Syst. Biol. 57(6):920–938, 2008
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150802570791

Resolving Arthropod Phylogeny: Exploring Phylogenetic Signal within 41 kb of Protein-Coding Nuclear Gene Sequence

Jerome C. Regier,[1] Jeffrey W. Shultz,[2] Austen R. D. Ganley,[3,6] April Hussey,[1] Diane Shi,[1] Bernard Ball,[3] Andreas Zwick,[1] Jason E. Stajich,[3,7] Michael P. Cummings,[4] Joel W. Martin,[5] and Clifford W. Cunningham[3]

[1]*Center for Biosystems Research, University of Maryland Biotechnology Institute, College Park, Maryland 20742, USA; E-mail: regier@umbi.umd.edu (J.C.R.).*
[2]*Department of Entomology, University of Maryland, College Park, Maryland 20742, USA*
[3]*Department of Biology, Duke University, Durham, North Carolina 27708, USA*
[4]*Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland 20742, USA*
[5]*Natural History Museum of Los Angeles County, Los Angeles, California 90007, USA*
[6]*Current Address: Division of Cytogenetics, National Institute of Genetics, Mishima, Shizuoka, 411-8540, Japan*
[7]*Current Address: Department of Plant and Microbial Biology, University of California, Berkeley, California 94720, USA*

Abstract.— This study attempts to resolve relationships among and within the four basal arthropod lineages (Pancrustacea, Myriapoda, Euchelicerata, Pycnogonida) and to assess the widespread expectation that remaining phylogenetic problems will yield to increasing amounts of sequence data. Sixty-eight regions of 62 protein-coding nuclear genes (approximately 41 kilobases (kb)/taxon) were sequenced for 12 taxonomically diverse arthropod taxa and a tardigrade outgroup. Parsimony, likelihood, and Bayesian analyses of total nucleotide data generally strongly supported the monophyly of each of the basal lineages represented by more than one species. Other relationships within the Arthropoda were also supported, with support levels depending on method of analysis and inclusion/exclusion of synonymous changes. Removing third codon positions, where the assumption of base compositional homogeneity was rejected, altered the results. Removing the final class of synonymous mutations—first codon positions encoding leucine and arginine, which were also compositionally heterogeneous—yielded a data set that was consistent with a hypothesis of base compositional homogeneity. Furthermore, under such a data-exclusion regime, all 68 gene regions individually were consistent with base compositional homogeneity. Restricting likelihood analyses to nonsynonymous change recovered trees with strong support for the basal lineages but not for other groups that were variably supported with more inclusive data sets. In a further effort to increase phylogenetic signal, three types of data exploration were undertaken. (1) Individual genes were ranked by their average rate of nonsynonymous change, and three rate categories were assigned—*fast*, *intermediate*, and *slow*. Then, bootstrap analysis of each gene was performed separately to see which taxonomic groups received strong support. Five taxonomic groups were strongly supported independently by two or more genes, and these genes mostly belonged to the *slow* or *intermediate* categories, whereas groups supported only by a single gene region tended to be from genes of the *fast* category, arguing that *fast* genes provide a less consistent signal. (2) A sensitivity analysis was performed in which increasing numbers of genes were excluded, beginning with the fastest. The number of strongly supported nodes increased up to a point and then decreased slightly. Recovery of Hexapoda required removal of *fast* genes. Support for Mandibulata (Pancrustacea + Myriapoda) also increased, at times to "strong" levels, with removal of the fastest genes. (3) Concordance selection was evaluated by clustering genes according to their ability to recover Pancrustacea, Euchelicerata, or Myriapoda and analyzing the three clusters separately. All clusters of genes recovered the three concordance clades but were at times inconsistent in the relationships recovered among and within these clades, a result that indicates that the a priori concordance criteria may bias phylogenetic signal in unexpected ways. In a further attempt to increase support of taxonomic relationships, sequence data from 49 additional taxa for three *slow* genes (i.e., *EF-1α*, *EF-2*, and *Pol II*) were combined with the various 13-taxon data sets. The 62-taxon analyses supported the results of the 13-taxon analyses and provided increased support for additional pancrustacean clades found in an earlier analysis including only *EF-1α*, *EF-2*, and *Pol II*. [Arthropoda, Cambrian Explosion, Chelicerata, data partitioning, Mandibulata, Myriapoda, nuclear genes, Pancrustacea, Paradoxopoda, Pycnogonida.]

Determining the phylogenetic relationships among the major arthropod lineages has been a long-standing goal of systematic biology, but the successes and failures of this research program have implications that extend well beyond the phylum itself. In many ways, the temporal structure of diversification among the traditionally defined arthropod groups (crustaceans, chelicerates, hexapods, myriapods) mirrors that of the metazoan phyla and other major lineages (Fortey and Thomas, 1998). Specifically, there was an ancient and apparently rapid radiation followed by a long period of independent evolution. This pattern of diversification presents significant challenges to systematists, as phylogenetic signal would have accumulated over a relatively brief interval and then would have been degraded by extinction and hundreds of millions of years of subsequent evolution (Rokas et al., 2005; Whitfield and Lockhart, 2007). Despite these apparently unfavorable conditions, the largely unanticipated but well-supported findings that hexapods and crustaceans form a clade called Pancrustaea and that myriapods are monophyletic have been major successes of molecular systematics and have fostered optimism that other major phylogenetic problems within Arthropoda and elsewhere will eventually yield under the weight of more molecular sequence data (e.g., Friedrich and Tautz, 1995; Boore et al., 1995, 1998; Mallatt et al., 2004; Giribet et al., 2005; Regier et al., 2005a, 2005b; see also Glenner et al., 2006, and references therein). Still, many important phylogenetic problems remain within Arthropoda, including the relationships among and within the four basal lineages: Euchelicerata, Pycnogonida, Myriapoda, and Pancrustacea (e.g., Martin and Davis, 2001).

Previous efforts to collect large amounts of DNA sequence data to resolve deep arthropod phylogeny fall into three major approaches, each yielding about the same number of characters: full sequences of nuclear 18S and 28S ribosomal DNA (3852 conservatively aligned base pairs; Mallatt et al., 2004; Mallatt and Giribet, 2006); whole mitochondrial genomes (3555 concatenated amino acids from 11 genes; e.g., Hwang et al., 2001; Podsiadlowski et al., 2006, 2008); and three single-copy nuclear protein-coding genes—elongation factor-1α (*EF-1α*), elongation factor-2 (*EF-2*), and the largest subunit of RNA polymerase II (*Pol II*; 3626 basepairs (bp) combined after excluding third codon positions; Regier et al., 2005a).

The most recent publications from these three approaches (Regier et al., 2005a; Mallatt and Giribet, 2006; Podsiadlowski et al., 2008) all independently support a monophyletic Pancrustacea (Hexapoda plus Crustacea) and a monophyletic Euchelicerata (Chelicerata minus Pycnogonida). On the other hand, the three approaches have not reached a consensus on the monophyly of Chelicerata. Most importantly, these approaches disagree on the resolution of the deepest arthropod relationships. Mitochondrial amino acids support Chelicerata plus Myriapoda (a.k.a. Paradoxopoda; Hwang et al., 2001; Podsiadlowski et al., 2008); nuclear protein-coding genes are ambiguous (Regier et al., 2005a); and the addition of more taxa has caused 18S and 28S ribosomal genes to shift from modest support for the Parodoxopoda (Mallatt et al., 2004) to weak support of Pancrustacea plus Myriapoda (a.k.a. Mandibulata; Mallatt and Giribet 2006).

Now that ribosomal genes and mitochondrial genomes have been sequenced in their entirety, only the third approach—single-copy protein-coding nuclear genes—remains as a source of new data. Previous studies have shown that data sets comprising >30 kb of protein-coding nuclear gene sequence showed great success in resolving the rapid radiation of eutherian mammals (Murphy et al., 2001; Madsen et al., 2001; Johnson et al., 2006). More recently, Savard et al. (2007) concatenated 185 single-copy nuclear genes from genome sequencing projects to amass over 100 kb from eight holometabolous insects (see Pennisi, 2007, for a description of similar efforts with primate genomes).

Here we report results from an attempt to resolve deep arthropod relationships by adding primer-amplified sequences from 62 new single-copy genes, adding nearly 36 kb of new sequences for each of 12 arthropod taxa plus a tardigrade outgroup. When added to *EF-1α, EF-2*, and *Pol II*, this data matrix includes nearly 41 kb of single-copy nuclear sequences. We turned to testing character partitioning and exclusion as a means to enhance phylogenetic signal, an approach rendered practical only by large data sets like the one generated here (Naylor and Brown, 1998; Arisue et al., 2005; Philippe et al., 2005; Rodríguez-Ezpeleta et al., 2007). Our analyses compare four approaches as regards partitioning and excluding data: (1) progressively excluding sites with potentially synonymous substitutions, beginning by excluding third codon positions, and then first codon positions coding for leucine and arginine (Regier and Shultz, 2001a); (2) progressively excluding gene regions with the most rapid rates of nonsynonymous substitution; (3) excluding genes that are missing the tardigrade outgroup; and (4) analyzing sets of gene regions that individually support well-accepted groups such as Euchelicerata, Myriapoda, and Pancrustacea. Our analyses, which included analyses of nucleotide base homogeneity, showed strong justification for removing potentially synonymous first codon positions and for removing the most rapidly evolving gene regions.

Our data set—which comes to approximately 20 kb after removing 10 *fast* gene regions, synonymous sites, and ambiguously aligned regions—is still more than five times larger than data sets from the earlier approaches using ribosomal DNA, mitochondrial amino acids, and protein-coding nuclear genes. Our various analyses not only converged on strong support for the monophyly of Euchelicerata, Myriapoda, and Pancrustacea but also for a major clade within Pancrustacea (Copepoda, (Thecostraca, Malacostraca)) and for joining Myriapoda and Pancrustacea to form the Mandibulata. Support for the Mandibulata was maintained following the addition of 49 taxa sequenced only for three of the 68 gene regions (i.e., *EF-1α, EF-2*, and *Pol II*), plus there was increased resolution within the Pancrustacea. Our results tell a cautionary but hopeful tale about the degree to which more sequence data can contribute towards resolving arthropod phylogeny, particularly in the context of ever-improving models of nucleotide change and expanding taxon sampling.

Materials and Methods

Selecting Candidate Genes for Primer Development

Our aim was to identify single-copy, orthologous nuclear protein-coding regions for phylogenetic analysis across Arthropoda and Tardigrada. We started the project during the fall of 2001, when only three metazoan genomes were available, from *Homo sapiens*, *Drosophila melanogaster*, and *Caenorhabditis elegans*. We began with the 5275 putative *Drosophila/Homo* orthologs identified by Sonnhammer and colleagues using the InParanoid method (Remm et al., 2001), which performs reciprocal pairwise alignments on two genomes using BLAST (Altschul et al., 1990) to discover possible orthologs. A Perl script was written to display the amino acid alignments of these putative orthologs to identify potential regions for primer design. For promising ortholog pairs, the *Caenorhabditis* genome was analyzed through pairwise sequence alignments to the *Drosophila* ortholog. The *Drosophila* ortholog was also compared back to its own genome by BLAST to check for paralogs.

Using this approach, we discovered 289 orthologous sequences in all three species with amino acid similarities >55% and no apparent paralogs in *Drosophila* (named

1fin to *289fin, aspec*). We also found 213 sequences that were present in all three species but that had a distant paralog (16% to 44% identity) in *Drosophila* (named *3001fin* to *3213fin, acc*). Ninety-five sequences had orthologs in *Drosophila* and *Homo* but not in *Caenorhabditis*, no paralogs in *Drosophila*, and a sequence similarity >55% (named *8001fin* to *8095fin*). In addition, we developed primers for 12 sequences that had orthologous sequences from at least three of the four major arthropod groups; that is, *Drosophila* (Hexapoda) and any species from Crustacea, Myriapoda, and Chelicerata (named *5001fin* to *5012fin*). Alignments were created using Clustal W (Thompson et al., 1994).

