ABI3700 sequencer at the Yale University DNA sequencing service. Resulting reads were bidirectional sequencing did not show any variability. Ambiguous positions in dnaB were replaced with an “N”.

2.3. Alignments and phylogenetic analysis

Accession numbers of the sequences obtained in this study and those retrieved from GenBank for the phylogenetic analyses are listed in Table S3. Matrices for individual genes were compiled from the data obtained in this study and from data retrieved from GenBank, including corresponding Buchnera sequences and sequences for selected outgroup taxa. Outgroup sequences were chosen from the Ishikawaella capsulata genome (Hosokawa et al., 2006), which shows the lowest pairwise divergence from Buchnera, as well as other members of the Gamma proteobacteria, including Serratia symbiotica, Salmonella enterica, and Escherichia coli. Complete datasets for groEL, trpB, dnaB, and ilvD were aligned using the ClustalW algorithm as implemented in SeaView software (Gouy et al., 2010). 16S rDNA sequences were aligned in server-based program MAFFT, http://mafft.cbrc.jp/alignment/server/index.html, using the Q-INS-i algorithm with default parameters, which allows for a more accurate alignment of variable loops of the rRNA molecule. The raw alignments were manually corrected in SeaView (Gouy et al., 2010) and further processed in GBLOCKS (Castresana, 2000) to remove unreliably aligned positions.

The resulting matrices were analyzed using maximum likelihood (ML) and Bayesian inference (BI). To account for the possibility of unreconcilable substitutional saturation in complete datasets, all sequences were analyzed using only the first two codon positions versus all three codon positions for DNA-based analyses, and were also analyzed using amino acid translations for protein-coding genes. The evolutionary models best fitting the data, as well as parameters for ML analyses, were selected using programs jModelTest (Posada, 2008) and ProtTest 3 (Darriba et al., 2011). ML analyses and 100 bootstrap replicates were performed for all nucleotide and protein matrices with PhyML 3.0 (Guindon and Gascuel, 2003) with substitution models and parameters estimated from the data. Bayesian analyses were performed with the closest approximation of best-fit evolutionary models (Table 1) as implemented in MrBayes 3.1.2 (Ronquist and Huelsenbeck, 2003) in two independent runs with four chains. The analyses were run for three to twelve million generations with tree sampling every 100 generations. Exploration of MCMC convergence and burn-in determination was based on values of the average standard deviation of split frequencies, potential scale reduction factor (PSRF), and the log likelihood of the cold chain. Burn-in corresponded to 20–25% of the sampled trees. All the presented topologies are majority-rule consensus trees calculated from the post-burn-in tree samples.

The approximately unbiased (AU) test (Shimodaira and Hasegawa, 2001) was used to compare topologies in two ways. First, we used the AU test to determine whether hypothetical alternative topologies were significantly different from those retrieved by our analyses based on individual genes. In particular, we tested the monophyly/paraphyly of the subfamily Eriosomatinae. Second, we evaluated robustness of the phylogenetic signal by applying the AU test to the best ML topology for each gene and a topology with an obvious introduced artifact, i.e., polyphyly of Aphidinae. To obtain the alternative topologies, we constrained particular artifacts and constructed the tree under the ML criterion in PAUP* 4.0 (Swofford, 2003). To optimize computational demands for these analyses, we compiled simplified matrices for each gene consisting of one representative for each cluster consistently found in all ML and BI derived topologies (not shown). Significant differences between likelihood site patterns for each topology pair were tested using CONSEL version 0.11 (Shimodaira and Hasegawa, 2001). The results of the AU tests were used to identify the set of genes with sufficient information for phylogenetic inference (i.e., rejecting the artificial topology). Two concatenated matrices were constructed from the selected genes. The protein concatenate comprised amino acid sequences of the coding regions for groEL, dnaB, trpB, and ilvD. The nucleotide concatenate also included 16S rRNA sequences. Both of the matrices were analyzed as described above.

