The LINNEAN SOCIETY of London BIOLOGICAL Journal of the Linnean Society

Biological Journal of the Linnean Society, 2011, **104**, 770–785. With 7 figures

Phylogenetic information of genes, illustrated with mitochondrial data from a genus of gastropod molluscs

SIMON F. K. HILLS*, STEVEN A. TREWICK and MARY MORGAN-RICHARDS

Ecology Group, Institute of Natural Resources, Massey University, Private Bag 11 222, Palmerston North, New Zealand

Received 10 February 2011; revised 11 June 2011; accepted for publication 11 June 2011

A critical assessment of sequencing markers is desirable to ensure that they are appropriate for the specific questions that are to be addressed. This consideration is particularly important where the data set will be used in highly sensitive analyses such as molecular clock studies. However, there is no standard practice for marker assessment. We examined the mitochondrial DNA sequences of a genus of marine molluscs to assess the relative phylogenetic signal of a number of genes using an extension of splits-based spectral analysis. With a data set of almost 8 kb of DNA sequences from the mitochondrial genome of a lineage of marine molluscs, we compared the phylogenetic information content of six protein coding, two ribosomal DNA, and 12 transfer RNA genes. Split-support graphs were used to identify which genes contributed a relatively low signal-to-noise ratio of phylogenetic information. We found that *cox*2 and *atp*8 did not perform well for reconstruction at the within-genus level for this lineage. Consideration of nested subsets of taxa improved the resolution of relationships among closely related species by reducing the time frame over which evolutionary processes have occurred, allowing a better fit for models of DNA substitution. Through this fine-tuning of available data it is possible to generate phylogenetic reconstructions of increased robustness, for which there is a greater understanding of the underlying signals in the data. We recommend a suitable mitochondrial DNA fragment and new primers for intergeneric studies of molluscs, and outline a general pipeline for phylogenetic analysis. *Biological Journal of the Linnean Society*, 2011, **104**, 770–785.

ADDITIONAL KEYWORDS: *Alcithoe* – data exploration – marker assessment – molecular evolution – phylogenetic splits – spectral analysis.

INTRODUCTION

As phylogenetic analysis has become more sophisticated, the molecular evolution of the genes used to infer species evolution has required increased scrutiny. Down-stream analyses, such as molecular-clock studies, are increasingly a common focus in molecular phylogenetics, and such techniques tend to be highly sensitive to incongruent signals in the underlying data (Ho & Phillips, 2009). To ensure the accuracy of inferences made with such techniques it is necessary to have data that will infer robust phylogenetic trees. In order to confidently build robust phylogenies one needs to critically assess sequence data to create molecular data sets that are best suited to different levels of divergence.

*Corresponding author. E-mail: s.f.hills@massey.ac.nz

Comparative approaches allow the phylogenetic utility of markers to be determined (Graybeal, 1994). It is desirable to know whether there is sufficient data for the phylogenetic estimation to reflect the evolutionary history of the entire genome (and therefore the organism), rather than the evolutionary history of one or a small set of genes. For example, given that the genes in the mitochondrial genome are contained as a single linkage group it might be expected that the individual gene trees would agree with one another, but this is not always the case (Cummings, Otto & Wakeley, 1995).

The phylogenetic information content of genes has generally been scrutinized from two extreme viewpoints – deep phylogeny (e.g. between orders of vertebrates; Cummings *et al*., 1995) and within species (e.g. human samples; Non, Kitchen & Mulligan, 2007) – but many studies that occupy the

middle ground, i.e. phylogenies of related species, within genera or among sister genera, do not usually address the question of the phylogenetic information content of the markers used. The great majority of phylogenetic studies of animals in this intermediate range have concentrated on mitochondrial *cox1*, *16S*, *cytB*, and nuclear *18S* and *28S* sequencing markers. This reliance is largely based on the availability of universal PCR primers, but the reliability of DNA extraction to recover mitochondrial sequences has also contributed. However, universal markers have some limitations. Because of the need to anneal across broad taxonomic ranges, most universal primers target highly conserved DNA fragments. The consequence of this requirement is that the sequencing markers obtained are limited in the taxonomic depths for which they have robust phylogenetic resolution. For example, the universal nuclear markers, such as *18S* and *28S*, tend to lack resolution for shallow intragenus-level divergence. Conversely, the three mitochondrial genes commonly used for intraspecies analysis (*cox1*, *12S*, and *16S*) are more rapidly evolving sequences, but *12S* and *16S* can be difficult to align for deeper relationships, and *cox1* rapidly becomes saturated at third codon positions, and therefore loses resolution (Simon *et al*., 1994; Roe & Sperling, 2007).

The relative information content of mitochondrial genes has been investigated, and a range of signals has been found in different genes (e.g. Corneli & Ward, 2000; Mueller, 2006; Paton & Baker, 2006). Some studies separate mitochondrial genes into classes based on the level of phylogenetic usefulness (e.g. Zardoya & Meyer, 1996); however, the majority of such studies deal with vertebrate lineages or very broad evolutionary distances (Simon *et al*., 1994). It is therefore likely that the patterns of gene variability observed are not the same in all data sets. An analysis of the utility and critical selection of the markers to be used to resolve a phylogeny would lend greater confidence to the resulting phylogenetic hypothesis, and would provide a foundation from which to assess challenging phylogenetic relationships. Such an analysis is expected to aid marker choice for studies of similar organisms. As the ease and cost-effectiveness of DNA sequencing increases, the reliance on universal primers should diminish. Thus targeting genes suitable for a given type of analysis will be a more feasible strategy, rather than marker selection by convenience. An additional benefit of characterizing the phylogenetic utility of markers is to provide information as to the most cost-effective regions to sequence from poor quality DNA samples, such as ancient DNA and extractions from poorly preserved museum specimens.

Assessing the robustness of molecular data sets is not a trivial problem. Robustness can be judged by both congruence among different tree-building methods (where a more robust signal in the data is likely to result in more consistent results from disparate methods) and by the support for inferred clades. High bootstrap values and Bayesian posterior probabilities are often considered to be indicative of 'true' tree topology. These measures are only indicative of accuracy if the evolutionary model is accurate; however, this is rarely the case for biological data. As such, misleading signals can occur, high bootstrap values can be obtained for incorrect topologies (Phillips, Delsuc & Penny, 2004), and Bayesian support can be inflated and not representative of the probability of the correct resolution of clades (Simmons, Pickett & Miya, 2004). It is preferable to assess the robustness of a given phylogeny by exploring the signal in the underlying data, independently of the tree. This allows for the assessment of the validity of bootstrap and Bayesian support values, and also for the evaluation of the signals behind clades with low bootstrap and Bayesian support values.

One method of doing this is through the examination of phylogenetic splits, which represent bipartitions of taxa in the DNA data set (Bandelt & Dress, 1992). Any molecular data set will contain one or more sets of compatible splits, and for each compatible set there will be a set of incompatible splits. Any branch in a phylogenetic tree represents a split dividing the represented taxa into two sets. A set of splits is compatible if, when combined, they describe all or part of a fully resolved phylogenetic tree for the taxa involved; if not, they are incompatible (Bryant & Moulton, 2004). Any given phylogeny derived from a data set can be described as a set of compatible splits, and any signal in the data that conflicts with that phylogeny can be described by a set of incompatible splits, with reference to the compatible split set that describes the tree. For any given split, a split-weight (or support) value can be derived from an underlying sequence alignment or distance matrix. This support value will reflect the level of evidence in the underlying data for a given bipartition of the taxa, and is analogous to a branch length separating the two sets of taxa. Conflict values for individual splits are calculated from the sum of the support for splits that contradict a given split, normalized by the ratio of total support for all splits over the total of all conflict values (Lento *et al*., 1995).

Analysis of splits has proven to be a powerful tool for visualizing signal and conflict in phylogenetic data (Holland *et al*., 2004; Huson & Bryant, 2006), but has not realized its full potential as a tool for scrutinizing molecular data sets. Spectral analysis allows the

visualization of conflict and support for all signals in a data set, independently of a tree. When referenced to a tree generated from the same data the spectral analysis can be used to diagnose weaknesses in that tree, and reinforce likely true signals. Identifying genes that provide poor phylogenetic information is achieved by a comparison of signal and conflict for splits provided by individual genes. Previous studies have shown the potential of spectral analysis for this purpose (Lento *et al*., 1995; Wagele & Mayer, 2007). Furthermore, when large sequence data sets exist it is likely that the selection of a subset of the genes that maximize the signal-to-noise ratio will result in phylogenies of greater robustness (Jeffroy *et al*., 2006).

Here we compare mitochondrial genes from the New Zealand marine gastropod genus *Alcithoe*. These snails are benthic, direct-developing, carnivorous neogastropod molluscs. Molluscs represent a group for which there has been limited molecular phylogenetic analysis considering the extensive species diversity. As such, molluscan phylogenetics has relied heavily on universal markers and, to date, there has been little consideration of the relative information content of these genes. In the *Alcithoe* a prevalence of large intraspecific and low interspecific morphological variation has made the phylogeny of the genus difficult to estimate using morphological characters, and to date there has been no molecular treatment. The taxonomy of both extant and extinct *Alcithoe* is well described, although its stability is subject to the vagaries of morphological characters. Based on shell characteristics, 17 living species are recognized, three of which are subdivided into either two or three subspecies (Bail & Limpus, 2005). There has been a recent increase in the number of extant taxa recognized as a result of the development of new commercial fisheries and research trips that have yielded new specimens from deeper waters. It is possible that several of these new putative taxa represent local forms of known species. Although it is possible that widespread variable taxa may represent species complexes, the recent history of *Alcithoe* taxonomy is dominated by the synonymy of species, as new samples have bridged apparent morphological gaps.

