Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences

Article *in* Nature · February 2010

DOI: 10.1038/nature08742 · Source: PubMed

8 authors, including:

Jeffrey W Shultz
University of Maryland, College Park
90 PUBLICATIONS **3,661** CITATIONS
SEE PROFILE

Bernard Ball
University College Dublin
19 PUBLICATIONS **1,305** CITATIONS
SEE PROFILE

Regina Wetzer
Natural History Museum of Los Angeles County
55 PUBLICATIONS **1,673** CITATIONS
SEE PROFILE

Clifford W Cunningham
Duke University
223 PUBLICATIONS **11,064** CITATIONS
SEE PROFILE

doi:10.1038/nature08742

LETTERS

Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences

Jerome C. Regier[1], Jeffrey W. Shultz[1,2,3], Andreas Zwick[1], April Hussey[1], Bernard Ball[4], Regina Wetzer[5], Joel W. Martin[5] & Clifford W. Cunningham[4]

The remarkable antiquity, diversity and ecological significance of arthropods have inspired numerous attempts to resolve their deep phylogenetic history, but the results of two decades of intensive molecular phylogenetics have been mixed[1–7]. The discovery that terrestrial insects (Hexapoda) are more closely related to aquatic Crustacea than to the terrestrial centipedes and millipedes[2,8] (Myriapoda) was an early, if exceptional, success. More typically, analyses based on limited samples of taxa and genes have generated results that are inconsistent, weakly supported and highly sensitive to analytical conditions[7,9,10]. Here we present strongly supported results from likelihood, Bayesian and parsimony analyses of over 41 kilobases of aligned DNA sequence from 62 single-copy nuclear protein-coding genes from 75 arthropod species. These species represent every major arthropod lineage, plus five species of tardigrades and onychophorans as outgroups. Our results strongly support Pancrustacea (Hexapoda plus Crustacea) but also strongly favour the traditional morphology-based Mandibulata[11] (Myriapoda plus Pancrustacea) over the molecule-based Paradoxopoda (Myriapoda plus Chelicerata)[2,5,12]. In addition to Hexapoda, Pancrustacea includes three major extant lineages of 'crustaceans', each spanning a significant range of morphological disparity. These are Oligostraca (ostracods, mystacocarids, branchiurans and pentastomids), Vericrustacea (malacostracans, thecostracans, copepods and branchiopods) and Xenocarida (cephalocarids and remipedes). Finally, within Pancrustacea we identify Xenocarida as the long-sought sister group to the Hexapoda, a result confirming that 'crustaceans' are not monophyletic. These results provide a statistically well-supported phylogenetic framework for the largest animal phylum and represent a step towards ending the often-heated, century-long debate on arthropod relationships.

The molecular phylogeny of Arthropoda has proven difficult to resolve. In an attempt to overcome this, we increased both taxon and gene sampling relative to earlier studies. Our broad taxon sample includes the basal lineages of Hexapoda, every class of traditional 'Crustacea', every class in Myriapoda, every order in Arachnida and multiple representatives from Xiphosura, Pycnogonida and the outgroups Onychophora and Tardigrada.

Until recently, arthropod molecular phylogenetics relied mainly upon nuclear ribosomal DNA and mitochondrial sequences. Our data come from the complementary DNA of single-copy nuclear protein-coding genes, which represent the largest source of data for phylogenetics. Three phylogenomic studies of single-copy nuclear genes in the past year included 9 (ref. 13), 32 (ref. 14) and 6 (ref. 15) arthropod taxa, respectively (165 kilobases (kb), 319 kb and 432 kb of DNA sequences, not including missing data). The present study of 75 arthropods brings to bear 2.6 megabases of aligned arthropod DNA from 62 single-copy protein-coding genes (not including 18% missing data). This data set builds on a previous study that sequenced the same gene regions for 12 arthropods and one tardigrade outgroup (446 kb)[16]. With the exception of three nodes, that study showed unconvincing bootstrap support, which improved only modestly when we added 44 arthropods previously sequenced for only three genes (an additional 227 kb)[16]. The present study enlarges our earlier data set nearly fourfold, greatly improving support throughout the entire phylogeny.

Sequences for each of the 62 genes from 80 taxa were obtained using PCR primers designed to amplify genes determined a priori to be single-copy orthologues in *Drosophila melanogaster* when compared with *Caenorhabditis elegans* and *Homo sapiens*[16]. Alignment of these orthologues was based on translated amino-acid sequences, which are highly conserved. For example, the most divergent pair of protein sequences in the entire data set is still identical at 73% of their amino-acid sites. Uncertainty about homology at sites bracketing small insertion–deletion (indel) regions resulted in exclusion of approximately 6.5% of all sites[16].