2.4. Tests of heterogeneity in evolutionary rates of symbiont genes

In Buchnera genes, rates of evolution at sites under purifying selection, such as nonsynonymous sites in protein-coding genes, are affected by the host population size and by the number of cells transferred from mother to offspring (Rispe and Moran, 2000). In addition to providing new data useful for inferring aphid phylogeny, our Buchnera dataset provides the opportunity to assess whether symbiont population structure differs among aphid taxa. Differences among aphid groups in life cycles, ecology, and transmission of symbionts are expected to produce different levels of purifying selection, leading to rate heterogeneity among aphid groups. Specifically, we tested for differences among groups in the rate of evolution at nonsynonymous sites (dN) relative to the rate at synonymous sites (dS).

We computed dN/dS for pairs of sister taxa in each lineage. Average dN/dS values were then compared among subfamilies, and the data were tested for the presence of statistically deviant values using Dixon’s Q test (Rorabacher, 1991). Because dN/dS ratios are meaningless if changes at synonymous sites are saturated, the saturation index for synonymous positions was determined for each dataset, as calculated in DAMBE (Xia and Xie, 2001), with a proportion of invariable positions assessed in JModelTest (Posada, 2008). Both saturation pattern tests and dN/dS calculations were performed only with the sites available for all analyzed taxa. These analyses were not performed for subfamilies represented by a single species or by a single pair of species.

2.5. Long branch attraction (LBA)

Preliminary results indicated that some relationships may have been affected by LBA, i.e., artificial clustering of rapidly evolving taxa regardless of their true relationship (for review see Bergsten,
2005). We focused on Aphidinae, where clustering of Aspidophorodon longicaudus, Capitophorus hudsonicus and two Pterocoma species would, if confirmed, contradict monophyly of the tribes Macrosiphini and Pterocommatini. We first tested for effects of LBA on positions of A. longicaudus and C. hudsonicus in the context of the entire tree, particularly their possible affinity to long-branched taxa of long-branched lineages from other subfamilies (e.g., Saltusaphidiinae, Calaphidiinae). Additional analyses were performed exclusively within Aphidinae between the four taxa mentioned above. For each analysis, a single branch was removed from the Aphidinae data sets.

3. Results

3.1. Datasets

Altogether, this study generated 255 new sequences from 70 Buchnera lineages. Characteristics of the resulting alignments are detailed in Table 1. Because individual single-gene matrices differ in taxon sampling, all of the concatenated matrices are partially incomplete. Missing sequences were either not available in GenBank (for taxa from previous studies), or, for our samples, we were not able to amplify the genes due to PCR failures, most likely caused by substitutions in the primer sites, loss of the gene or even possible absence of Buchnera in certain samples. For example, a PCR failure apparently related to primer specificity was experienced for combinations of dnaB primers, which failed to amplify sequences for all samples from Calaphidinae and Saltusaphidiinae subfamilies. Also, only a few representatives of the subfamily Lachninae yielded PCR products for trpB.

Individual matrices of the protein-coding genes contained comparable proportions of variable characters, with the largest proportion found in trpB and ilvD (64% and 65%, respectively; Table 1). AU tests with constrained topologies indicated different strengths of phylogenetic signal for different matrices. Sequences for trpB, groEL, ilvD and the 16S rRNA gene contained sufficient information to reject all of the tested artificial topologies (see Section 2). For the dnaB dataset, the test was significant only when a major topological artifact was introduced (data not shown). In contrast, the mitochondrial COI dataset contained few informative characters (Table 1), and the AU tests were repeatedly insignificant. This dataset was therefore not used for phylogenetic inference; COI sequences were used only to verify species determination based on nucleotide similarity with sequences already available in several copies in GenBank. COI sequences recovered in this study shared similarity above 99% with GenBank homologs from the same species, and thus as redundant records were not deposited in the database.

Nucleotide composition, measured as % GC content, varied by 10% across taxa. The lowest GC levels were found in Buchnera from Pseudoregma koshuenis, which averaged 27.7% for protein-coding genes. A bias towards higher AT content was also observed in Buchnera sequences from Drepanosiphinae, Greenideinae and Lachninae. On the contrary, most of the Buchnera sequences isolated from Macrosiphini showed a higher GC content, averaging 34.0% for coding genes.