We sequenced more than 7 kb of mitochondrial DNA from eleven *Alcithoe* species, covering nine genes. Although we aim to infer a robust phylogeny for the *Alcithoe*, our main goal here is to demonstrate a technique for the assessment of the suitability of the mitochondrial genes comprising this data set for intrageneric phylogenetic studies. To do this we will:

1. Explore the signal in each of the genes separately using summary statistics and tree-building methods.
2. Assess the comparative phylogenetic utility of each gene, using splits to examine the relative contribution of signal and noise in a novel approach using spectral analysis to compare the combined spectra of all genes.
3. Make recommendations as to the suitability of the genes comprising this data set for molluscan phylogenetic studies, and recommend a phylogenetic analysis pipeline.
4. Infer a robust phylogeny for the *Alcithoe* using the identified genes.

MATERIAL AND METHODS

Taxon sampling

We sampled ten of the 17 extant species of *Alcithoe* recognized in Bail & Limpus (2005) from New Zealand waters (Table 1). With the exception of *Alcithoe larochei tigrina* Bail & Limpus, 2005, all putative subspecies have been excluded. *Alcithoe* species not sampled here all have restricted ranges, largely in the far north and far south of New Zealand, and have eluded sampling efforts. The only member of the genus not found in New Zealand, *Alcithoe aillaudorum* Bouchet & Poppe, 1988, has been sourced from New Caledonia. Eight putative out-group species for *Alcithoe* were obtained from New Zealand, Australia, and South America (Table 1).

DNA extraction and amplification

DNA was extracted from foot tissue from both frozen and ethanol-preserved specimens using a high-salt buffered extraction method (Norman, Olsen & Christidis, 1998), modified as described in Appendix S1. Initially short-range polymerase chain reaction (SR-PCR) was carried out to amplify fragments of between 300 and 1000 bp of mitochondrial cytochrome *c* oxidase 1 (*cox1*) and *16S* ribosomal DNA (*16S*) using universal primers (Table S1). Products generated by SR-PCR were sequenced with both forward and reverse primers using BigDye Terminator v3.1 and an ABI 3730. From these short sequences primers were designed in *cox1* and *16S* to amplify longer mitochondrial DNA (mtDNA) fragments (Table S1), of approximately 6 kb in length. LR-PCR products were sequenced by primer walking (Kusukawa *et al*., 1990). The PCR protocols used are described in Appendix S1. From these DNA fragments and complete mitochondrial sequences available on GenBank [*Ilyanssa obsoleta* (Say, 1822) NC_007781; *Lophiotoma cerithiformis* (Powell, 1964) NC_008098; *Conus textile* Linnaeus, 1758 NC_008098; *Littorina saxatilis* (Olivi, 1792) AJ132137] additional primers were designed for highly conserved regions (Table S1), in order to extend the ends of the sequence fragment

Table 1. Volute species used to study the phylogenetic information in 9 mitochondrial genes

Genus	Species	Voucher number	Sample Location		GenBank accessions
Alcithoe	*aillaudorum*	NB 1024	Isle des Pins	New Caledonia	JN379020 JN379030
Alcithoe	*arabica*	M.279684	Wellington	New Zealand	JN182223
Alcithoe	*benthicola*	M.183806	Coromandel	New Zealand	JN182217
Alcithoe	*fissurata*	M.183785	Coromandel	New Zealand	JN182225
Alcithoe	*flemingi*	M.183833	Chatham Rise	New Zealand	JN182218
Alcithoe	*fusus*	M.279683	Nelson	New Zealand	JN182220
Alcithoe	*jaculoides*	M.274972	North Island East Coast	New Zealand	JN182221
Alcithoe	*larochei*	M.274116	North Island East Coast	New Zealand	JN182227
Alcithoe	*larochei tigrina*	M.183799	Coromandel	New Zealand	JN182224
Alcithoe	*lutea*	NIWA 30452	Challenger Plateau	New Zealand	JN182219
Alcithoe	*pseudolutea*	M.183802	Coromandel	New Zealand	JN182222
Alcithoe	*wilsonae*	M.190062	South Island	New Zealand	JN182228
Cymbiola	*pulchra subelongata*	M.273459	Queensland	Australia	JN182216
Odontocymbiola	*simulatrix*	MZSP44320	Cabo Santa Marta	Brasil	JN379019 JN379027
Athleta	*studeri*	M.273462	Queensland	Australia	JN379024 JN379025
Amoria	*hunteri*	M.273463	Queensland	Australia	JN182226
Adelomelon	*beckii*		Mar del Plata	Argentina	JN379023 JN379029
Adelomelon	*brasiliana*	MACN-In39336	Mar del Plata	Argentina	JN379021 JN379026
Adelomelon	*riosi*	MZSP32971	Cabo Frio	Brasil	JN379022 JN379028

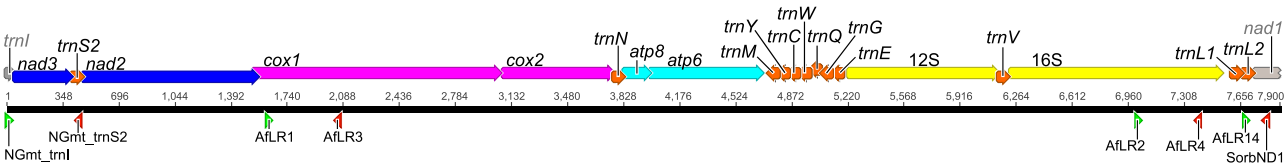

Figure 1. Mitochondrial gene arrangement in the New Zealand marine mollusc genus *Alcithoe*. Genes that comprise the sequenced DNA fragment from Volutidae are labelled, and include: *12S*, short subunit rRNA; *16S*, long subunit rRNA; *atp6*, ATP synthase F0 subunit 6; *atp8*, ATP synthase F0 subunit 8; *cox1*, cytochrome *c* oxidase subunit 1; *cox2*, cytochrome *c* oxidase subunit 2; *nad2*, NADH dehydrogenase subunit 2; *nad3*, NADH dehydrogenase subunit 3; *trnC*, cystine tRNA; *trnE*, glutamic acid tRNA; *trnG*, glycine tRNA; *trnL1*, leucine (CUN) tRNA; *trnL2*, leucine (UUR) tRNA; *trnM*, methionine tRNA; *trnN*, asparagine tRNA; *trnQ*, glutamine tRNA; *trnS2*, serine (AGN) tRNA; *trnV*, valine tRNA; *trnW*, tryptophan tRNA; *trnY*, tyrosine tRNA. Arrowheads indicate the direction of transcription for each gene. Binding sites for the primers designed to generate this DNA fragment are indicated (see Table S1).

by up to 1 kb in each direction. The binding sites of primers developed here are shown in Figure 1.

Sequence analysis and phylogenetic reconstruction

Sequences were edited using SEQUENCHER v4.6 (Gene Codes Corporation, Ann Arbor, MI, USA). Alignments were generated in SEQUENCHER and exported in nexus format. SE-AL 2.0a11 (Rambaut, 2002) was used to infer protein sequences from the nucleotide sequences and to refine alignments, as appropriate. Ribosomal DNA genes were aligned based on secondary structure. The ribosomal RNA gene *16S* was aligned using a published molluscan consensus structure (Lydeard *et al.*, 2000), although we found domain 1 to be too variable to unambiguously align based on this consensus structure, and was aligned based on common secondary structures for volute species returned by Mfold (Zuker, Mathews

& Turner, 1999). As no consensus structure of *12S* is available for molluscs, this alignment was based on similarity to structures on the Comparative RNA Web Site (http://www.rna.ccbb.utexas.edu; Cannone *et al*., 2002) using the secondary structures of *Paracentrotus lividus* (Lamarck, 1816) and *Artemia franciscana* Kellogg, 1906. Because of alignment ambiguity with the 5′ and 3′ ends of the chosen model sequences, Mfold was used to infer structures of the volute sequences to use as an alignment guide for these regions. Transfer RNA genes were compared with structures reported for *L. cerithiformis* (Bandyopadhyay *et al*., 2006) in order to identify putative stem and loop regions for accurate alignment.

For the purpose of phylogenetic analysis several partitioned subsets of the sequence data were created. In addition to the complete data set a concatenated data set was generated with all intergenic spacers removed, and where an overlap exists the relevant nucleotide positions were included for both genes separately. Each protein coding gene and the two ribosomal RNA (rRNA) genes were each given individual partitions, and the transfer RNA (tRNA) genes were partitioned as a single concatenated set.