TABLE 1. Taxa sampled and their classification.[a]

- ARTHROPODA (56)
- Pancrustacea (34)
  - Hexapoda (12)
    - *Ctenolepisma lineata* (Cli)
    - *Nicoletia meinerti* (Nme)
    - *Machiloides banksi* (Mba)
    - *Pedetontus saltator* (Psa)
    - *Forficula auricularia* (Fau)*
    - *Periplaneta americana* (Pam)
    - *Hexagenia limbata* (May)
    - *Podura aquatica* (Paq)*
    - *Tomocerus* sp. (Tom)
    - *Orchesella imitari* (Oim)
    - *Eumesocampa frigilis* (Efr)
    - *Metajapyx subterraneus* (Jap)
  - Branchiopoda (5)
    - *Limnadia lenticularis* (Lle)
    - *Lynceus* sp. (Lyn)
    - *Triops longicaudatus* (Tlo)*
    - *Artemia salina* (Asa)
    - *Streptocephalus seali* (ufs)
  - Copepoda (3)
    - *Acanthocyclops vernalis* (A369)
    - *Mesocyclops edax* (Meso)*
    - *Eurytemora affinis* (Eaf)
  - Thecostraca (4)
    - *Semibalanus balanoides* (Bba)
    - *Chthalamus fragilis* (Cfr)
    - *Lepas anserifera* (Lean)
    - *Loxothylacus texanus* (Lox)
  - Malacostraca (4)
    - *Armadillidium vulgare* (Avu2)
    - *Neogonodactylus oerstedii* (Neo)
    - *Libinia emarginata* (Lem)
    - *Nebalia hessleri* (Nhe)*
  - Cephalocarida (1)
    - *Hutchinsoniella macracantha* (Hma)
  - Remipedia (1)
    - *Speleonectes tulumensis* (Stu)*
  - Ostracoda: Podocopa (1)
    - *Cypridopsis vidua* (Ost)*
  - Ostracoda: Myodocopa (2)
    - *Harbansus paucichelatus* (Hapa)
    - *Skogsbergia lerneri* (Skle)
- Branchiura (1)
    - *Argulus* sp. (Arg)
- Myriapoda (15)
  - Chilopoda (6)
    - *Anopsobius neozelanicus* (Ane)
    - *Paralamyctes grayi* (Para)
    - *Bothropolys multidentatus* (Bmu)
    - *Lithobius forficatus* (Lfo)*
    - *Scolopendra polymorpha* (Spo)
    - *Thereuonema* sp. (The)
  - Diplopoda (5)
    - *Abacion magnum* (Ama2)
    - *Trachyiulus nordquisti* (Tnor)
    - *Rhinotus purpureus* (Rpur)
    - *Narceus americanus* (Nam)*
    - *Polyxenus fasciculatus* (Pol)
  - Symphyla (2)
    - *Hanseniella* sp. (Han)
    - *Scutigerella* sp. (Scu2)
  - Pauropoda (2)
    - *Allopauropus proximus* (Apr)
    - *Eurypauropus spinosus* (Eury)
- Chelicerata (7)
  - Euchelicerata (4)
  - Xiphosura (2)
    - *Carcinoscorpius rotundicauda* (Cro)

(Continued on next page)

Identifying PCR Primer Sites on Candidate Genes

The amino acid alignments were visually scanned for highly conserved hexapeptide (or longer) sequences for use as PCR primers. We assumed complete degeneracy with respect to synonymous changes and aimed to design primers with degeneracy levels less than 128-fold, although this was relaxed for longer primers. We further constrained our search to suitable forward-reverse primer pairs separated by about 300 to 1000 bp; nested primer sites were identified whenever available. *M13REV* and *M13(−21)* 18-mer sequences were added to the 5′ ends of all forward and reverse PCR primers, respectively, to increase primer effectiveness for amplification (Regier and Shi, 2005; see online Appendix 1 [http://www.systematicbiology.org] for the actual *M13* sequences) and for ease of subsequent sequencing. Candidate primers were synthesized and gel purified commercially. A list of primer sequences and useful amplification strategies can be found in online Appendix 1. Primer sequences follow IUPAC conventions, and I = deoxyinosine. Each sequence's amplification strategy always began with a reverse transcription reaction followed by PCR (RT-PCR). In most cases (but not all; e.g., *6fin*), this was then followed by one or more heminested PCR's ("heminested" indicates that only one of the two primers is nested), using the gel-isolated RT-PCR product as template. These sequential reactions are represented in "Amplification Strategy" (see online Appendix 1) by separating the primer pairs used in each reaction with a forward slash (e.g., 1F_4R/2F_4R for *25fin*). For some individual genes (e.g., *aspec*), there were multiple RT-PCR products, each with its own amplification strategy. *EF-1α* and *EF-2* contained overlapping segments, which were co-assembled and analyzed as single sequences. Not all of the *Pol II* sequences overlapped, but they were analyzed as single, concatenated sequences. All other sequences came from single PCR fragments.

Testing PCR Primers on Candidate Genes

The efficacy of primer pairs was assessed using five diverse test species, namely, *Limulus polyphemus* (Chelicerata), *Narceus americanus* (Myriapoda), *Nebalia hessleri* (Pancrustacea: Malacostraca), *Podura aquatica* (Pancrustacea: Hexapoda), and *Thulinia stephaniae* (Tardigrada). When amplification was successful in at least three test

TABLE 1. Taxa sampled and their classification.[a] *(Continued)*

Limulus polyphemus (Lpo)*
Arachnida (2)
Mastigoproctus giganteus (Mga)*
Nipponopsalis abei (Nab)
Pycnogonida (3)
Endeis laevis (Ele)
Tanystylum orbiculare (Tor)*
Colossendeis sp. (Col)
TARDIGRADA (6)
Isohypsibius elegans (Iso)
Thulinia stephaniae (Thul)*
Macrobiotus islandicus (Mis)
Richtersius coronifer (Rco)
Milnesium tardigradum (Hyp)
Echiniscus viridissimus (Evi)

[a] After each higher-level taxon name, the number of species sampled (62 species total) is shown in parentheses. Genus-species names are followed by lab code names within parentheses. Those followed by an asterisk were also included in the 13-taxon study.

species, amplifications were performed for eight additional species, bringing the total to 13 (see Table 1). Gene regions were sequenced when at least 9 of the 13 taxa yielded amplicons for a given gene region (Table 2).

Amplification of Gene Regions and Specimen Storage Conditions

An RT-PCR strategy was chosen to amplify mRNA sequences. Laboratory procedures used to generate amplicons are in online Appendix 2 (http://www.systematicbiology.org) and, in a more expansive version, at http://www.umbi.umd.edu/users/jcrlab/PCR_primers.pdf.

Taxon and Gene Sampling

We attempted to sequence the entire set of 68 gene regions for 12 arthropod and one tardigrade species (taxa are identified in Table 1; genes are listed in Table 2). Another 49 species were sequenced for only three gene regions—*EF-1α*, *EF-2*, and *Pol II* (Table 1). Many, but not all, of these latter sequences were already available (Regier et al., 2005a). All GenBank numbers are listed or referenced in online Appendix 3 (http://www.systematicbiology.org).

Sequencing, Contig Assembly, and Data Set Assembly.—PCR amplicons were directly sequenced on a 3730 DNA Analyzer (Applied Biosystems). Sequences were edited and assembled using programs in the Staden package (Staden et al., 1998). Nucleotide polymorphisms in sequences of individual taxa were widespread but typically represented less than 1% to 2% of total characters, and a large majority of these implied synonymous change only. Polymorphisms were coded as ambiguous. A few highly polymorphic sequences were assembled and included in the analyses. Multiple sequence alignments were made in Genetic Data Environment (Smith et al., 1994). Two data-exclusion masks were applied. The less conservative *mask1* excluded characters immediately surrounding ambiguous regions of overlapping indels (approximately 2.9% of the 40,935 characters from all gene regions combined were excluded). The more conservative *mask2* excluded additional surrounding characters to increase the certainty of character homology further (approximately 4.7% excluded). All analyses were run separately under each mask to test whether characters immediately surrounding indel regions noticeably affected phylogenetic conclusions. Because the results are generally quite similar, we present only those using the more conservative *mask2*. Nucleotide data sets were constructed in Nexus format using PAUP* 4b10 (Swofford, 2002). Nucleotide sequences were translated and amino acid data sets constructed using MacClade 4.08 (Maddison and Madison, 2002).

Two basic data sets were constructed, one consisting of 13 taxa and up to 68 gene regions for each taxon (~39,759 nucelotides [nt]/taxon) and the other that adds 49 arthropods sequenced only for *EF-1α*, *EF-2*, and *Pol II* sequences (~5433 nt combined/taxon). This 62-taxon/68-gene-region data set (1–68 in Table 2) has approximately 71% missing data (versus approximately 11% missing data for the 13-taxon/68-sequence data set, mostly due to PCR failures). Other data subsets for both 13 and 62 taxa were constructed that sequentially excluded the 10 fastest gene regions in Table 2 (called *11–68*), the 20 fastest genes regions (*21–68*), the 30 fastest gene regions (*31–68*), and the 37 fastest gene regions (*38–68*). The percentage of missing data in these data subsets decreases to approximately 58 when the 62-taxon set includes only the 31 slowest gene regions (*38–68* in Table 2). Levels of missing data in the corresponding *11–68*, *21–68*, and *31–68* gene region sets for 62 taxa were between 71% and 58%.

In constructing data sets for analysis, total nucleotide sequences (*nt123*) were at times partitioned by codon position (*nt1*, *nt2*, *nt3*) and by gene region. In addition, because leucine (*L*) and arginine (*R*) codons are unique in their ability to undergo synonymous change at the first position, *nt1* characters were subdivided into one bin (*noLR1*) that included no leucine or arginine codons and another (*LR1*) that included one or more leucine or arginine codons across the entire set of taxa under consideration (Regier and Shultz, 2001a, 2001b). In the current study, *LR1* and *noLR1* character sets were calculated separately for the 13- and 62-taxon data sets. A computer script was written in Perl (see online Appendix 4, http://www.systematicbiology.org) that generated *LR1* and *noLR1* character sets. In an effort to analyze nonsynonymous change independently of synonymous change, *noLR1* was combined with *nt2* to generate a *noLR1+nt2* data set with no synonymous changes whatsoever. In Bayesian phylogenetic analyses, *LR1* and *noLR1+nt2* were modeled separately but analyzed together, as were *nt1* + *nt2* (referred to henceforth as *nt12*) and *nt3*.

Two Nexus-formatted files containing the *nt123* data matices for 13 and 62 panarthropod taxa (online Appendix 5 and Appendix 6, respectively), along various character set definitions, including data-exclusion masks, can be downloaded in the online Appendix section of this article (http://www.systematicbiology.org). These same files are also downloadable at http://www.umbi.umd.edu/users/jcrlab/Arthropod_13tx62gn-2008.

TABLE 2. Information about genes sampled for this study.[a]

No.	Gene region	No. taxa: no. aligned nucleotides	Putative protein function	No. nt2 changes/ position •tree: standard error
1	265fin2_3	13 : 441[c]	H-tRNA synthetase	3.86 : 2.02
2	226fin1_2	11 : 549[c]	gln amidotransferase	3.68 : 1.28
3	192fin1_2	11 : 402	E+P-tRNA synthetase	2.78 : 1.21
4	197fin1_2	10 : 435[c]	triosephosphate isomerase	2.30 : 0.98
5	3059fin1_3	13 : 738[c]	arg methyltransferase	2.23 : 0.84
6	aspec11_12[b]	12 : 594	α-spectrin	2.06 : 0.73
7	247fin1_2	11 : 390[c]	L-tRNA synthetase	2.03 : 1.13
8	3089fin1_3	13 : 306	acetyltransferase	2.02 : 1.34
9	62fin2_3	11 : 765[c]	protein phosphatase	1.96 : 0.65
10	42fin1_2	13 : 834[c]	GTP-binding protein	1.95 : 0.73
11	3007fin1_2	10 : 606[c]	glucose phosphate dehydrogenase	1.85 : 0.75
12	8028fin1_2	10 : 324	nucleolar cysteine-rich protein	1.81 : 1.79
13	40fin2_3	13 : 744[c]	phosphogluconate dehydrogenase	1.77 : 0.66
14	3017fin1_2	12 : 594[c]	tetrahydrofolate synthase	1.74 : 0.67
15	270fin2_3	11 : 423[c]	”hypothetical protein”	1.70 : 0.85
16	268fin1_2	10 : 759[c]	AMP deaminase	1.68 : 0.60
17	3031fin2_3[b]	10 : 579[c]	myosin	1.62 : 0.61
18	149fin2_3	12 : 927[c]	protein kinase	1.61 : 0.48
19	267fin2_3	12 : 600	pyrimidine biosynthesis	1.60 : 0.68
20	3114fin1_2	11 : 378	Q-tRNA synthetase	1.59 : 0.76
21	3121fin2_3	10 : 435[c]	protein kinase	1.53 : 0.71
22	262fin1_2	11 : 453[c]	proteasome subunit	1.48 : 0.74
23	3070fin4_5	9 : 705[c]	A-tRNA synthetase	1.43 : 0.56
24	109fin1_2	11 : 537[c]	gelsolin	1.40 : 0.61
25	69fin2_3	13 : 624[c]	clathrin coat assembly protein	1.36 : 0.55
26	8053fin2_3	11 : 459[c]	phosphatidylinositol kinase	1.32 : 0.66
27	acc2_4	13 : 471[c]	acetyl-coA carboxylase	1.31 : 0.94
28	113fin1_2	13 : 975	glycogen synthase	1.27 : 0.42
29	3006fin1_2	13 : 222	dynamin	1.24 : 0.78
30	8091fin1_2	11 : 666	glucose phosphate isomerase	1.22 : 0.46
31	aspec2_6[b]	10 : 297	α-spectrin	1.20 : 0.63
32	2fin3_4[b]	13 : 531	pre-mRNA splicing factor	1.17 : 0.93
33	73fin2_3	12 : 852[c]	acetylglucosaminyl-transferase	1.08 : 0.52
34	8018fin1_2	7 : 303	proteasome non-ATPase regulatory subunit	1.07 : 0.76
35	127fin1_2	12 : 468	methylmalonate semialdehyde dehydrogenase	1.03 : 0.61
36	3202fin1_3	12 : 504[c]	ATP synthase	1.03 : 0.54
37	8029fin6_7	9 : 387[c]	neurofibromin	1.02 : 0.58
38	3012fin1_2	12 : 525	DNA replication licensing factor	0.97 : 0.42
39	3094fin2_3	10 : 390[c]	ATPase	0.97 : 0.54
40	EF-2	62 : 2178[c]	translational elongation factor	0.95 : 0.24[d]
41	274fin1_2	7 : 537[c]	methionine aminopeptidase	0.87 : 0.44
42	Pol II	62 : 2025[c]	RNA polymerase, largest subunit	0.87 : 0.23[d]
43	8070fin1_3	10 : 531[c]	SH2 domain binding protein	0.86 : 0.35
44	3066fin1_3	10 : 744[c]	RNA helicase	0.85 : 0.35
45	44fin2_3	13 : 528	glucosamine phosphate isomerase	0.85 : 0.51
46	EF-1α	62 : 1092	translational elongation factor	0.81 : 0.30[d]
47	58fin3_6[b]	13 : 927	clathrin heavy chain	0.79 : 0.32
48	96fin1_3	13 : 459[c]	ATP synthase	0.78 : 0.44
49	aspec19_21[b]	13 : 450[c]	α-spectrin	0.71 : 0.46
50	26fin3_4	9 : 729[c]	spliceosome-associated protein	0.69 : 0.46
51	58fin7_9[b]	11 : 339	clathrin heavy chain	0.68 : 0.55
52	63fin2_3	13 : 501[c]	α-adaptin	0.68 : 0.34
53	3064fin6_7	13 : 606	transmembrane protein	0.65 : 0.51
54	3196fin5_6[b]	13 : 702[c]	RNA polymerase, subunit 1	0.65 : 0.37
55	3031fin4_5[b]	13 : 426	myosin	0.60 : 0.36
56	2fin7_8[b]	11 : 618	pre-mRNA splicing factor	0.58 : 0.39
57	3196fin1_3[b]	13 : 543	RNA polymerase, subunit 2	0.54 : 0.66
58	3055fin2_3	11 : 246	protein kinase	0.42 : 0.40
59	3136fin1_2	13 : 849[c]	histone deacetylase	0.39 : 0.20
60	36fin1_2	11 : 471	syntaxin	0.39 : 0.22
61	3153fin1_2	13 : 573	RNA helicase	0.37 : 0.28
62	25fin2_4	12 : 303[c]	signal recognition particle	0.35 : 0.23