3.2. Phylogenetic analyses of single-gene matrices

For all genes, analyses restricted to first and second codon positions yielded results similar to those obtained from full nucleotide sequences. However, measures of statistical support (i.e. bootstrap and posterior probabilities) as well as phylogenetic resolution became weaker with decreasing number of positions (i.e., complete DNA sequence, two positions only, and translation into amino acids). This result, together with tests based on a few amino acid matrices (results not shown), indicates that amino acid sequences would yield only poorly supported and largely unresolved topologies. Consequently, amino acid sequences were not used in the single-gene matrices.

All topologies derived from the single-gene datasets placed most aphid species into clades corresponding to tribes of the current classification. For higher taxonomic levels, results were less conclusive, and branch support values were weaker. However, monophyly of Aphidinae, Lachninae, Greenideinae and Chaitophorinae was recovered across different analyses and was well supported. The topology inferred from the ilvD single-gene matrix (the dataset with the highest proportion of informative sites) is shown in Fig. 1. Within Aphidinae, Aphidini was monophyletic; however, two species of Macrosiphini, Aspidophorodon longicaudus and Capitophorus hudsonicus, grouped with Pterocoma in most of the single-gene analyses. This grouping was recovered repeatedly in all LBA-avoiding tests, performed on a simplified dataset with one of the long branches removed in each analysis (an example topology shown in Fig. S1).

Monophyly of Hormaphidinae was recovered only in ML analysis of the trpB dataset. Two tribes, Hormaphidini and Cerataphidini, did not form a monophyletic group in any other single-gene analysis. Sequences derived from Pseudoregma koshuenis (Hormaphidinae: Cerataphidini) had the lowest GC content and clustered artifactualy with other sequences of markedly biased nucleotide content. For instance, in the dnaB based topology, Pseudoregma fell within the same cluster as Greenideinae and Lachninae, which displayed the lowest GC content values for dnaB (Figs. S2 and S3). The same pattern was found in the ilvD dataset, where Pseudoregma clustered as a close relative to significantly AT-biased sequences of the subfamily Drepanosiphinae (Fig. 1). Similar to results in Ortiz-Rivas and Martínez-Torres (2010), Hormaphidini and Thelaxini were supported as sister taxa. This relationship was supported by all analyses based on single genes, except for ilvD.

In most single-gene analyses, Calaphidinae was paraphyletic with respect to Saltusaphidinae. The only exception was the topology derived from trpB (Fig. S4), in which Saltusaphidinae was not represented and Calaphidinae was monophyletic. In contrast to the results of Ortiz-Rivas and Martínez-Torres (2010), none of

### Table 1

<table>
<thead>
<tr>
<th>Alignment designation</th>
<th>Number of taxa</th>
<th>Number of characters</th>
<th>Variable characters</th>
<th>Selected model AICc</th>
<th>Mr. Bayes</th>
</tr>
</thead>
<tbody>
<tr>
<td>trpB</td>
<td>60</td>
<td>489</td>
<td>311</td>
<td>GTR + I + Γ</td>
<td>GTR + I + Γ</td>
</tr>
<tr>
<td>dnaB</td>
<td>52</td>
<td>1209</td>
<td>677</td>
<td>GTR + I + Γ</td>
<td>GTR + I + Γ</td>
</tr>
<tr>
<td>groEL</td>
<td>85</td>
<td>1135</td>
<td>666</td>
<td>GTR + I + Γ</td>
<td>GTR + I + Γ</td>
</tr>
<tr>
<td>ilvD</td>
<td>77</td>
<td>801</td>
<td>520</td>
<td>GTR + Γ</td>
<td>GTR + Γ</td>
</tr>
<tr>
<td>16S rRNA</td>
<td>65</td>
<td>1494</td>
<td>1087</td>
<td>GTR + Γ</td>
<td>GTR + Γ</td>
</tr>
<tr>
<td>trpB_dnaB_groEL_ilvD_16S</td>
<td>94</td>
<td>5079</td>
<td>2700</td>
<td>TIM2 + Γ</td>
<td>TIM3 + Γ</td>
</tr>
<tr>
<td>trpB_dnaB_groEL_ilvD_1st2nd</td>
<td>94</td>
<td>2390</td>
<td>1021</td>
<td>GTR + I + Γ</td>
<td>GTR + I + Γ</td>
</tr>
<tr>
<td>trpB_dnaB_groEL_ilvD_AA</td>
<td>94</td>
<td>1195</td>
<td>662</td>
<td>MREV + I + Γ + F</td>
<td>WAG + I + Γ</td>
</tr>
</tbody>
</table>
our single-gene derived topologies supported monophyly of Drep-
osiphidae sensu Heie (1980), i.e., including Calaphidinae, Chaito-
phorinae and Drepanosiphinae.

