



# Mitochondrial genome data alone are not enough to unambiguously resolve the relationships of Entognatha, Insecta and Crustacea *sensu lato* (Arthropoda)

Stephen L. Cameron<sup>1,\*</sup>, Kelly B. Miller<sup>1</sup>, Cyrille A. D'Haese<sup>3</sup>, Michael F. Whiting<sup>1</sup>  
and Stephen C. Barker<sup>2</sup>

<sup>1</sup>Department of Integrative Biology, Brigham Young University, Provo, UT 84602, USA; <sup>2</sup>Department of Microbiology and Parasitology, and Institute for Molecular Biosciences, The University of Queensland, Brisbane, 4072, Australia; <sup>3</sup>FRE 2695 CNRS, Département Systématique et Evolution, Muséum National d'Histoire Naturelle, 45 rue Buffon, Paris, 75005, France

Accepted 27 October 2004

---

## Abstract

An analysis of the relationships of the major arthropod groups was undertaken using mitochondrial genome data to examine the hypotheses that Hexapoda is polyphyletic and that Collembola is more closely related to branchiopod crustaceans than insects. We sought to examine the sensitivity of this relationship to outgroup choice, data treatment, gene choice and optimality criteria used in the phylogenetic analysis of mitochondrial genome data. Additionally we sequenced the mitochondrial genome of an archaeognathan, *Nesomachilis australica*, to improve taxon selection in the apterygote insects, a group poorly represented in previous mitochondrial phylogenies. The sister group of the Collembola was rarely resolved in our analyses with a significant level of support. The use of different outgroups (myriapods, nematodes, or annelids + mollusks) resulted in many different placements of Collembola. The way in which the dataset was coded for analysis (DNA, DNA with the exclusion of third codon position and as amino acids) also had marked effects on tree topology. We found that nodal support was spread evenly throughout the 13 mitochondrial genes and the exclusion of genes resulted in significantly less resolution in the inferred trees. Optimality criteria had a much lesser effect on topology than the preceding factors; parsimony and Bayesian trees for a given data set and treatment were quite similar. We therefore conclude that the relationships of the extant arthropod groups as inferred by mitochondrial genomes are highly vulnerable to outgroup choice, data treatment and gene choice, and no consistent alternative hypothesis of Collembola's relationships is supported. Pending the resolution of these identified problems with the application of mitogenomic data to basal arthropod relationships, it is difficult to justify the rejection of hexapod monophyly, which is well supported on morphological grounds.

© The Willi Hennig Society 2004.

---

Mitochondrial (mt) genomes are seeing wider use as phylogenetic markers in studies aiming to resolve the relationships of distantly related groups (Saccone et al., 1999). Many of these studies have, however, produced results which are difficult to reconcile with trees produced using other markers. For example, in mammal

phylogeny, mitochondrial analyses suggest (monotremes + marsupials) + eutherians (Janke et al., 1996, 1997, 2002), whereas most nuclear and morphological/physiological studies support (marsupials + eutherians) + monotremes (Lee et al., 1999; Lou et al., 2002). These studies have highlighted the need for sophisticated analysis of the different groups of signals found within the mitochondrial genome (e.g., protein coding versus ribosomal genes, first and second versus third codon position and DNA versus amino acid sequence data) and the effect these differing data sets/treatments may

---

\*Correspondence: Department of Integrative Biology, Brigham Young University, Provo, UT 84602, USA.  
Tel.: +1 801 422 1396; Fax: +1 801 422 0090.  
E-mail address: slc236@email.byu.edu

have on resulting phylogenetic hypotheses. A recent example of an analysis using mitochondrial genomes for phylogenetics that produced a result which was at odds with previous morphological results is by Nardi et al. (2003a). They found that collembolans, or springtails, are not closely related to the insects, as has been traditionally accepted, but are instead closely related to one of the crustacean groups, the Branchiopoda. Whilst this idea of two independent invasions of land by six-legged arthropods is not new and has previously been found using nuclear gene analyses (Giribet et al., 1996; Spears and Abele, 1997; Giribet and Ribera, 1998, 2000; Giribet and Wheeler, 1999; Giribet, 2002) and mitochondrial data (Nardi et al. 2001<sup>1</sup>), it runs contrary to the traditional classification of arthropods where collembolans are part of a paraphyletic grade, the “Entognatha”, which is sister to the Insecta (Wheeler et al., 2001). Several studies using nuclear ribosomal genes (mostly *ssu-rRNA* genes) have reported sister group relationships between Collembola and crustacean groups. These studies vary significantly as to which crustacean group is sister to Collembola: Cirripedia + Decapoda (Giribet et al., 1996), Branchiopoda (Spears and Abele, 1997); (Maxillopoda + Pentostomida) + Ectognatha (Giribet and Ribera, 1998; Giribet and Wheeler, 1999) and a variety of groups depending upon different parameter sets used for nucleotide alignment including Myodocopa + Pentostomida, Branchiura, Copepoda, Podocopa, Cirripedia or the entirety of Pancrustacea (Giribet and Ribera, 2000; Giribet, 2002). Mitochondrial data favors the grouping of Collembola with Branchiopoda (Nardi et al., 2001, 2003a,b). So whilst there are many studies challenging the monophyly of the Hexapoda, there is also a lack of consensus about which crustacean lineage they should be grouped.

Conversely there is also molecular support for the traditional placement of collembolans within Hexapoda. Edgecombe et al. (2000) found in a combined analysis of morphology and the genes Histone H3 and small nuclear RNAase U2 that collembolans group with the other entognathous classes Protura and Diplura in a paraphyletic clade at the base of the insects. Morphological synapomorphies which group Collembola + Protura to Insecta + Diplura included blastodermal cuticle present, primary pigment cells of the ommatidium with pigment granules, eyes of the mandibulate type (two corneagenous cells, four Semper cells, cone with four parts, retinula with eight cells), the gene *Distal-less* not expressed in the mandible in any ontological stage, posterior tentorial apodemes with tentorial

arms, opening of the maxillary salivary glands a median opening at the base of the second maxillae, maxillary plate present, Mx1 divided into cardo, stipes, lacinia and galeae, thorax with three limb bearing segments, coxal vesicles present on numerous trunk segments, patella/tibia joint fused and the embryonic gonoduct arises in association with the splanchnic mesoderm (Edgecombe et al., 2000).

Nardi et al.’s (2003a) finding was challenged by Delsuc et al. (2003) on the methodological grounds that if the data were transformed in a different way (conversion of the primary nucleotide sequence to purine–pyrimidine (R–Y) coding rather than conversion to amino acid coding), a tree more in line with traditional classifications is recovered. Specifically, collembolans resolve as sister to the insects, and, thus, Hexapoda is monophyletic, albeit with a fairly low Bayesian posterior probability (0.57). The authors have previously used this method to reconcile mammal phylogenies based on mitochondrial genome sequences with those based on nuclear sequences (Phillips and Penny, 2003). Nardi et al. (2003b) responded to Delsuc et al. (2003) by arguing against the validity of R–Y coding and modeling the likelihoods of competing topologies. Nardi et al. (2003b) preferred amino acids to R–Y coding because amino acids have a greater number of potential states (20 versus 2), are more conservative regardless of mutation saturation, and are a readily accepted method of phylogenetic reconstruction (for supporting references see Nardi et al., 2003b). Likelihood analyses also showed a consistent preference for their topology over the alternative advanced by Delsuc et al. (2003), and thus they felt confident in their initial hypothesis that hexapods are polyphyletic.

However, Delsuc et al. (2003) only examined a few of the problems associated with the analytical methodology used by Nardi et al. (2003a). Specifically we think five areas of this study need to be examined to determine their possible effects on the phylogenetic outcome.

### 1. Outgroup selection

Nardi et al. (2003a,b) used the majority of the mitochondrial genomes available for arthropods at the time and employed as outgroup taxa representatives of the Mollusca and Annelida because of the hypothesis that they collectively represent the sister-group to the Arthropoda. A diversity of opinions about the sister group of the arthropods exists, but few molecular studies of arthropod phylogeny have seriously addressed the issue of outgroup selection and its possible effect on recovered phylogenies (Giribet and Ribera, 1998). Early studies placed the arthropods in the protostomes (Field et al., 1988; Lake, 1990) but more recent work has included the arthropods in the “Ecdysozoa”, which share the characteristic of shedding

<sup>1</sup>It should be noted that this study used almost identical taxon sampling and phylogenetic methods as their 2003 study and so cannot really be considered independent confirmation. Similarly the Giribet papers have considerable taxic overlap and all use the same gene, the small ribosomal subunit (18S).

their outer body wall in a hormonally controlled periodic event, ecdysis, and includes the nematodes, tardigrades and onychophorans (Aguinaldo et al., 1997). We sought to evaluate the effect of outgroup choice on the analysis by also using nematodes, the only member of the ecdysozoa for which the mt genome sequence is currently available.

## 2. Ingroup taxon selection and alignment methodology

Inclusion of distant outgroups can lead to highly ambiguous alignment, i.e., a preponderance of gaps: the random outgroup effect noted by Wheeler (1990). This effect was further compounded in the Nardi et al. (2003a) study by the inclusion of highly divergent ingroup taxa such as the louse *Heterodoxus* and the bee *Apis*. This is entirely needless as none of these taxa bears on the question of the apterygote sister group but still introduced many gaps into the alignment. The next step taken in Nardi et al.'s (2003a) alignment procedure was to remove all sites where gaps occur prior to phylogenetic analysis. Given that phylogenies were then inferred using likelihood and Bayesian methods which automatically exclude sites with gaps this step was redundant, but the effect of excluding these sites on the phylogeny is non-trivial. One can readily imagine situations where the inclusion of a highly divergent taxon would introduce gaps into an alignment which otherwise would not have any. Furthermore, at least some of these sites may be informative, regardless of your preferred optimality criterion, for the remaining taxa. Thus the inclusion of taxa, which should have no bearing on the relationships of interest, can have a negative effect on phylogenetic reconstruction by imposing gaps which result in the loss of informative sites. The taxon selection used by Nardi et al. (2003a) may therefore have an effect on the phylogenetic reconstruction, even after those taxa have been removed as was done in their second, reduced taxa tree.<sup>2</sup>

## 3. Amino acid sequence analyses

Many mitochondrial genome phylogenies have translated sequence data into amino acid data and performed phylogenetic analyses on these data rather than the nucleotide sequence data as collected. The advantages of this approach are simply that amino acids are much

easier to align in these datasets than the constituent nucleotide sequence data. The disadvantage of amino acids is that they are subject to homoplasy, which does not occur when DNA sequence data are used directly, due to the degeneracy of the genetic code, i.e., multiple triplets coding for the same amino acid (Simmons and Freudenstein, 2002; Simmons et al., 2002). Therefore, the possibility of homoplasy in the amino-acid dataset producing a different topology from one produced by DNA sequence datasets needs exploration.

## 4. Choice of genes and their contribution to a total evidence tree

Contrary to suggestions in their published paper, Nardi et al. (2003a) did not use the entire mitochondrial genome in their analyses. Instead they used just four genes: *cox1*, *cox2*, *cox3* and *cytB*, chosen on the basis of “low gaps, high invariant sites”. As outlined above, the proportion of gaps in their analysis was a direct result of the inclusion of highly divergent taxa which, despite not being of interest in the study, probably contributed to the character ambiguities of some genes and therefore their exclusion from the analysis. Furthermore, there is no independent analysis of the individual genes (either those included or excluded) to determine whether they produce different topologies or what their contribution might be to a total evidence tree, such as using partitioned Bremer (Baker and DeSalle, 1997) or partitioned likelihood (Lee and Hugall, 2003) supports. Even if one wishes to exclude particular genes, at a minimum each gene within the mitochondrial genome must be evaluated thoroughly for its contribution to the phylogeny of the arthropods, and areas of agreement and conflict determined before data can be excluded as unreliable.

## 5. Analytical techniques

Nardi et al. (2003a,b) evaluated the phylogeny of the arthropods using two very similar model-based approaches to phylogenetic inference, likelihood (in ProtML) and Bayesian analysis (in MrBayes). There has been no investigation of the influence that the choice of optimality criteria may have on the recovered topology, and it seems odd that parsimony was entirely overlooked. Whilst the debates about the relative merits of optimality criteria are almost endless, it has been clearly shown that there are strengths and weaknesses to each approach and comparison of results from different methods is valuable in determining whether data are sensitive to analytical method. Data that behave differently under different optimality criteria may not be the most reliable data to use for the question at hand.

Given these concerns we undertook a reanalysis of the phylogeny of arthropods using mitochondrial genome sequences to see if the conclusion of

<sup>2</sup>The preceding argument applies just as thoroughly to the issue of taxon selection in all studies using likelihood or Bayesian analysis or parsimony analyses where the workers are in the habit of excluding gap sites before final analyses. Highly divergent taxa, which introduce large numbers of gaps into an alignment, should probably be either pre-excluded if not pertinent to the question under study or at least the effect of their presence/absence on tree topology evaluated as part of the study by taxon jackknifing.

Nardi et al. (2003a,b)—that hexapods are not monophyletic—is robust to the composition of the taxa used and to outgroup choice, alignment methodology, type of dataset analyzed (amino acids versus DNA), the genes used in the phylogenetic reconstruction and the types of analytical and support statistics used in the analysis. Furthermore, to improve the data we also sequenced the mitochondrial genome of the archaeognathan, *Nesomachilis australica*, to determine whether it is possible that the reported relationship between the collembolans and brachiopods is simply a long branch effect that can be mediated by the inclusion of more basal insect taxa than the zygentoman, *Tricholepidion*, the sister group to pterygote insects (Edgecombe et al., 2000).

## Materials and methods

### Mitochondrial genome sequencing

Specimens of the archaeognathan, *Nesomachilis australica*, Tillyard, 1924 were collected from Palms National Park (near Kingaroy), Queensland, Australia, by Michael Rix (UQ), snap-frozen in liquid nitrogen and stored at  $-80^{\circ}\text{C}$  in the insect tissue collection of the Department of Integrative Biology, Brigham Young University. Identification was by S.L.C., following the method of Sturm (1980); a voucher specimen has been deposited in the Bean Museum, Brigham Young University reference collection, accession no. IGC-AR04.

Whole genomic DNA was extracted from muscle tissue with a DNeasy Tissue kit (Qiagen). Short regions of the *12S* and *16S* genes were amplified using general insect primers (Simon et al., 1994), and sequenced. Short sequenced regions were used to design specific primers which in combination with a general insect primer allowed us to link those regions by long PCR: *12S*  $\rightarrow$  *cox1*; *cox1*  $\rightarrow$  *cox3*; *cox3*  $\rightarrow$  *cytB*; *cytB*  $\rightarrow$  *16S*; *16S*  $\rightarrow$  *12S*. Primer sequence and location for each short and long PCR is listed in Table 1. Within each long PCR product the full, double stranded sequence was determined by primer walking (primers available from S.L.C. upon request). Short PCRs were performed using Elongase (Invitrogen) with the following cycling conditions:  $95^{\circ}\text{C}$  for 12 min, 40 cycles of  $94^{\circ}\text{C}$  for 1 min,  $40^{\circ}\text{C}$  for 1 min,  $72^{\circ}\text{C}$  for 1 min, and a final elongation of  $72^{\circ}\text{C}$  for 7 min. Long PCRs were performed using Elongase with the following cycling conditions:  $92^{\circ}\text{C}$  for 2 min, 40 cycles of  $92^{\circ}\text{C}$  for 30 s,  $50^{\circ}\text{C}$  for 30,  $68^{\circ}\text{C}$  for 12 min, and a final run out step of  $68^{\circ}\text{C}$  for 20 min. Sequencing was performed using ABI BigDye version 3 dye terminator sequencing technology and run on ABI 3770 or ABI 3740 capillary sequencer. Sequencing PCR conditions were 28 cycles of  $94^{\circ}\text{C}/10$  s,  $50^{\circ}\text{C}/5$  s,  $60^{\circ}\text{C}/4$  min.