The phylogeny shown in Figs 1 and 2 has largely robust bootstrap support from four methods of analysis that are well suited to inferring deep-level phylogenies (Fig. 1). Additional likelihood, Bayesian and parsimony analyses shown in Supplementary Figs 1–6 also support all major conclusions described here. With the major exception of ordinal relationships within Arachnida, bootstrap values above 80% and posterior probabilities of 1.0 pertain in Fig. 1 and Supplementary Figs 3–5. With few exceptions, we also recovered clades widely accepted by morphologists, which is an informal criterion that supports our conclusions based on bootstrap analyses. Furthermore, there was little evidence of strong conflict among 68 gene regions from 62 nuclear protein-coding genes. Single-gene analyses showed that 95% of all relationships with >70% bootstrap support agreed with the phylogeny from the concatenated genes shown in Fig. 1 (Supplementary Table 3). This paucity of strong conflict is consistent with the orthology of these sequences. Of the six newly named groups shown in bold in Fig. 1, only two had more than 70% bootstrap support in single-gene analyses (one gene each; Supplementary Table 3). This suggests that the strong support in Fig. 1 is the result of the cumulative phylogenetic signal across numerous gene regions. Bootstrap values derived for amino acids were sometimes lower than those derived from nucleotides. This result is due, at least in part, to the failure of amino-acid models to distinguish between serine residues encoded by TCN and AGY codons (A.Z. and J.C.R., manuscript in preparation), and is consistent with recent work questioning the performance of widely implemented amino-acid models in comparison with nucleotide and codon models[17,18].

At the deepest level, our phylogeny strongly supports Mandibulata[11] (Pancrustacea plus Myriapoda), a controversial result that

[1]Center for Biosystems Research, University of Maryland Biotechnology Institute, [2]Department of Entomology, [3]Institute for Bioscience and Biotechnology Research, University of Maryland, College Park, Maryland 20742, USA. [4]Department of Biology, Duke University, Durham, North Carolina 27708, USA. [5]Natural History Museum of Los Angeles County, Los Angeles, California 90007, USA.