(Continued on next page)

TABLE 2. Information about genes sampled for this study.[a] *(Continued)*

No.	Gene region	No. taxa: no. aligned nucleotides	Putative protein function	No. nt2 changes/ position •tree: standard error
63	6fin2_3	10 : 342	casein kinase	0.35 : 0.31
64	166fin2_3	13 : 324[c]	CDC 5-related protein/cell division	0.33 : 0.33
65	220fin1_2	10 : 552	F-box protein	0.32 : 0.30
66	3009fin2_3	13 : 369	G protein-coupled receptor kinase	0.32 : 0.23
67	3044fin1_2	13 : 324	prohormone convertase	0.25 : 0.21
68	3152fin1_2	12 : 279	protein kinase	0.21 : 0.16
	All genes	: 39,759[c]		1.07 : 0.06[d]

[a] Column 1, sequential numbers; column 2, gene region names; column 3, number of taxa successfully sequenced (out of 13 total, except for EF2, PolII, and EF1a, which was 62 total), followed by number of aligned nucleotides, not including primer sequences; column 4, putative protein function; column 5, average number of nucleotide changes at the second codon position (nt2) per nt2 position when mapped under the preferred ML model on a constrained 13-taxon tree and adjusted for missing taxa, followed by the standard error on the rate estimate, also adjusted for missing taxa. For purposes of subsequent analysis, individual gene regions were assigned to one of three rate categories based on their relative ranking in this table—*fast,* gene regions 1 to 20; *intermediate,* 21 to 37; *slow,* 38 to 68—and thin horizontal lines partition these rate categories.

[b] Other nonoverlapping sequences from the same gene are separately listed in this table.

[c] A small fraction of the data for this sequence has problematic across-taxon alignments and is not included in this length calculation or in the calculations presented in column 5.

[d] Estimated from 13 taxa only.

txt and http://www.umbi.umd.edu/users/jcrlab/Arthropod_62tx62gn-2008.txt.

Data for each gene region were analyzed separately and a group of three (overlapping) character sets for 13 taxa were constructed from those gene regions that individually recovered in their bootstrap consensus diagrams (LE50 option = yes in PAUP*4.0) one of three non-controversial a priori taxonomic groups—Euchelicerata, Pancrustacea, or Myriapoda. These combined-gene character sets were then subjected to likelihood analysis and bootstrap analysis. A fourth character set was constructed from individual gene regions that each recovered all three taxonomic groups. The point of this type of analysis was to determine if data sets biased to recover a well-established group(s) might also reveal enhanced node support for otherwise less well-supported groups (Philip and Creevey, 2005). A second category of character set included only those gene regions that included the outgroup taxon *Thulinia,* in order to test the effect of its absence from some data sets on rooting of Arthropoda.

Phylogenetic Analysis

Equally weighted parsimony analysis of amino acids and nucleotides (*nt12, nt123*) were conducted with PAUP* (Swofford, 2002) and consisted of heuristic searches with tree bisection followed by reconnection (TBR) and 100 random sequence-addition replicates. Nonparametric bootstrap analyses (Felsenstein, 1985), with 1000 bootstrap replications, differed from parsimony searches of the original data sets in having fewer sequence-addition replicates (10) per bootstrap replicate.

Before undertaking any model-based phylogenetic searches, including those on single and multiple gene regions, a preferred model was selected using MrModelTest 2.2 (Nylander, 2004) and the AIC criterion (discussed in Felsenstein, 2004). For multigenic data sets, a "general time-reversible (GTR) + gamma + invariant" (GTR+I+G) model was inevitably chosen, but less complex models were chosen for some individual-gene analyses. Rates of nucleotide substitution at *nt2* for individual gene regions were estimated under likelihood using PAUP* and a constraint topology of 13 taxa that was consistent with previous investigations (Regier et al., 2005a): ((((*Mesocyclops*, *Nebalia*), *Cypridopsis*), (((*Podura*, *Forficula*), *Triops*), *Speleonectes*)), ((*Narceus*, *Lithobius*), ((*Mastigoproctus*, *Limulus*), *Tanystylum*))), *Thulinia*). All branch lengths were summed to provide an estimate of the average number of substitutions and standard error per character across the entire 13-taxon tree. In order to make the gene rates approximately comparable, the data were adjusted for the number of missing branches by multiplying the branch length sums by a correction factor, namely, the number of branches for 13 taxa (= 23 branches) divided by the number of branches in the gene under consideration; e.g., for 226fin1_2 there are 11 taxa and 19 branches.

Bayesian phylogenetic analysis was performed using MrBayes 3.1 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003), with each analysis consisting of two runs of four chains each (three hot, one cold) and using random starting trees. AWTY online (Wilgenbusch et al., 2004) was used dynamically with MrBayes to assess when chain convergence had occurred. One of the AWTY online tools called Showsplits, which provides a node-by-node numerical comparison of posterior probabilities for each node from the two independent runs, was particularly useful for estimating in real time when each node of the two independent runs had reached both stationarity and convergence. For analysis of *nt123* (minus data-exclusion masks), we applied either a single GTR+I+G model or separate models (described in Swofford et al., 1996) to *nt12 and nt3*. For analysis of *nt12* only, either a single GTR+I+G model was applied or separate models were applied to *LR1* and *noLR1+nt2*.

Likelihood analysis of the 13-taxon data sets under the GTR+I+G model was performed using GARLI, version 0.951 (Genetic Algorithm for Rapid Likelihood Inference; Zwickl, 2006), with its default parameters. The starting topology was randomly determined. To confirm the optimal topology, analyses were repeated until two independently derived topologies of highest likelihood were identical and their lnL values were nearly identical.

Likelihood analysis of the 13-taxon data set under a codon model and of the 62-taxon data sets under a GTR+I+G model used BOINC GARLI, version 0.96 beta 8 (March 2008), with grid computing (Cummings and Huskamp, 2005) through The Lattice Project (Bazinet and Cummings, 2008), which includes clusters and desktops in one encompassing system (Myers et al., 2008). Briefly, a grid service for GARLI was developed using a special programming library and associated tools (Bazinet et al., 2007). Following the general computational model of a previous phylogenetic study (Cummings et al., 2003), which used an earlier grid computing system (Myers and Cummings, 2003), required files were distributed among hundreds of computers, where the analyses were conducted asynchronously in parallel. For the 62-taxon data sets, 100 GARLI runs were performed and the tree of highest lnL was chosen. For the 13-taxon data set (under the codon model), the best of five runs (favored topology found twice) was chosen. We also attempted to run MrBayes under a codon model, but the two independent runs (four chains each) failed to converge.

Nonparametric bootstrap likelihood analyses of the 13- and 62-taxon data sets were performed using BOINC GARLI with grid computing. Under the GTR+I+G model, 500 replications were performed. Under the codon model, 269 replications were performed.

Likelihood and bootstrap analyses of recoded amino acids were performed using GARLI, version 0.951, and sequences from 13 taxa (Rodríguez-Ezpeleta et al., 2007). In brief, conceptually translated amino acids were placed in one of four functional categories (group 1: AGPST; group 2: DENQ; group 3: HKR; group 4: FILMVWY; cysteine was coded as missing) and run under a GTR+I+G model.

To test the potential effect of base compositional heterogeneity on the recovered topology for 13 taxa, logDet-corrected distance matrices (Lockhart et al., 1994) were generated from three data sets (*nt123* minus uniformative, *LR1* minus uniformative, *noLR1nt2* minus uniformative) and the best topologies were selected under a minimum evolution optimality criterion as implemented in PAUP*. These topologies were compared to the corresponding minimum-evolution topologies generated from distance matrices under a GTR model without LogDet correction.

Nucleotide base compositions of various character sets and chi-square tests of compositional homogeneity were performed using PAUP*.

Results

Summary Statistics of the Gene Search

Segments from 595 nuclear protein-coding genes from *Homo sapiens*, *Caenorhabditis elegans*, and *Drosophila melanogaster* were aligned and visually scanned for suitable PCR primer sites. Possible PCR primer pairs (572 primers total) were identified for 159 of the 595 genes. After testing over 1800 primer combinations by RT-PCR, sometimes followed by heminested reamplification, 176 primers were identified that could amplify 67 noncontiguous gene regions (two to four primers/gene region) from 60 genes for 13 panarthropod taxa. Two gene segments were too polymorphic to permit assembly of their forward and reverse strands and were excluded, leaving 65 gene regions from 59 genes available for phylogenetic analysis (Table 2). We also used previously developed primers to amplify *EF-1α*, *EF-2*, and *Pol II*, resulting in a total count of 68 gene regions from 62 genes (Table 2).

Characteristics of the Genes

Gene names and putative protein functions are listed in Table 2, along with the number of taxa that were successfully sequenced, individual sequence lengths in nucleotides (excluding primer sequences), and average nonsynonymous (i.e., *nt2*) rate estimates with standard errors across each gene region. With all gene regions concatenated and data-exclusion *mask2* enforced, the total characters per taxon was 39,000 bp. The average number of taxa successfully sequenced for each gene region was 11.4 out of 13 total, except for *EF-1α, EF-2*, and *Pol II*, which were sequenced for 62 taxa, including all 13 focal taxa. Based on GenBank annotations, the 62 genes include 18 intermediary metabolism genes; 9 structural protein genes; 9 protein synthesis, modification, and degradation genes; 7 protein kinase genes; 7 replication/transcription/post-transcription genes; 5 tRNA synthetase genes; and 7 others.