Maximum parsimony reconstruction, ModelTest (Posada & Crandall, 1998) and partition homogeneity tests were implemented using PAUP* 4.0 (Swofford, 1998). Consistency indices and partitioned homogeneity tests were also generated in PAUP*. Neighbour-joining trees were constructed using the GENEIOUS tree builder in the GENEIOUS software package (Drummond *et al*., 2007). Maximum likelihood reconstruction and Bayesian analysis were carried out using PHYML 2.4.4 (Guindon & Gascuel, 2003) and MR BAYES 3.1.2 (Huelsenbeck & Ronquist, 2001), respectively, as implemented in GENEIOUS. Maximum parsimony reconstruction was carried out with default parameters, with the exception that 1000 bootstrap replicates were performed. Alignments that contained gaps were analysed with gaps as missing and with gaps as a fifth state, in order to assess the effect of gaps on phylogenetic reconstruction within this data set. Maximum likelihood reconstruction was carried out under three sets of modelling parameters for each data set. A simple model (HKY with default parameters in PHYML), an intermediate model (HKY with the averaged parameters for all models, as returned by ModelTest), and the specific model and parameters (or as close as possible using PHYML settings) returned by ModelTest using Akaike's information criterion (AIC) (generally the most complex model).

Bayesian reconstruction was carried out separately for each data partition using both GTR and HKY models with default parameters and with four heated chains of length 1 000 000, sampling every 1000 generations, with a 10% burn-in. Visualization of conflict in the data through analysis of splits and networks was carried out in Splits Tree (Huson & Bryant, 2006). Splits were derived from nucleotide alignments in Splits Tree 4. These splits were transferred into SpectroNet (Huber *et al*., 2002) to calculate split weight and conflict values. These values were used to generate Lento plots (Lento *et al*., 1995) for individual genes and split-support graphs for the collected data.

Spectral analysis of the splits data using Lento plots allows a detailed examination of the relative signal-to-noise level in a data set. The sum of support and conflict of all gene partitions for each split, calculated by Neighbor-Net from uncorrected p distances, illustrates the distribution of information in the data set. The significance of observed splits was judged not only by the level of support or conflict that any given gene partition has for a set of splits, but also on the number of genes that support or conflict with a given split. Therefore splits that have little support in any given gene partition become significant when most or all considered genes exhibit some support for that split. Phylogenetically problematic genes can be identified where splits are supported or show conflict from only one gene.

RESULTS

Out-group selection

Short sequences were obtained from a range of available volute samples to identify which are the most closely related to *Alcithoe*, and therefore the most appropriate out-group. Initial phylogenetic analyses using 313 bp of *16S* and 354 bp of *nad3* resolved New Zealand *Alcithoe* as monophyletic, but *A. aillaudorum* from New Caledonia was not part of this clade, and nor was it sister to it (Fig. 2). Constraining the tree topology to include all currently recognized *Alcithoe* species in a monophyletic clade resulted in a significantly less-likely tree (Shimodaira–Hasegawa, SH, test; $P < 0.05$). Two Australian volute species [*Cymbiola pulchra* (Sowerby, 1825) and *Amoria hunteri* (Iredale, 1931)] were chosen as an out-group for the New Zealand *Alcithoe*, being the most closely related volute taxa in the phylogeny, with bootstrap support of 90 for this sister relationship. The two out-group species and 11 New Zealand *Alcithoe* species were then used to generate the full sequence data set.

Sequence data

DNA sequences of between 7681 and 7733 bp were generated for each taxon, and include the following genes: *nad3*, *trnS*(AGN), *nad2*, *cox1*, *cox2*, *trnD*, *atp8*, *atp6*, *trnM*, *trnY*, *trnC*, *trnW*, *trnQ*, *trnG*, *trnE*, *12S*, *trnV*, *16S*, *trnL*(CUN) and *trnL*(UUR) (Fig. 1). A

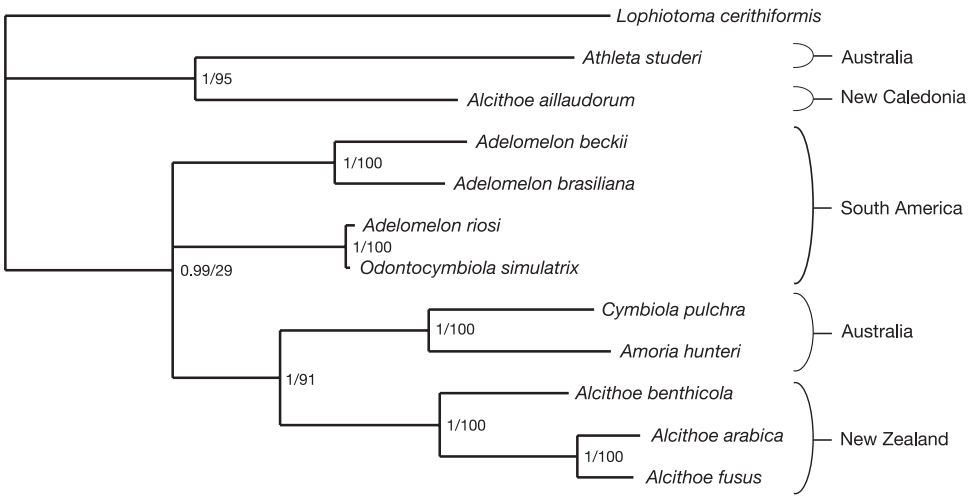

Figure 2. Phylogeny to establish the molecular context of *Alcithoe* within Volutidae derived from a concatenated alignment of 313 bp of *16S* and 354 bp of *nad3* from 19 volute taxa. *Lophiotoma cerithiformis* (NC_008098), a conoidean gastropod, was used as an out-group for the Volutidae. Bayesian posterior probabilities and maximum-likelihood bootstrap support are given for each node (B/ML). *Cymbiola* and *Amoria* are revealed as the most appropriate out-group taxa for phylogenetic reconstruction of *Alcithoe* species within this data set. *Alcithoe aillaudorum* from New Caledonia is not sister to the New Zealand *Alcithoe*.

summary of the details of these sequences is given in Table S2. This DNA fragment represents approximately half the neogastropod mitochondrial genome, and the gene arrangement and order is identical to the 13 neogastropod mollusc mitochondrial genomes published on GenBank to date (e.g. Bandyopadhyay *et al*., 2006; Simison, Lindberg & Boore, 2006; Cunha, Grande & Zardoya, 2009; McComish *et al*., 2010). Variability in the length of rRNA and tRNA genes, and intergenic spacer regions, primarily between the two out-group taxa (*Cymbiola pulchra* and *Amoria hunteri*) and the *Alcithoe* in-group, required the inclusion of gaps in the alignment of these sequences. Additionally, *nad2* from *Cymbiola* contained a single amino acid insertion, with respect to the other taxa. However, there was insufficient variability to warrant the exclusion of any coding sequence on the grounds of ambiguity.

Summary statistics

Alignment length, summaries of variability, consistency index, and ModelTest results for the complete data set, each of the nine gene partitions, and the concatenated data set are presented in Table 2. These statistics provide useful general information about the data, and identify genes that might be problematical for phylogenetic reconstruction. Although relatively short (159 bp), *atp8* exhibits high variability, but third codon position variability is lower than other genes, whereas second position variability is twice that of any other gene. Overall, low variability is seen in *cox1* and *12S*, but the most conserved partition is the set of 12 tRNA genes. The genes *cox1* and *cox2* show an accumulation of variability in the third codon positions (85 and 77%, respectively), but low amino acid variability (3 and 15%, respectively), suggesting a high rate of synonymous substitution. However, consistency indices for all subsets of the data are very similar and do not indicate high levels of saturation in any genes (Table 2). The highest consistency indices are seen where the lowest variation is recorded (tRNAs and *12S*), but these genes also have the highest ratio of parsimony-uninformative to parsimony-informative sites.

Four DNA substitution models are recovered as best fits for the individual genes using ModelTest. Interestingly, the two most complete subsets of data (complete and concatenated) are best modelled by the nine-parameter TVM+I+G model, even though individual genes, such as *cox2* and *atp6*, recover more complex models. Partition homogeneity tests were carried out on all pairwise combinations of the nine individual gene partitions, both excluding and including gap information. Only *cox2*/*12S* and *cox2*/*nad2* gene combinations showed significant partition heterogeneity. However, when corrected for multiple tests there is no significant partition heterogeneity among the set of nine genes (data not shown).

Phylogenetic reconstruction

Tree reconstruction using a range of methods (maximum likelihood, maximum parsimony and

Table 2. Summary statistics from alignments of mitochondrial DNA sequence data for 13 volute species

Gene	Align. length	pVAR* total	pVAR* 1st codon position	pVAR* 2nd codon position	pVAR* 3rd codon position	Amino acid pVAR*	ci†	Parsimony uninform./ inform.‡	DNA substitution model§
nad3	354	0.40	0.26	0.09	0.68	0.25	0.65	51/89	TVM+G
nad2	1089	0.41	0.29	0.10	0.61	0.32	0.68	171/273	TVM+I+G
cox1	1536	0.27	0.14	0.01	0.85	0.03	0.64	154/260	HKY+I+G
cox2	687	0.32	0.19	0.05	0.77	0.15	0.62	81/140	GTR+I+G
atp8	159	0.44	0.27	0.20	0.53	0.40	0.66	27/43	K81uf+G
atp6	696	0.38	0.24	0.10	0.65	0.23	0.64	89/173	GTR+I+G
12S	898	0.25	N/A	N/A	N/A	N/A	0.80	101/108	HKY+I+G
16S	1375	0.31	N/A	N/A	N/A	N/A	0.76	181/208	K81uf+I+G
tRNAs	829	0.21	N/A	N/A	N/A	N/A	0.79	72/77	K81uf+G
Concat	7623	0.31	N/A	N/A	N/A	N/A	0.69	927/1371	TVM+I+G
Comp	7822	0.33	N/A	N/A	N/A	N/A	0.69	1002/1434	TVM+I+G

*Proprotion of observed variable sites.
†Consistency index.
‡Number of parsimony uninformative sites/number of parsimony informative sites.
§Infered using Akaike's information criterion (AIC) in ModelTest.