Raw sequence files were edited and assembled into contigs in Sequencher versions 3 and 4 (Gene Codes Corp., Ann Arbor, MI). Transfer RNA analysis was conducted using tRNAscan-SE (Lowe and Eddy, 1997) using mitochondrial predictors and a cove score cut-off of 1. Reading frames between tRNAs were identified in

Table 1

Primers used in the amplification of the *Nesomachilis australica* mt genome, sequence and location. Sequences used in primer walking are available from S.L.C. on request

Region	Primer pair (F and R)	Location*	Sequence (5' $\rightarrow$ 3')
Short PCRs			
<i>12S</i>	12SA‡	14110	TACTATGTTACGACTTAT
	12SB‡	14508	AAACTAGGATTAGATACCC
<i>16S</i>	LR-J-12887†	12787	CCGGTCTGAACTCAGATACCGT
	LR-N-13398†	13304	CGCCTGTTTAAACAAAAACAT
Long PCRs			
<i>12S</i> $\rightarrow$ <i>cox1</i>	12SA3	14110	TACTATGTTACGACTTAT
	mtd6R¶	1682	GGAAC TAATCAATTTCCAAATCCTCC
<i>cox1</i> $\rightarrow$ <i>cox3</i>	C1-J-1718†	1707	GGAGGATTTGGAAATTGATTAGTTCC
	C3-N-5460†	5444	TCAACAAAGTGTCAGTATCA
<i>cox3</i> $\rightarrow$ <i>cytB</i>	ARs8§	5375	GGAACCACCTTCTTATTGG
	ARs15§	11005	TAGGTGAATTAGAATAGCTCTTGC
<i>cytB</i> $\rightarrow$ <i>16S</i>	ARs3§	11357	CCTCCTAACTTCACCTTTTCTC
	ARs4§	12508	GCTTTTTTAACATTGCTTGAAAGG
<i>16S</i> $\rightarrow$ <i>12S</i>	ARs13§	13582	TATTCACAAATGGTTGGG
	12SB‡	14508	AAACTAGGATTAGATACCC

\*Location of the 3' base of the primer.

†Primers taken from Simon et al. (1994).

‡Primer taken from Skerratt et al. (2002).

§Primers specifically designed for sequencing this genome.

¶Primer the reverse complement of Simon et al. (1994) primer C1-J-1718.

Sequencher and identified using translated BLAST searches (BlastX) (Altschul et al., 1997) as implemented by the NCBI website (<http://www.ncbi.nlm.nih.gov/>).

#### Taxon selection

All the currently available arthropod mitochondrial genomes on GenBank were used (see Table 2),<sup>3</sup> with the following exceptions, the two hymenopteran sequences for *Apis* (Crozier and Crozier, 1993) and *Melipona* (Arias and Silvestre, unpubl. data) and the paraneopteran sequences for *Heterodoxus* (Shao et al., 2001), *Thrips* (Shao and Barker 2003) and Lepidopsocid sp. (Shao et al., 2003). These three genomes were excluded, as they have previously been shown to have highly divergent sequences, are highly unstable within tree topologies and never group within Insecta without the removal of large quantities of data (Anna Murrell, UQ, pers comm.; Shao et al., 2003). In addition, the paraneopterans also all have moderate (Lepidopsocid sp.) to extreme (*Thrips* and *Heterodoxus*) levels of gene rearrangements which, although having an unknown effect on a phylogenetic analysis, contribute to their degree of nucleotide divergence from other insect sequences (Saccone et al., 1999). Given that they have no bearing on the question of apterygote relationships they were omitted to reduce ambiguities in our alignments. Four datasets were generated using different outgroups—nematodes, annelids and mollusks (ALL), nematodes only (NEM), mollusks and annelids (ANMOL), and arthropods only (ARTH) using myriapods as an outgroup. All mt genomes available on GenBank for nematodes and annelids were used, all mollusk genomes except those of bivalves were used due to the aberrant mitochondrial inheritance in bivalves which is believed to have unpredictable effects on phylogenetic reconstruction (Serb and Lydeard, 2003). Furthermore, bivalve sequences were not used by Nardi et al. (2003a,b) so this exclusion does not preclude valid comparisons between our study and theirs. Simply removing taxa from an aligned dataset could result in an alignment, and a phylogenetic hypothesis, that is dependent on the removed taxon, so alignments were made independently for each of the datasets rather than generating datasets by removing taxa from the aligned ALL dataset (as was done by Nardi et al., 2003a in their second, reduced taxon tree). Each of the 13 protein coding genes were used in the alignments of the datasets ANMOL and ARTH, *atp8* was omitted from the NEM

and ALL datasets due to the absence of this gene in all available nematode mitochondrial genomes except *Trichinella*.

#### Alignment and phylogenetic inference

DNA sequences for each of the 12/13 protein coding genes were converted to their amino acid sequences and these sequences were aligned using ClustalW (Thompson et al., 1994) with the following parameters: Pair-wise alignment gap opening penalty = 10 and extension penalty = 0.1; Multiple alignment gap opening penalty = 10 and extension penalty = 0.2; Protein weight matrix = Gonnet; Residue specific penalties: On; Hydrophilic penalties: On; Gap separation distance = 4; End gap separation: Off; Negative matrix = Off; Delay divergent cutoff = 30%. This is identical to the default amino acid alignment parameters of ClustalX used by Nardi et al. (2003a) to generate their amino acid alignment. The amino acid alignment was used to generate a DNA sequence alignment with the program CodonAlign 1.0 (<http://www.rochester.edu/College/BIO/labs/HallLab/CodonAlign.html>) which inserts gaps into the nucleotide sequences in triplets to match those inferred from the amino acid alignment. The alignments of individual genes were then concatenated in MacClade 4.06 (Maddison and Maddison, 2003) and data partitions delimited on the basis of each included gene and for each codon position. Amino acid alignments were concatenated in a similar fashion.

Phylogenetic analysis was performed with PAUP 4.0b10 (Swofford, 2002), for each of the four datasets (ALL, NEM, ANMOL, ARTH). Heuristic and bootstrap replicate trees were constructed for each of three matrix types—DNAALL (all codons), DNA12 (first and second codons) and PROT (amino acid sequences) using parsimony. Bootstrap supports were calculated with PAUP 4.0b10 from 1000 replicates. Tree statistics were calculated in PAUP 4.0b10. The relative contribution of each dataset to the total evidence topology was calculated with partitioned Bremer (Baker and DeSalle, 1997), using a batch file generated by TreeRot version 2 (Sorenson, 1999) and implemented in PAUP. To compare parsimony and model based approaches, Bayesian analyses were performed with MrBayes version 3.0b4 (Huelsenbeck and Ronquist, 2001). Partitions for parameter calculation were made for each of the 12/13 genes, and four chains were run for 1 million generations with sampling every 1000 generations. Completed analyses were examined for the asymptotic behavior of each parameter and total tree likelihood; trees collected prior to this asymptotic point were treated as burn-in and discarded. Models were chosen using ModelTest (Posada and Crandall, 1998). GTR + I + G was always the favored model. Finally, gene “quality” was assessed by replicating our analyses on a restricted

<sup>3</sup>This statement was true at the time the analyses were done in early December 2003, subsequent to those analyses several additional ticks (Shao et al., 2004), a spider (Masta and Boore, 2004) and crustaceans (Lavrov et al., 2004) have been deposited. As none of these groups are currently considered close to either Collembola or Branchipoda they are not critical to the current investigation.

Table 2  
Taxon sampling and availability

Species	Order	Class	Accession no.	Availability
<i>Nesomachilis australica</i>	Archaeognatha	Insecta	AY_793551	Present study
<i>Tricholepidion gertschi</i>	Zygentoma	Insecta	AY_191994	Nardi et al., 2003a
<i>Locusta migratoria</i>	Orthoptera	Insecta	NC_001712	Flook et al., 1995
<i>Triatoma dimidiata</i>	Hemiptera	Insecta	NC_002609	Dotson and Beard, 2001
<i>Tribolium castaneum</i>	Coleoptera	Insecta	NC_003081	Friedrich and Muqim, 2003
<i>Crioceris duodecimpunctata</i>	Coleoptera	Insecta	NC_003372	Stewart and Beckenbach, 2003
<i>Pyrocoelia rufa</i>	Coleoptera	Insecta	NC_003970	Bae et al., 2004
<i>Anopheles pernyi</i>	Lepidoptera	Insecta	NC_004622	Liu et al., unpublished
<i>Bombyx mori</i>	Lepidoptera	Insecta	NC_002355	Lee et al., unpublished
<i>Bombyx mandarina</i>	Lepidoptera	Insecta	NC_003395	Yukuhiro et al., 2002
<i>Ostrinia furnacalis</i>	Lepidoptera	Insecta	NC_003368	Coates et al., unpublished
<i>Ostrinia nubilalis</i>	Lepidoptera	Insecta	NC_003367	Coates et al., unpublished
<i>Anopheles quadrimaculatus</i>	Diptera	Insecta	NC_000875	Mitchell et al., 1993
<i>Anopheles gambiae</i>	Diptera	Insecta	NC_002084	Beard et al., 1993
<i>Drosophila yakuba</i>	Diptera	Insecta	NC_001322	Clary and Woolstenholme, 1985
<i>Drosophila melanogaster</i>	Diptera	Insecta	NC_001709	Lewis et al., 1995
<i>Chrysomya putoria</i>	Diptera	Insecta	NC_002697	Junqueira et al., 2004
<i>Cochliomyia hominivorax</i>	Diptera	Insecta	NC_002660	Lessinger et al., 2000
<i>Ceratitis capitata</i>	Diptera	Insecta	NC_000857	Spanos et al., 2000
<i>Gomphiocephalus hodgsoni</i>	Arthropoda	Collembola	AY_191994	Nardi et al., 2003a
<i>Tetrodonotophora bielaniensis</i>	Arthropoda	Collembola	NC_002735	Nardi et al., 2001
<i>Daphnia pulex</i>	Branchiopoda	Crustacea	NC_000844	Crease, 1999
<i>Artemia franciscana</i>	Branchiopoda	Crustacea	NC_001620	Perez et al., 1994
<i>Triops cancriformis</i>	Branchiopoda	Crustacea	NC_004465	Umetsu et al., 2002
<i>Tigriopus japonicus</i>	Copepoda	Crustacea	NC_003979	Machida et al., 2002
<i>Panulirus japonicus</i>	Decapoda	Crustacea	NC_004251	Yamauchi et al., 2002
<i>Portunus trituberculatus</i>	Decapoda	Crustacea	NC_005037	Yamauchi et al., 2003
<i>Pagurus longicarpus</i>	Decapoda	Crustacea	NC_003058	Hickerson and Cunningham, 2000
<i>Penaeus monodon</i>	Decapoda	Crustacea	NC_002184	Wilson et al., 2000
<i>Limulus polyphemus</i>	Xiphosura	Chelicerata	NC_003057	Lavrov et al., 2000a
<i>Varroa destructor</i>	Acari	Chelicerata	NC_004454	Navajas et al., 2002
<i>Ornithodoros moubata</i>	Ixodida	Chelicerata	NC_004357	Shao et al., 2004
<i>Ixodes hexagonus</i>	Ixodida	Chelicerata	NC_002010	Black and Roehrdanz, 1998
<i>Ixodes persulcatus</i>	Ixodida	Chelicerata	NC_004370	Shao et al., 2004
<i>Rhipicephalus sanguineus</i>	Ixodida	Chelicerata	NC_002074	Black and Roehrdanz, 1998
<i>Thyropygus</i> sp.	Diplopoda	Myriapoda	NC_003344	Lavrov et al., 2002
<i>Narceus annularis</i>	Diplopoda	Myriapoda	NC_003343	Lavrov et al., 2002
<i>Lithobius forficatus</i>	Chilopoda	Myriapoda	NC_002629	Lavrov et al., 2000
<b>OUTGROUPS</b>				
<i>Platyneris dumerilii</i>	Polychaeta	Annelida	NC_000931	Boore and Brown, 2000
<i>Lumbricus terrestris</i>	Oligochaeta	Annelida	NC_001673	Boore and Brown, 1995
<i>Katharina tunicata</i>	Polyplacophora	Mollusca	NC_001636	Boore and Brown, 1994
<i>Loligo bleekeri</i>	Cephalopoda	Mollusca	NC_002507	Tomita et al., 2002
<i>Albinaria coerulea</i>	Gastropoda	Mollusca	NC_001761	Hatzoglou et al., 1995
<i>Pupa strigosa</i>	Gastropoda	Mollusca	NC_002176	Kurabayashi and Ueshima, 2000
<i>Roboastra europaea</i>	Gastropoda	Mollusca	NC_004321	Grande et al., 2002
<i>Cepaea nemoralis</i>	Gastropoda	Mollusca	NC_001816	Yamazaki et al., 1997
<i>Onchocerca volvulus</i>	Spirurida	Nematoda	NC_001861	Keddie et al., 1998
<i>Brugia malayi</i>	Spirurida	Nematoda	NC_004298	Daub et al., unpublished
<i>Caenorhabditis elegans</i>	Rhabditida	Nematoda	NC_001328	Okimoto et al., 1992
<i>Strongyloides stercoralis</i>	Rhabditida	Nematoda	NC_005143	Hu et al., 2003
<i>Cooperia oncophora</i>	Rhabditida	Nematoda	NC_004806	van der Veer and de Vries, unpublished
<i>Necator americanus</i>	Rhabditida	Nematoda	NC_003416	Hu et al., 2002
<i>Ancylostoma duodenale</i>	Rhabditida	Nematoda	NC_003415	Hu et al., 2002
<i>Ascaris suum</i>	Ascaridida	Nematoda	NC_001327	Okimoto et al., 1992
<i>Trichinella spiralis</i>	Trichocephalida	Nematoda	NC_002681	Lavrov and Brown, 2001

dataset and by determining levels of homoplasious difference between gene partitions. A restricted dataset was generated for each dataset and treatment by the exclusion of all genes except *cox1*, *cox2*, *cox3* and *cytB*

to generate a dataset that most closely matched those used by Nardi et al. (2003a,b). The distribution of homoplasy amongst the genes was estimated by comparing the proportional size of each gene, the length of

the aligned gene partition divided by the total alignment length, with its proportional contribution to tree length, tree length for that gene partition determined by TreeRot divided by the total tree length for the combined analysis:  $nl_g/nl_t : tl_g/tl_t$  where  $nl_g$  is the nucleotide length of a particular gene,  $nl_t$  is the nucleotide length of the total alignment,  $tl_g$  is the partitioned tree length for that gene and  $tl_t$  is the total tree length. If homoplasy is distributed randomly within the dataset then these proportions should be the same, if homoplasy is clustered in particular genes these should contribute extra tree length to the combined analysis disproportionate to the length of the gene. Extremely high levels of homoplasy in a particular gene may be grounds for its exclusion from an analysis and is a more rigorous criterion for gene exclusion than the alignment based methods proposed by Nardi et al. (2003a).