Nucleotide Base Compositions and Character Set Selection

Most phylogenetic reconstruction algorithms assume homogeneity in nucleotide base compositions and can give rise to phylogenetic inaccuracies when this assumption is violated (e.g., Blanquart and Lartillot, 2006). In our experience, highly significant compositional heterogeneity is common at sites capable of undergoing synonymous substitution, for example, *nt3* characters (see Table 3), especially when comparing taxa with very deep divergences. Indeed, the present total data set for 13 taxa was significantly nonhomogeneous ($P < 0.0001$). Given this, we followed Regier and Shultz's (2001a) example of identifying first-codon-position characters with the potential for synonymous change (*LR1*: sites that include one or more leucine or arginine codons across all taxa) and those that include no leucine or arginine codons (*noLR1*). Homogeneity tests on these character sets are reported in Table 3. Although the potentially synonymous *LR1* and *nt3* sets were heterogeneous (both P values <0.0001), the two entirely nonsynonymous sets (i.e., *noLR1* and *nt2*) did not depart from homogeneity ($P =$ 0.54 for *noLR1*; $P = 0.73$ for *nt2*). Even the base compositions of each and every individual gene region for the *noLR1+nt2* character set was indistinguishable from homogeneity (all P values >0.99). These results support

TABLE 3. Nucleotide compositional statistics of four character subsets from 68 gene regions combined.[a]

	LR1				*noLR1*				*nt2*				*nt3*			
Taxon	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
Mesocyclops	.30	.48	.08	.14	.29	.12	.41	.18	.33	.21	.15	.30	.11	.46	.29	.14
Nebalia	.31	.46	.09	.15	.28	.12	.42	.18	.32	.21	.16	.30	.20	.26	.24	.29
Triops	.25	.40	.07	.28	.28	.13	.42	.18	.32	.21	.17	.30	.20	.29	.23	.28
Forficula	.28	.45	.07	.21	.28.	.12	.42	.18	.33	.21	.16	.30	.29	.18	.15	.38
Podura	.32	.43	.08	.18	.29	.12	.41	.18	.33	.20	.17	.30	.30	.14	.17	.39
Speleonectes	.34	.41	.07	.17	.29	.13	.41	.18	.33	.20	.17	.30	.25	.20	.24	.30
Cypridopsis	.30	.49	.09	.12	.27	.13	.42	.18	.32	.21	.17	.30	.06	.47	.35	.12
Lithobius	.29	.43	.07	.21	.29	.12	.42	.17	.33	.20	.17	.30	.24	.23	.23	.30
Narceus	.31	.41	.08	.20	.29	.12	.41	.17	.33	.20	.17	.30	.21	.27	.26	.26
Limulus	.33	.41	.07	.19	.29	.12	.41	.17	.33	.20	.16	.30	.29	.16	.19	.37
Mastigoproctus	.35	.35	.08	.22	.29	.12	.41	.17	.33	.20	.17	.30	.32	.11	.16	.40
Tanystylum	.34	.34	.07	.25	.29	.12	.41	.17	.33	.21	.16	.30	.27	.21	.19	.33
Thulinia	.24	.49	.09	.17	.29	.12	.42	.17	.32	.21	.17	.30	.18	.29	.28	.25
Highest/lowest:	1.46	1.44	1.29	2.00	1.07	1.08	1.02	1.06	1.03	1.05	1.13	1.00	5.33	4.27	2.33	3.33
No. characters:		3164				9836				13,000				13,000		
Chi-square test:	$P < 0.0001$ (heterogeneous)				$P = 0.38$ (homogeneous)				$P = 0.41$ (homogeneous)				$P < 0.0001$ (heterogeneous)			

[a] Taxa listed in Table 1 are abbreviated to genus name. *LR1*, the complete subset of first-codon-position nucleotide characters that encode one or more leucine or arginine residues across the 13 taxa. *noLR1*, the complete subset of first-codon-position nucleotide characters that encode no leucine or arginine residues across the 13 taxa. *nt2*, the complete set of second-codon-position characters across the 13 taxa. *nt3*, the complete set of third-codon-position characters across the 13 taxa. A, adenine; C, cytosine; G, guanine; T, thymine. Highest/lowest, the ratio of the highest to the lowest compositional values on a column-by-column basis.

partitioning *LR1* and *noLR1* first-codon positions in addition to the standard practice of partitioning and excluding third-codon positions.

Phylogenetic Analysis of 13 Taxa and 68 Gene Regions in Combination

The data generated from 13 taxa and all 68 gene regions in Table 2 were evaluated using various data subsets and methods of phylogenetic analysis (Fig. 1). Parsimony analysis of all nucleotides (not shown) yielded the highly unlikely result of a paraphyletic Pancrustacea at the base of the Arthropoda (the wide range of support for the Pancrustacea is summarized in Glenner et al., 2006). Bayesian analyses run under two different models (GTR+I+G with and without partitioning of *nt12* and *nt3*) recovered a more conventional rooting (Fig. 1a, b). However, although the posterior probabilities (PP) were mostly 1.0, including within Pancrustacea, there were strongly supported (i.e., PP $\geq$ 0.95) conflicts, confirming that high precision of posterior probabilities cannot be equated with phylogenetic accuracy. Likelihood analysis under a codon model yielded a third topology (Fig. 1c), but only four nodes received strong bootstrap support (i.e., BP $\geq$ 80%). Parsimony and likelihood analyses of amino acids (Fig. 1d, e) showed strong bootstrap support at only two nodes (Myriapoda, Euchelicerata), both of which are widely accepted as monophyletic. Pancrustacea was recovered in the optimal amino acid parsimony topology, although with low bootstrap support (61%), whereas it was not recovered in the optimal amino acid likelihood topology, although it still received 62% bootstrap support. Hexapoda was recovered under likelihood (67% BP) but not under parsimony (<50% BP).

With only first and second codon positions, parsimony, Bayesian, and likelihood methods recovered the aforementioned Myriapoda, Euchelicerata, and Pancrustacea (Fig. 1f to i). Outside Pancrustacea, the only strongly supported conflict was the placement of Pycnogonida, either as sister group to all other arthropods under parsimony (Fig. 1f) or as part of a monophyletic Chelicerata within the Paradoxopoda (= Chelicerata + Myriapoda; Fig. 1g to i). Within the Pancrustacea, all four methods recovered exactly the same relationships, often with strong support. Confidence in these congruent analyses of the *nt12* character set is tempered by the fact that Hexapoda (*Forficula* + *Podura*) was strongly contradicted by all of these analyses (Fig. 1f to h), even when the *LR1* and *noLR1* characters were allowed different models in a Bayesian analysis (Fig. 1i).

As mentioned above, the *nt123* and *nt12* character sets departed significantly from base compositional homogeneity, whereas the *noLR1+nt2* character set did not (Table 3). Bayesian and likelihood methods yielded identical *noLR1+nt2* topologies (*cf.* Fig. 1j, k), although the Bayesian results were much more decisive, with each and every node having a posterior probability >95% and a paraphyletic Hexapoda. In contrast, the likelihood analysis only strongly supported the unsurprising Pancrustacea, Myriapoda, and Euchelicerata. And, although Hexapoda was not recovered, neither was it strongly disconfirmed. Taken together, these results suggest the greater sensitivity of Bayesian posterior probabilities to misspecification of the model of nucleotide change. A comparison of likelihood inferences for *nt12* (Fig. 1h) and *noLR1+nt2* (Fig. 1k) and of Bayesian inferences for *nt12* (Fig. 1g) and *noLR1+nt2* (Fig. 1j) provided a ready means for evaluating the effect of the *LR1* data on topology and node support (see Discussion). Taken together, these analyses suggest that the *noLR1+nt2* character set—analyzed under likelihood bootstrap—is a conservative approach to deep arthropod phylogenetic inference. Likelihood analysis of *nt123* under a codon model also looked promising in that strong support was

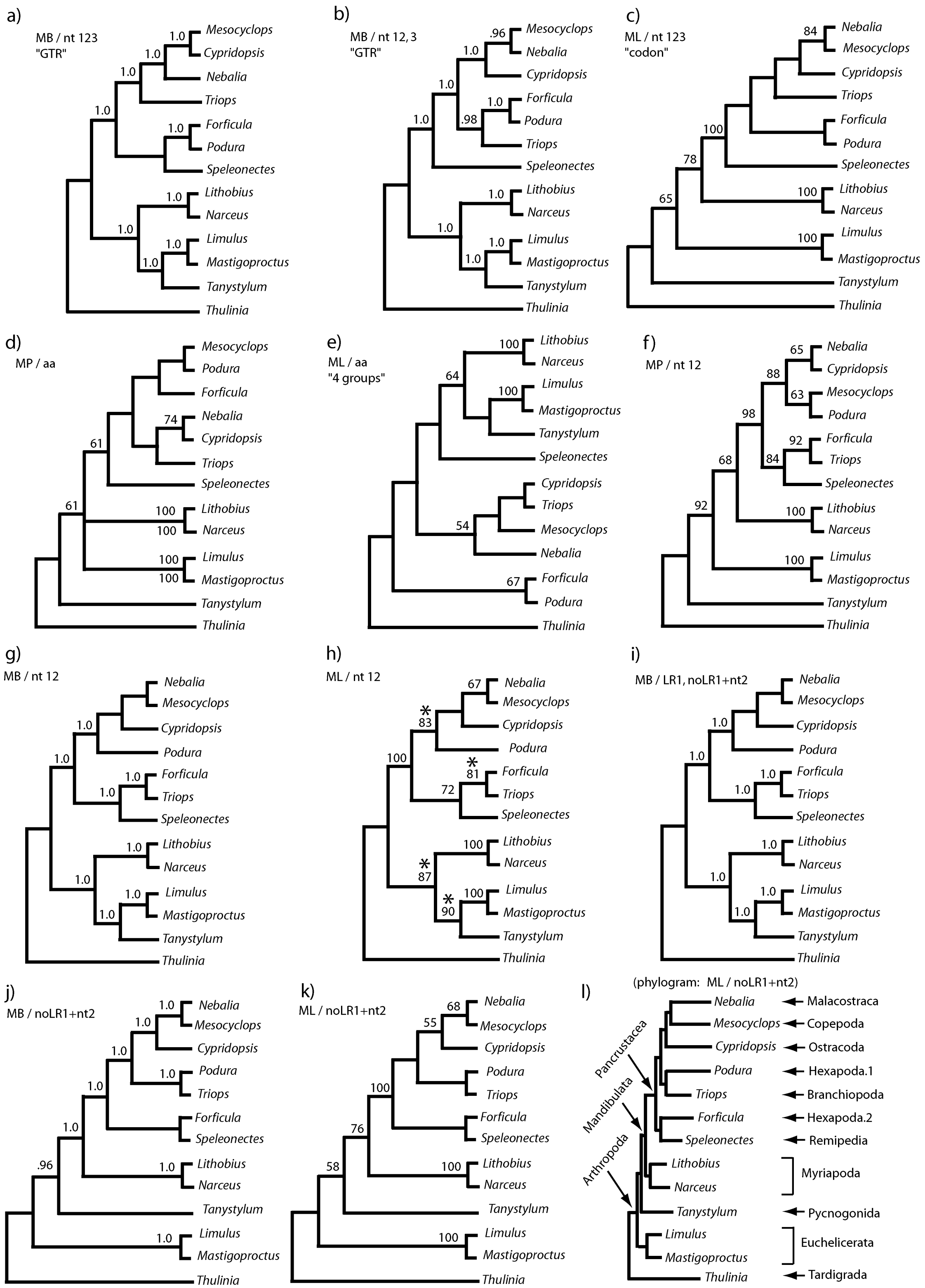

FIGURE 1. Analysis of 13-taxon, 68-sequence data sets using different character partitions and phylogenetic methods. (a) Bayesian analysis of *nt123*. (b) Bayesian analysis with separate models applied to *nt12* and *nt3*. (c) Likelihood analysis of *nt123* using a codon model. (d) Parsimony analysis of amino acids. (e) Likelihood analysis of amino acids assigned to four functional groups. (f) Parsimony analysis of *nt12*. (g) Bayesian analysis of *nt12*. (h) Likelihood analysis of *nt12*. (i) Bayesian analysis with separate models applied to *LR1* and *noLR1+nt2*. (j) Bayesian analysis of *noLR1+nt2*. (k) Likelihood analysis of *noLR1+nt2*. (l) Likelihood analysis of *ntLR1+nt2*—phylogram format. In each part, taxa are labeled by their generic names. Higher-level names are shown only in part l. Node support values (bootstrap percentages ≥50% for parsimony and likelihood analyses, posterior probabilities ≥0.95 for Bayesian analyses) are shown above internal branches in all parts except part l. In part h, four internal nodes with strong bootstrap support values are identified with an asterisk for reference to part k, where these same four nodes are not recovered.

limited to well-established taxonomic groups, but unfortunately the computational demands of running under a codon model, even with access to grid computing (see Materials and Methods), were too great for data exploration beyond that shown in Figure 1c.

A phylogram of the likelihood analysis of *noLR1+nt2* (Fig. 1l) showed that terminal branches were considerably longer than internal ones. The three strongly supported clades (i.e., Euchelicerata, Myriapoda, Pancrustacea) had the lowest ratios of terminal-to-internal branch lengths on the tree (i.e., minimum-maximum ranges of 2.5 to 2.6, 2.9 to 3.4, and 1.7 to 3.8, respectively). All other ratios varied from 8.1 to 19.8. In our experience, such high ratios are correlated with ambiguity in the phylogenetic resolution of these groups.

To determine whether compositional heterogeneity at *LR1* and *nt3* could explain the across–data set differences in groups recovered, the logDet correction was applied. The *nt123*- and *LR1*-alone topologies were very different with and without the logDet correction, whereas the *noLR1+nt2* topologies were identical whether or not the logDet correction was applied. These results are consistent with compositional heterogeneity at *LR1+nt3* introducing topological variability.

Phylogenetic Analysis of 13 Taxa for Individual Gene Regions (68 Total)

Bootstrap analyses of individual gene regions (*noLR1+nt2* data sets) were performed under optimal likelihood models to assess the frequency that various taxonomic groups received bootstrap support >80% (Table 4). The three groups with strongest combined support (Euchelicerata, Myriapoda, Pancrustacea; see Fig. 1k) were strongly supported by the largest number of individual gene regions (12, 9, and 4, respectively). The pairing of *Mesocyclops* and *Nebalia* (= Copepoda + Malacostraca), also present in Figure 1k, was strongly supported by two gene regions (Table 4). Nine additional groups (see Table 4) were supported by one gene region each. Overall, 10 instances of strong support occurred amongst 7 of the 20 genes in the *fast* rate category (no. 1 to 20 in Table 2), 3 instances occurred among 2 of 17 genes in the *intermediate* rate category (no. 21 to 37 in Table 2), and 23 instances occurred amongst 13 of 31 genes in the *slow* rate category (no. 38 to 68 in Table 2). However, six of the nine singleton groups conflict with groups that are strongly supported in the combined analyses presented below, so their accuracy is questionable (see Table 4, Discussion).

TABLE 4. Tabulation of taxonomic groups recovered by ML analysis of individual gene regions with strong BP support (13 taxa, *noLR1+nt2* data sets).