Bayesian analyses) and each of the individual data partitions produced a range of topological solutions. From a total of 88 combinations of tree-building methods, models, and data sets, 25 alternative tree topologies were returned. The complete data set returned the same tree topology under each reconstruction method, each with high support. However, none of the data subsets were as consistent under the alternative tree-building strategies. Each of the individual gene partitions produced several tree topologies. Even the concatenated data set, which only omits 167 bp of intergenic spacer, and where nucleotides in overlapping regions appear twice, produces two different tree topologies. For the individual gene partitions the least conflict in tree topology is seen in *cox1* and *atp6*, which each return only two alternative tree topologies, whereas *16S* returns a different tree topology for each of the tree estimation methods and parameter sets used. SH tests of the trees returned for each data subset showed that the different topologies did not give a significantly better fit to the DNA sequences. The main areas of conflict in the tree topology are around the placement of *Alcithoe wilsonae* (Powell, 1933) and the resolution of four closely related taxa: *Alcithoe lutea* (Watson, 1882), *Alcithoe larochei* Marwick, 1926, *Alcithoe fusus* (Quoy and Gaimard, 1833), and *Alcithoe fissurata* (Dell, 1963). In addition, two less prevalent inconsistencies were observed: the *cox2* gene recovered a sister relationship of *Alcithoe jaculoides* Powell, 1924 and *Alcithoe Arabica* (Gmelin, 1791), but with low support. The tRNA set *12S*, *atp8*, and *cox2* consistently place *Alcithoe tigrina* Bail & Limpus, 2005 and *Alcithoe pseudolutea* Bail & Limpus, 2005 in a clade that is the most recently derived in the phylogeny.

A Neighbor-Net network, derived from the complete data set, illustrates the areas in the *Alcithoe* phylogeny that are problematic (Fig. 3). This network shows the two regions that are the primary cause of alternative trees. The first is that little resolution is seen regarding the positions of the four most recently diverged *Alcithoe* species (*A. lutea*, *A. larochei*, *A. fusus*, and *A. fissurata*). These taxa differ by between 0.7 and 6.9% in pairwise comparisons, and are responsible for the majority of the alternative topologies observed. The second is that there is a mixed signal in reference to the divergence of *A. wilsonae*. Two topologies have a similar quantity of signal: one in which *A. wilsonae* is derived from a lineage leading to the *Alcithoe benthicola* (Dell, 1963)/ *Alcithoe flemingi* Dell, 1978 clade; the other where *A. wilsonae* is independent of this clade. This inconsistency leads to the recovery of four (of a possible 105) topologies that differ in the placement of *A. wilsonae*, depending on the data subset and reconstruction method used. The two additional topological

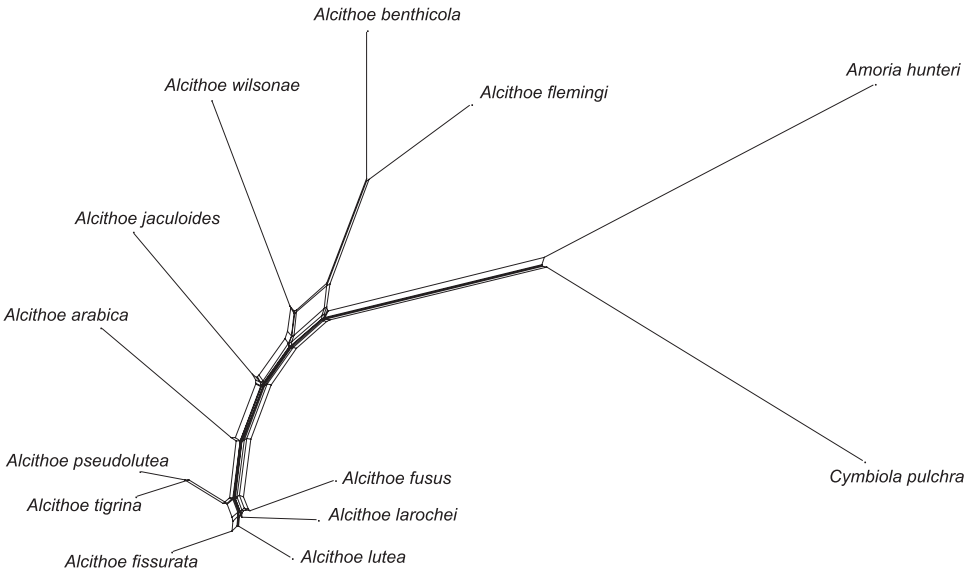

Figure 3. Spilts network of *Alcithoe* species based on the maximum sequence data set (7822 bp). Splits were generated using the Neighbor-Net algorithm in SPLITSTREE 4. Alternative phylogenetic relationships are apparent in the large box structure at the base of the *A. wilsonae* and *A. benthicola*/*A. flemingi* branches. The phylogenetic noise and short branch lengths separating six closely related and recently derived species is clear. This network visualizes an alternative signal, but only quantifies it in a general way and does not indicate the source of the conflicting signals in the data.

inconsistencies (monophyly of *A. arabica* and *A. jaculoides*, and the inconsistent placement of the *A. tigrina*/*A. pseudolutea* clade) are revealed to be relatively minor signals that are easily discarded when the whole data set is considered.

Spectral analysis

In order to gain a better understanding of the contribution of signal and noise from each of the gene partitions, we explored the phylogenetic information contained in the nucleotide data by visualizing support of taxa splits using networks and Lento plots for individual genes. These splits represent a summary of the total signal in an alignment, and are not generated assuming any given tree topology. As such they represent a description of the phylogenetic information in the data set that is independent of any reconstruction method or model of DNA evolution. A graph of the summed split support of the gene partitions illustrates the contribution of each of the genes to the total split support and conflict for the total concatenated data (Fig. 4). It is useful to compare the splits depicted in the graph with a reference tree, in this case the tree recovered from the complete data set (see the split key in Fig. 4).

The majority of the splits compatible with the complete data set tree exhibit significant support, in most cases with contributions from all genes. Many splits representing clades not present in the complete data set tree (incompatible splits) have very little support, large conflict, and tend to rely on signals from only a few genes. These splits are likely to result from homoplasy, and represent noise in the data. The most important splits that are incompatible with the complete data set reference tree describe the alternative topologies identified in the network (Fig. 3).

A single incompatible split refers to a taxon set grouping *A. wilsonae*, *A. benthicola* and *A. flemingi* separate from the remaining taxa. This split (× in the split key in Fig. 4) has considerable support, but also exhibits significant conflict. Support for this split is found in all genes except *nad2* and *12S*, but in all cases the quantities of conflict are much greater then the support values.

Several splits are associated with the topological uncertainty for the most recently derived *Alcithoe* depicted in the network (Fig. 3). Split H (see Fig. 4), which refers to a separation of the four most closely related taxa (*A. lutea*, *A. larochei*, *A. fusus*, and *A. fissurata*), and is compatible with the reference tree, is well supported, only lacking a contribution from *cox2* and the tRNA set. Although there is considerable conflict for this split, the majority of this conflicting signal comes from *atp8* only. Ignoring the signal from *atp8* in this case leads to a support/conflict ratio that favours the support of this split. In addition, splits that contradict split H all have much lower support, representing only a few genes, and generally have