## Results

*The genome of the archaeognathan, Nesomachilis australica* Tillyard, 1924.

The entire mitochondrial genome of *Nesomachilis australica* was sequenced on both strands by multiple overlapping of long PCR fragments and primer walking. The genome is 15 474 bp long and contains the usual metazoan complement of 13 protein genes, 2 ribosomal RNA genes and 21 of the 22 tRNA genes in the proposed ancestral insect arrangement (Boore, 1999). COVE analysis failed to find tRNA-Arg and tRNA-Asn; tRNA-Arg was found by manual comparison to published arthropod genomes, but tRNA-Asn could not be found in this way. A 160 bp stretch of non-coding sequence between tRNA-Arg and tRNA-Ser (GCT) genes, which is the normal location of the tRNA-Asn gene, contains several tRNA-like motifs but we could not fold a completed tRNA. This stretch of the genome most likely codes for tRNA-Asn but it is secondarily modified by a form of RNA editing such as that proposed by Lavrov et al. (2000b) to account for the lack of tRNAs in the genome of the centipede *Lithobius*.

### *Parsimony analysis of arthropod mitochondrial genomes*

The results of analyzing the ARTH and ALL datasets are depicted in Figs 1–6. A comparison of all datasets and treatments is presented in Table 3, indicating which taxon groups were found to be monophyletic in each analysis. Tree scores for each analysis are shown in Table 4. A full complement of the tree data (alignments, plus tree figures) is available from the Whiting lab website (<http://whitinglab.byu.edu/>). It is difficult to determine the significance of the differences between the different datasets and, thus, of varying outgroup choice.

The NEM datasets yielded the shortest trees and had the highest consistency and retention statistics (Table 4) of any of the three nonarthropod outgrouping options tried, which is to be expected given other lines of evidence for the “ectysozoa” hypothesis. However, each of the three datasets which had nonarthropod outgroups—ANMOL, NEM and ALL—tended to produce an unresolved bush of arthropod ingroup relationships with a greater or lesser tendency to resolve as monophyletic groups such as the Ixodida or the Myriapoda (see Table 3). This is due to two factors: insufficient taxon sampling within the putative outgroups (e.g., the nematodes are mostly represented by parasites and a single free-living species *Caenorhabditis* which is closely related to parasitic nematodes), and the distance of each of these phyla from the ingroup. Mt genomes from additional minor invertebrate phyla such as Tardigrada, Nemertea and Onychophora are needed before the question of properly rooting the arthropods can be adequately addressed. For now, however, we consider it sufficient to use a single, clearly monophyletic arthropod lineage, such as the myriapods or the chelicerates, as an outgroup for resolving relationships among the remaining arthropod groups. The main trend in the data treatments is of increased resolution using amino acids instead of nucleotides. We expected the exclusion of the third codon position to result in trees similar to the amino acid trees, since most amino acids are redundant at this position, however, we did not observe this effect. Different results were found for each of the three data treatments. However, it is clear that there is considerable homoplasy in the third codon position, as the DNA12 trees are only half as long as the comparable DNAALL tree. The DNA12 trees are usually better resolved and lack the obviously artefactual relationships found in the DNAALL trees (e.g., the pairing of the nematode *Trichinella* with the copepod *Tigriopus* in the NEM-DNAALL tree which is not supported in either the NEM-DNA12 or NEM-PROT trees). The failure of the DNA12 trees to replicate the PROT trees may be putative evidence for multiple hits within the first and second codon positions obscuring the phylogenetic signal from DNA in general. Equally, it could be due to convergent evolution within the amino acids, as feared by authors such as Simmons et al. (2002). Either way, the data treatment is not trivial and needs to be explored as a part of phylogenetic analyses.

The most notable result of our study is that Hexapoda was not monophyletic in any of our analyses. Collembola, whilst strongly supported as monophyletic in all datasets and treatments, never groups with any of the insect groups. Interestingly it never groups with the brachiopod crustaceans *Artemia*, *Daphnia* or *Triops* as was suggested by Nardi et al. (2003a,b) either. The position of Collembola is highly variable between datasets and within them, depending upon the treatment

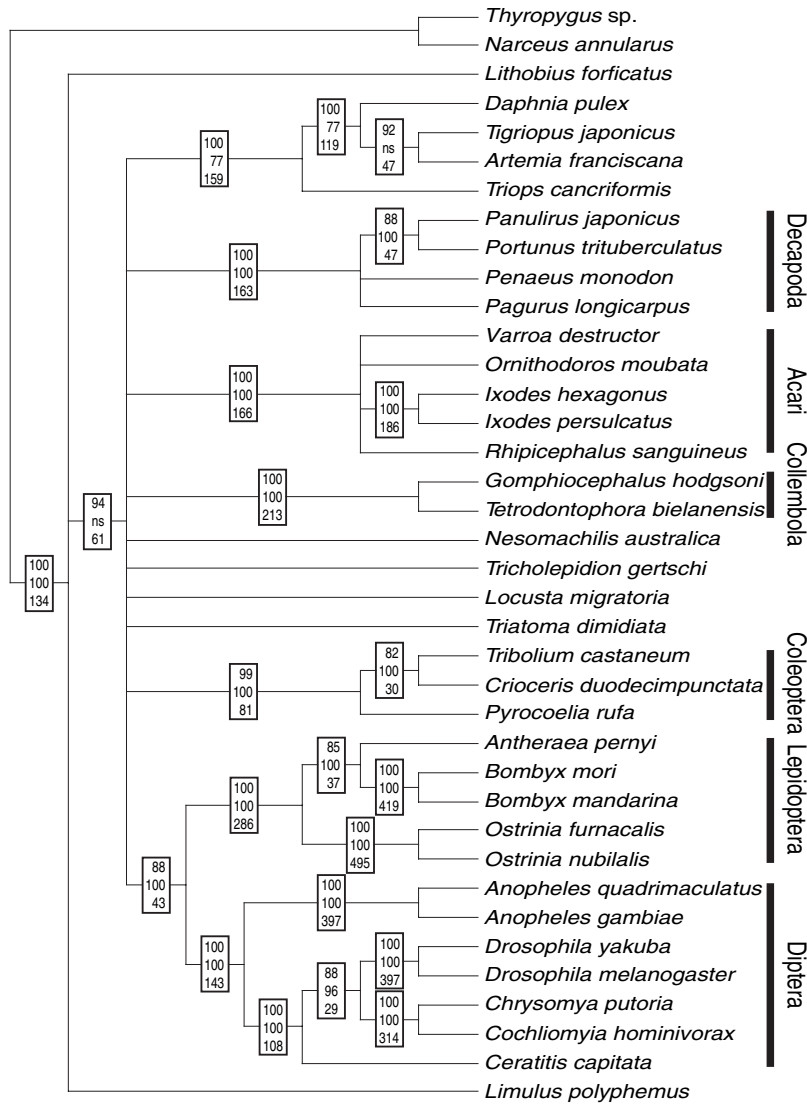


Fig. 1. Cladogram of arthropod relationships inferred from the DNAALL-ARTH dataset. Support values for each node are placed in a box on the branch subtending the node, the top number is the percentage bootstrap support from parsimony analysis, the second is the Bayesian posterior probability, and the bottom number is the Bremer support from the combined parsimony analysis. Supraspecific taxa supported by this analysis are indicated at the right by bars. ns: not supported.

used: sister to Pancrustacea (ARTH-PROT; ARTH-DNA12), sister to Pancrustacea + Myriapoda + *Limulus* (ANMOL-DNAALL), one of 4 (ALL-DNA-ALL), 9 (NEM-DNAALL), 10 (ALL-DNA12) or 11 (NEM-DNA12) groups in a clade including all arthropods except the Acari, and one of 8 (NEM-PROT) 9 (ANMOL-PROT), 10 (ARTH-DNAALL; ALL-PROT) or 15 (ANMOL-DNA12) lineages in the ingroup polytomy of all arthropods. Clearly Collembola cannot be unambiguously grouped with any other arthropod lineage at this time on the basis of mitochondrial genome data alone.

Many other relationships are consistent across datasets and treatments and strongly supported regardless of

how the matrix was constructed—monophyly of the insect orders Diptera, Lepidoptera and Coleoptera, and the crustacean group Decapoda were recovered in all analyses, and the Acari, Ixodida and Myriapoda in most analyses. Conversely several taxa are highly mobile within the trees depending upon outgroup choice and data treatment. The copepod *Tigriopus* was found in many different locations, including as sister to *Artemia*, which is responsible for most instances where brachiopod monophyly breaks down (all ARTH sets, ANMOL-DNAALL), as sister to all arthropods (ANMOL-PROT; NEM-PROT; NEM-DNA12), as an independent arthropod lineage (ANMOL-DNA12), as sister group to *Trichinella* (NEM-DNAALL), and as an

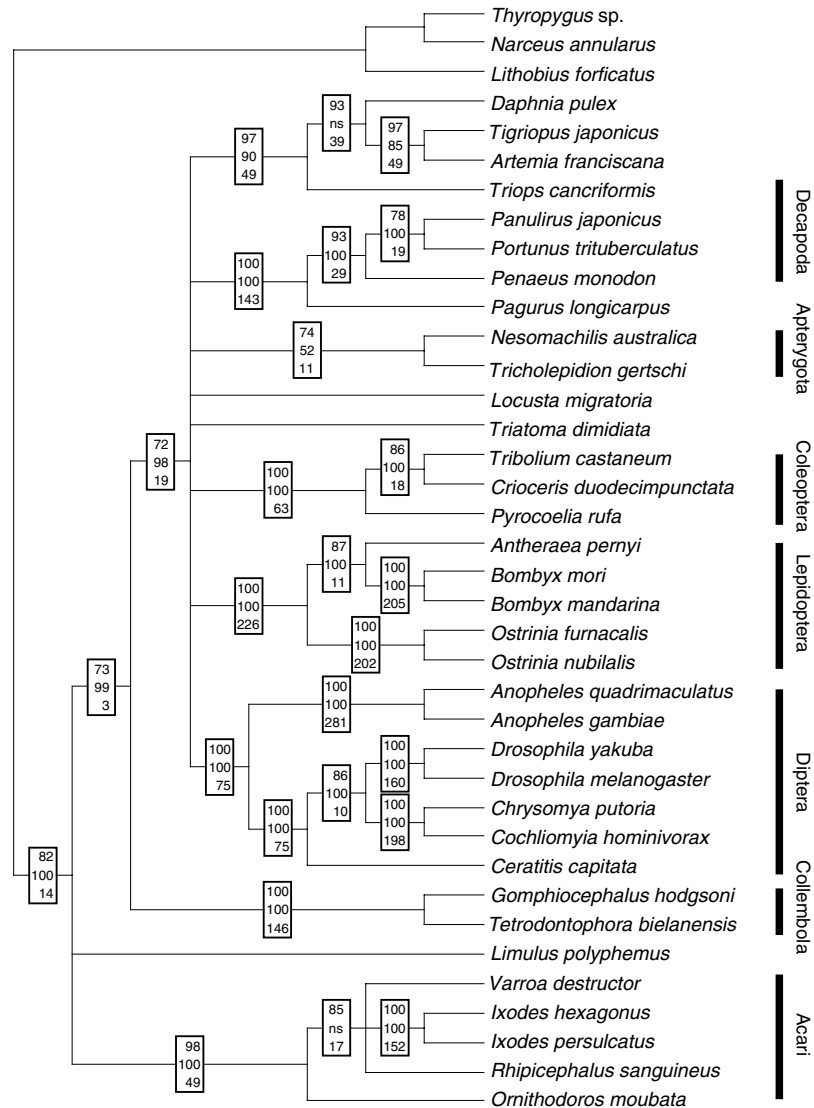


Fig. 2. Cladogram of arthropod relationships inferred from the DNA12-ARTH dataset. Support values for each node are placed in a box on the branch subtending the node, the top number is the percentage bootstrap support from parsimony analysis, the second is the Bayesian posterior probability, and the bottom number is the Bremer support from the combined parsimony analysis. Supraspecific taxa supported by this analysis are indicated at the right by bars. ns: not supported.

independent ingroup lineage (ALL-PROT; ALL-DNA-ALL; ALL-DNA12). The mt genome of *Tigriopus* is highly rearranged relative to other arthropods (Machida et al., 2002), and this may be a complicating factor in its use in phylogenetics, as similar effects have been noted using the highly rearranged genome of the louse *Heterodoxus* (Nardi et al., 2003a; Shao et al., 2003). A second problematic taxon is the horseshoe crab *Limulus*, which is currently considered to be the most basal chelicerate (Giribet et al., 2001). The only analyses that recovered a monophyletic Chelicerata were ARTH-PROT, ARTH-DNAALL, ARTH-DNA12, ANMOL-PROT, NEM-PROT and ALL-PROT, but the relationship only had significant bootstrap support

in the ARTH-PROT set. Far more frequent is the grouping of *Limulus* with the myriapods either as sister to the entire Myriapoda or to the Diplopoda. This preference to group with myriapods instead of the remaining chelicerates may be due to the large phyletic distances between *Limulus* and the remaining chelicerates which are all members of the comparatively highly derived Acari. It would be interesting to see if this apparent long branch attraction would survive the addition of earlier branching chelicerate groups.