Eriosomatinae was polyphyletic or paraphyletic in all single-
gene analyses. For ilvD-derived topologies, Fordini and Pemphigini
formed a monophyletic group while Eriosomatini species were
placed at a distant position (Fig. 1). A similar result was previously
reported for aphid nuclear gene-derived topologies by Ortiz-Rivas
and Martínez-Torres (2010). All other single-gene analyses recov-
ered topologies in which the tribes were distant from each other
e.g., Figs. S2 and S4). However, when a constrain t of Eriosomatinae
monophyly was tested on individual gene sets, all AU test values

![Fig. 1. BI topology based on ilvD dataset. Solid vertical lines refer to monophyletic tribes and subfamilies. Dashed vertical lines designate paraphyletic taxa. The branch of Pseudoregma koshuensis is scaled to ½ of its actual length. The values at nodes show posterior probability support.](image-url)
were not significant; thus, the monophyly of Eriosomatinae could not be confidently rejected ($p = 0.08–0.47$).

In most of the rooted single-gene topologies, Pemphigini formed the sister group to all other taxa. However, branching of the other deep nodes was inconsistent, varying with datasets and analytical methods, particularly for subfamilies represented by a single specimen, namely *Tamalia coveni* (Tamalinae), *Anoea oenotherae* (Anoeicinae), *Neophyliphis totarae* (Neophyliphidinae) and *Mindarus obliquus* (Mindarinae). The overall topological characteristics sorted into three different arrangements. The first supported three distinct lineages: Pemphigini, a monophyletic Aphidinae + Calaphidinae, and a cluster composed of all remaining subfamilies and tribes. In the other two scenarios, aphid diversity was distributed into two major clusters, either Aphidinae versus a lineage of all other taxa, or (Aphidinae + Calaphidinae) versus a lineage of all remaining taxa.

### 3.3. Phylogenetic analyses of concatenated data sets

In order to test the potential impact of data incompleteness in concatenated matrices, an additional groEL + trpB dataset was constructed from the taxa for which sequences of both genes were available. These genes were chosen because they were available for the largest number of taxa. The topology obtained from both BI and ML corresponded to the scenario described above with Aphidinae versus a lineage of all other taxa, and did not differ substantially from the results obtained from analyses of the incomplete concatenated matrices.

Similarly to the single-gene topologies, most results from concatenated datasets placed Pemphigini as sister to the remaining taxa. Two main types of topology were repeatedly retrieved from the concatenated data. While most BI analyses tended to produce topologies with three main clusters, e.g. Pemphigini, Aphidinae + Calaphidinae, and a clade comprising the rest of taxa (Figs. 2 and 3), a few ML analyses supported grouping into two major lineages with one consisting exclusively of the subfamily Aphidinae (not shown).

The topologies recovered in this study are partially congruent with the recently published aphid phylogeny inferred from two nuclear and one mitochondrial genes (Ortiz-Rivas and Martinez-Torres, 2010). The overall similarity between the aphid-derived and *Buchnera*-derived trees provides sufficient background for detailed comparisons and evaluation of particular phylogenetic and evolutionary patterns. Ortiz-Rivas and Martinez-Torres (2010) recovered three major aphid clades, namely Lachninae, a clade containing Eriosomatinae + Anoeicinae + Thelaxinae + Hormaphidinae + Mindarinae, and a clade containing Aphidinae + Calaphidinae + Drepanosiphinae + Chaitophorinae. Lachninae formed the sister group to a lineage comprising the latter two clades in their results. In contrast, we consistently recovered Pemphigini as the sister group to an including all other taxa (Figs. 2–4). None of our analyses produced a topology in which Lachninae branched from the base of the tree, although this position was not rejected by the AU tests. Furthermore, only the ML analysis based on concatenation of nucleotides for 1st and 2nd positions produced a cluster containing Aphidinae + Calaphidinae + Drepanosiphinae + Chaitophorinae, as in Ortiz-Rivas and Martinez-Torres (2010). In this topology, the cluster also included Saltusaphidinae and Phyllaphidinae, which were not sampled by Ortiz-Rivas and Martinez-Torres (2010). For other recovered topologies, only the close relationship between Aphidinae and Calaphidinae was highly supported. Thus, the third cluster from the Ortiz-Rivas and Martinez-Torres study, Eriosomatinae + Anoeicinae + Thelaxinae + Hormaphidinae + Mindarinae, was never recovered (Figs. 2–4).