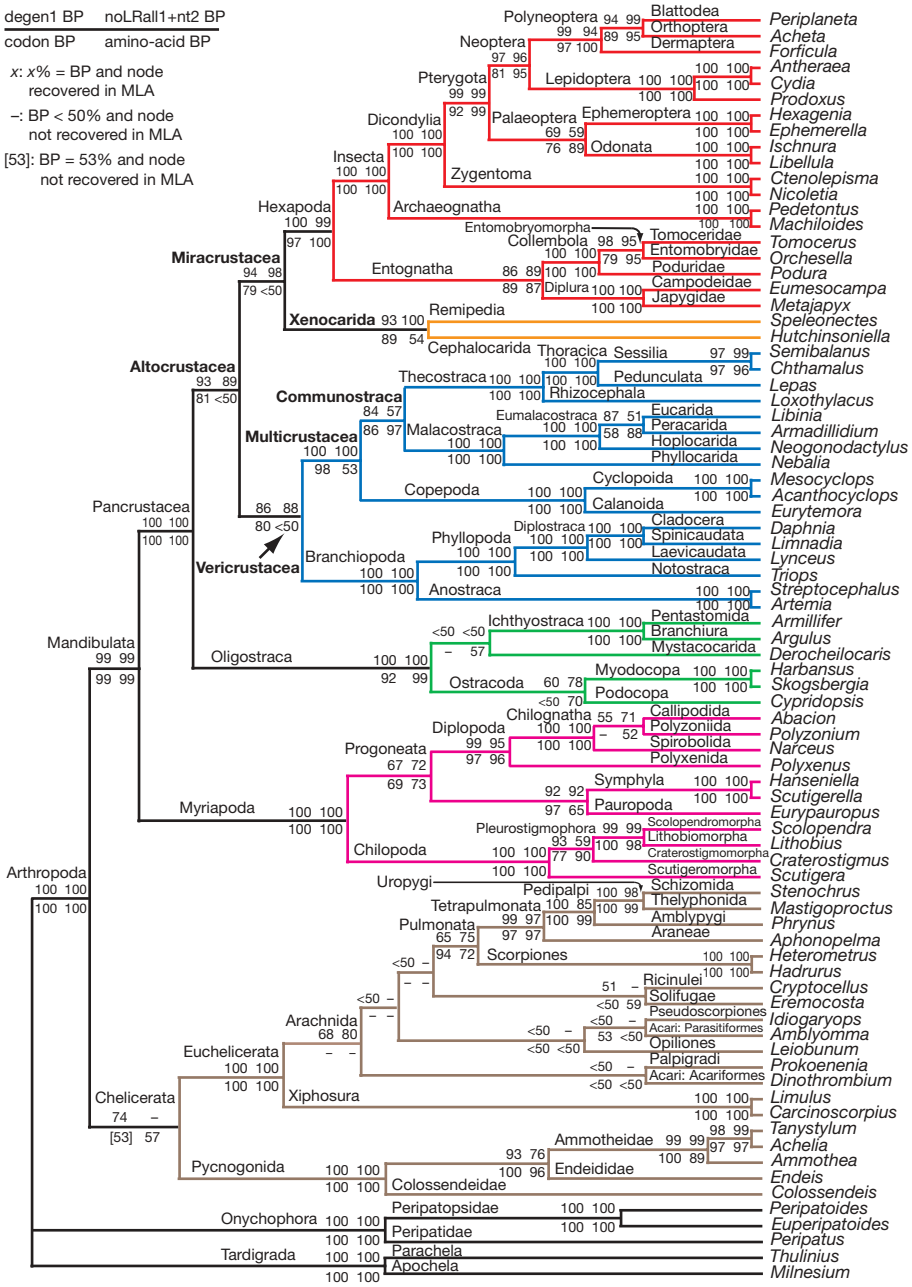

Figure 1 | Phylogenetic relationships of 75 arthropod and five outgroup species. Aligned sequences for 62 nuclear protein-coding genes were analysed under the likelihood criterion[29] using four strategies. Each strategy is designed to minimize deleterious effects of rapid sequence evolution and heterogeneous base composition: degen1, which fully degenerates all codons encoding the same amino acid; noLRall1+nt2, which excludes all third-codon positions and those first-codon positions encoding one or more leucine or arginine codons[16]; codon model, by which in-frame triplets of nucleotides are analysed directly under a model of codon change[30]; and amino acid. The degen1 maximum-likelihood topology is shown. Each group of four numbers shows the respective bootstrap percentages calculated using (clockwise from top left) degen1, noLRall1+nt2, amino acid and codon model (see figure key). MLA, maximum-likelihood analysis.

is robust to expanded outgroup sampling. Specifically, the addition of five more outgroup taxa (two nematodes, one priapulid, one mollusc and *H. sapiens*) to an amino-acid analysis had virtually no effect on the support for Mandibulata (reducing bootstrap support from 99% to 93%) and very little effect elsewhere in the phylogeny (Supplementary Fig. 7). Our strong support for Mandibulata contradicts several molecular studies that have placed Myriapoda as sister group to the Chelicerata (Euchelicerata plus Pycnogonida), a grouping so contrary to morphology that it was recently dubbed Paradoxopoda[2,5,12]. Broadly speaking, Paradoxopoda has received its strongest support from nuclear ribosomal genes[2,6,12] and mitochondrial genome sequences[5,19]. Recent phylogenomic studies have disagreed on support for[16] and against[13,15] the Mandibulata. Given our broad taxon and gene sampling, and stability with regard to outgroup choice, we now consider this issue to be resolved in favour of Mandibulata.

Although our phylogeny resolves many problems within Mandibulata, it does not resolve the status of Chelicerata, the group including Pycnogonida (sea spiders) and Euchelicerata (horseshoe crabs, scorpions and spiders). Of the four methods described in Fig. 1, only degen1 recovered Chelicerata with modest support (bootstrap percentage, BP = 74%), whereas amino acid (BP = 57%), noLRall1+nt2 (BP = 49%) and codon model (BP = 53%) only marginally favoured Chelicerata over the alternative of a more basal placement of Pycnogonida (BP = 41%, 48% and 43%, respectively). Monophyly of Euchelicerata (Xiphosura plus Arachnida) is strongly recovered, as