Partitioned Bremer values were used to assess the relative contributions of individual genes to tree support. In general, support is split amongst the 13 analyzed genes, and no single gene or set of genes

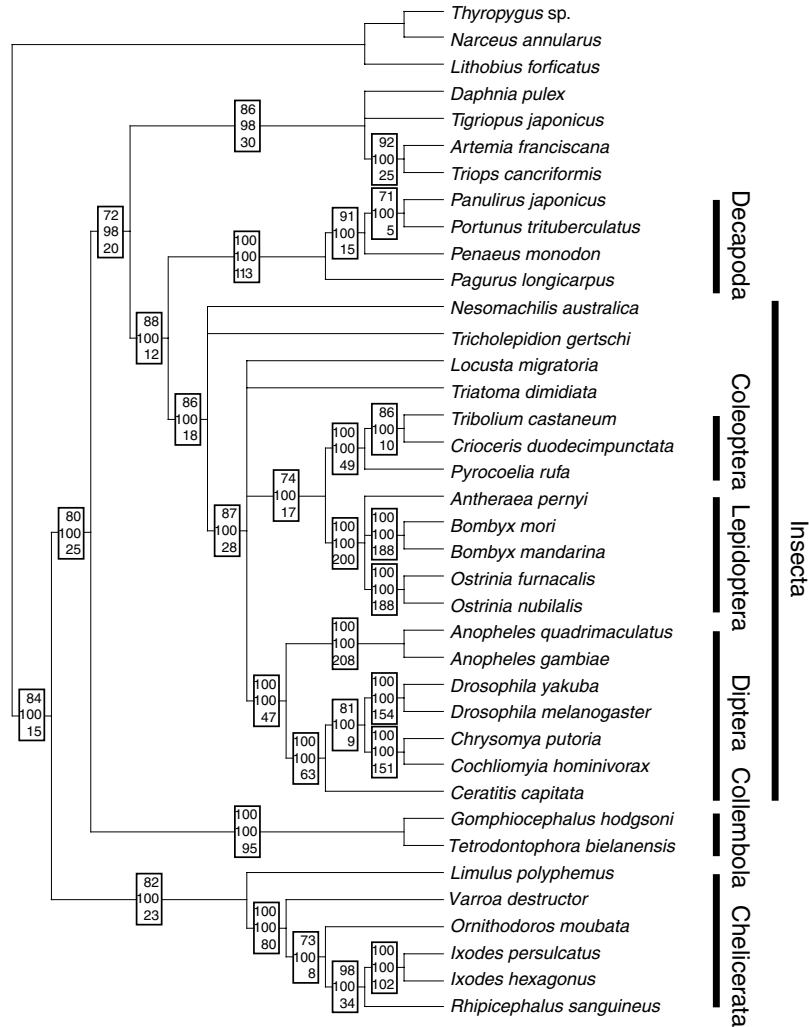


Fig. 3. Cladogram of arthropod relationships inferred from the PROT-ARTH dataset. Support values for each node are placed in a box on the branch subtending the node, the top number is the percentage bootstrap support from parsimony analysis, the second is the Bayesian posterior probability, and the bottom number is the Bremer support from the combined parsimony analysis. Supraspecific taxa supported by this analysis are indicated at the right by bars. ns: not supported.

appears to “drive” the analysis. The genes that support a particular node vary. No genes consistently conflict with the majority result either by providing negative support values for all nodes, which would suggest a radically different phylogenetic history for that gene, or by strongly conflicting with a node across all datasets and treatments. Interestingly the magnitude of Bremer supports varies considerably among datasets and between treatments within a given dataset, such that one gene may strongly support a node when coded as amino acids, weakly as DNA with all codon positions and conflict as DNA first and second codon positions only. This is to be expected considering the high levels of homoplasy in these datasets as suggested by the very low CI, RI and RC scores for all trees (Table 4). There are noticeable differences between the comparable DNA-ALL and DNA12 treatments for many nodes, probably

resulting from considerable homoplasy in the third codon position. In nodes which group congeneric pairs such as the two *Drosophila* species, the supports are consistently weaker in the DNA12 treatment than in the DNAALL treatment suggesting that the third codon position has a phylogenetic signal. Contrary to the suggestion by Nardi et al. (2003a) that only some of the mitochondrial protein coding genes are useful in phylogenetic reconstruction of the arthropods, these results suggest all genes are compatible with each other and the resolving power produced by their inclusion greatly outweighs the possible noise caused by their inclusion.

#### Bayesian analyses

Bayesian analyses of the four datasets by three data treatments produced results which are not radically

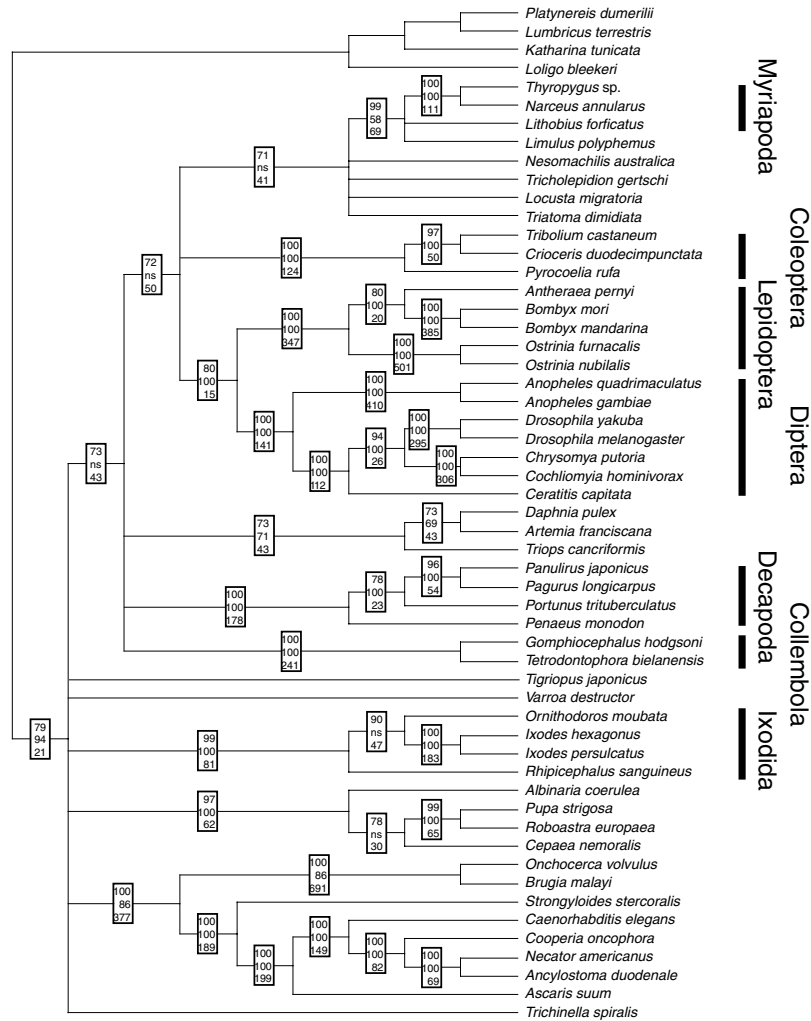


Fig. 4. Cladogram of arthropod, annelid, mollusk and nematode relationships inferred from the DNAALL-ALL dataset. Support values for each node are placed in a box on the branch subtending the node, the top number is the percentage bootstrap support from parsimony analysis, the second is the Bayesian posterior probability, the bottom number is the Bremer support from the combined parsimony analysis. Supraspecific taxa supported by this analysis are indicated at the right by bars. ns: not supported.

different from those obtained by parsimony analysis (Table 5; Figs 1–6). In general, for any given dataset and treatment, the topology with the highest posterior probability is very similar to the topology of the lowest cost heuristic parsimony tree. The posterior probabilities for individual nodes are often higher than the bootstrap supports recovered from the same node in parsimony bootstrapping, suggesting that this is another case where Bayesian posteriors are a less conservative measure of branch support than bootstrapping. The variation among data treatments is far less pronounced than for the parsimony analyses—similar topologies were recovered from each of the three treatments—however, posterior probabilities are consistently highest in the PROT treatments, less in DNA12 and lowest in DNAALL. As in the parsimony analyses, Hexapoda is

not found to be monophyletic in any dataset or treatment. The placement of the collembolans varied almost as much as in the parsimony analysis and was unresolved in the majority of instances: sister group to Pancrustacea (ARTH-PROT; ARTH-DNA12, ALL-PROT), sister group to Myriapoda + Chelicerata (ANMOL-DNA12), or one of 5 (ALL-DNA12), 7 (ANMOL-PROT), 8 (ALL-DNAALL), 9 (ARTH-DNAALL, ANMOL-DNAALL) or 17 (NEM-PROT) lineages in the ingroup polytomy of all arthropods. Analytical methodology is not responsible for the differences between our analyses and those of Nardi et al. (2003a,b). However, the inflation of a node's posterior probability in the amino acid analyses may have led to excessive confidence in the accuracy of their topology.

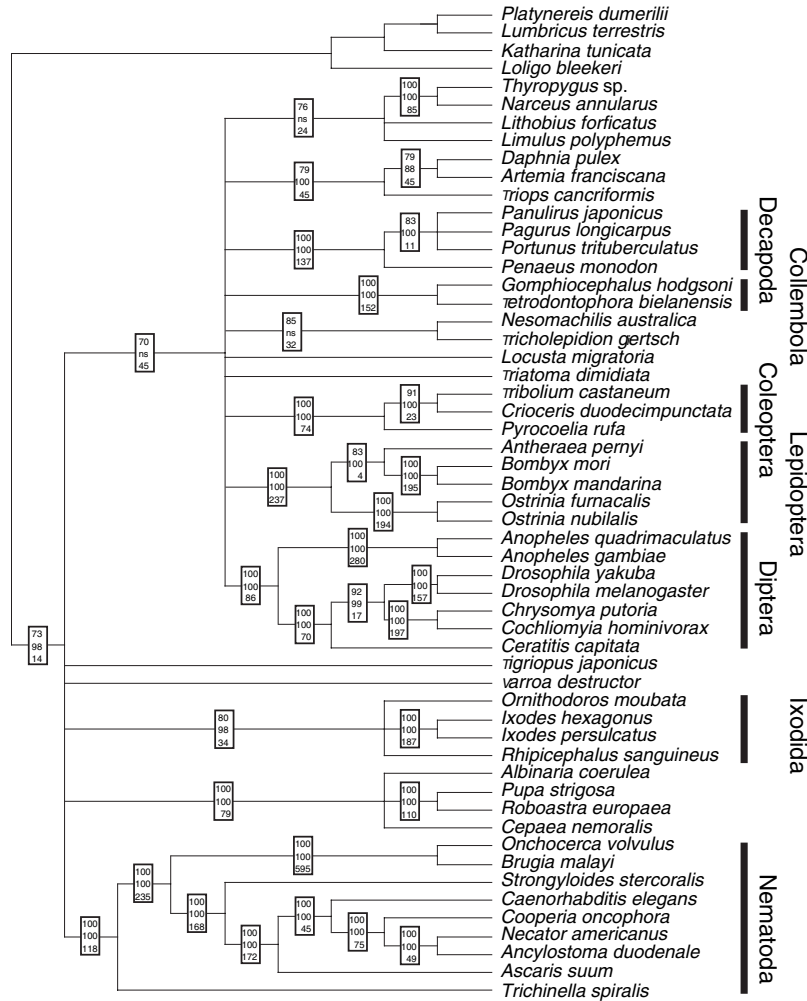


Fig. 5. Cladogram of arthropod, annelid, mollusk and nematode relationships inferred from the DNA12-ALL dataset. Support values for each node are placed in a box on the branch subtending the node, the top number is the percentage bootstrap support from parsimony analysis, the second is the Bayesian posterior probability, and the bottom number is the Bremer support from the combined parsimony analysis. Supraspecific taxa supported by this analysis are indicated at the right by bars. ns: not supported.

### Reduced gene comparisons

To investigate the proposition that certain mitochondrial genes give better estimates of phylogeny than others, analyses were undertaken with the four genes used by Nardi et al. (2003a,b): *cox1*, *cox2*, *cox3* and *cytB*, for each of the datasets and treatments described above. If their result is a function of the signal in these four genes, which may be obscured by noise from the other nine, we would expect reduced gene analyses to recover a similar result, at least for the ANMOL-PROT matrix as this most closely resembles the data used previously. In parsimony analysis none of the 12 datasets, as assessed by bootstrap resampling, significantly link Collembola to any other arthropod taxon. In each case Collembola is one of several independent arthropod lineages in a basal polytomy. Furthermore,

since no hexapod groups more inclusive than Diptera, Lepidoptera, Acari or Decapoda are found to be monophyletic in any of these analyses, it appears that reducing the number of genes results in a significant loss of phylogenetic signal throughout the tree. The most reasonable conclusion is that there is insufficient signal in these data alone to resolve the relationships of hexapod groups. In general, the reduced datasets are inferior in their resolving power to the comparable all gene datasets. This is also evident from the partitioned Bremer values, which are spread evenly across the genome as opposed to concentrated in any particular gene.

### Assessing gene “quality”

As *a priori* assessments of gene quality based on alignment parameters such as percentage gaps or

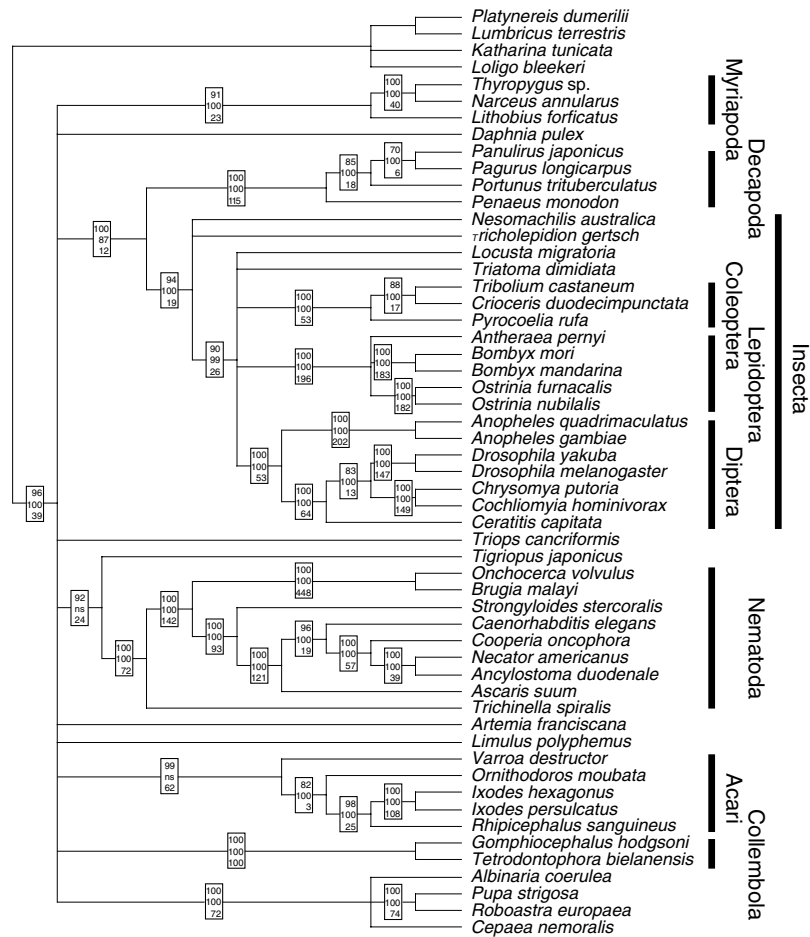


Fig. 6. Cladogram of arthropod, annelid, mollusk and nematode relationships inferred from the PROT-ALL dataset. Support values for each node are placed in a box on the branch subtending the node, the top number is the percentage bootstrap support from parsimony analysis, the second is the Bayesian posterior probability, and the bottom number is the Bremer support from the combined parsimony analysis. Supraspecific taxa supported by this analysis are indicated at the right by bars. ns: not supported.