Taxonomic group	Number of gene regions strongly supporting	Rate category[a]	Group in Figure 2a–e?	Strong conflict with Figure 2a–e?[b]
Euchelicerata	8	Slow	Yes	No
	4	Fast		
Myriapoda	8	Slow	Yes	No
	1	Fast		
Pancrustacea	2	Slow	Yes	No
	2	Intermediate		
Mesocyclops +*Nebalia*	2	Slow	Yes	No
Pancrustacea excluding *Cypridopsis*	1	Intermediate	Yes	No
Chelicerata	1	Fast	No	No
Cypridopsis + *Triops* + *Speleonectes*	1	Slow	No	No
Cypridopsis + *Nebalia*	1	Fast	No	Yes
Cypridopsis + *Mesocyclops*	1	Slow	No	Yes
Forficula + *Triops*	1	Fast	No	Yes
Mesocyclops + *Podura*	1	Fast	No	Yes
Arthropoda excluding *Cypridopsis*	1	Fast	No	Yes
Tanystylum + *Mesocyclops*	1	Slow	No	Yes

[a] Individual gene regions were assigned to one of three rate categories based on their relative ranking in Table 2. *fast*, gene regions 1 to 20; *intermediate*, 21 to 37; *slow*, 38 to 68. The number of gene regions that individually strongly recover (i.e., BP ≥ 80%) the corresponding taxonomic group in column 1 is shown in column 2.

[b] *Strong* conflict is identified if (by definition) the corresponding taxonomic group in column 1 conflicts with any other group in Figure 2 that has BP support ≥79%.

Phylogenetic Analysis of 13 Taxa Using Nested Sets of Gene Regions Partitioned by Average Rate of Nonsynonymous Change

It is widely assumed that ancient divergences are best resolved using slowly evolving characters, an assumption that inspired the tactic of removing those sites with a high propensity for synonymous change (i.e., *nt3* and *LR1*) prior to analysis. This reasoning also suggests that ancient divergences can be resolved more reliably by slowly evolving nonsynonymous change than by fast-evolving nonsynonymous change. Although reasonable in principle, this strategy presents a challenge for most studies because reliance on rarer changes would likely require more sequence data to generate a significant signal. However, the abundance of characters and range of rates present in our data set allowed us to test the assumption that slower evolution is better when attempting to resolve ancient divergences using nonsynonymous substitutions. Specifically, there is an approximately 18-fold difference in the average rate of nonsynonymous change across the 68 gene regions listed in Table 2. Starting with the all-68 gene region set (i.e., the *1–68* data set, Fig. 2a), we created four hierarchically nested sets of gene regions; we excluded the 10 fastest gene regions (data set called *11–68* in Fig. 2b), the 20 fastest gene regions (*21–68* in Fig. 2c), the 30 fastest gene regions (*31–68* in Fig. 2d), and the 37 fastest gene regions (*38–68* in Fig. 2e). Although the data sets reflected progressively lower average rates of nucleotide substitution, even the slowest set (*38–68*) still had 19,275 alignable characters in its *nt123* data set and 11,503 characters in its *noLR1+nt2* data set. Likelihood analyses for each of the nested sets of gene regions are shown in Figure 2 and explained further in Discussion.

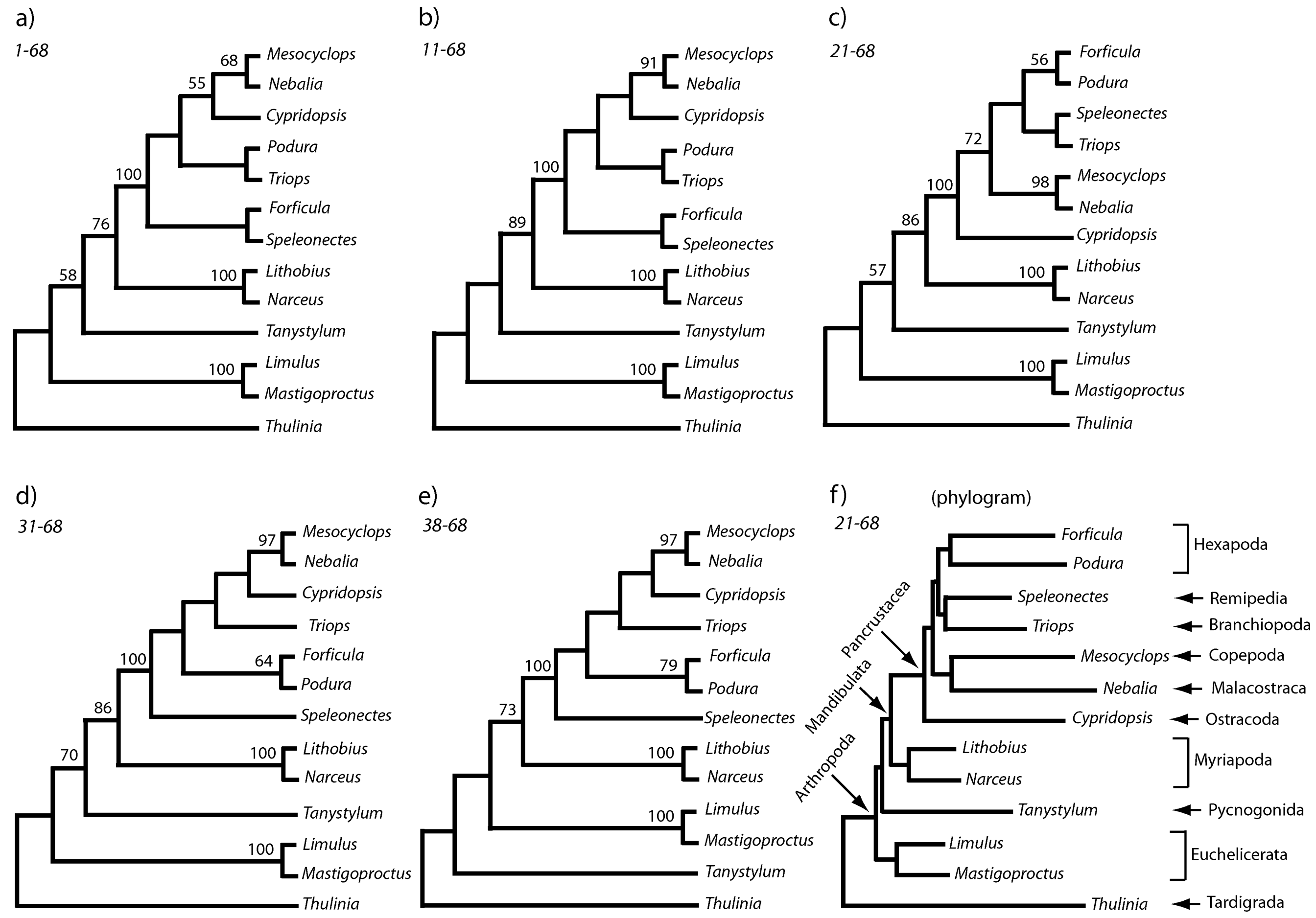

FIGURE 2. Likelihood analysis of 13-taxon, *noLR1+nt2*-character data sets with nested gene sets that evolve at different average rates. (a) All 68 sequences (Table 2). (b) The fastest 10 sequences are removed, leaving sequences *11–68*. (c) The fastest 20 sequences are removed, leaving sequences *21–68*. (d) The fastest 30 sequences are removed, leaving sequences *31–68*. (e) The fastest 37 sequences are removed, leaving sequences *38–68*. The topology in part f is identical to that part c but is shown in phylogram format. In each part, taxa are labeled by their generic names. Suprageneric names are shown only in part f. Bootstrap percentages ≥50% are shown above internal branches in all parts except part f.

Phylogenetic Analysis of 62 Taxa Using Nested Sets of Gene Regions Partitioned by Average Rate of Nonsynonymous Change

To assess the effect of taxon sampling, we combined the 13-taxon/68-gene-region data set with *EF-1α*, *EF-2*, and *Pol II* from 49 additional panarthropods for which all other gene regions were missing (Table 1). Then, we performed likelihood analyses with bootstrapping using the hierarchically nested, 62-taxon, *noLR1+nt2* character sets (Table 5). As reference, we show the ML topology in phylogram format of the *21–68* data set (Fig. 3), with the original 13 taxa shown with asterisks, and bootstrap values described in Table 5. Bootstrap percentages (BP) for all tardigrade subgroups under all analyses in Table 5 are ≥95% (see also Regier et al., 2004), although this includes only one taxon (*Thulinia*) that is sampled for more than *EF1α*, *EF-2*, and *Pol II*. Table 5 also provides a comparison of group support levels between *noLR1+nt2* and *nt12* for data set *21–68* and between data sets *38–68* and *EF-1α+ EF-2 + Pol II* (i.e., *3genes*) for *noLR1+nt2*. As with the 13-taxon analyses of the nested character sets, the results are explained further in Discussion.

Phylogenetic Analysis of 13 Taxa Using Gene Regions Concordant with a Priori Clades

We analyzed three additional character sets that grouped those gene regions recovering one of three test clades considered by many researchers to be very well supported. These overlapping gene regions are concordant with Euchelicerata (35 gene regions; *Euchel:genes*), Pancrustacea (34 gene regions; *Pancrust:genes*), and Myriapoda (30 gene regions; *Myria:genes*).

In a separate test, we identified two nested sets of gene regions: those for which the tardigrade outgroup was successfully sequenced (51 gene regions; *Thuliana:genes*) and the subset of *Thuliana:genes* concordant with all three groups just mentioned—Euchelicerata, Pancrustacea, and Myriapoda (four gene regions; *EuMy-Pan:genes*). The results for these concordance tests are shown in Table 6 and explained in the Discussion.

TABLE 5. Clade recovery assessment from phylogenetic analyses of 62 taxa using nested sequence sets of decreasing average rates of substitution.

			Bootstrap percentage[d]						
			noLR1+nt2[e]					*nt12*	*noLR1+nt2*
Node[a]	13tx[b]	Taxon[c]	1–68[f]	11–68	21–68	31–68	38–68	21–68	3 genes
1	12	*ARTHROPODA*	100	100	100	100	100	100	100
2	10	Arthropoda excluding Euchelicerata	51	54	59	59	[–]	[–]	[–]
•	11	Arthropoda excluding Pycnogonida	[–]	[–]	[–]	[–]	[–]	[–]	[–]
3	9	Pancrustacea + Myriapoda (= Mandibulata)	64	77	75	69	[–]	[–]	[–]
•	5	Chelicerata + Myriapoda (= Paradoxopoda)	[–]	[–]	[–]	[–]	—	77	63
4	7	*Pancrustacea*	100	100	100	100	100	100	100
•	5	Crustacea	[–]	[–]	[–]	[–]	—	[–]	[–]
5	2	*Hexapoda*	82	87	92	91	91	80	86
6	1	*Branchiopoda*	97	95	94	94	94	99	82
7	1	*Copepoda*	100	100	100	100	100	100	100
8	1	*Thecostraca*	100	100	100	100	100	100	100
9	1	*Malacostraca*	100	100	100	100	100	100	100
•	1	Ostracoda	[–]	[–]	[–]	[–]	[–]	[–]	[–]
10	1	Ostracoda + Branchiura	—	—	—	—	—	53	—
11	6	Pancrustacea excluding (Ostracoda + Branchiura)	[–]	—	—	[–]	[–]	—	—
•	6	Pancrustacea excluding (Cephalocarida + Remipedia)	—	[–]	[–]	[–]	[–]	[–]	[–]
•	5	Pancrustacea excluding Hexapoda	[–]	[–]	[–]	[–]	—	[–]	[–]
12	1	*Cephalocarida + Remipedia*	81	81	78	83	79	77	76
•1(+1)		Copepoda + Thecostraca + Branchiura (+ Ostracoda) (= Maxillopoda)	[–]	[–]	[–]	[–]	[–]	[–]	[–]
13	1	*Malacostraca + Thecostraca*	99	98	99	99	99	99	98
14	2	*Copepoda + Malacostraca + Thecostraca*	94	97	98	97	97	97	81
15	3	Hexapoda + Remipedia + Cephalocarida	[–]	[–]	—	—	[–]	[–]	[–]
•	3	Hexapoda + Branchiopoda	[–]	—	[–]	[–]	[–]	—	[–]
•	4	Hexapoda + Branchiopoda + Remipedia + Cephalocarida	[–]	—	[–]	[–]	[–]	—	[–]
16	3	Branchiopoda + Copepoda + Malacostraca + Thecostraca	[–]	[–]	—	[–]	[–]	[–]	—
•	4	Branchiopoda + Ostracoda + Branchiura + Copepoda + Malacostraca + Thecostraca	—	[–]	[–]	—	[–]	[–]	[–]
•	3	Chelicerata	[–]	[–]	[–]	[–]	53	96	83
17	2	*Euchelicerata*	100	100	100	100	100	100	100
18	1	*Arachnida*	100	99	100	99	98	97	90
19	1	*Xiphosura*	100	100	100	100	100	100	100
20	1	*Pycnogonida*	100	100	100	100	100	100	100
21	2	*Myriapoda*	99	99	99	99	98	100	96
22	1	*Chilopoda*	100	100	99	99	100	100	99
23	0	*Pauropoda*	100	100	100	100	100	100	100
24	0	*Symphyla*	100	100	100	100	100	100	100
25	1	Diplopoda	69	67	67	63	61	74	57

[a] Node numbers refer to the corresponding nodes in Figure 3; •identifies groups not present in Figure 3.