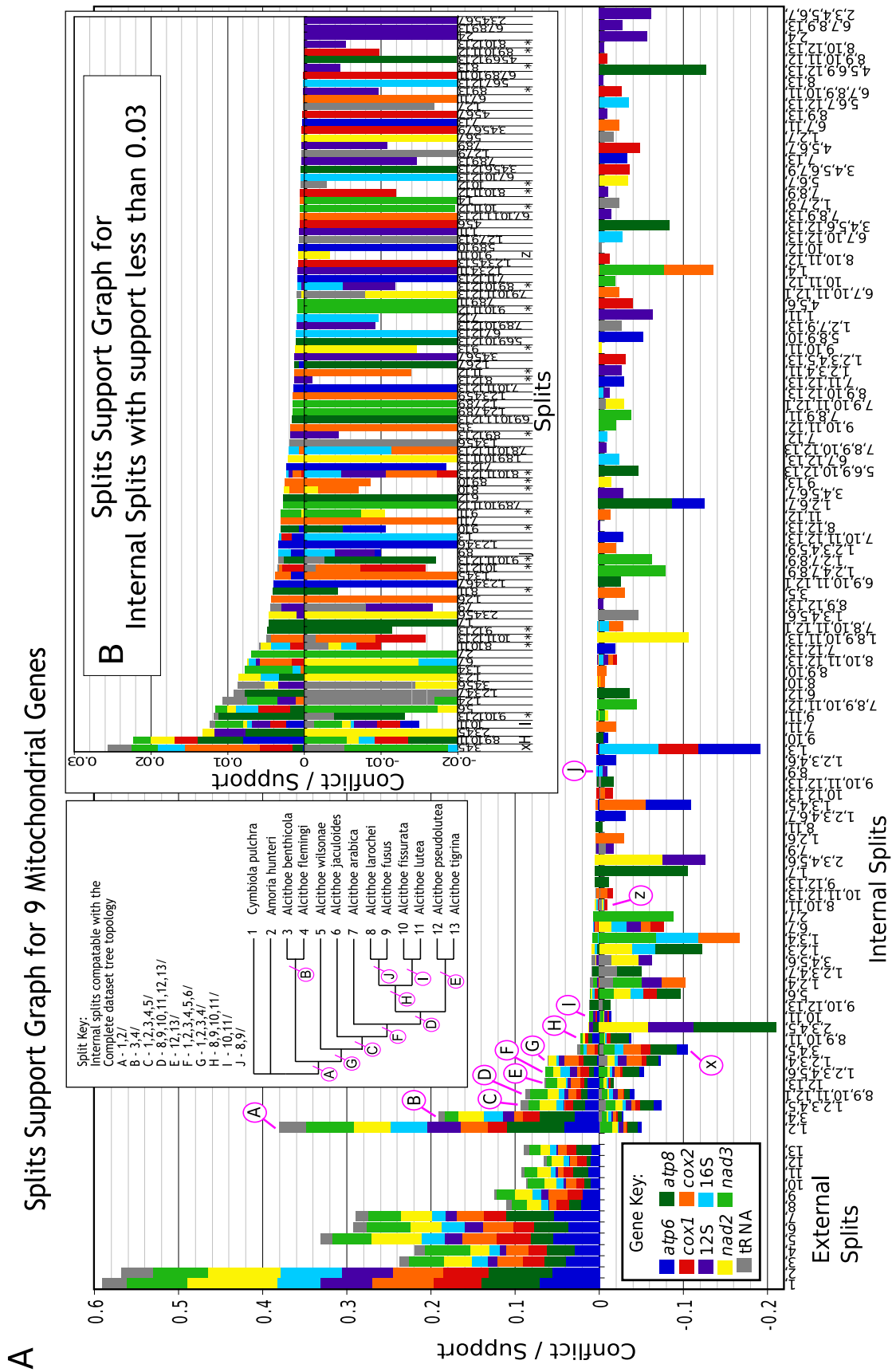
A
Splits Support Graph for 9 Mitochondrial Genes
Conflict / Support
External Splits
Internal Splits
Gene Key:
atp6
atp8
cox1
cox2
12S
16S
nad2
nad3
tRNA
Split Key:
Internal splits compatable with the Complete dataset tree topology
A - 1,2/
B - 3,4/
C - 1,2,3,4,5/
D - 8,9,10,11,12,13/
E - 12,13/
F - 1,2,3,4,5,6/
G - 1,2,3,4/
H - 8,9,10,11/
I - 10,11/
J - 8,9/
1 Cymbiola pulchra
2 Amoria hunteri
3 Alcithoe benthicola
4 Alcithoe ftemingi
5 Alcithoe wilsonae
6 Alcithoe jaculoides
7 Alcithoe arabica
8 Alcithoe larochei
9 Alcithoe fusus
10 Alcithoe fissurata
11 Alcithoe lutea
12 Alcithoe pseudolutea
13 Alcithoe tigrina
B
Splits Support Graph for
Internal Splits with support less than 0.03
Conflict / Support
Splits

Figure 4. Summed split support from partitioned data reveals the relative levels of signal and noise contained in each partition of the mitochondrial DNA sequences from *Alcithoe* species. A, support and conflict for all gene-based partitions of the data (Table 2) are graphed for the all splits. Columns represent the sum of support (above 0) or conflict (below 0) for splits derived from the alignments of each individual gene partition. Single taxa split from the others are external branches in the tree and have no conflict: these 13 splits are clustered on the left. Internal splits are ordered by their level of support in the data. In general, splits representing branches in the complete data set have high support and are clustered to the left. However, splits between recently diverged, closely related species are more ambiguous, often with low support and little conflict, and can be distributed further to the right of the graph. Splits are listed by the group of taxa represented on one side of the division. The number code for each taxon is given in the split key inset. Internal splits compatible with the complete data set tree are labelled with capital letters. Lower case letters (x, z) mark the splits representing conflicting topologies found for individual genes. B, an extracted section of the graph shows the splits with support of less than 0.03 in greater detail. *Splits representing signals relating to the six most recently diverged taxa that are inconsistent with the complete data set tree. Many of these splits exhibit a similar degree of support and conflict, indicating a paucity of discriminating signal.

large conflict. The relationships of the four most closely related taxa can potentially be resolved by considering all splits referring to subsets of these taxa (i.e. any subsets of the taxon sets 8, 9, 10, and 11 in Fig. 4). Of the ten possible combinations of these taxa, one (8, 9, and 11) is not observed in the data. Most of the remainder have very small support, similar or more conflict, and only have support from one or two genes, often *cox2* or *atp8*. The two exceptions to this are the splits I and z, for which support is observed in at least four genes. The general lack of resolution for these four species is unsurprising given the small levels of sequence difference between them (0.7–6.9%). Homoplasy between these derived taxa and separate lineages, particularly the out-groups, could be confounding the signal at this level.

A reduced taxon set clarifies six closely related species

In order to clarify the evolutionary relationships of the four closely related *Alcithoe* species, a taxon-reduced data set was created (*A. fusus*, *A. larochei*, *A. fissurata*, and *A. lutea*, plus *A. larochei tigrina*, *A. pseudolutea*, and *A. arabica* as out-group taxa) and the splits spectrum for the nucleotide data was generated (Fig. 5A). Using this pruned taxon set reduces the level of noise produced by homoplasy with more divergent species, this can be seen in the significant reduction in split conflict compared with Figure 4. In addition to the nine gene-based data partitions, a partition for the intergenic spacer regions was added, as these rapidly evolving regions could be useful in resolving closely related taxa. Low phylogenetic resolution is observed, but, as predicted, a better signal-to-noise ratio is achieved. This allows for a more clear interpretation of the split support for alternative topologies. This is important as three different topological solutions are consistently returned from three phylogenetic reconstruction methods: Bayesian (Fig. 5B), maximum likelihood (Fig. 5C), and neighbour joining (Fig. 5D).

As in the analysis of the complete taxon set, consideration of the number of genes that contribute signals to these low-resolution relationships is informative, in addition to the levels of support and conflict. The alternative tree topologies recovered in different reconstructions are defined by the four splits C, D, E, and F (letters refer to designations in the split key of Fig. 5A). The most well-supported of these, split D, only lacks contribution from *atp8* and *cox2*, and the majority of the conflict comes from the intergenic spacer. Split D appears consistently in all trees built from this reduced taxon data set, and indicates confidence in the monophyletic grouping of *A. fissurata* and *A. lutea*.

The number of genes contributing signal drops sharply for the remaining splits, which define the positions of the two species *A. fusus* and *A. larochei*. The maximum-likelihood tree (Fig. 5C) contains split C, which is supported by *atp6*, *cox1*, *16S*, and the intergenic spacer; however, the majority of the signal comes from the intergenic spacer. The Bayesian tree has split F, which is supported by only *atp8*, but also has significantly more conflict than support. The neighbour-joining tree includes split E, which has some support from *atp6*, *cox1*, *nad2*, and *16S*, but has slightly more conflict from the same genes. This reduced taxa data set highlights some gene-based problems not apparent in the complete data set. Inclusion of the intergenic spacer provided some additional support for some in-tree splits (e.g. split C), but also introduced considerable conflict for others (e.g. splits B and D). The splits support graph shows that *atp8* and *cox2* are problematic, providing little support and greater conflict for internal splits. Additionally, in this data set, *atp8* shows no support for the external branches for *A. pseudolutea* and *A. fusus*. This means that these two taxa are indistinguishable from other taxa (*A. pseudolutea* from *A tigrina* and