invariable sites is likely to be unduly influenced by taxon selection and choice of alignment parameters, we chose to assess gene quality *a posteriori* by examining the evenness of homoplasy within the parsimony analyses. If homoplasy was spread evenly amongst the 13 genes then each gene should contribute to tree length in proportion to its own length. Genes with higher or lower levels of homoplasy should contribute a greater or lesser proportion of tree length, respectively, than would be expected from the length of the gene. The proportion of total and variable sites of each gene in the alignment and the proportion of total tree length inferred for that gene are presented in Table 6. In most datasets and treatments, *atp6*, *cox1*, *cox2*, *cox3*, *cytB* and *nadh1* contribute less to combined tree length than would be expected from gene length, and the effect for *cox1* is especially dramatic (7.6% less). In contrast, support from *nadh2*, *nadh4* and *nadh5* is consistently greater than expected from gene length (up to 4.39% greater) and the remaining genes contribute slightly less or more depend-

ing on the analysis. The magnitude of differences is least for the DNAALL treatment, more pronounced for DNA12 and most extreme for PROT, whereas outgroup choice appears to have a negligible effect. Comparable results were found for each of the four datasets, again suggesting that data treatment is a more significant source of variation than outgroup choice. The extent of this bias is also correlated with the proportion of invariant sites in an individual gene. This is to be expected, as invariant sites contribute nothing to tree length whilst increasing the size of the gene. If the invariant sites are excluded from the calculations of proportionate gene length the magnitudes of the differences are greatly reduced in all cases. In at least one case (*cox1*, ARTH-DNAALL) it is actually reversed such that compared to total gene length the gene has lower homoplasy but has higher homoplasy with invariants excluded. In general, when the proportionate tree lengths are compared to proportionate gene lengths with invariant sites excluded, the magnitude of the

Table 3

Comparisons of relationships recovered in the different datasets and treatments by parsimony analysis. A single tick denotes the relationship is supported in the heuristic search alone, a double tick that it is also supported with greater than a 70% bootstrap support, a cross indicates that it is not supported, a D indicates that monophyly is defined *a priori* by the taxon's use as an outgroup, and NA means that the taxon was not included in this analysis. Ar-P: Arthropod protein; Ar-A: Arthropod all codons; Ar12: Arthropod first and second codon positions; M-P: Mollusca + Annelida protein; M-A: Mollusca + Annelida all codons; M12: Mollusca + Annelida first and second codon positions; N-P: Nematode protein; N-A: Nematode all codons; N12: Nematode first and second codon positions; A-P: All outgroups protein; A-A: All outgroups all codons; A12: All outgroups first and second codon positions

Groups supported as monophyletic	Ar-P	Ar-A	Ar12	M-P	M-A	M12	N-P	N-A	N12	A-P	A-A	A12
Diptera	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓
Coleoptera	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓
Lepidoptera	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓
Holometabola	✓	×	×	✓	×	×	✓	×	×	✓	×	×
Eumetabola	×	×	×	×	×	×	×	×	×	×	×	×
Apterygota	×	✓	✓✓	×	✓✓	✓	×	✓	✓✓	×	✓	✓✓
Insecta	✓✓	✓	✓✓	✓	×	✓✓	✓✓	×	×	✓✓	×	✓✓
Hexapoda	×	×	×	×	×	×	×	×	×	×	×	×
Collembola	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓
Decapoda	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓
Branchiopoda	×	×	×	×	×	×	×	✓✓	✓✓	✓	✓✓	✓✓
Crustacea	×	✓	✓	×	✓	×	×	×	×	×	×	×
Chelicerata	✓✓	✓	✓	✓	×	×	✓	×	×	✓	×	×
Acari	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	×	×	✓✓	×	✓
Ixodida	✓✓	✓	×	✓✓	✓	✓✓	✓✓	✓✓	✓	✓✓	✓✓	✓✓
Myriapoda	D	D	D	✓	✓	✓	✓✓	✓✓	✓✓	✓✓	✓	✓✓
Annelida	NA	NA	NA	D	D	D	NA	NA	NA	D	D	D
Mollusca	NA	NA	NA	D	D	D	NA	NA	NA	D	D	D
Nematoda	NA	NA	NA	NA	NA	NA	D	D	D	D	D	D

Table 4

Parsimony tree statistics. CI: consistency index; RI: retention index; RC: rescaled consistency index

	No. sites	Tree length	CI	RI	RC
ARTH-PROT	3919	32 151	0.48	0.44	0.21
ARTH-DNAALL	11 792	80 700	0.26	0.34	0.09
ARTH-DNA12	7862	39 681	0.30	0.40	0.12
ANMOL-PROT	4027	42 883	0.45	0.42	0.19
ANMOL-DNAALL	12 120	107 212	0.22	0.33	0.07
ANMOL-DNA12	8080	53 757	0.26	0.39	0.10
NEM-PROT	3956	40 117	0.47	0.51	0.24
NEM-DNAALL	11 904	98 712	0.24	0.41	0.10
NEM-DNA12	7936	49 005	0.29	0.49	0.14
ALL-PROT	4047	50 641	0.43	0.48	0.21
ALL-DNAALL	12 177	124 735	0.21	0.38	0.08
ALL-DNA12	8118	62 829	0.24	0.45	0.11

differences are quite small (not more than 5% for PROT, 4% for DNA12 and 2% for DNAALL). Such low variances suggest to us that homoplasy is not sufficiently concentrated in any particular genes to warrant exclusion of genes from phylogenetic analysis, and, instead, the spread of support among all genes argues for their inclusion to improve the resolving power of the analysis.

## Discussion

Historically it has often been the case that when major new classes of data become available for

phylogenetic analysis they have initially produced results that are at odds with previous notions of phylogeny or classification. Sometimes these conflicting results have exposed deficiencies in the previous data, e.g., protozoan classifications based on light microscopy versus electron microscopy (Corliss, 1979), bacterial classifications based on biochemical data versus DNA hybridization or 16S sequencing (Boone and Castenholz, 2001), and metazoan classifications based on adult morphology versus sequence data (Zrzavy et al., 1998). At other times they have revealed the inapplicability of the new data type to the phylogenetic questions under investigation, e.g., using DNA-hybridization to resolve bird phylogeny (Sibley et al., 1988; Sibley and Ahlquist, 1990). Nardi et al. (2003a,b) proposed a novel grouping of springtails (Collembola) with the branchiopod crustaceans which would suggest an independent evolution of the six-legged, hexapod, condition in arthropods and two independent invasions of the terrestrial habitats by insects and ectognaths (or some portion of the ectognaths). This result is at odds with previously accepted scenarios of arthropod evolution. However, does it represent a novel type of data illuminating a previously obscure topic or a case of taking analyses beyond the capacity of the data to answer the questions being asked? Or put more simply, are mitochondrial genome sequences alone sufficient to resolve the phylogeny of arthropods and, if so, using what methods? To answer these questions we need to consider two factors: first, does the novel result explain

Table 5

Comparisons of relationships recovered from the different datasets and treatments using Bayesian analysis. A single tick denotes the relationship is supported in the tree with the highest posterior probability, a double tick that it is also supported with greater than 90% posterior probability, a cross indicates that it is not supported, a D indicates that monophyly is defined by the groups use as an outgroup, and NA means that the group was not included in this analysis. Ar-P: Arthropod protein; Ar-A: Arthropod all codons; Ar12: Arthropod first and second codon positions; M-P: Mollusca + Annelida protein; M-A: Mollusca + Annelida all codons; M12: Mollusca + Annelida first and second codon positions; N-P: Nematode protein; N-A: Nematode all codons; N12: Nematode first and second codon positions; A-P: All outgroups protein; A-A: All outgroups all codons; A12: All outgroups first and second codon positions

Groups supported as monophyletic	Ar-P	Ar-A	Ar12	M-P	M-A	M12	N-P	N-A	N12	A-P	A-A	A12
Diptera	√√	√√	√√	√√	√√	√√	√√	√√	√√	√√	√√	√√
Coleoptera	√√	√√	√√	√√	√√	√√	√	√√	√√	√√	√√	√√
Lepidoptera	√√	√√	√√	√√	√√	√√	√√	√√	√√	√√	√√	√√
Holometabola	√√	√	×	√√	×	×	√	×	×	√√	×	×
Eumetabola	√√	×	√√	√√	×	√√	√	×	√√	√√	×	√√
Apterygota	√√	×	×	×	×	√	√√	×	√√	√√	×	×
Insecta	√√	√√	√√	√√	√√	√√	√√	√√	√√	√√	√√	√√
Hexapoda	×	×	×	×	×	×	×	×	×	×	×	×
Collembola	√√	√√	√√	√√	√√	√√	√√	√√	√√	√√	√√	√√
Decapoda	√√	√√	√√	√√	√√	√√	√√	√√	√√	√√	√√	√√
Branchiopoda	×	×	×	×	×	√√	×	×	×	√√	√	√√
Crustacea	×	×	√√	×	×	×	×	√√	×	×	×	×
Chelicerata	√√	√√	√	√	√√	√√	√	√	√√	×	×	√√
Acari	√√	√√	√√	√√	√√	√√	√	√	√√	×	×	√√
Ixodida	√√	√√	√√	√√	√√	√√	√	√	√√	√√	√√	√√
Myriapoda	D	D	D	√√	×	×	√	×	×	√√	√√	×
Annelida	NA	NA	NA	D	D	D	NA	NA	NA	D	D	D
Mollusca	NA	NA	NA	D	D	D	NA	NA	NA	D	D	D
Nematoda	NA	NA	NA	NA	NA	NA	D	D	D	D	D	D

other types of data which may not have received sufficient attention previously, and second, how robust is the new dataset to alternative methods of phylogenetic analysis? Nardi et al. (2003a) offer no alternative lines of evidence to support the exclusion of collembolans from Hexapoda or their grouping with Crustacea. In their reply paper Nardi et al. (2003b) offered evidence of a recently discovered marine hexapod fossil, which challenges the idea that the hexapod condition is associated with the transition to a terrestrial existence. In contrast, however, the vast majority of alternative evidence based on morphology and molecules favors a grouping of collembolans with insects (Pritchard et al., 1993; Edgecombe et al., 2000; D'Haese, 2002, 2003; Bitsch et al., 2004). A rigorous examination of the morphological, developmental and ecological evidence bearing on collembolan relationships is greatly needed but is currently beyond the scope of this study. Instead we have chosen to focus on the issue of conclusion robustness. Specifically, we sought to examine the robustness of the proposed relationship Collembola + Branchiopoda to outgroup choice, data alignment and treatment, gene choice and optimality criteria. We found that the recovered sister group of Collembola was highly variable depending upon these factors, so variable in fact that the best conclusion is that phylogenetic resolution of basal arthropod relationships with currently available mitochondrial genomes alone is not possible.

### Outgroup

Outgroup choice carries with it two burdens in phylogenetic analysis. First, the outgroup must be outside the group of interest if it is to have more than arbitrary resolving power. Second, it must not be too distantly related, or the accumulated differences between the outgroup and ingroup will make analysis more difficult (alignment issues) or obscure relationships (“random outgroup effect” Wheeler, 1990). Outgroup choice is often confounded by the lack of sufficiently closely related taxa. In this study we sought to examine variation in outgroups to measure the effect on ingroup relationships. Nardi et al. (2003a,b) employed as outgroup taxa representatives of the Mollusca and Annelida because of the hypothesis that they collectively represent the sister-group to the Arthropoda. Interestingly, they do not consider using a nematode outgroup in line with the predictions of the Ecdysozoa hypothesis. Considerable evidence in favor of the Ecdysozoa hypothesis has accumulated, and, indeed, all the recent comprehensive phylogenies of the Metazoa have supported this rather than the traditional placement of arthropods among the protostomes (Giribet and Ribera, 1998; Zrzavy et al., 1998; Giribet and Wheeler, 1999; Giribet, 2002). The consensus of these studies has been to identify Tardigrada (water bears) or Onychophora (velvet worms) as the sister group to the Arthropoda. This has been reflected in the

Table 6

Relative contributions made by each gene to the whole phylogeny. Differences in proportional length of each aligned gene versus tree steps and variable sites only versus tree steps. Negative values indicate that a gene's contribution to overall tree length is less than expected based on its aligned length

Gene	ARTH-PROT		ARTH-DNA		ARTH-DNA12	
	Comp-total	Comp-var	Comp-total	Comp-var	Comp-total	Comp-var
<i>atp6</i>	-0.53%	-0.68%	0.05%	-0.09%	-0.07%	-0.21%
<i>atp8</i>	0.74%	0.75%	0.10%	0.11%	0.55%	0.56%
<i>cox1</i>	-7.63%	-2.93%	-3.54%	1.15%	-6.77%	-2.07%
<i>cox2</i>	-1.61%	-1.47%	-0.76%	-0.62%	-1.34%	-1.20%
<i>cox3</i>	-1.57%	-1.30%	-0.66%	-0.39%	-1.45%	-1.18%
<i>cytB</i>	-2.53%	-1.78%	-0.95%	-0.19%	-2.15%	-1.40%
<i>nadh1</i>	-0.74%	-1.11%	-0.97%	-1.34%	-1.24%	-1.61%
<i>nadh2</i>	4.39%	2.97%	2.48%	1.07%	4.06%	2.65%
<i>nadh3</i>	0.29%	0.18%	0.07%	-0.04%	0.20%	0.09%
<i>nadh4</i>	2.44%	1.49%	1.11%	0.16%	2.11%	1.16%
<i>nadh4L</i>	0.61%	0.03%	0.17%	-0.41%	0.56%	-0.02%
<i>nadh5</i>	3.60%	2.46%	1.57%	0.43%	3.25%	2.12%
<i>nadh6</i>	2.55%	1.39%	1.32%	0.16%	2.28%	1.12%
Gene	ANMOL-PROT		ANMOL-DNA		ANMOL-DNA12	
	Comp-total	Comp-var	Comp-total	Comp-var	Comp-total	Comp-var
<i>atp6</i>	-0.35%	-0.69%	0.00%	-0.35%	0.08%	-0.27%
<i>atp8</i>	0.76%	0.61%	0.24%	0.09%	0.60%	0.45%
<i>cox1</i>	-7.34%	-3.89%	-3.40%	0.05%	-6.58%	-3.13%
<i>cox2</i>	-1.28%	-1.25%	-0.58%	-0.55%	-1.10%	-1.08%
<i>cox3</i>	-1.86%	-1.55%	-0.81%	-0.50%	-1.63%	-1.33%
<i>cytB</i>	-2.33%	-1.63%	-0.91%	-0.21%	-2.14%	-1.44%
<i>nadh1</i>	-0.66%	-0.86%	-0.79%	-0.99%	-0.98%	-1.19%
<i>nadh2</i>	4.13%	3.06%	2.25%	1.18%	3.78%	2.70%
<i>nadh3</i>	0.05%	-0.05%	-0.12%	-0.22%	0.05%	-0.05%
<i>nadh4</i>	2.00%	1.70%	0.64%	0.34%	1.58%	1.29%
<i>nadh4L</i>	0.60%	0.24%	0.20%	-0.16%	0.58%	0.22%
<i>nadh5</i>	3.93%	2.91%	2.14%	1.12%	3.71%	2.68%
<i>nadh6</i>	2.35%	1.42%	1.13%	0.21%	2.06%	1.13%
Gene	NEM-PROT		NEM-DNA		NEM-DNA12	
	Comp-total	Comp-var	Comp-total	Comp-var	Comp-total	Comp-var
<i>atp6</i>	-1.10%	-0.78%	-0.78%	-0.46%	-0.72%	-0.40%
<i>cox1</i>	-7.46%	-4.53%	-3.68%	-0.75%	-6.66%	-3.73%
<i>cox2</i>	-1.38%	-1.48%	-0.57%	-0.67%	-1.10%	-1.20%
<i>cox3</i>	-1.08%	-1.31%	-0.28%	-0.52%	-0.94%	-1.17%
<i>cytB</i>	-1.84%	-1.67%	-0.47%	-0.30%	-1.55%	-1.38%
<i>nadh1</i>	-0.57%	-0.72%	-0.69%	-0.84%	-0.96%	-1.10%
<i>nadh2</i>	4.00%	3.21%	2.05%	1.26%	3.57%	2.78%
<i>nadh3</i>	0.37%	0.39%	0.12%	0.14%	0.26%	0.28%
<i>nadh4</i>	2.55%	1.96%	1.30%	0.71%	2.17%	1.58%
<i>nadh4L</i>	0.57%	0.18%	0.16%	-0.23%	0.49%	0.10%
<i>nadh5</i>	3.50%	3.00%	1.59%	1.10%	3.21%	2.72%
<i>nadh6</i>	2.45%	1.74%	1.24%	0.54%	2.22%	1.51%
Gene	ALL-PROT		ALL-DNA		ALL-DNA12	
	Comp-total	Comp-var	Comp-total	Comp-var	Comp-total	Comp-var
<i>atp6</i>	-0.78%	-0.58%	-0.61%	-0.41%	-0.47%	-0.28%
<i>cox1</i>	-7.32%	-5.23%	-3.56%	-1.48%	-6.56%	-4.48%
<i>cox2</i>	-1.17%	-1.25%	-0.44%	-0.53%	-0.96%	-1.04%
<i>cox3</i>	-1.61%	-1.46%	-0.71%	-0.56%	-1.44%	-1.29%
<i>cytB</i>	-1.79%	-1.61%	-0.50%	-0.32%	-1.60%	-1.42%
<i>nadh1</i>	-0.49%	-0.59%	-0.57%	-0.67%	-0.77%	-0.88%
<i>nadh2</i>	4.11%	3.21%	2.20%	1.30%	3.73%	2.83%