Similar to the topologies derived from aphid genes (Ortiz-Rivas and Martinez-Torres, 2010), we obtained only low support for several deep nodes (Fig. 2). Particularly, the position of the subfamilies represented by a single specimen, namely *Tamalia coveni* (Tamalinae), *Anoea oenotherae* (Anoeicinae), and *Neophyliphis totarae* (Neophyliphidinae) appeared to be unstable under different analysis settings. Bootstrap support for topologies calculated under the ML criterion was low, possibly due to the incompleteness of concatenated matrices (Fig. 4).

Because the outgroups selected for all analyses are relatively distant to *Buchnera*, unrooted analyses (i.e. analyses excluding outgroups) were also performed with concatenated datasets to avoid potentially misleading information introduced by very distant sequences. An example of topologies consistently retrieved, regardless of method or gene composition of concatenates, is shown in Fig. 5. Taxa in this unrooted trichotomy fell into three main clusters: (1) Lachninae, (2) Neophyliphidinae, Chaitophorinae, Anoeicinae, Cerataphidinae, Drepanosiphinae, Thelaxinae, Greenideinae, and (3) Aphidinae, Calaphidinae, Saltusaphidinae, and Hormaphidinae, which included species with complex, host-alternating life cycles.

### 3.4. Tests of heterogeneity in evolutionary rate of symbiont genes

Comparisons of average dN/dS indicated heterogeneity in evolutionary rates across subfamilies and tribes. Elevated values of dN/dS were found for Calaphidinae, Chaitophorinae, Greenideinae and Saltusaphidinae (Table 2), all groups with simple life cycles containing few distinct female morphs and narrow host-plant ranges. Rates at nonsynonymous sites were not elevated in the Lachninae, nor in Aphidini, Pemphigini, or Hormaphidinae, which include species with complex, host-alternating life cycles.

### 4. Discussion

#### 4.1. Phylogenetic inference: comparison of *Buchnera* and aphid data

The results obtained here show that the *Buchnera* genes selected for this study are useful sources of data for corroborating or rejection of particular phylogenetic hypotheses, although they cannot completely resolve aphid phylogeny. Here we summarize how our results compare to the most recent phylogeny for Aphididae based on aphid genes (Ortiz-Rivas and Martinez-Torres, 2010). First, relationships strongly supported by both *Buchnera* and aphid sequences include monophyly of individual tribes and monophyly of Lachninae. Second, relationships congruent with those from aphid genes but more strongly supported by *Buchnera* sequences include paraphyly of Macrosiphini with respect to Pterocommatini and monophyly of Aphidinae (including Aphidini, Macrosiphini, Pterocommatini). Third, relationships incongruent between analyses based on *Buchnera* versus aphid genes, with either providing strong or stable resolution, were concentrated at the deeper nodes. An example is the monophyly versus polyphyly of Eriosomatinae: while Fordini was repeatedly recovered as a sister group to Eriosomatinae in *Buchnera*-based topologies, the aphid data supported a sister relationship of Fordini and Pemphigini. Other instances of unstable resolution by either aphid or *Buchnera* genes involve relationships between Calaphidinae and Aphidinae, and among the Calaphidinae, Chaitophorinae, and Drepanosiphinae, and relationships between Anoeicinae and Mindarinae. Most notably, we did not find any clear example of strongly supported discrepancies between analyses based on aphid genes versus *Buchnera* genes. We note that differences or lack of differences between the studies also could be affected by the differences in taxon sampling.
This overview demonstrates that, despite particular differences among the many trees recovered in this study and in Ortiz-Rivas and Martinez-Torres (2010), both data sources are highly congruent for strongly supported relationships, even though they may differ for weakly supported nodes. This suggests that the inconclusiveness of previous studies based on aphid DNA reflects real evolutionary processes (e.g., rapid diversification in some lineages) rather than methodological artifacts.
Apart from a general assessment of the suitability of Buchnera data for resolving aphid phylogeny, our study also found unexpected relationships, suggesting the need for taxonomic reassessment in some groups. For example, two genera classified as Macrosiphini, Aspidophorodon longicaudus and Capitophorus hudsonicus, grouped significantly with Pterocomma (Pterocommatini)
rather than with other Macrosiphini, thus rendering Macrosiphini paraphyletic. Clustering of these three species was supported by analyses of single-gene topologies and by concatenated matrices. Furthermore, a possible effect of LBA on the position of these taxa was tested and discounted, suggesting that their relationship reflects evolutionary history rather than phylogenetic artifact. A
previous study (von Dohlen et al., 2006) also recovered a paraphyletic Macrosiphini, due to the sister relationship of Pterocomma with Cavariella (a genus not included in the present study); a later study (Kim et al., 2011) duplicated this finding. Together, these studies indicate the need for a re-examination and probable taxonomic revision of Macrosiphini and Pterocommatini.