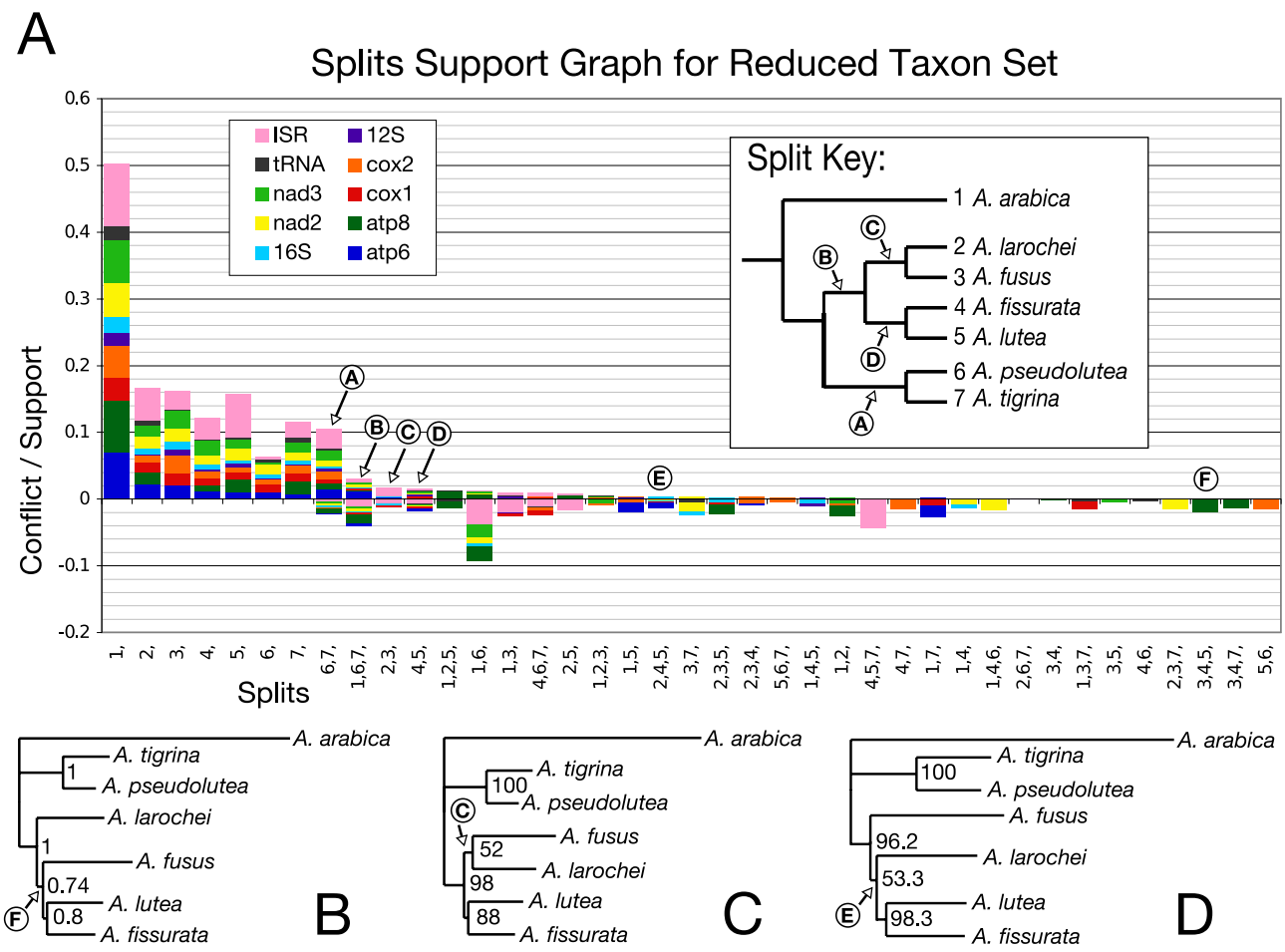

Figure 5. Refinement of phylogenetic inference for New Zealand *Alcithoe* species by consideration of a subtree only. A, support and conflict for splits generated from seven *Alcithoe* species are summarized in a splits support graph. Splits representing external branches are cluttered on the left and internal splits are then ordered by decreasing split support. In addition to the nine gene partitions an intergenic spacer region (ISR) sequence partition is included, as these non-coding regions are likely to be more informative at the level of the closely related taxa examined here. The splits key shows the taxa included and the expected splits (based on the complete data set tree): these splits are labelled on the graph. Three topological solutions for this data set are returned from different tree-building methods: (B) Bayesian inference; (C) maximum likelihood; (D) neighbor joining. The key splits that differentiate these topologies are indicated.

A. fusus from *A. larochei*), and that data from this gene leads to increased splits conflict.

Refinement of analysis

Taking into account the accumulated information about variability, compatibility, signal and noise now generated for this data set, an informed decision can be made as to the most appropriate genes to include to maximize the ratio of signal to noise for robust phylogenetic reconstruction of the *Alcithoe*. In the consideration of the splits data two genes, *atp8* and *cox2*, were consistently associated with conflicting signals, despite the partition homogeneity tests being non-significant. In addition, the level and distribution of sequence variation in *cox2* indicated that the pattern of DNA substitution in this gene is unusual in the context of the total data set, and thus the best substitution model for *cox2* is unlikely to be a good fit to the rest of the data. Also, *atp8* is the shortest (159 bp) but most variable of the nine genes, again with a different pattern of sequence variation compared with the other genes. As a result, *atp8* seems to introduce a high level of conflict in the data while not providing a large level of additional support for any splits. In order to reduce noise associated with these genes they were trimmed from the data set. Three additional gene sets, *nad3*, *12S* and the tRNA set, were considered to offer little information at this phylogenetic level, based on the splits data. However, these genes tended to lack resolution for recent

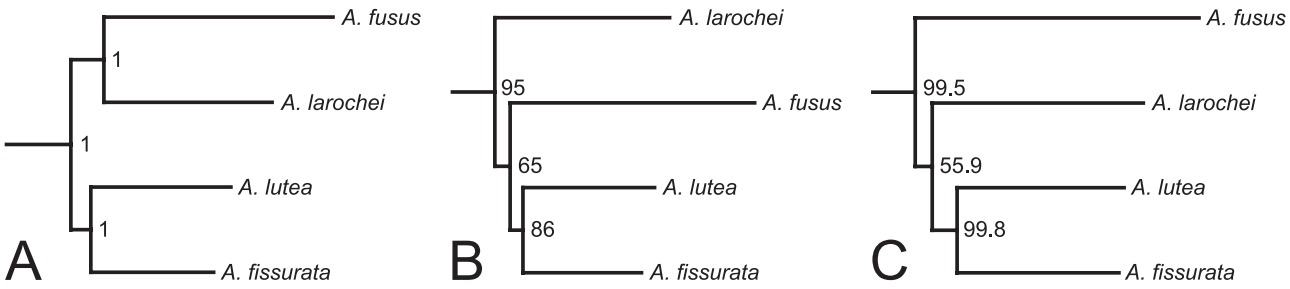

Figure 6. Alternative phylogenetic topologies in a quality-controlled data set show that the relationships of *Alcithoe fusus* and *Alcithoe larochei* cannot be resolved with these data. Phylogenetic reconstructions for the 'best gene' data set (*nad3*, *nad2*, *cox1*, *atp6*, *16S*, *12S*, and the tRNA set) returned from Bayesian (A), maximum-likelihood (B), and neighbour-joining (C) analyses illustrate that the different reconstruction methods interpret low-support, low-conflict splits in different ways. Only the four most derived taxa are shown, as the topology of the remainder of the tree is identical for all analyses.

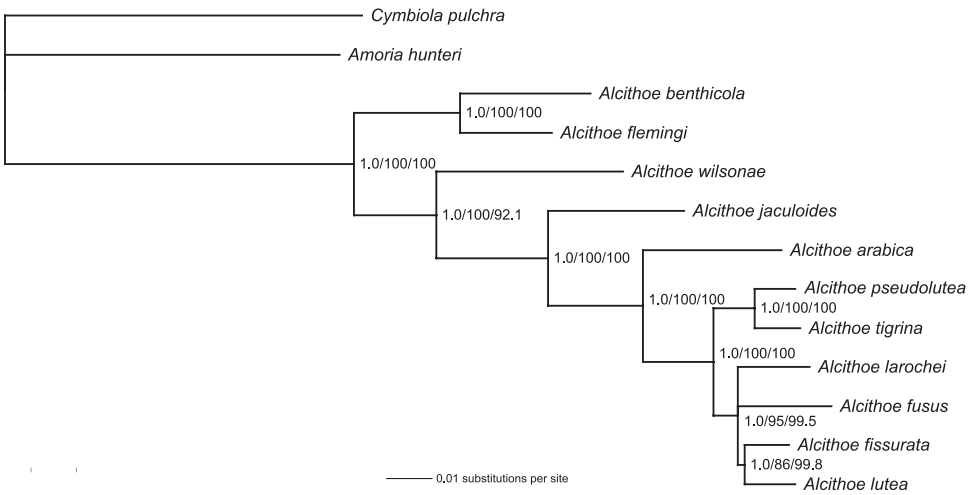

Figure 7. Molecular phylogeny of the gastropod genus *Alcithoe* based on a reduced gene data set with maximized signal and minimized noise. The data set is 6777 bp in length (including *nad3*, *nad2*, *cox1*, *atp6*, *16S*, *12S* and tRNA). Bayesian posterior probability/maximum-likelihood bootstrap support/neighbour-joining bootstrap support is given for each node. The unresolved relationship for *Alcithoe larochei* and *Alcithoe fusus* has been enforced on this phylogeny, although phylogenetic reconstruction methods cannot return a topology with an unresolved three-way polytomy.

divergences rather than generate significant conflict. Therefore, they were retained in the data set as they contain consequential signal for deeper relationships.