[b] Lists number of the 13 taxa (i.e., those labeled with asterisk in Table 1 and analyzed in Figures 1 and 2) present in the particular taxon.

[c] Taxa that receive strong node support in at least one analysis of the noLR1+nt2 data set are italicized.

[d] Bootstrap percentages ≥50% are listed for likelihood analyses of *noLR1+nt2* and *nt12* data sets. *1–68* refers to the 68 sequences listed in Table 2. *11–68* refers to sequences 11 to 68, etc. *3genes* refers to sequences numbered 40 (*EF-2*), 42 (*Pol II*), and 46 (*EF-1α*) in Table 2. Brackets (with or without bootstrap percentages enclosed) indicate that that clade was not recovered in the favored topology for that data set. A dash indicates BP < 50%.

[e] Character set.

[f] Gene sequences.

DISCUSSION

Project Design

Resolving the most challenging phylogenetic questions requires that careful attention be paid to the number and quality of characters (e.g., Rokas et al., 2003; Gatesy et al., 2007), to the density and distribution of taxa (e.g., Mitchell et al., 2000), and to the effectiveness and practicality of computer algorithms for capturing and assessing phylogenetic signal (e.g., Pagel and Meade, 2004; Lartillot et al., 2007). Even then, contingencies of evolution may confound the most determined investigation, as exemplified by rapid, ancient radiations like the "Cambrian Explosion." In systematic language, this problem can be expressed as the difficulty of resolving multiple short internodes that subtend long terminal branches. Deep-level arthropod phylogeny is a prime example of such a problem, and, despite many years of morphological and molecular studies, important nodes remain ambiguous or only weakly resolved (Martin and Davis, 2001; Mallatt et al., 2004; Giribet et al., 2005; Regier et al., 2005a,b; Mallatt and Giribet, 2006). Our study is an ambitious attempt to expand the number of slowly evolving nuclear protein-coding genes available for use in resolving arthropod phylogeny and offers a tangible product—primer sequences that amplify regions of 62 genes from 13 diverse panarthropods (Tables 1, 2). The complete data set includes almost 41 kb/taxa and allows a renewed investigation of arthropod phylogeny.

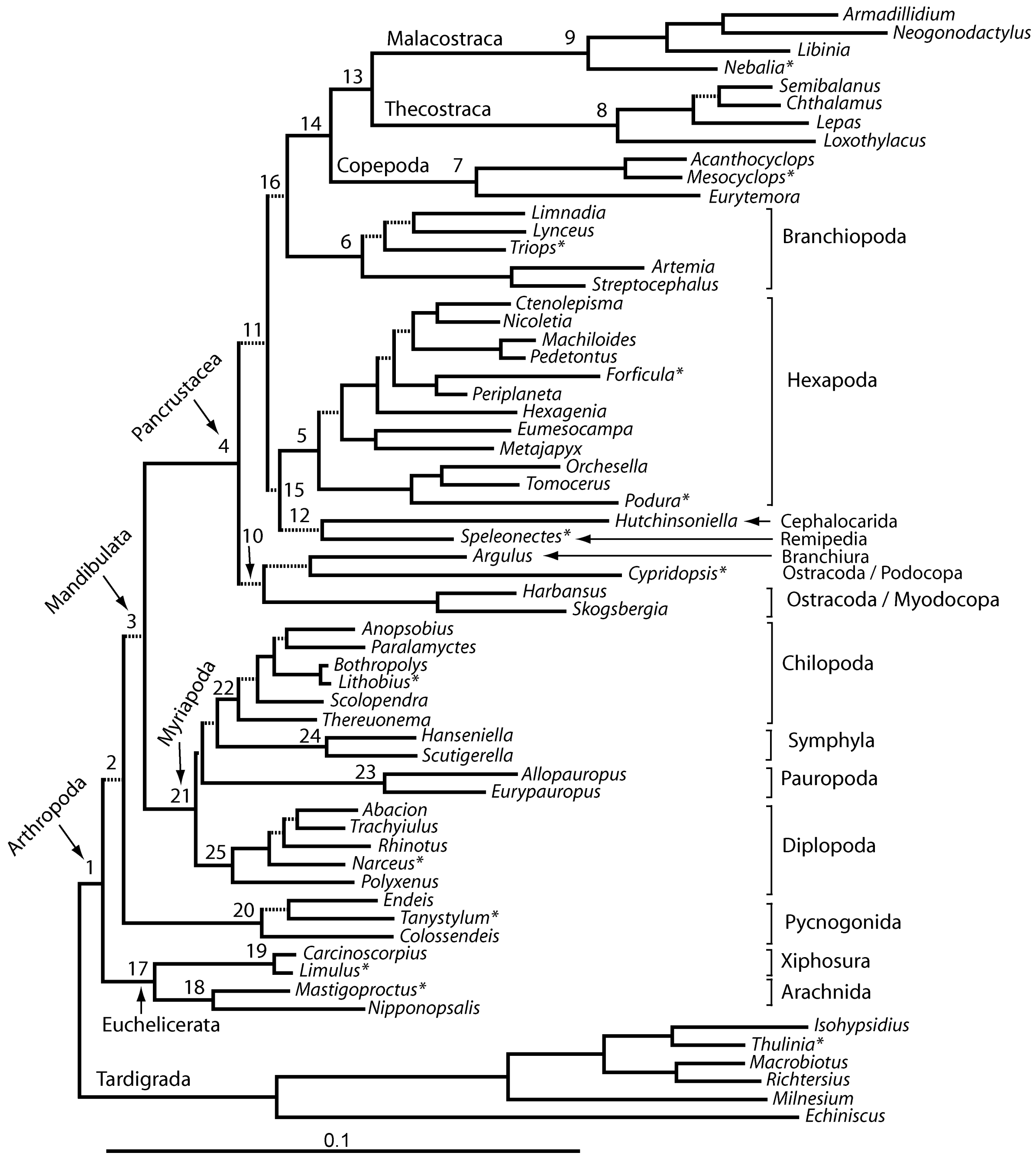

FIGURE 3. Likelihood analysis of the 62-taxon, *21–68*-gene sequence, *noLR1+nt2*-character data set. All nodes with bootstrap percentages <80% have their subtending branches represented as dashed lines. Selected nodes are numbered for reference to Table 5, where bootstrap values for this and related analyses are listed for comparison. The 13 taxa represented in Figures 1 and 2 are identified with an asterisk following their generic name.

Our focus on protein-coding nuclear genes capitalizes upon their relative conservation and ease of alignment. It is likely that there will soon be fast, inexpensive methods for generating whole nuclear genome sequences, and these will provide more sequence data as well as new types of genomic characters (e.g., *Drosophila* 12 Genomes Consortium, 2007). Nevertheless, the current study immediately brings to bear substantial new data for arthropod phylogeny and addresses the important question as to what may actually be required to solve outstanding problems.

The manner in which the new genes used in the current study were identified—initially based on alignments across *Homo sapiens*, *Caenorhabditis elegans*, and *Drosophila melanogaster*—ensures a generally high level of sequence conservation, which is likely to be useful for resolving

TABLE 6. Data exploration: Examining sequence sets that support specific taxonomic groups.[a]

Combined sequence data set name	Number of characters	Number and type of sequence sets			Bootstrap percentage (part)					Bootstrap percentage (part)						
		Total	+*Thul*	−*Thul*	Mandib	Paradox	Arthro excluding Pycno	Arthro excluding Euchel	Euchel plus Myria	Chel	Euchel	Myria	Pancr	Pancr excluding Ostra	Copepoda plus Malaco	Hexapod
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Test 1[b]:																
Euchel:genes	13,877	35	25	10	21	70	21	8	51	25	100∗	100	100	45	59	11
Pancrust:genes	12,436	34	23	11	95	<5	12	86	<5	<5	100	100	100∗	82	88	23
Myria:genes	11,732	30	19	11	63	32	25	41	20	13	100	100∗	100	49	58	42
Test 2[c]:																
Thulinia:genes	17,740	51	51	0	71	27	24	48	12	18	100	100	100	37	84	34
EuMyPan:genes	3,618	4	4	0	48	32	17	35	8	35	100∗	100∗	100∗	95	81	<5

[a] All analyses are likelihood-based using the preferred model of nucleotide substitution. Data set names in column 1 refer to the test group recovered by each of the sequence sets in the data set (see below). Groups recovered in the tree of highest likelihood have their bootstrap percentages underlined. Bootstrap values followed by asterisks identify groups whose recovery has been preselected in constructing the various data sets. Column 1: name of data set; column 2: number of characters in data set; column 3: total number of gene-sequence sets of sequences (maximum of 13 sequences/set, see Table 2); column 4: number of sets of sequences for which the *Thulinia* (*i.e.*, the nonarthropod outgroup species) sequence is not missing; column 5: number of sets of sequences for which the *Thulinia* sequence is missing; column 6: Mandibulata; column 7: Paradoxopoda; column 8: Arthropoda excluding Pycnogonida; column 9: Arthropoda excluding Euchelicerata; column 10: Euchelicerata + Myriapoda; column 11: Chelicerata; column 12: Euchelicerata; column 13: Myriapoda; column 14: Pancrustacea; column 15: Pancrustacea excluding Ostracoda; column 16: Copepoda + Malacostraca; column 17: Hexapoda.

[b] Test 1: Does biasing toward recovery of one test group improve overall tree topology? To test this, data sets are constructed by combining sequence sets (selected from *1–68* in Table 2) that recover one of three basal test groups—Euchelicerata, Myriapoda, and Pancrustacea—in their individual-sequence likelihood bootstrap consensus diagrams (LE50 option = yes in PAUP*4.0). Some sequence sets are lacking the sequence for the outgroup *Thulinia* (compare values in columns 3 and 4), and in these cases basal group recovery (e.g., Mandibulata and Paradoxopoda) was assumed if any rooting could yield that result. Likelihood + bootstrap analyses are performed on the combined-sequence data sets, and results are shown for selected groups (columns 6 to 17). Combined sequence data set names in column 1: *Euchel:genes*: all sequence sets that recover Euchelicerata; *Pancrust:genes*: all sequence sets that recover Pancrustacea; *Myria:genes*: all sequence sets that recover Myriapoda.

[c] Test 2: Does exclusion of sequence sets that are missing the sequence of the outgroup taxon *Thulinia* bias the outcome of Test 1? To test this, data sets are constructed by combining sequence sets in Table 2 for which the sequence of the outgroup taxon *Thulinia* is not missing. Combined sequence data set names in column 1: *Thulinia:genes*: all sequence sets that include *Thulinia* sequences; *EuMyPan:genes*: that subset of *Thulinia:genes* sequence sets that recover Euchelicerata, Myriapoda, and Pancrustacea in each of the constituent individual-gene likelihood bootstrap consensus diagrams (LE50 option = yes in PAUP*4.0). Likelihood + bootstrap analyses are performed on these combined-sequence data sets, and results are shown for selected groups (columns 6 to 17).

ancient nodes, although the reported accelerated evolution of *C. elegans* and *D. melanogaster* genomes may have limited the discovery of new genes. Additionally, the vast majority of genes had only one detectable homolog, so it was reasonable to expect that inter-species comparisons would be of orthologs that track speciation events, rather than unrelated gene duplication events. Our decision to test PCR primers using reverse-transcription coupled with PCR (RT-PCR) resulted in the successful amplification of segments from 62 genes. Although there are tradeoffs for both RT-PCR and direct-gene-sequencing approaches, the current approach is satisfactory until more rapid sequencing technologies are developed.

The PCR primers listed in online Appendix 1 are likely to be useful for other investigations within Arthropoda and protosomes more broadly, but use of RT-PCR rather than direct gene amplification is likely to be necessary to maximize the utility of many of these primers.

Selecting a Method of Analysis and a Conservative Character Set for 13 Taxa

Our analyses began by analyzing amino acid and DNA sequence data for all 68 gene regions using parsimony, likelihood, and Bayesian approaches (Fig. 1). These analyses are presented in detail in Results, but the main conclusions are straightforward:

1. The nucleotide composition of the entire data set deviated significantly from homogeneity even when only the first two codon positions (character set *nt12*) were included (see Table 3). In contrast, significant deviation from compositional homogeneity disappeared when potentially synonymous first codon positions additionally were excluded from the data set (i.e., the *LR1* character set: sites coding for leucine or arginine).
2. Analysis of the two significantly heterogeneous character sets (*nt123* and *nt12*) yielded troubling results when analyzed by parsimony, likelihood under a GTR+I+G model, and partitioned Bayesian approaches (Fig. 1a, b, f to i). The *nt123* character set yielded very strongly supported phylogenies that conflict just as strongly with each other depending on the method of analysis. Although the *nt12* character set yielded very strongly supported phylogenies that were (except for parsimony) identical regardless of method of analysis, every one of these phylogenies strongly contradicted the monophyly of the Hexapoda (Fig. 1f to i).
3. Bayesian partitioning of *nt12* and *nt3* did not correct the troublesome appearance of overconfident and strongly conflicting nodes within and between analyses (compare the linked and unlinked substitution models in Fig. 1a, b, respectively). Even when partitioning took into account potentially synonymous change at the first codon position (*LR1*) after removal of *nt3*, Bayesian analysis still contradicted the monophyly of the Hexapoda at posterior probabilities of 1.0 (Fig. 1i). This illustrates the well-documented tendency for MrBayes, v. 3.1, to inflate posterior probabilities in the presence of long terminal branches and short internodes (Huelsenbeck et al., 2002; Cummings et al., 2003; Lewis et al., 2005; Yang and Rannala, 2005).
4. Overall nodal support in likelihood analysis dropped sharply when all potentially synonymous (*LR1* and *nt3*) sites were excluded from the data (Fig. 1k). This is apparent from the observation that multiple nodes strongly supported by the likelihood analysis of *nt12* (marked by asterisks in Fig. 1h) fall to <50% in the likelihood analysis of the *noLR1+nt2* character set (Fig. 1k). Only three strongly supported nodes remained (plus near-strong support for Mandibulata), but they happened to be the three groups with overwhelming morphological and/or molecular support over the past two decades—Euchelicerata, Myriapoda, and Pancrustacea (Fig. 1j to l). This suggests that the *noLR1+nt2* character set—the only one that did not deviate from nucleotide homogeneity—yields conservative phylogenetic inference at deeper nodes. Interestingly, likelihood analysis under a codon model also looked very encouraging (Fig. 1c). In particular, the same three groups (and only those) were again recovered with 100% bootstrap support. Unfortunately, further data exploration under a codon model proved impractical even with access to grid computing (see Materials and Methods). However, we note that the codon-model analysis (Fig. 1c) also strongly supported *Nebalia* + *Mesocyclops* (84% BP) and that it recovered Hexapoda (<50% BP), two groups strongly supported under certain conditions by *noLR1+nt2* (e.g., see Fig. 2e), as discussed elsewhere in this article.