Some of the relationships found in this study have implications for the evolution of host-plant relationships and other interesting features of aphid biology. One example concerns Lachninae. Nor- mark (2000) found previously that Tramini was nested within Lachnini, rendering the latter tribe paraphyletic. Our analysis recovers the same pattern using Buchnera genes, and with high support. Not only do these results indicate that the tribal classification does not reflect phylogeny, they have implications for the interpretation of feeding strategies within Lachninae. Ortiz-Rivas and Martínez-Torres, 2010 proposed that a single transition from gymnosperm hosts to angiosperms may have occurred in an ancestor of Lachnini + Tramini, with Euclachnini retaining the ancestral

Fig. 5. Unrooted BI topology based on the concatenated matrix including 1st and 2nd codon positions. Three main clusters are defined (shaded areas) to facilitate comparison with the Ortiz-Rivas and Martínez-Torres (2010) scheme, and are resolved as a trichotomy in this unrooted analysis. Solid lines refer to monophyletic taxa. Dashed lines designate paraphyletic/polyphyletic taxa. Numbers are posterior probabilities.
state of gymnosperm feeding. However, they acknowledged that this scenario was not fully supported by their data, because a lineage comprising Lachnini + Tramini was only rarely present in their trees. The nesting of Tramini within Lachnini, supported strongly by our data, is compatible with the hypothesis of a single transition between gymnosperms and angiosperms within the Lachninae. Another example of a phylogenetic result with implications for the evolutionary history of host plant associations concerns the position of Saltusaphidinae. This subfamily was recovered previously as nested within Calaphidinae (von Dohlen and Moran, 2000); with greater taxon sampling in our study, this nested position was even more strongly supported. This relationship suggests that ancestors of sedge- and rush-feeding Saltusaphidinae shifted directly from feeding on trees (as in extant Calaphidinae) to feeding on herbaceous monocots. In contrast, most other aphid species living on herbaceous plants seem to have acquired these hosts through a host-alternating intermediate life-cycle stage including both trees and herbs (Moran, 1992).

One final example of a phylogenetic result with implications for the evolution of an unusual phenotype relates to the monophyly versus paraphyly of Eriosomatinae, whose members all produce dwarf sexual forms. Two of the three tribes (Eriosomatini and For-}

Table 2

Saturation pattern and evolutionary rates found for major aphid lineages.