Table 6  
Continued

Gene	ALL-PROT		ALL-DNA		ALL-DNA12	
	Comp-total	Comp-var	Comp-total	Comp-var	Comp-total	Comp-var
<i>nadh3</i>	0.20%	0.17%	−0.02%	−0.05%	0.17%	0.14%
<i>nadh4</i>	2.07%	1.93%	0.84%	0.70%	1.73%	1.59%
<i>nadh4L</i>	0.58%	0.31%	0.16%	−0.11%	0.51%	0.24%
<i>nadh5</i>	3.78%	3.31%	2.00%	1.53%	3.54%	3.07%
<i>nadh6</i>	2.41%	1.79%	1.21%	0.59%	2.14%	1.52%

outgroup choices used in recent studies of arthropod phylogeny: Mollusca, Annelida, Lophophorata, Nemertinea and Tardigrada (Giribet et al., 1996); Nematoda, Onychophora and Tardigrada (Giribet and Ribera, 2000); and Onychophora and Tardigrada (Giribet et al., 2001). We used four outgroup sets—Annelida, Mollusca and Nematoda (ALL), Annelida and Mollusca (ANMOL), Nematoda alone (NEM) and rooting within Arthropoda to Myriapoda (ARTH). We found that by including more distantly related outgroups the resolving power was worse. Comparing between datasets for the same data treatment, ALL performed the worst, had the longest trees, the worst tree statistics (Table 4) and the greatest lack of resolution (Tables 3 and 5). NEM was better than ANMOL as was expected since a monophyletic Ecdysozoa is well supported. The concern that high rates of gene rearrangement in nematodes preclude the use of their mitochondrial genomes in phylogenetic analysis is unfounded. With their inclusion, alignments and trees were shorter and the tree statistics better, suggesting lower homoplasy than those found in the ANMOL dataset. Of the nonarthropod outgroups currently available, Nematoda produces the best result but even this dataset only weakly resolves relationships within the Arthropoda. Much better resolution of collembolan relationships was found in the ARTH data treatments. Rooting to well supported, monophyletic groups such as Chelicerata or Myriapoda (but see Negrisol et al., 2004) appears to be a much better strategy for resolving pancrustacean relationships. This is to be expected since other arthropod groups are closer to Pancrustacea and thus more likely to retain the phylogenetic signal. Resolution of monophyletic arthropod lineages and their relative branching order is necessary to define monophyletic lineages to be used as outgroups to Hexapoda. Therefore, since mitochondrial genome data also seems to support the Ecdysozoa hypothesis, the sequencing of representatives of Onychophora, Tardigrada and Nemertinea is urgently required, as these are more closely related to arthropods than Nematoda. For the same reasons, ingroup sampling need to be increased in order to fill the gap between Collembola and the rest of Hexapoda, namely mitochondrial genomes of representatives of Protura

and Diplura have to be sequenced to seriously assess the position of Collembola.

#### Data transformations

A phylogenetic hypothesis is a product of both the empirical observations and the auxiliary assumptions used to analyze those observations (Grant and Kluge, 2003). Auxiliary assumptions include factors such as the optimality criteria used in the alignment or the analysis, partition strategies, and data treatment. In this context the transformation of data, such as the transformation of DNA sequence to amino acids, is an assumption the validity of which must be examined. None of the genes used in our, Nardi et al.'s (2003a,b) or any other phylogenetic analysis of mitochondrial genomes were sequenced as proteins. All were sequenced as DNA using conventional PCR and translated into putative amino acids as part of the analytical process. Furthermore, none of these genes evolved as amino acid sequences, mutations did not change one amino acid into another directly but rather nucleotide mutations changed the DNA sequence which presumably changed the amino acid sequence (the absence of mRNA editing is another assumption). This, combined with degeneracy of the genetic code, such that each amino acid is coded by 2–8 different triplet codons, means that shared amino acid changes cannot be simply assumed to be synapomorphic. The probability of homoplasy in amino acid sequence data is very high. These form the basis of the objections to the use of amino acid sequence in phylogenetic analysis (Simmons and Freudenstein, 2002; Simmons et al., 2002). Additionally, there is an assumption of a uniform genetic code across the taxa examined. GenBank currently recognizes seven different mitochondrial genetic codes (vertebrate, yeast, protozoan/coelenterate, invertebrate, echinoderm, ascidian and flatworm) and mitochondrial genomes submitted to GenBank as DNA sequences are translated according to one of these codes. However, it is known that minor divergences from a taxon's "normal" mitochondrial genetic code are not infrequent (Yokobori et al., 2001). Unless workers scrupulously test every taxon investigated to determine the exact genetic code for that

species, they will be, by necessity, making assumptions about the genetic code used. This is safest with smaller analyses, which consider members of a single phylum, but becomes increasingly hard to justify as more and more phyla are added to the analysis. Indeed, genetic code differences may be responsible for the apparently high divergence rates of particular taxa such as bivalve mollusks, lice or bees. Each of these assumptions, that amino acids are not homoplasious and that the translated sequences are accurate, add a significant level of uncertainty to the phylogenetic analyses performed upon them. This is reflected in the differences between PROT trees and DNA12 trees for each dataset. If these assumptions are approximately true these trees should be basically the same, but they are not.

### Alignment

Despite these criticisms, amino acid sequence-based analyses have a definite attraction—they are convenient and easier to perform than DNA sequence analysis due to the ease of aligning amino acids. We were unable to divorce our analysis from amino acids since we were unable to generate reasonable alignments from the DNA data alone. DNA alignments performed in Clustal had indels inserted which broke up almost all the reading frames, and trees inferred on these alignments conflicted strongly with previous analyses, even failing to resolve congeneric species (trees not shown). This is due to the highly noisy nature of mitochondrial sequence data over the time frames being examined. For the DNA trees the consistency index varied from 0.21 to 0.30, indicating how non-conserved these data are. By contrast the amino acid sequence was much more conservative, with CIs from 0.43 to 0.48. We therefore used amino acid alignments as a guide to the generation of DNA alignments, which ensured that all indels occurred as triplets maintaining the reading frame. The trees produced by this approach were overall much more congruent with previous phylogenetic hypotheses. However, we are concerned that this methodology, whilst an improvement on using the amino acid sequence alone, is not completely indicative of historical events. Are single base indels impossible? Maintaining intact reading frames is no doubt important, but do all the compensatory indels have to all occur at the same site? One could readily envision portions of the protein which were less critical to overall function in which frame shift mutations would be unproblematic, as long as compensatory mutations occurred upstream of more critical areas. Since such situations are being found more frequently even in length-invariant genes (Grant and D'Haese, 2004), how much more plausible is it for length-variable protein coding genes such as those found in mitochondrial

genomes? Currently amino acid and DNA based alignment strategies represent the extremes against which good arguments can be made. Amino acids are susceptible to homoplasious mutations and do not account for local frame shifts but produce better trees. DNA is closer to evolutionary reality but is highly noisy and almost certainly suffers from some degree of substitution saturation, especially at third codon positions. New approaches, which are more sophisticated than basic multiple alignment methods such as Clustal, are therefore needed, and this critical issue needs to be resolved as part of the process of determining how deep into phylogenetic history mitochondrial genomes can reach.

### Gene choice

Nardi et al. (2003a,b) *a priori* exclude 9 of the 13 mitochondrial genes, arguing that these genes have a lower phylogenetic utility due to the proportions of gaps and invariable sites in the alignment. The genes were ranked according to the proportion of sites which were invariant (a good thing) and the proportion which were gaps (a bad thing), which resulted in the choice of *cox1*, *cox2*, *cox3* and *cytB* as “sufficiently reliable” (that is > 25% invariant sites and < 15% gap sites, with these figures apparently chosen arbitrarily) to analyze further. Statistics on individual genes for each dataset and treatment used in this study are shown in Table 7. Despite using the same alignment parameters as Nardi et al. (2003a,b), the only genes that met this standard in our analysis were *cox1*, *cox2*, *cox3* and *cytB* in the ARTH-PROT analysis, only *cox1* in the ANMOL-PROT and NEM-PROT analysis and no genes in the ALL-PROT analysis. In the DNA based datasets the only genes which met the standard were *cox1*, *cox2*, *cox3*, *CytB* and *ndh1*, and then only in the ARTH-DNA and ARTH-DNA12 datasets. If one accepts this reasoning for ranking gene quality, it is possible that one would also apply less rigorous standards to DNA alignments since they are less conservative than the amino acid alignments and always have fewer invariant sites than the comparable amino acid alignments (Table 6). Gap proportion does not vary between the data treatments, since we generated the alignments by aligning amino acid translations. If, however, the standard proposed by Nardi et al. (2003a) were to be applied universally virtually no gene would ever be used to infer phylogenies. Furthermore, both parameters, invariant sites and gaps, are highly susceptible to taxon choice. The inclusion of highly divergent taxa or a wider taxonomic range will alter these statistics. As gaps can be an artifact of taxon selection we do not feel that they are really an independent variable, which can be used for the *a priori* exclusion of data from an

Table 7

Proportion of Gap and Constant sites for each of the datasets and treatments. Note due to the way the alignment was constructed the proportion of gaps for each treatment is the same whereas the proportion of constant sites varies

ARTH	Gaps	Const-PROT	Const-DNA12	Const-DNAALL
<i>atp6</i>	20.60%	18.88%	18.09%	26.07%
<i>atp8</i>	66.67%	21.215	22.89%	25.37%
<i>cox1</i>	2.30%	48.75%	38.31%	56.71%
<i>cox2</i>	7.95%	22.59%	23.06%	33.13%
<i>cox3</i>	6.34%	23.88%	22.80%	33.46%
<i>cytB</i>	8.25%	26.80%	24.42%	35.60%
<i>nadh1</i>	16.86%	17.43%	19.75%	27.78%
<i>nadh2</i>	29.64%	8.59%	9.21%	12.57%
<i>nadh3</i>	26.51%	18.18%	20.25%	25.76%
<i>nadh4</i>	25.16%	14.50%	15.46%	21.28%
<i>nadh4L</i>	30.19%	3.77%	9.03%	12.15%
<i>nadh5</i>	19.74%	14.93%	14.57%	20.86%
<i>nadh6</i>	51.37%	1.09%	5.80%	7.88%
ANMOL	Gaps	Const-PROT	Const-DNA12	Const-DNAALL
<i>atp6</i>	36.03%	12.55%	13.98%	18.75%
<i>atp8</i>	73.44%	9.38%	11.79%	14.62%
<i>cox1</i>	5.92%	39.12%	32.44%	47.90%
<i>cox2</i>	11.72%	17.57%	17.92%	25.63%
<i>cox3</i>	11.11%	20.79%	21.07%	29.29%
<i>cytB</i>	10.66%	23.10%	21.60%	30.89%
<i>nadh1</i>	22.60%	15.25%	16.90%	23.52%
<i>nadh2</i>	42.86%	7.55%	8.43%	10.75%
<i>nadh3</i>	38.10%	14.97%	18.69%	21.62%
<i>nadh4</i>	29.11%	15.25%	15.68%	19.96%
<i>nadh4L</i>	40.54%	6.31%	9.52%	11.61%
<i>nadh5</i>	25.46%	11.48%	11.19%	16.03%
<i>nadh6</i>	46.07%	1.05%	5.03%	6.77%
NEM	Gaps	Const-PROT	Const-DNA12	Const-DNAALL
<i>atp6</i>	44.04%	18.41%	20.62%	22.66%
<i>cox1</i>	7.79%	32.43%	29.29%	41.95%
<i>cox2</i>	11.07%	13.11%	16.05%	22.45%
<i>cox3</i>	9.33%	11.57%	13.75%	19.89%
<i>cytB</i>	10.31%	15.98%	16.54%	23.91%
<i>nadh1</i>	21.37%	13.11%	15.91%	21.88%
<i>nadh2</i>	39.62%	7.28%	8.69%	11.02%
<i>nadh3</i>	31.58%	15.04%	17.41%	21.27%
<i>nadh4</i>	26.65%	10.23%	10.99%	14.89%
<i>nadh4L</i>	33.33%	1.90%	8.18%	10.85%
<i>nadh5</i>	26.96%	11.76%	11.15%	15.09%
<i>nadh6</i>	48.39%	1.61%	4.99%	6.42%
ALL	Gaps	Const-PROT	Const-DNA12	Const-DNAALL
<i>atp6</i>	50.00%	15.60%	17.79%	19.26%
<i>cox1</i>	11.03%	26.40%	25.57%	36.37%
<i>cox2</i>	15.98%	11.89%	13.47%	18.78%
<i>cox3</i>	16.03%	14.98%	17.01%	21.88%
<i>cytB</i>	12.44%	14.72%	16.03%	22.53%
<i>nadh1</i>	24.79%	12.11%	14.14%	19.24%
<i>nadh2</i>	46.09%	4.58%	5.9%	7.12%
<i>nadh3</i>	39.31%	12.41%	16.67%	18.84%
<i>nadh4</i>	36.06%	12.15%	12.52%	16.00%
<i>nadh4L</i>	44.14%	4.50%	8.04%	9.82%
<i>nadh5</i>	32.25%	10.42%	9.86%	13.25%
<i>nadh6</i>	51.85%	1.59%	4.39%	5.53%

analysis. Questions of gene choice need to be justified on the basis of metrics that show levels of incongruence between datasets or the demonstration that different genes are showing different evolutionary histories. Simply stating that some genes are more variable and thus “noisier” than others and concluding that they are therefore less reliable is not justified on either theoretical or empirical grounds. Simulation studies have shown that simple noise in phylogenetic datasets is unlikely to overwhelm the signal even in extreme circumstances (Wenzel and Siddall, 1999). Total evidence studies have shown that additional genes improve support for phylogenetic hypotheses, even when the additional genes individually do not support that combined hypothesis (Gatesy et al., 1999). Furthermore neither of these conclusions is restricted to one optimality criterion and forms the basis for the current general acceptance of total evidence approaches to phylogenetic study (Kluge, 1989).