Finally, in terms of providing an estimate with no false positives (Euchelicerata, Myriapoda, and Pancrustacea are hereafter *assumed* to be correct), our analyses suggest that likelihood analysis of the *noLR1+nt2* character set is the most practical, conservative means of analysis. The distribution of support values derived from the *noLR1+nt2* data sets (Fig. 1j, k) is most similar to those of parsimony and likelihood for amino acids (Fig. 1d, e), which by definition have no synonymous substitutions.

Further Optimization of the Compositionally Homogeneous noLR1+nt2, *13-Taxon Data Set for Phylogenetic Analysis: Factoring in Rates*

The failure of even our preferred, conservative method of analysis (i.e., likelihood) and character subset (i.e., *noLR1+nt2*) to strongly recover more than three groups led us to test the additional effect of sequentially excluding faster-evolving genes (Table 2, Fig. 2). In addition to the strongly supported Euchelicerata, Myriapoda, and Pancrustacea, three of the remaining seven nodes receive strong support as sets of rapidly evolving genes were sequentially removed:

1. Hexapoda is recovered with the removal of the 10 fastest genes, and bootstrap support levels rise to 79% when the 37 fastest gene regions are excluded *(38–68)*.
2. Malacostraca + Copepoda (= *Mesocyclops* + *Nebalia*) was recovered with all gene regions *(1–68)* but support increases from 68% to 91% after only the 10 fastest gene regions are removed *(11–68)* and up to 98% with the *21–68* character set.
3. Mandibulata (= Pancrustacea + Myriapoda) is consistently recovered, with support values ranging from 73% to 89%.

The remaining four nodes differ between analyses and consistently receive low support values. An examination of support values in Figure 2 reveals that removing the 20 fastest gene regions produces the most decisive character set, with 8 of 10 nodes supported at greater than 55% and 6 of 10 supported at >71% (Fig. 2c). Of course, decisiveness (or precision) and accuracy are not necessarily coincident (see Hillis and Bull, 1993; Alfaro and Holder, 2006), but at least nucleotide compositional heterogeneity, among others, has been removed as a source of misleading signal. Interestingly, decisiveness decreases as more genes are removed (Fig. 2d, e; Yang, 1998).

Evaluating Concordance as a Criterion for Optimizing Phylogenetic Analysis

Concordance with previously established, higher-level relationships has long been proposed as a criterion for favoring those genes that provide supporting phylogenetic signal and for disfavoring those that conflict (Friedlander et al., 1992; Graybeal, 1994). Concordance tests were carried out for individual genes with the *noLR1+nt2* character set. In Test 1, bootstrap trees were inspected for recovery of three strongly supported and noncontroversial clades: Euchelicerata, Myriapoda, and Pancrustacea. These yielded overlapping sets of gene regions supporting each group (*Euchel:genes, Myria:genes,* and *Pancrust:genes* in Table 6; note that bootstrap support could be <50% and that these are *not* the same individual gene regions identified in Table 4, for which the criterion was >80% BP support). In Test 2, we considered two nested sets of genes: those for which the tardigrade outgroup was successfully amplified (*Thuliana:genes*) and the subset of those *Thuliana:genes* that are concordant with all three groups in Test 1 (*EuMyPan:genes*).

The results are encouraging and almost entirely consistent with the results from the previous analysis in which rapidly evolving genes were progressively excluded (Fig. 2).

1. Four of the five character sets identified in Tests 1 and 2 recovered the Mandibulata (= Pancrustacea + Myriapoda), albeit with varying levels of bootstrap support (48 to 95% BP; see Table 6). These include gene regions concordant with Pancrustacea or Myriapoda and both sets of gene regions with the tardigrade outgroup. Mandibulata was also recovered with all 68 gene regions (Fig. 2a) and in all cases when sets of rapidly evolving genes were progressively excluded (Fig. 2b to e).
2. These same four of five character sets also supported, but only weakly except for *Pancrust:genes* (see Table 6), a sister group relationship between Euchelicerata and all other Arthropoda (including Pycnogonida). Again, this relationship was recovered in four of the five character exclusion sets in Figure 2, although again generally weakly.
3. These same four character sets supported Malacostraca + Copepoda, another group also supported by the character exclusion sets in Figure 2.
4. The fifth character set—*Euchel:genes* (concordant with the Euchelicerata)—serves as a somewhat cautionary tale. This set moderately supports (70% BP) Euchelicerata + Myriapoda + Pycnogonida (the aforementioned Paradoxopoda, although without a monophyletic Chelicerata) over Mandibulata. This character set only weakly supports Malacostraca + Copepoda (59% BP), which is strongly supported (81% to 88% BP) in three of the other four character sets (and recovered in all) and in the character exclusion sets in Figure 2b to e.
5. Although concordance is a risky criterion, if well-established groups are known, then preselecting gene regions that recover them *all* might be appropriate, although this would certainly dramatically reduce the size of the data set (e.g., see *EuMyPan:genes* in Test 2).

Increased Taxon Sampling

In a final set of analyses, we assess the effect of adding 49 taxa from across the Arthropoda and Tardigrada that are only sequenced for 3 of the 68 gene regions: *EF-1α*, *EF-2,* and *Pol II*. This allowed us to assess the effects of increased taxon sampling on the conclusions reached for the full 13-taxon data set. Conversely, this also allowed us to assess the effect of the additional 65 gene regions on relationships previously supported by *EF-1α*, *EF-2*, and *Pol II* by themselves (Regier et al., 2005a). It should be noted that the pattern of missing data in the current study is highly ordered; either most gene regions are present for the same 13 taxa or they are present for all 62 taxa (Wiens, 2006).

The earlier study (Regier et al., 2005a) of 62 taxa with three genes supports many of the conclusions in the current 62-taxon study but most significantly maintains strong support for Hexapoda and for ((Malacostraca + Thecostraca) + Copepoda). Support for Hexapoda, in particular, increases in the current study as the taxon sample increases from 13 to 62, including 10 additional hexapods (Figs. 2, 3; Table 5). Furthermore, Thecostraca, which is missing in the 13-taxon analyses (Fig. 2), remains strongly grouped with Malacostraca and Copepoda in the current 62-taxon study, even though Thecostraca is represented by only three genes.

Of course, the current study evaluates far more sequence data than the earlier 3-gene/62-taxon study, and this must provide the explanation for the substantial support for Mandibulata in the current study versus the inability to decide between Mandibulata and Paradoxopoda in the earlier study. Further supporting this conclusion is the observation that Mandibulata receives substantial support in the 13-taxon analyses (see particularly Fig. 2b to d).

Additional data are also responsible for a more consistent, albeit at times still weakly supported, placement of Ostracoda + Branchiura as sister group to all other Pancrustacea, something that was only suggested in the earlier study (Regier et al., 2005a; Fig. 2c; Tables 5, 6).

However, there are at least two meaningful differences between this and the earlier study. First, in the earlier study, a group including Hexapoda, Cephalocarida, Branchiopoda, and Remipedia was consistently recovered, albeit with only moderate bootstrap support (BP ≤ 71%); whereas in the current study, this group is recovered only once (i.e., with the 62-taxon *11–68* data set), even when analyzing the same three genes, as long as *LR1* (and *nt3*) characters are not included (Table 5, Fig. 1). Because *LR1* characters were included in the earlier study, this suggests that misleading synonymous change may underlie the previous, more consistent grouping of these taxa. However, we note that in the current study Hexapoda + Remipedia + Cephalocarida are still recovered by three of six 62-taxon *noLR1+nt2* character sets (once together with Branchiopoda), although with low bootstrap support (Table 6). In contrast, Branchiopoda is now closer to the well-supported (Copepoda, (Thecostraca, Malacostraca)) clade in five of six cases, so the major difference in the two studies may be largely in the placement of Branchiopoda.

A second difference is that the earlier three-gene study revealed strong support for Chelicerata (82% to 87% BP in various analyses), even when analyzing amino acids, whereas the current study of 62 taxa and *noLR1+nt2* character sets only recovers Chelicerata with the slowest gene regions (i.e., *38–68*, which includes *EF-1α*, *EF-2*, and *Pol II*) and 62 taxa. With more inclusive character sets, Euchelicerata is consistently, but weakly (53% BP), recovered as sister group to Pycnogonida + Mandibulata for both 13 and 62 taxa (Fig. 2, Table 5). Although synonymous change does favor Chelicerata (e.g., Fig. 1, Table 5), this cannot be the entire explanation. Instead, the *slow* genes, particularly *EF-1α*, *EF-2*, and *Pol II* (see penultimate column in Table 5), have a definite nonsynonymous signal favoring Chelicerata. Interestingly, the various concordance tests (Table 6) provide some evidence for a sister group relationship between Euchelicerata and all other arthropods over Chelicerata, but support is seldom strong (see also Fig. 2). Hopefully, a more definitive consideration of arthropod phylogeny, including the question of Chelicerata and its alternatives, should be possible when the 68 gene regions are available for many more than 13 taxa (Simmons et al., 2004). For now, we present our best estimate of arthropod phylogeny in summary Figure 4.

Analytical Conclusions and Recommendations

Exclude All Potentially Synonymous Sites

We make a strong case earlier in the Discussion that synonymous substitutions at both *nt3* and *nt1* can mislead phylogenetic inference in the 13-taxon data set. Specifically, synonymous changes at *nt1* positions coding for leucine and arginine codons can cause the same analytical problems typically associated with *nt3*, especially with regard to base compositional heterogeneity. Excluding *LR1* eliminated the significant deviation from compositional heterogeneity present in the *nt12* data set. This is true even though the *LR1* positions only account

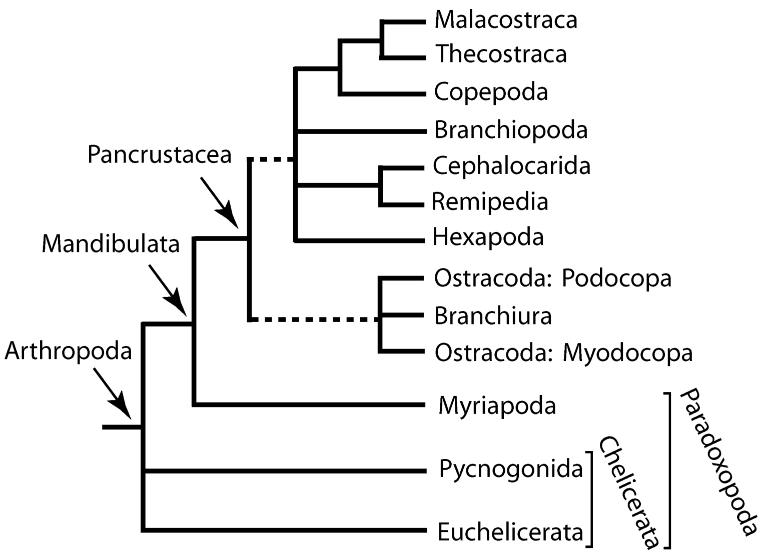

FIGURE 4. Summary diagram of higher-level arthropod relationships. Taxonomic groups that are connected by solid lines are well supported in the current and/or previous studies. Groups connected by dashed lines receive increased support in this study (relative to Regier et al., 2005a) and represent our best estimate of relationships.

for 12% of the *nt12* data set. We illustrate this with a data set *(21–68)* that includes no *fast* genes, our largest taxon sample (62 taxa), and our most conservative analytical method (likelihood) by comparing results of *nt12* and *noLR1+nt2* character sets (*cf.* BP values for *nt12/21–68* and *noLR1+nt2/21–68* in Table 5). Analyses of the two data sets yield largely similar results with two important exceptions. Bootstrap values decreased from 77% to <50% for Paradoxopoda and from 96% to <50% for Chelicerata when using *nt12* and *noLR1+nt2*, respectively, thus illustrating that strong support for Paradoxopoda and Chelicerata disappears when the *LR1* characters are excluded. One obvious interpretation is that Paradoxopoda and Chelicerata receive their strong support from less reliable synonymous change. That the codon model (Fig. 1c) recovers many of the same groups favored by the *noLR1+nt2* analyses can also be interpreted to support our presumption that synonymous change can be misleading. Because synonymous change generally occurs much more rapidly—but also happens to be compositionally heterogeneous—the codon model, like *noLR1+nt2*, effectively and specifically downweights synonymous characters in assessing the contributions of the various characters to the overall topological estimate.