<table>
<thead>
<tr>
<th>Lineage</th>
<th>Iss(^a) all sites</th>
<th>Iss(^c) all sites</th>
<th>Iss(^d) resolved sites</th>
<th>Iss(^c) resolved sites</th>
<th>Iss(^c)</th>
<th>dN/dS average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aphidinae</td>
<td>0.6124</td>
<td>0.8006</td>
<td>0.1488</td>
<td>0.7577</td>
<td>0.4817</td>
<td>0.0646</td>
</tr>
<tr>
<td>Calaphidinae</td>
<td>0.7304</td>
<td>0.8112</td>
<td>0.2120</td>
<td>0.7282</td>
<td>0.5740</td>
<td>0.1263(^a)</td>
</tr>
<tr>
<td>Fordini</td>
<td>1.0924</td>
<td>0.8426</td>
<td>0.1680</td>
<td>0.8368</td>
<td>0.8310</td>
<td>0.0691</td>
</tr>
<tr>
<td>Greenedeinae</td>
<td>0.7071</td>
<td>0.8305</td>
<td>0.0358</td>
<td>0.8081</td>
<td>0.7079</td>
<td>0.1191(^a)</td>
</tr>
<tr>
<td>Hornaphidinae</td>
<td>0.5282</td>
<td>0.8426</td>
<td>0.1925</td>
<td>0.8402</td>
<td>0.8391</td>
<td>0.0864</td>
</tr>
<tr>
<td>Chaitophorinae</td>
<td>1.1997</td>
<td>0.8426</td>
<td>0.1447</td>
<td>0.8229</td>
<td>0.8133</td>
<td>0.1053(^a)</td>
</tr>
<tr>
<td>Lachninae</td>
<td>0.6455</td>
<td>0.8267</td>
<td>0.1816</td>
<td>0.7626</td>
<td>0.6145</td>
<td>0.0609</td>
</tr>
<tr>
<td>Pemphigini</td>
<td>0.7388</td>
<td>0.8426</td>
<td>0.1902</td>
<td>0.8252</td>
<td>0.8152</td>
<td>0.0558</td>
</tr>
<tr>
<td>Saltusaphidinae</td>
<td>1.0317</td>
<td>0.8405</td>
<td>0.1530</td>
<td>0.8067</td>
<td>0.7985</td>
<td>0.1362(^a)</td>
</tr>
<tr>
<td>Thelaxinae</td>
<td>0.5449</td>
<td>0.8426</td>
<td>0.2386</td>
<td>0.8418</td>
<td>0.8445</td>
<td>0.0886</td>
</tr>
</tbody>
</table>

\(^a\) Index of substitution saturation (Iss) calculated using DAMBE software.

\(^b\) Critical Iss values for calculations including all sites and resolved sites only.

\(^c\) Iss\(^c\) stands for critical Iss values calculated for an asymmetrical tree.

\(^d\) Lineages demonstrated to have elevated dN/dS.

The results obtained from Buchnera-derived sequences confirm the overall suitability of this source of phylogenetic information, and further suggest that Buchnera genes may outperform host genes for solving particular phylogenetic problems. A potential drawback that must be considered when using these markers is the well-known tendency of symbiont sequences to suffer from various phylogenetic artifacts, mostly due to LBA and compositional bias. However, in many animal groups, these issues also affect mitochondrial and nuclear sequences; in aphids in particular, most nuclear genes are affected by extensive paralogy. Fortunately, new computational techniques are in development that can avoid artifacts due to LBA and compositional bias, provided that suffi-
cienct amounts of data are available (Lartillot et al., 2009; Husnik et al., 2011).

Codiversification with obligate bacterial symbionts has now been documented for many invertebrate taxa, including a wide range of insects such as cockroaches, ants, tsetse flies, weevils, lice, and almost all insect groups feeding on plant sap (Allen et al., 2007; Conord et al., 2008; Moran et al., 2008), as well as a diversity of other animal clades such as marine flatworms and some marine clams (Gruber-Vodika et al., 2011; Peek et al., 1998). Considering the rapidly expanding bacterial genome databases, symbiont-derived data are likely to become a powerful tool for reconstructing host phylogenies in these strictly coexisting associations.

Acknowledgments

We thank S. Aoki, T. Fukatsu, U. Kurosu, J. Ortego, D. A. J. Teunol and E. Maw for donating samples. This work was supported by the United States National Science Foundation (DEB-0723472, DEB-1106195 to N.A.M.). E.N. was supported on a Fulbright fellowship and by the Grant Agency of the University of South Bohemia (206/1106195 to N.A.M.). E.N. was supported on a Fulbright fellowship.

Supplementary data associated with this article can be found, in the online version, at https://dx.doi.org/10.1016/j.ympev.2013.03.016.

References