The spectrum of splits for the data set containing the retained seven gene partitions (*nad3*, *nad2*, *cox1*, *atp6*, *16S*, *12S*, and tRNAs) is similar to that for the complete data set, but conflict is reduced. However, trees generated from this reduced data set are not consistent under different phylogenetic reconstruction methods, with alternative topologies being found for the two closely related species, *A. fusus* and *A. larochei*. Bayesian analysis returns a sister relationship between *A. fusus* and *A. larochei* (Fig. 6A), whereas maximum likelihood returns a tree with *A. larochei* as the first of the four taxa to diverge (Fig. 6B), and neighbour joining returns a tree with *A. fusus* as the first taxon to diverge (Fig. 6C). The split information for this data indicates little signal for any of the possible topologies (Fig. 5). SH tests of four possible topologies of *A. fusus* and *A. larochei* (unresolved, sister, *A. fusus* diverging first, and *A. larochei* diverging first) show that these phylogenetic solutions are equally good explanations of the data. Our results indicate that the branching order of these two taxa is sensitive to the tree-building method used. Given that the alternative topologies of these taxa are equally likely, the correct hypothesis of the phylogeny, based on this data, is a three-way polytomy of *A. fusus*, *A. larochei*, and the *A. fissurata*/*A. lutea* clade (Fig. 7), even though no phylogenetic reconstruction method independently returns this result.

DISCUSSION

Exploration of sequence data

Eleven gene partitions returned a range of DNA substitution models (from the six-parameter HKY+I+G model to the ten-parameter GTR+I+G model), reflecting different complexities in the substitution dynamics of the different genes. Individually each gene was too short for robust phylogenetic reconstruction. In combination, partition homogeneity tests and variability data only highlighted one gene as potentially problematic for phylogenetic analysis (*cox2*), but the partition homogeneity test has been criticized as an inadequate measure of the combinability of data sets (Barker & Lutzoni, 2002).

Assessment of phylogenetic signal with splits

Splits data allow the assessment of phylogenetic signals that may not be apparent when only trees are considered. Using splits, the source of conflict leading to low bootstrap or Bayesian support values can be examined in detail, and the validity of high support values for reconstructed nodes can be tested (e.g. Esser *et al*., 2004; Wagele *et al*., 2009). Analysis of splits identified conflicting signal for some deeper relationships for the *Alcithoe*, despite high node support. Furthermore, consideration of the splits data allowed the determination of a polyphyletic relationship of four closely related species (*A. fusus*, *A. larochei*, *A. fissurata*, and *A. lutea*), and implies that phylogenetic reconstruction methods handle splits with low support and low conflict in different ways, leading to alternative tree topologies. In our examination of splits data, we have built on previous uses of spectral analysis by considering the contribution of each individual data partition (in this case genes) to the total signal. In this way we have identified two genes, *cox2* and *atp8*, that contain anomalous signal when compared with the majority. Given the departure from the consensus signal we assume that these genes reflect signals peculiar to their particular constraints and mutational model, rather than the evolutionary history of the species. Based on this assumption we justify the removal of these partitions from the data set in order to reduce phylogenetic noise. This finding is consistent with studies that have shown *cox2* and *atp8* to have limited phylogenetic utility in other taxonomic groups (Zardoya & Meyer, 1996; Corneli & Ward, 2000; Mueller, 2006). The methodological approach of this examination of phylogenetic signal in a sequence data set is similar to the deep evolutionary analysis of mitochondrial origins by Esser *et al*. (2004). However, our study differs significantly in the evolutionary timescale under consideration. This illustrates that the consideration of splits is informative at different levels of evolutionary divergence, and shows that splits analysis has significant utility as a tool for phylogenetic data exploration.

Taxon subsets reduce noise resulting from homoplasy

We recommend analysing subsections of taxa independently to resolve relationships among closely related species and populations, thus reducing noise from homoplasy. By analysing taxa that differ in their degree of divergence separately, models of DNA substitution should more realistically describe DNA evolution within each data set. Relationships among closely related taxa can then be constrained in larger scale analyses. This will prevent these relationships from being disrupted by the addition of more divergent taxa (and accompanying homoplasy) and poorly fitting models of DNA evolution.

Marker selectivity

Our refined data set with fewer genes was not consistent under different phylogenetic reconstruction methods, but generated trees that differed only in the placement of two of the most closely related taxa (*A. larochei* and *A. fusus*). Unlike the complete data set, the sensitivity to the model of DNA evolution is, we think, a more biologically accurate result. This is a better data set because we know from separate analysis of recently derived taxa that relationships can be distorted by homoplasy, and by an inappropriate model of DNA evolution. Thus our refined set of genes provided us with a data set that does not hide the very real difficulty in accurately estimating a phylogeny of this group of volutes. Furthermore, awareness of the specific source of this uncertainty in the data will allow for it to be accounted for in down-stream applications, such as molecular-clock analysis.

To date the majority of phylogenetic analyses with molluscan species have included only a few genes. Recent studies of other taxonomic groups have shown that it is of great value to perform multi-gene analyses in order to account for idiosyncrasies in individual genes that might otherwise mislead the phylogeny. We go further to suggest that exclusion of genes that can be clearly shown to have anomalous signals or adhere to disparate substitution models is desirable in order to increase the robustness of final evolutionary hypotheses. Ultimately it is likely that the consideration of multi-scale analysis will be appropriate, particularly for large taxon sets. In such an analysis each gene would only be considered up to the level at which the signal-to-noise ratio for that gene remained acceptable. For example, some genes in a data set

may be informative at the species level but not at the genus level, and the decision as to which level they are informative will be based on the signal-to-noise ratio. Additionally, such genes could be used to resolve some subtrees but not others, based on their information content.

For the analysis of gastropod taxa separated by between approximately 1 and 50 Myr, and not overly species dense, we recommend the combination of the genes in our reduced set: *nad2*, *cox1*, *atp6*, and *16S*. The genes *nad3* and *12S* are also suitable, but *nad3* is preferable for a more shallow divergence and *12S* is preferable for a deeper divergence. In neogastropod molluscs an appropriate continuous mitochondrial fragment to target would be an approximately 3-kb section spanning the genes *nad3*, *nad2*, and *cox1* (see Table S1).

Finally, we suggest an analysis pipeline based on the work described here. Once suitable sequence alignments have been constructed summary statistics can be easily generated that will give an initial measure of the distribution and a rough idea of the pattern of signals in the data (such as those shown in Table 2). We then recommend network-based analysis. Early consideration of the network structure of the data allows the general assessment of the phylogenetic signal-to-noise ratio, which will then inform how subsequent analysis should proceed. In addition, this approach does not assume a tree-like structure in the data, and will therefore show when the signal is not tree-like and should not be treated with standard phylogenetic methods. Where significant or particularly interesting conflict is observed, an analysis of the splits spectra can then be carried out to further examine these signals. Splits spectra can also be used to assess the prevalence of various phylogenetic errors in the data (Wagele *et al*., 2009). With results from the previous steps one will have the information to critically assess the resulting phylogenetic reconstruction. Such an assessment can, for instance, determine the credibility of node support values and diagnose elements of the data that are prone to misrepresentation in bifurcating trees. If, after these steps there is still inconsistency in the phylogenetic inference, the source of inconsistency can be accounted for in subsequent analyses that might be sensitive to inconsistent phylogenetic signal.

ALCITHOE SYSTEMATICS

In the selection of out-group taxa for this analysis it was clear that *A. aillaudorum* from New Caledonia is not closely related to the New Zealand *Alcithoe*. If the history of the mitochondria of these volutes is representative of the species history, then the genus *Alcithoe* is not monophyletic and we suggest the current placement of the New Caledonian species *A. aillaudorum* should be re-examined.

Phylogenetic analysis of nine mitochondrial genes from 13 volute taxa resolved a stable evolutionary hypothesis for the group. Within the New Zealand *Alcithoe* there is one major point of difference in the assignment of species between our data and the work of Bail & Limpus (2005). Bail and Limpus treat *A. tigrina* as a subspecies of *A. larochei* based on shell morphology. This molecular data set, however, supports the clear separation of these two species. Our molecular data support the close morphological relationship of *A. larochei*, *A. lutea*, and *A. pseudolutea* recognized by Bail & Limpus (2005). The possibility of a common origin of *A. fusus* and *A. jaculoides* suggested by Bail and Limpus can be discarded, as the molecular data clearly shows that these two species are not closely related. Indeed, the close relationship of *A. fusus*, *A. larochei*, *A. fissurata*, and *A. lutea* is novel and somewhat unexpected. However, this is consistent with fossil evidence for *A. fusus* and *A. larochei*, which indicates a relatively recent origin of both these species around 1.6 Ma (Beu & Maxwell, 1990).

CONCLUSIONS

Although this study focuses on a specific molluscan example, it serves as a general demonstration of the utility of careful examination of sequence data before, or in parallel with, phylogenetic reconstruction. We show that consideration of networks will highlight potential inconsistencies in the underlying sequence data, and that analysis of splits data will help to diagnose the causes of these signals. Awareness of the value of such analysis is growing (e.g. White *et al*., 2007; Morrison, 2010), and a few recent studies serve as important demonstrations of the benefits of exploratory data analysis (e.g. Roberts, Sargis & Olson, 2009; Wagele *et al*., 2009; Dabert *et al*., 2010). Our work illustrates the value of exploratory data analysis even in relatively consistent data sets, particularly where more complex downstream analyses are planned. For example, if one plans to estimate divergence times from a molecular phylogeny, more accurate estimates will be obtained when topological anomalies arising from the underlying sequence data are accounted for.

It is important to note the phylogenetic utility of a marker is dependent on the distance of evolutionary relatedness and taxonomic context. Additionally, it is likely to be beneficial to select genes that can be modelled by the same or similar substitution models. Until such time as more realistic models of DNA substitution are available, and particularly in the current environment of ever larger nucleotide data sets, it makes sense to identify and analyse sets of

genes that better obey the conditions of an existing model, rather then attempting to fit a less well suited model or over parameterize by using multiple models.