In the present study we found that there is no reason for the *a priori* exclusion any of the mitochondrial genes from phylogenetic analysis. Given that both simulation and empirical studies support total evidence approaches, the only remaining justification for gene exclusion would be financial, if resources were limited and the signal in the mitochondrial genome were concentrated in a few genes it might make sense to concentrate sequencing efforts on those genes to the exclusion of others. Our analysis of partitioned support across all the data treatments showed that no gene dominates support in the mitochondrial genome of arthropods (Table 6). The distribution of branch support was more or less proportional to the size of the gene, suggesting that the signal in this dataset is evenly spread throughout the genome, and therefore the addition of more genes adds more signal to the analysis and improves resolution. Conversely, the notion that some genes “mislead” the analysis with excessive levels of noise was not supported. Whilst mt genomes as a whole have high levels of noise, as is evident in the low consistency index values in all analyses, this did not translate into incompatibilities between the different genes in a combined analysis. Thus, given that additional genes improve the analysis and are not incongruent with each other there is no reason to use only four of the 13 mitochondrial genes in a phylogenetic analysis. Indeed, restricting the analysis to just the four genes used by Nardi et al. (2003a,b) results in the loss of many nodes which were resolved in analyses of all 13 and fails to resolve any additional nodes which were not recovered in analyses of all 13. Excluding data without a justified reason (such as, perhaps, demonstrated incongruence) is a worse approach than total evidence for the analysis of mt genomes.

### *Optimality criteria*

The sensitivity of a dataset to the optimality criteria used in tree reconstruction is one of the oldest sources of debate in phylogenetics. Fortunately it does not seem to be a factor in the current analysis. Parsimony and Bayesian approaches yielded very similar results for each of the dataset/treatment combinations and both were equally poor at unambiguously determining the sister group of the Collembolans. Bayesian posteriors tended to be less conservative indicators of branch support than bootstrapping. This finding has been made previously, based on both simulations (Cummings et al., 2003; Erixon et al., 2003) and empirical data (Leaché and Reeder, 2002; Whittingham et al., 2002). Although most of these studies compared support values within a common optimality criterion (e.g., Bayesian posteriors versus Likelihood bootstraps) rather than across criteria (Bayesian versus Parsimony) as we have done here. The radically different means of estimating support in the two approaches is possibly responsible for the wide divergences between posteriors and bootstraps that we observed. For datasets with high levels of noise, such as mitochondrial genomes, bootstrap resampling is less likely to recover all nodes at each replicate since the signal in the data set is stretched thinly. Thus bootstrap values will be lower than comparable datasets (equivalent numbers of informative sites) with lower levels of noise. In contrast, since Bayesian analyses utilize the whole dataset at each generation a strongly favored topology will receive high posterior probabilities, regardless of noise. Bayesian analysis may therefore be a solution to estimating branch support for datasets with low signal to noise ratios.

More interesting in the current context is the decline in posterior probabilities from the PROT to DNA12 to DNAALL data treatments, an effect that has not been previously demonstrated. Unlike in the parsimony analyses of different treatments, where different topologies were found with significant bootstrap support, the Bayesian analyses are more consistent topologically but varied at the level of posterior probabilities. This can be interpreted two ways: using amino acids corrects for the saturation of nucleotide substitutions and therefore the higher posteriors demonstrate the lower homoplasy of this data type; or, inflated posteriors may lead workers to place too much confidence in the results of amino acid based analyses. If the first proposition were true with our data, then the DNA12 results should match those from PROT as the vast majority of substitutions will occur at the third codon position and correspondingly the chances of saturation at these sites is much higher. This is not the case. The DNA12 results do not perfectly match those for the PROT topologically for corresponding datasets. Most critically, the placement of *Collembola* varied between different data treatments

just as much with Bayesian analysis as it did with parsimony. Effects other than simple substitution saturation differentiate the DNA based analyses from the amino acid based analyses, suggesting that the problems of homoplasious change to amino acids are real and are not totally compensated for by the use of model-based approaches such as Bayesian analysis. We caution against the exclusive use of amino acid based analyses within a Bayesian framework, as the posterior probabilities so estimated will almost certainly be inflated.

## Conclusions

In conclusion, we regard the basal relationships of the arthropods and, in particular, the sister-group of Collembola, to be unresolved at this time. Existing mitochondrial genome data alone is insufficient to resolve this issue and this dataset appears to be highly vulnerable to taxon selection, outgroup choice, data manipulations and gene exclusion but apparently not phylogenetic reconstruction methodology. Given that the topology is so variable, depending on such a wide variety of factors, rejection of the null hypothesis of hexapod monophyly (which is supported by morphology) is not justified at this time. However as a monophyletic Hexapoda was not recovered by these analyses it is clear that the relationships of the Collembola require additional attention both from morphological and molecular investigations, and the mitochondrial genomes of additional key taxa (Protura and Diplura) need to be sequenced before this issue can be resolved. Mitochondrial genomes are not a “magic bullet”, which will solve any phylogenetic problem. Indeed the limits and applicability of these data have not yet been thoroughly explored. The most glaring limitation on their use currently is the lack of taxa for which genomes are available. Fewer than 60 arthropod mitochondrial genomes are currently available and those are highly concentrated in a few relatively derived groups such as Diptera, Ixodida and Decapoda. Will mitochondrial genomes be able to resolve deep level relationships within the arthropods? The answer is as yet unknown but the indications are that each additional taxon improves our understanding of this data type and its use in phylogenetics. Targeted sequencing of key taxa pertinent to the phylogenetic questions being posed is the surest way forward, coupled with rigorous testing of how these data behave in phylogenetic analyses.

## Acknowledgments

The authors would like to thank Michael Rix (UQ) for collecting the Archaeognatha specimens used in this analysis, Cath Covacin, Renfu Shao and Anna Murrell

(UQ) for assisting S.L.C. in sequencing methodology and Katharina Dittmar de la Cruz, Heath Ogden, Matt Terry and Gavin Svenson (BYU) for discussion of how best to apply phylogenetic methodology to mitochondrial genomes. This work was supported by NSF grant DEB0120718.

## References

- Aguinaldo, A.M.A., Turbeville, J.M., Lindford, L.S., Rivera, M.C., Garey, J.R., Raff, R.A., Lake, J.A., 1997. Evidence for a clade of nematodes, arthropods and other molting animals. *Nature*, 387, 489–493.
- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* 25, 3389–3402.
- Bae, J.S., Kim, I., Sohn, H.D., Jin, B.R., 2004. The mitochondrial genome of the firefly, *Pyrocoelia rufa*: complete DNA sequence, genome organisation, and phylogenetic analysis with other insects. *Mol. Phylogen. Evol.* 32, 978–985.
- Baker, R.H., DeSalle, R., 1997. Multiple sources of character information and the phylogeny of Hawaiian Drosophilids. *Syst. Biol.* 46, 654–673.
- Beard, C.B., Hamm, D.M., Collins, F.H., 1993. The mitochondrial genome of the mosquito *Anopheles gambiae*: DNA sequence, genome organization, and comparisons with mitochondrial sequences of other insects. *Insect Mol. Biol.* 2, 103–124.
- Bitsch, J., Bitsch, C., Bourgoin, T., D’Haese, C., 2004. The phylogenetic position of basal hexapod lineages: morphological data contradict molecular data. *Syst. Entomol.* 29, 433–440.
- Black, W.C., Roehrdanz, R.L., 1998. Mitochondrial gene order is not conserved in arthropods: prostriate and metastriate tick mitochondrial genomes. *Mol. Biol. Evol.* 15, 1772–1778.
- Boone, D.R., Castenholz, R.W., 2001. *Bergey’s Manual of Systematic Bacteriology*, 2nd edn. Springer, New York, NY.
- Boore, J.L., 1999. Animal mitochondrial genomes. *Nucl. Acids Res.* 27, 1767–1780.
- Boore, J.L., Brown, W.M., 1994. Complete DNA sequence of the mitochondrial genome of the black chiton, *Katharina tunicata*. *Genetics*, 138, 423–443.
- Boore, J.L., Brown, W.M., 1995. Complete sequence of the mitochondrial DNA of the annelid worm *Lumbricus Terrestris*. *Genetics*, 141, 305–319.
- Boore, J.L., Brown, W.M., 2000. Mitochondrial genomes of *Galathealinum*, *Helobdella*, and *Platynereis*: sequence and gene arrangement comparisons indicate that Pogonophora is not a phylum and Annelida and Arthropoda are not sister taxa. *Mol. Biol. Evol.* 17, 87–106.
- Cary, D.O., Wolstenholme, D.R., 1985. The mitochondrial DNA molecular of *Drosophila yakuba*: nucleotide sequence, gene organization, and genetic code. *J. Mol. Evol.* 22, 252–271.
- Corliss, J.O., 1979. *The Ciliated Protozoa: Characterisation, Classification, and Guide to the Literature*, 2nd edn. Pergamon Press, Oxford, UK.
- Crease, T.J., 1999. The complete sequence of the mitochondrial genome of *Daphnia pulex* (Cladocera: Crustacea). *Gene*, 233, 89–99.
- Crozier, R.H., Crozier, Y.C., 1993. The mitochondrial genome of the honeybee *Apis mellifera*: complete sequence and genome organisation. *Genetics*, 133, 97–117.
- Cummings, M.P., Handley, S.A., Myers, D.S., Reed, D.L., Rokas, A., Winka, K., 2003. Comparing bootstrap and posterior probability values in the four-taxon case. *Syst. Biol.* 52, 477–487.

- D'Haese, C., 2002. Were the first springtails semi-aquatic? A phylogenetic approach by means of 28S rDNA and Optimization Alignment. *Proc. R. Soc. Lond. B*, 269, 1143–1151.
- D'Haese, C., 2003. Morphological appraisal of Collembola phylogeny with special emphasis on Poduromorpha and a test of the aquatic origin hypothesis. *Zool. Scrip.* 32, 563–586.
- Delsuc, F., Phillips, M.J., Penny, D., 2003. Comment on “Hexapod origins: Monophyletic or Paraphyletic?” *Science*, 301, 1482d.
- Dotson, E.M., Beard, C.B., 2001. Sequence and organization of the mitochondrial genome of the Chagas disease vector, *Triatoma dimidiata*. *Insect Mol. Biol.* 10, 205–215.
- Edgecombe, G.D., Wilson, G.D.F., Colgan, D.J., Gray, M.R., Cassis, G., 2000. Arthropod cladistics: combined analysis of Histone H3 and U2 snRNA sequences and morphology. *Cladistics*, 16, 155–203.
- Erixon, P., Sennblad, B., Britton, T., Oxelman, B., 2003. Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. *Syst. Biol.* 52, 665–673.
- Field, K.G., Olsen, G.J., Lane, D.J., Giovannoni, S.J., Ghiselin, M.T., Raff, E.C., Pace, N.R., Raff, R.A., 1988. Molecular Phylogeny of the animal kingdom. *Science*, 239, 748–753.
- Flook, P.K., Rowell, C.H., Gellissen, G., 1995. The sequence, organization, and evolution of the *Locusta migratoria* mitochondrial genome. *J. Mol. Evol.* 41, 928–941.
- Friedrich, M., Muqim, N., 2003. Sequence and phylogenetic analysis of the complete mitochondrial genome of the flour beetle *Tribolium castaneum*. *Mol. Phylogenet. Evol.* 26, 502–512.
- Gatesy, J., O'Grady, P., Baker, R., 1999. Corroboration among data sets in simultaneous analysis: Hidden support for phylogenetic relationships among higher level arthropod taxa. *Cladistics*, 15, 271–313.
- Giribet, G., 2002. Relationships among metazoan phyla as inferred from 18S rRNA sequence data: a methodological approach. In: DeSalle, R., Giribet, G., Wheeler, W.C. (Eds.) *Molecular Systematics and Evolution: Theory and Practice*. Birkhauser-Verlag, Basel, pp. 85–101.
- Giribet, G., Carranza, S., Baguna, J., Riutort, M., Ribera, C., 1996. First molecular evidence for the existence of a Tardigrada + Arthropoda clade. *Mol. Biol. Evol.* 13, 76–84.
- Giribet, G., Edgecombe, G.D., Wheeler, W.C., 2001. Arthropod phylogeny based on eight molecular loci and morphology. *Nature*, 413, 157–161.
- Giribet, G., Ribera, C., 1998. The position of arthropods in the animal kingdom: a search for a reliable outgroup for internal arthropod phylogeny. *Mol. Phylogenet. Evol.* 9, 481–488.
- Giribet, G., Ribera, C., 2000. A review of arthropod phylogeny: new data based on ribosomal DNA sequences and direct character optimization. *Cladistics*, 16, 204–231.
- Giribet, G., Wheeler, W.C., 1999. The position of arthropods in the animal kingdom: Ecdysozoa, islands, trees and the ‘parsimony ratchet’. *Mol. Phylogenet. Evol.* 13, 619–623.
- Grande, C., Templado, J., Cervera, J.L., Zardoya, R., 2002. The Complete Mitochondrial Genome of the Nudibranch *Roboastra europaea* (Mollusca: Gastropoda) Supports the Monophyly of Opisthobranchs. *Mol. Biol. Evol.* 19, 1672–1685.
- Grant, T., D'Haese, C.A., 2004. Insertions and deletions in the evolution of equal-length DNA fragments. In: Stevenson, D.W.M. (Ed.), *Abstracts of the 22th Annual Meeting of the Willi Hennig Society*. *Cladistics* 20, 84.
- Grant, T., Kluge, A.G., 2003. Data exploration in phylogenetic inference: scientific, heuristic or neither. *Cladistics*, 19, 379–418.
- Hatzoglou, E., Rodakis, G.C., Lecanidou, R., 1995. Complete sequence and gene organization of the mitochondrial genome of the land snail *Albinaria coerulea*. *Genetics*, 140, 1353–1366.
- Hickerson, M.J., Cunningham, C.W., 2000. Dramatic mitochondrial gene rearrangements in the hermit crab *Pagurus longicarpus* (Crustacea, anomura). *Mol. Biol. Evol.* 17, 639–644.
- Hu, M., Chilton, N.B., Gasser, R.B., 2002. The mitochondrial genomes of the human hookworms, *Ancylostoma duodenale* and *Necator americanus* (Nematoda: Secernentea). *Int. J. Parasitol.* 32, 145–158.
- Hu, M., Chilton, N.B., Gasser, R.B., 2003. The mitochondrial genome of *Strongyloides stercoralis* (Nematoda) – idiosyncratic gene order and evolutionary implications. *Int. J. Parasitol.* 33, 1393–1408.
- Huelsenbeck, J.P., Ronquist, F.R., 2001. MrBayes: Bayesian inference of phylogeny. *Biometrics*, 17, 754–755.
- Janke, A., Gemmel, N.J., Feldmaier-Fuchs, G., von Haeseler, A., Pääbo, S., 1996. The mitochondrial genome of a monotreme – the platypus (*Ornithorhynchus anatinus*). *J. Mol. Evol.* 42, 153–159.
- Janke, A., Magnell, O., Weiczorek, G., Westerman, M., Arnason, U., 2002. Phylogenetic analysis of 18S rRNA and the mitochondrial genomes of the wombat, *Vombatus ursinus*, and the spiny anteater, *Tachyglossus aculeatus*: increased support for the Marsupionta hypothesis. *J. Mol. Evol.* 54, 71–80.
- Janke, A., Xu, X., Arnason, U., 1997. The complete mitochondrial genome of the wallaroo (*Macropus robustus*) and the phylogenetic relationship among Monotremata, Marsupialia and Eutheria. *Proc. Natl. Acad. Sci. USA*, 94, 1276–1281.
- Junqueira, A.C.M., Lessinger, A.C., Torres, T.T., da Silva, F.R., Vettore, A.L., Arruda, P., Azeredo Espin A.M.L., 2004. The mitochondrial genome of the blowfly *Chrysomya chloropyga* (Diptera: Calliphoridae). *Gene* 339, 7–15.
- Keddie, E.M., Higazi, T., Unnasch, T.R., 1998. The mitochondrial genome of *Onchocerca volvulus*: sequence, structure and phylogenetic analysis. *Mol. Biochem. Parasitol.* 95, 111–127.
- Kluge, A.G., 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae: Serpentes). *Syst. Zool.* 38, 7–25.
- Kurabayashi, A., Ueshima, R., 2000. Complete sequence of the mitochondrial DNA of the primitive opisthobranch gastropod *Pupa strigosa*: systematic implication of the genome organization. *Mol. Biol. Evol.* 17, 266–277.
- Lake, J.A., 1990. Origin of the Metazoa. *Proc. Natl. Acad. Sci. USA*, 87, 763–766.
- Lavrov, D.V., Boore, J.L., Brown, W.M., 2000a. The complete mitochondrial DNA sequence of the horseshoe crab *Limulus polyphemus*. *Mol. Biol. Evol.* 17, 813–824.
- Lavrov, D.V., Boore, J.L., Brown, W.M., 2002. Complete mtDNA sequences of two millipedes suggest a new model for mitochondrial gene rearrangements: Duplication and non-random loss. *Mol. Biol. Evol.* 19, 163–169.
- Lavrov, D.V., Brown, W.M., 2001. *Trichinella spiralis* mtDNA. A nematode mitochondrial genome that encodes a putative ATP8 and normally structured tRNAs and has a gene arrangement relatable to those of coelomate metazoans. *Genetics*, 157, 621–637.
- Lavrov, D.V., Brown, W.M., Boore, J.L., 2000b. A novel type of RNA editing occurs in the mitochondrial tRNAs of the centipede *Lithobius forficatus*. *Proc. Natl. Acad. Sci. USA*, 97, 13738–13742.
- Lavrov, D.V., Brown, W.M., Boore, J.L., 2004. Phylogenetic position of the Pentastomatida and (pan)crustacean relationships. *Proc. R. Soc. London B*, 271, 537–544.
- Leaché, A.D., Reeder, T.W., 2002. Molecular systematics of the eastern fence lizard *Sceloporus undulatus*: a comparison of parsimony, likelihood and Bayesian approaches. *Syst. Biol.* 51, 44–68.
- Lee, M.S.Y., Hugall, A.F., 2003. Partitioned likelihood support and the evaluation of data set conflict. *Syst. Biol.* 52, 15–22.
- Lee, M.H., Shroff, R., Cooper, S.J.B., Hope, R., 1999. Evolution and molecular characterisation of a  $\beta$ -globin gene from the Australian echidna, *Tachyglossus aculeatus* (Monotremata). *Mol. Phylogenet. Evol.* 12, 205–214.
- Lessinger, A.C., Martins Junqueira, A.C., Lemos, T.A., Kemper, E.L., Da Silva, F.R., Vettore, A.L., Arruda, P., Azeredo-Espin, A.M.L.,