Pay Close Attention to Rapidly Evolving Genes

Although our methods of gene selection presumably excluded many very rapidly evolving genes by our a priori criteria (see Materials and Methods), many rapidly evolving gene regions remained in our final set of 68 regions. Removing the most rapidly evolving gene regions led to more accurate inference (e.g., recovering a monophyletic Hexapoda) while still being conservative (Fig. 2).

Acknowledgments

We thank N. Takebayashi for making a Perl script to facilitate handling of sequences during primer design and the Associate Editor and two anonymous reviewers for their many helpful comments. The research described herein was supported by the National Science Foundation under grant number DEB-0120635 (C.W.C., J.C.R., J.W.S., J.W.M.) and the Maryland Agricultural Experiment Station (J.W.S.).

References

Alfaro, M. E., and M. T. Holder. 2006. The posterior and the prior in Bayesian phylogenetics. Annu. Rev. Ecol. Evol. Syst. 37:19–42.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410.

Arisue, N., M. Hasegawa, and T. Hashimoto. 2005. Root of the Eukaryota tree as inferred from combined maximum likelihood analyses of multiple molecular sequence data. Mol. Biol. Evol. 22:409–420.

Bazinet, A. L., and M. P. Cummings. 2008. The Lattice Project: A grid research and production environment combining multiple grid computing models. *In* Distributed & grid computing—Science made transparent for everyone. Principles, applications and supporting communities (M. H. W. Weber, ed.). Rechenkraft.net, Marburg. In press.

Bazinet, A. L., D. S. Myers, J. Fuetsch, and M. P. Cummings. 2007. Grid services base library: A high-level, procedural application program interface for writing Globus-based grid services. Future Generation Computer Systems 23:517–522.

Blanquart, S., and N. Lartillot. 2006. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. Mol. Biol. Evol. 23:2058–2071.

Boore, J. L., T. M. Collins, D. Staton, L. L. Daehler, and W. M. Brown. 1995. Deducing the pattern of arthropod phylogeny from mitochondrial gene rearrangements. Nature 376:163–165.

Boore, J. L., D. V. Lavrov, and W. M. Brown. 1998. Gene translocation links insects and crustaceans. Nature 392:667–668.

Cummings, M. P., S. A. Handley, D. S. Myers, D. L. Reed, A. Rokas, and K. Winka. 2003. Comparing bootstrap and posterior probability values in the four-taxon case. Syst. Biol. 52:477–487.

Cummings, M. P., and J. C. Huskamp. 2005. Grid computing. Educause Rev. 40:116–117.

Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450:203–218.

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783–791.

Felsenstein, J. 2004. Choosing among nonnested hypotheses: AIC and BIC. Pages 315–318 *in* Inferring phylogenies. Sinauer Associates, Sunderland, Massachusetts.

Fortey, R. A., and R. H. Thomas (eds.). 1998. Arthropod relationships. Chapman & Hall, London.

Friedlander, T. P., J.C. Regier, and C. Mitter. 1992. Nuclear gene sequences for higher level phylogenetic analysis: 14 promising candidates. Syst. Biol. 41:483–490.

Friedrich, M. and D. Tautz. 1995. Ribosomal DNA phylogeny of the major extant arthropod classes and the evolution of myriapods. Nature 376:165–167.

Gatesy, J., R. DeSalle, and N. Wahlberg. 2007. How many genes should a systematist sample? Conflicting insights from a phylogenomic matrix characterized by replicated incongruence. Syst. Biol. 56:355–363.

Giribet, G., S. Richter, G. D. Edgecombe, and W. C. Wheeler. 2005. The position of crustaceans within Arthropoda—Evidence from nine molecular loci and morphology. *In* Crustacea and arthropod relationships (S. Koenemann and R. Jenner, eds.). Crustacean Issues 16:307–352.

Glenner, H., P. F. Thomsen, M. B. Hebsgaard, M. V. Sorensen, and E. Willerslev. 2006. Perspectives: The origin of insects. Science 314:1883–1884.

Graybeal, A. 1994. Evaluating the phylogenetic utility of genes: A search for genes informative about deep divergences among vertebrates. Syst. Biol. 43:174–193.

Hillis, D. M., and J. J. Bull. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. 42:182–192.

Huelsenbeck, J. P., B. Larget, R. E. Miller, and F. Ronquist. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. Syst. Biol. 51:673–688.

Huelsenbeck, J. P., and F. Ronquist. 2001. MrBayes: Bayesian inference of phylogeny. Bioinformatics 17:754–755.

Hwang, U. W., M. Friedrich, D. Tautz, C. J. Park, and W. Kim. 2001. Mitochondrial protein phylogeny joins myriapods with chelicerates. Nature 413:154–157.

Johnson, W. E., E. Eizirik, J. Pecon-Slattery, W. J. Murphy, A. Antunes, E. Teeling, and S. J. O'Brien, 2006. The Late Miocene radiation of modern Felidae: A genetic assessment. Science 311:73–77.

Lartillot, N., H. Brinkmann, and H. Philippe. 2007. BMC Evol. Biol. (Suppl 1):S4.

Lewis, P. O., M. T. Holder, and K. E. Holsinger. 2005. Polytomies and Bayesian phylogenetic inference. Syst. Biol. 54:241–253.

Lockhart, P. J., Steel, M. A., Hendy, M. D., and Penny, D. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. 11:605–612.

Maddison, W. P., and D. R. Maddison. 2002. MacClade 4: Analysis of phylogeny and character evolution. Sinauer Associates, Sunderland, Massachusetts.

Madsen, O., M. Scally, C. J. Douady, D. J. Kao, R. W. DeBry, R. Adkins, H. M. Amrine, M. J. Stanhope, W. W. D. Jong, and M. S. Springer. 2001. Parallel adaptive radiations in two major clades of placental mammals. Nature 409:610–614.

Mallatt, J. M., J. R. Garey, and J. W. Shultz. 2004. Ecdysozoan phylogeny and Bayesian inference: First use of nearly complete 28S and 18S rRNA gene sequences to classify the arthropods and their kin. Mol. Phyl. Evol. 31:178–191.

Mallatt, J., and G. Giribet. 2006. Further use of nearly complete 28S and 18S rRNA genes to classify Ecdysozoa: 37 more arthropods and a kinorhynch. Mol. Phyl. Evol. 40:772–794.

Martin, J. W., and G. E. Davis. 2001. An updated classification of the recent Crustacea. Natural History Museum of Los Angeles County, Science Series no. 39.

Mitchell, A., C. Mitter, and J. C. Regier. 2000. More taxa or more characters revisited: Combining data from nuclear protein-encoding genes for phylogenetic analysis of Noctuoidea (Insecta: Lepidoptera). Syst. Biol. 49:202–224.

Murphy, W. J., E. Eizirik, W. E. Johnson, Y. P. Zhang, O. A. Ryder, and S. J. O'Brien. 2001. Molecular phylogenetics and the origins of placental mammals. Nature 409:614–618.

Myers, D. S., and M. P. Cummings. 2003. Necessity is the mother of invention: A simple grid computing system using commodity tools. J. Parallel Distrib. Comput. 63:578–589.

Myers, D. S., A. L. Bazinet, and M. P. Cummings. 2008. Expanding the reach of Grid computing: Combining Globus- and BOINC-based systems. Pages 71–85 *in* Grids for bioinformatics and computational biology (E.-G. Talbi and A. Zomaya, eds.). Wiley Book Series on Parallel and Distributed Computing. John Wiley & Sons, New York.

Naylor, G. J. P., and W. M. Brown. 1998. Amphioxus mitochondrial DNA, chordate phylogeny, and the limits of inference based on comparisons of sequences. Syst. Biol. 47:61–76.

Nylander, J. A. A. 2004. MrModelTest v2.2. Program distributed by the author. Evolutionary Biology Centre, Uppsala University, Sweden.

Pagel, M., and A. Meade. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst. Biol. 53:571–581.

Pennisi, E., 2007. News: Genomicists tackle the primate tree. Science 316:218–221.

Philip, G., and C. J. Creevey. 2005. The Opisthokonta and the Ecdysozoa may not be clades: Stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the Coelomata than Ecdysozoa. Mol. Biol. Evol. 22:1175–1184.

Philippe, H., F. Delsuc, F., H. Brinkmann, and N. Lartillot. 2005. Phylogenomics. Annu. Rev. Ecol. Evol. Syst. 36:541–562.

Podsiadlowski, L., and A. Braband. 2006. The mitochondrial genome of the sea spider Nymphon gracile (Arthropoda: Pycnogonida). BMC Genomics 7:284.

Podsiadlowski, L., A. Braband, and G. Mayer. 2008. The complete mitochondrial genome of the onychophoran *Epiperipatus biolleyi* reveals a unique transfer RNA set and provides further support for the Ecdysozoa hypothesis. Mol. Biol. Evol.25:42–51.

Regier, J. C., and D. Shi, 2005. Increased yield of PCR product from degenerate primers with nondegenerate, nonhomologous 5′ tails. BioTechniques 38:34–38.

Regier, J. C., and J. W. Shultz, 2001b. A phylogenetic analysis of Myriapoda (Arthropoda) using two nuclear protein-coding genes. Zool. J. Linn. Soc. 132:469–486.

Regier, J. C., and J. W. Shultz, 2001a. Elongation factor-2: A useful gene for arthropod phylogenetics. Mol. Phyl. Evol. 20:136–148.

Regier, J. C., H. M. Wilson, and J. W. Shultz. 2005b. Phylogenetic analysis of Myriapoda using three nuclear protein-coding genes. Mol. Phyl. Evol. 34:147–158.

Regier, J. C., J. W. Shultz, and R. E. Kambic. 2005a. Pancrustacean phylogeny: Hexapods are terrestrial crustaceans and maxillopods are not monophyletic. Proc. R. Soc. Lond. 272:395–401.

Regier, M.C., J. W. Shultz, R. E. Kambic, and D. R. Nelson. 2004. Robust support for tardigrade clades and their ages from three protein-coding nuclear genes. Invertebr. Biol. 123:93–100.

Remm, M., C. E. V. Storm, and E. L. L. Sonnhammer. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J. Mol. Biol. 314:1041–1052.

Rodríguez-Ezpeleta, N., H. Brinkmann, B. Roure, N. Lartillot, B.F. Lang, and H. Philippe. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst. Biol. 56:389–399.

Rokas, A., D. Krüger, and S. B. Carroll. 2005. Animal evolution and the molecular signature of radiations compressed in time. Science 310:1933–1938.

Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–782.

Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406–425.

Savard, J., D. Tautz, S. Richards, G. M. Weinstock, R. A. Gibbs, J. H. Werren, H. Tettelin, and M. J. Lercher. 2007. Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of Holometabolous insects. Genome Res. 16:1334–1338.

Simmons, M. P., T. G. Carr, and K. O'Neill. 2004. Relative character-state space, amount of potential phylogenetic information, and heterogeneity of nucleotide and amino acid characters. Mol. Phylogenet. Evol. 32:913–926.

Smith, S. W., R. Overbeck, C. R. Woese, W. Gilbert, and P. M. Gillevet. 1994. The genetic data environment and expandable GUI for multiple sequence analysis. Comput. Appl. Biosci. 10:671–675.

Staden, R., K. F. Beal, and J. K. Bonfield. 1998. The Staden package. Pages 115–130 *in* Bioinformatics methods and protocols (S. Misener and S. A. Krawetz, eds.). The Humana Press, Totowa, New Jersey.

Swofford, D. L. 2002. PAUP*: Phylogenetic analysis using parsimony (*and other methods) 4.0 beta. Sinauer Associates, Sunderland, Massachusetts.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. Pages 407–514 *in* Molecular systematics, 2nd edition (D. M., Hillis, C. Moriz, and B. K. Mable, eds.). Sinauer Associates, Sunderland, Massachusetts.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Whitfield, J. B., and P. J. Lockhart. 2007. Deciphering ancient rapid radiations. Trends Ecol. Evol. 22:258–265.

Wiens, J. J. 2006. Missing data and the design of phylogenetic analyses. J. Biomed. Informatics 39:34–42.

Wilgenbusch, J. C., D. L. Warren, and D. L. Swofford. 2004. AWTY: A system for graphical exploration of MCMC convergence in Bayesian phylogenetic inference. http://ceb.csit.fsu.edu/awty.

Yang, Z. 1998. On the best evolutionary rate for phylogenetic analysis. Syst. Biol. 47:125–133.

Yang, Z., and B. Rannala. 2005. Branch-length prior influences Bayesian posterior probability of phylogeny. Syst. Biol. 54:455–470.

Zwickl, D. J. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence data sets under the maximum likelihood criterion. PhD dissertation, the University of Texas at Austin. (see also: http://www.bio.utexas.edu/faculty/antisense/GARLI/GARLI.html)

First submitted 12 November 2007; reviews returned 21 January 2008; final acceptance 6 March 2008
Associate Editor: Frank Anderson