ACKNOWLEDGEMENTS

This work was funded as part of the Marsden Fund contract 04 GNS 021, administered by the Royal Society of New Zealand. Additional support was given by the Allan Wilson Centre for Molecular Ecology and Evolution. We appreciatively thank our colleagues James Crampton, Alan Beu, and Bruce Marshall, whose knowledge of the palaeontological history and morphological taxonomy of the Volutidae was invaluable. We gratefully acknowledge those who have provided us with specimens for this study, particularly Bruce Marshall at the Museum of New Zealand, Te Papa Tongarewa. Also Sarah Samadi at the Muséum national d'Histoire naturelle, Paris, for samples of *A. aillaudorum*, and Margaret Richards for collecting our *A. arabica* specimen. Thanks also to Barbara Holland and Klaus Schliep for helpful discussion regarding the use of splits-based analysis, and to David Penny and two anonymous reviewers for comments on the article.

REFERENCES

Bail P, Limpus A, eds. 2005. *The recent volutes of New Zealand with a revision of the genus Alcithoe H. & A. Adams, 1853*. Hackenheim: ConchBooks.

Bandelt HJ, Dress AWM. 1992. A canonical decomposition-theory for metrics on a finite-set. *Advances in Mathematics* **92:** 47–105.

Bandyopadhyay PK, Stevenson BJ, Cady MT, Olivera BM, Wolstenholme DR. 2006. Complete mitochondrial DNA sequence of a Conoidean gastropod, *Lophiotoma* (*Xenuroturris*) *cerithiformis*: gene order and gastropod phylogeny. *Toxicon* **48:** 29–43.

Barker FK, Lutzoni FM. 2002. The utility of the incongruence length difference test. *Systematic Biology* **51:** 625–637.

Beu AG, Maxwell PA. 1990. *Cenozoic mollusca of New Zealand*. Lower Hutt: DSIR.

Bryant D, Moulton V. 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution* **21:** 255–265.

Cannone JJ, Subramanian S, Schnare MN, Collett JR, D'Souza LM, Du YS, Feng B, Lin N, Madabusi LV, Muller KM, Pande N, Shang ZD, Yu N, Gutell RR. 2002. The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *Bmc Bioinformatics* **3:** 2.

Corneli PS, Ward RH. 2000. Mitochondrial genes and mammalian phylogenies: increasing the reliability of branch length estimation. *Molecular Biology and Evolution* **17:** 224–234.

Cummings MP, Otto SP, Wakeley J. 1995. Sampling properties of DNA-sequence data in phylogenetic analysis. *Molecular Biology and Evolution* **12:** 814–822.

Cunha RL, Grande C, Zardoya R. 2009. Neogastropod phylogenetic relationships based on entire mitochondrial genomes. *Bmc Evolutionary Biology* **9:** 210.

Dabert M, Witalinski W, Kazmierski A, Olszanowski Z, Dabert J. 2010. Molecular phylogeny of acariform mites (Acari, Arachnida): strong conflict between phylogenetic signal and long-branch attraction artifacts. *Molecular Phylogenetics and Evolution* **56:** 222–241.

Drummond AJ, Cheung M, Heled J, Kearse M, Moir R, Stones-Havas S, Thierer T, Wilson A. 2007. Geneious v3.8. v3.8 ed. Available from http://www.geneious.com/

Esser C, Ahmadinejad N, Wiegand C, Rotte C, Sebastiani F, Gelius-Dietrich G, Henze K, Kretschmann E, Richly E, Leister D, Bryant D, Steel MA, Lockhart PJ, Penny D, Martin W. 2004. A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Molecular Biology and Evolution* **21:** 1643–1660.

Graybeal A. 1994. Evaluating the phylogenetic utility of genes – a search for genes informative about deep divergences among vertebrates. *Systematic Biology* **43:** 174–193.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* **52:** 696–704.

Ho SYW, Phillips MJ. 2009. Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. *Systematic Biology* **58:** 367–380.

Holland BR, Huber KT, Moulton V, Lockhart PJ. 2004. Using consensus networks to visualize contradictory evidence for species phylogeny. *Molecular Biology and Evolution* **21:** 1459–1461.

Huber KT, Langton M, Penny D, Moulton V, Hendy M. 2002. Spectronet: a package for computing spectra and median networks. *Applied Bioinformatics* **1:** 159–161.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics (Oxford, England)* **17:** 754–755.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* **23:** 254–267.

Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? *Trends in Genetics* **22:** 225–231.

Kusukawa N, Uemori T, Asada K, Kato I. 1990. DNA sequencing report – rapid and reliable protocol for direct sequencing of material amplified by the polymerase chain-reaction. *Biotechniques* **9:** 66.

Lento GM, Hickson RE, Chambers GK, Penny D. 1995. Use of spectral-analysis to test hypotheses on the origin of pinnipeds. *Molecular Biology and Evolution* **12:** 28–52.

Lydeard C, Holznagel WE, Schnare MN, Gutell RR. 2000. Phylogenetic analysis of molluscan mitochondrial LSU rDNA sequences and secondary structures. *Molecular Phylogenetics and Evolution* **15:** 83–102.

McComish BJ, Hills SFK, Biggs PJ, Penny D. 2010. Index-free de novo assembly and deconvolution of mixed mitochondrial genomes. *Genome Biology and Evolution* **2:** 410–424.

Morrison DA. 2010. Using data-display networks for exploratory data analysis in phylogenetic studies. *Molecular Biology and Evolution* **27:** 1044–1057.

Mueller RL. 2006. Evolutionary rates, divergence dates, and the performance of mitochondrial genes in Bayesian phylogenetic analysis. *Systematic Biology* **55:** 289–300.

Non AL, Kitchen A, Mulligan CJ. 2007. Identification of the most informative regions of the mitochondrial genome for phylogenetic and coalescent analyses. *Molecular Phylogenetics and Evolution* **44:** 1164–1171.

Norman J, Olsen P, Christidis L. 1998. Molecular genetics confirms taxonomic affinities of the endangered Norfolk Island Boobook Owl *Ninox novaeseelandiae undulata*. *Biological Conservation* **86:** 33–36.

Paton TA, Baker AJ. 2006. Sequences from 14 mitochondrial genes provide a well-supported phylogeny of the Charadriiform birds congruent with the nuclear RAG-1 tree. *Molecular Phylogenetics and Evolution* **39:** 657–667.

Phillips MJ, Delsuc F, Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. *Molecular Biology and Evolution* **21:** 1455–1458.

Posada D, Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics (Oxford, England)* **14:** 817–818.

Rambaut A. 2002. Se-Al: sequence alignment editor. Available at http://tree.bio.ed.ac.uk/software/seal/

Roberts TE, Sargis EJ, Olson LE. 2009. Networks, trees, and treeshrews: assessing support and identifying conflict with multiple loci and a problematic root. *Systematic Biology* **58:** 257–270.

Roe AD, Sperling FAH. 2007. Patterns of evolution of mitochondrial cytochrome c oxidase I and II DNA and implications for DNA barcoding. *Molecular Phylogenetics and Evolution* **44:** 325–345.

Simison WB, Lindberg DR, Boore JL. 2006. Rolling circle amplification of metazoan mitochondrial genomes. *Molecular Phylogenetics and Evolution* **39:** 562–567.

Simmons MP, Pickett KM, Miya M. 2004. How meaningful are Bayesian support values? *Molecular Biology and Evolution* **21:** 188–199.

Simon C, Frati F, Beckenbach A, Crespi B, Liu H, Flook P. 1994. Evolution, weighting, and phylogenetic utility of mitochondrial gene-sequences and a compilation of conserved polymerase chain-reaction primers. *Annals of the Entomological Society of America* **87:** 651–701.

Swofford DL. 1998. *PAUP*4.0- phylogentic analysis using parsimony (*and other methods)*. Sunderland, MA: Sinauer Associates Inc.

Wagele JW, Mayer C. 2007. Visualizing differences in phylogenetic information content of alignments and distinction of three classes of long-branch effects. *Bmc Evolutionary Biology* **7:** 147.

Wagele JW, Letsch H, Klussmann-Kolb A, Mayer C, Misof B, Wagele H. 2009. Phylogenetic support values are not necessarily informative: the case of the Serialia hypothesis (a mollusk phylogeny). *Frontiers in Zoology* **6:** 12.

White WT, Hills SF, Gaddam R, Holland BR, Penny D. 2007. Treeness triangles: visualizing the loss of phylogenetic signal. *Molecular Biology and Evolution* **24:** 2029–2039.

Zardoya R, Meyer A. 1996. Phylogenetic performance of mitochondrial protein-coding genes in resolving relationships among vertebrates. *Molecular Biology and Evolution* **13:** 933–942.

Zuker M, Mathews DH, Turner DH, eds. 1999. *Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide*. Dordrecht: Kluwer Academic Publishers.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Table S1. Primers used to amplify mtDNA of volute gastropods.
Table S2. Summary of sequenced DNA fragments from 13 marine molluscs of the family Volutidae.
Appendix S1. 1) Modified high Salt DNA extraction method; 2) PCR protocols.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.