2000. The mitochondrial genome of the primary screwworm fly *Cochliomyia hominivorax* (Diptera: Calliphoridae). *Insect Mol. Biol.* 9, 521–529.
- Lewis, D.L., Farr, C.L., Kaguni, L.S., 1995. *Drosophila melanogaster* mitochondrial DNA: completion of the nucleotide sequence and evolutionary comparisons. *Insect Mol. Biol.* 4, 263–278.
- Lou, Z.Z., Kielan-Jawrowska, Z., Cifelli, R., 2002. In quest for a phylogeny of Mesozoic mammals. *Acta Palaeont. Polonica*, 47, 1–78.
- Lowe, T.M., Eddy, S.R., 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucl. Acids Res.* 25, 955–964.
- Machida, R.J., Miya, M.U., Nishida, M., Nishida, S., 2002. Complete mitochondrial DNA sequence of *Tigriopus japonicus* (Crustacea: Copepoda). *Mar. Biotechnol.* 4, 406–417.
- Maddison, W., Maddison, D., 2003. MacClade, Version 4.06. Sinauer Associates, Sunderland, MA.
- Masta, S., Boore, J.L., 2004. The complete mitochondrial genome sequence of the spider *Habronattus oregonensis* reveals rearranged and extremely truncated tRNAs. *Mol. Biol. Evol.* 21, 893–902.
- Mitchell, S.E., Cockburn, A.F., Seawright, J.A., 1993. The mitochondrial genome of *Anopheles quadrimaculatus* species A: complete nucleotide sequence and gene organization. *Genome*, 36, 1058–1073.
- Nardi, F., Carapelli, A., Fanciulli, P.P., Dallai, R., Frati, F., 2001. The Complete mitochondrial DNA sequence of the basal hexapod *Tetradontophora bielensis*: evidence for heteroplasmy and tRNA translocations. *Mol. Biol. Evol.* 18, 1293–1304.
- Nardi, F., Spinsanti, G., Boore, J.L., Carapelli, A., Dallai, R., Frati, F., 2003a. Hexapod origins: monophyletic or paraphyletic? *Science*, 299, 1887–1889.
- Nardi, F., Spinsanti, G., Boore, J.L., Carapelli, A., Dallai, R., Frati, F., 2003b. Response to comment on “Hexapod Origins: Monophyletic or Paraphyletic?” *Science*, 301, 1482e.
- Navajas, M., Le Conte, Y., Solignac, M., Cros-Arteil, S., Cornuet, J.M., 2002. The complete sequence of the mitochondrial genome of the honeybee ectoparasite mite *Varroa destructor* (Acari: Mesostigmata). *Mol. Biol. Evol.* 19, 2313–2317.
- Negrisololo, E., Minelli, A., Valle, G., 2004. The mitochondrial genome of the house centipede *Scutigera* and the monophyly versus paraphyly of myriapods. *Mol. Biol. Evol.* 21, 770–780.
- Okimoto, R., Macfarlane, J.L., Clary, D.O., Wolstenholme, D.R., 1992. The mitochondrial genomes of two nematodes, *Caenorhabditis elegans* and *Ascaris suum*. *Genetics*, 130(3), 471–498 (1992).
- Perez, M.L., Valverde, J.R., Batuecas, B., Amat, F., Marco, R., Garesse, R., 1994. Speciation in the *Artemia* genus: mitochondrial DNA analysis of bisexual and parthenogenetic brine shrimps. *J. Mol. Evol.* 38(2), 156–168.
- Phillips, M.J., Penny, D., 2003. The root of the mammalian tree inferred from whole mitochondrial genomes. *Mol. Phylogenet. Evol.* 28, 171–185.
- Posada, D., Crandall, K.A., 1998. ModelTest: Testing the best-fit model of nucleotide substitution. *Bioinform.* 14, 817–818.
- Pritchard, Gordon and McKee, M.H., Pike, E.M., Scrimgeour, G.J., Zloty, J., 1993. Did the first insects live in water or air? *Biol. J. Linn. Soc.* 49, 31–44.
- Saccone, C., De Giorgi, C., Gissi, C., Pesole, G., Reyes, A., 1999. Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. *Gene*, 238, 195–209.
- Serb, J.M., Lydeard, C., 2003. Complete mt DNA sequence of the North American freshwater mussel, *Lampsilis ornata* (Unionidae): An example of evolution and phylogenetic utility of mitochondrial genome organisation in Bivalvia (Mollusca). *Mol. Biol. Evol.* 20, 1854–1866.
- Shao, R., Aoki, Y., Mitani, H., Tabuchi, N., Barker, S.C., Fukunaga, M., 2004. The mitochondrial genomes soft ticks have an arrangement of genes that has remained unchanged for over 400 millions of years. *Insect Mol. Biol.* 13, 219–224.
- Shao, R., Barker, S.C., 2003. The highly rearranged mitochondrial genome of the plague thrips, *Thrips imuginis* (Insecta: Thysanoptera): convergence of two novel gene boundaries and an extraordinary arrangement of rRNA. *Mol. Biol. Evol.* 20, 362–370.
- Shao, R., Campbell, N.J.H., Barker, S.C., 2001. Numerous gene rearrangements in the mitochondrial genome of the wallaby louse, *Heterodoxus macropus* (Phthiraptera). *Mol. Biol. Evol.* 18, 858–865.
- Shao, R., Dowton, M., Murrell, A., Barker, S.C., 2003. Rates of gene rearrangement and nucleotide substitution are correlated in the mitochondrial genomes of insects. *Mol. Biol. Evol.* 20, 1612–1619.
- Sibley, C.G., Ahlquist, J.E., 1990. *Phylogeny and Classification of Birds: a Study in Molecular Evolution*. Yale University Press, New Haven, CT.
- Sibley, C.G., Ahlquist, J.E., Monroe, B.L., 1988. A classification of the living birds of the world based on DNA-hybridization studies. *Auk*, 105, 409–423.
- Simmons, M.P., Freudenstein, J.V., 2002. Artifacts of coding amino acids and other composite characters for phylogenetic analysis. *Cladistics*, 18, 354–365.
- Simmons, M.P., Ochoterena, H., Freudenstein, J.V., 2002. Conflict between amino acid and nucleotide characters. *Cladistics*, 18, 200–206.
- Simon, C.F., Frati, A., Beckenbach, B., 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. *Ann. Entom. Soc. Am.* 87, 651–701.
- Skerratt, L.F., Campbell, N.J.H., Murrell, A., Walton, S., Kemp, D., Barker, S.C., 2002. The mitochondrial 12S gene is a suitable marker of populations of *Sarcoptes scabiei* from wombats, dogs and humans in Australia. *Parasitol. Res.* 88, 376–379.
- Sorenson, M.D., 1999. TreeRot, Version 2. Boston University, Boston, MA.
- Spanos, L., Koutroumbas, G., Kotsyfakis, M., Louis, C., 2000. The mitochondrial genome of the mediterranean fruit fly, *Ceratitis capitata*. *Insect Mol. Biol.* 9, 139–144.
- Spears, T., Abele, L.G., 1997. Crustacean phylogeny inferred from 18S rDNA. In: Fortey, R.A., Thomas, R.H. (Eds.), *Arthropod Relationships*. Chapman & Hall, London, UK, pp. 169–187.
- Stewart, J.B., Beckenbach, A.T., 2003. Phylogenetic and genomic analysis of the complete mitochondrial DNA sequence of the spotted asparagus beetle *Crioceris duodecimpunctata*. *Mol. Phylogenet. Evol.* 26, 513–526.
- Sturm, H., 1980. Redescription of *Nesomachilis* (Archaeognatha: Meinertellidae), with descriptions of new species from the Australian region. *NZ J. Zool.* 7, 533–550.
- Swofford, D.L., 2002. PAUP\* Phylogenetic Analysis Using Parsimony (\*and Other Methods), Version 4. Sinauer Associates, Sunderland, MA.
- Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. Clustal W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucl. Acids. Res.* 22, 4673–4680.
- Tomita, K., Yokobori Si, S., Oshima, T., Ueda, T., Watanabe, K., 2002. The cephalopod *Loligo bleekeri* mitochondrial genome: multiplied noncoding regions and transposition of tRNA genes. *J. Mol. Evol.* 54, 486–500.
- Umetsu, K., Iwabuchi, N., Yuasa, I., Saitou, N., Clark, P.F., Boxshall, G., Osawa, M., Igarashi, K., 2002. Complete mitochondrial DNA sequence of a tadpole shrimp (*Triops cancriformis*) and analysis of museum samples. *Electrophoresis*, 23, 4080–4084.
- Wenzel, J.W., Siddall, M.E., 1999. Noise. *Cladistics*, 15, 51–64.

- Wheeler, W.C., 1990. Nucleic acid sequence phylogeny and random outgroups. *Cladistics*, 6, 363–367.
- Wheeler, W.C., Whiting, M.F., Wheeler, Q.D., Carpenter, J.M., 2001. The phylogeny of the extant hexapod orders. *Cladistics*, 17, 113–169.
- Whittingham, L.A., Slikas, B., Winker, D.W., Sheldon, F.H., 2002. Phylogeny of the tree swallow genus *Tachycineta* (Aves: Hirundinidae) by Bayesian analysis of mitochondrial DNA sequences. *Mol. Phylogenet. Evol.* 22, 430–441.
- Wilson, K., Neville, V., Ballment, E., Benzie, J., 2000. The complete sequence of the mitochondrial genome of the crustacean *Penaeus monodon*: are malacostracan crustaceans more closely related to insects than to branchiopods? *Mol. Biol. Evol.* 17, 863–874.
- Yamauchi, M., Miya, M., Nishida, M., 2002. Complete mitochondrial DNA sequence of the Japanese spiny lobster, *Panulirus japonicus* (Crustacea: Decapoda). *Gene*, 295, 89–96.
- Yamauchi, M.M., Miya, M.U., Nishida, M., 2003. Complete mitochondrial DNA sequence of the swimming crab, *Portunus trituberculatus* (Crustacea: Decapoda: Brachyura). *Gene*, 311, 129–135.
- Yamazaki, N., Ueshima, R., Terrett, J.A., Yokobori, S., Kaifu, M., Segawa, R., Kobayashi, T., Numachi, K., Ueda, T., Nishikawa, K., Watanabe, K., Thomas, R.H., 1997. Evolution of pulmonate gastropod mitochondrial genomes: comparisons of gene organizations of *Euhadra*, *Cepaea* and *Albinaria* and implications of unusual tRNA secondary structures. *Genetics*, 145, 749–758.
- Yokobori, S., Suzuki, T., Watanabe, K., 2001. Genetic code variations in mitochondria: tRNA as a major determinant of genetic code plasticity. *J. Mol. Biol.* 53, 314–326.
- Yukuhiro, K., Sezutsu, H., Itoh, H., Shimizu, K., Banno, Y., 2002. Significant levels of sequence and gene rearrangements have occurred between the mitochondrial genomes of the wild mulberry silkworm, *Bombyx mandarina* and its close relative, the domesticated silkworm, *Bombyx mori*. *Mol. Biol. Evol.* 19, 1385–1389.
- Zrzavy, J., Mihulka, S., Kepka, P., Bezdek, A., Tietz, D., 1998. Phylogeny of the Metazoa based on morphology and 18S ribosomal evidence. *Cladistics*, 14, 249–285.