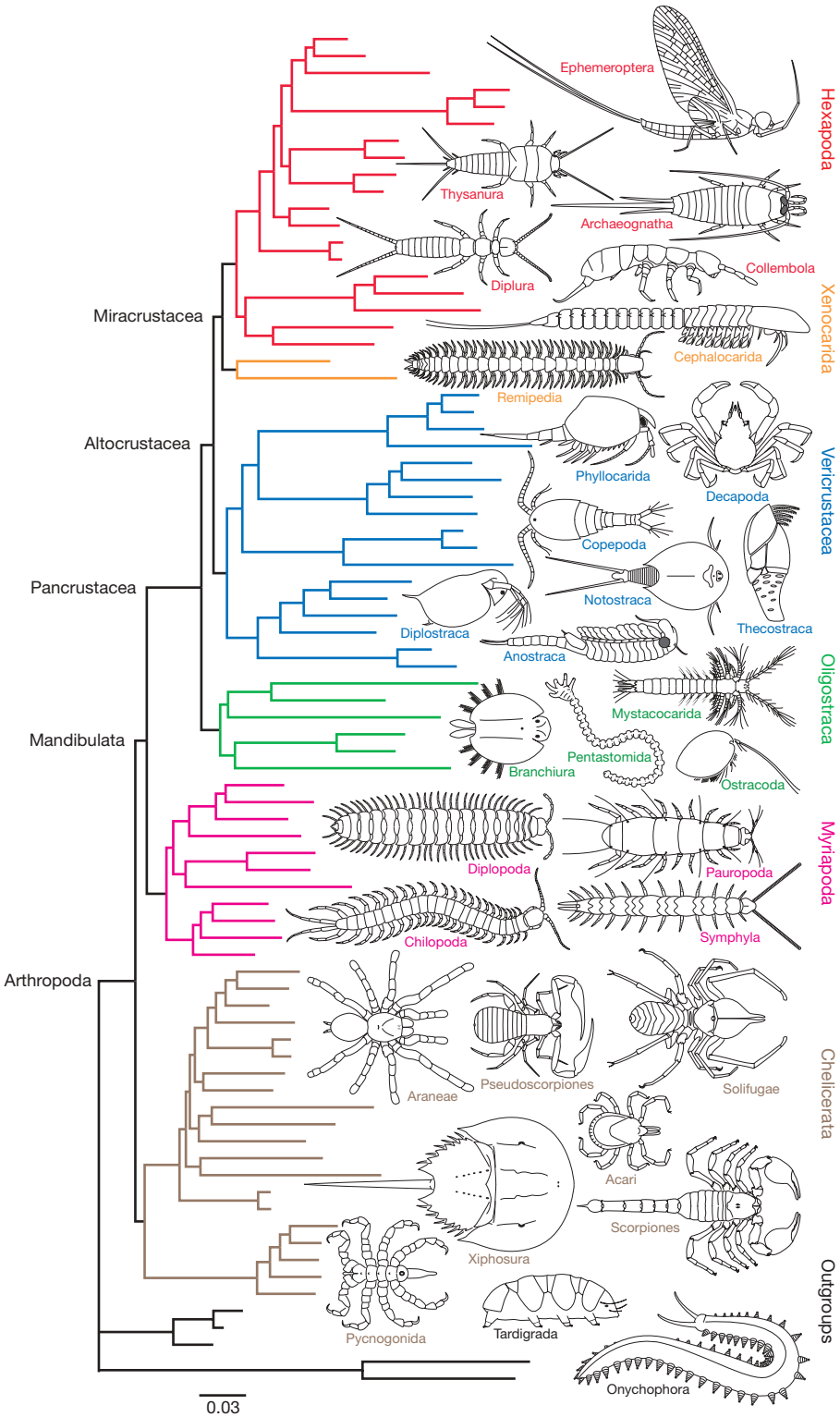

Figure 2 | Phylogram of relationships for 75 arthropod and five outgroup species. Based on likelihood analyses of 62 nuclear protein-coding genes. Branch lengths are proportional to the amount of inferred sequence change, with the topology and analytical conditions identical to the degen1 analysis in Fig. 1. Line drawings of representatives of the major taxonomic groups show the morphological disparity across Arthropoda. Scale bar, nucleotide changes per site.

is that of Arachnida. Among arachnids, our results strongly support the relationships among tetrapulmonate orders recovered in recent morphology-based studies[20], although our additional, strong placement of scorpions as the sister group to tetrapulmonates does not. Otherwise, there is a notable lack of robust resolution within Arachnida owing to a lack of phylogenetic information rather than intergene conflict (Supplementary Table 3).

Our phylogeny resolves the internal structure of Pancrustacea. Until now, it was unknown whether hexapods were the sister group to a monophyletic Crustacea or to some subset of crustaceans. Our phylogeny identifies Xenocarida ('strange shrimp') as the sister group to Hexapoda, thereby confirming the paraphyly of the Crustacea. Xenocarida includes two unusual and morphologically dissimilar classes of crustacean, Remipedia and Cephalocarida. We

place the xenocarids and hexapods in the newly named clade Miracrustacea ('surprising crustaceans'). Both Xenocarida and Miracrustacea are found in the maximum-likelihood trees for all four methods of analysis, although support varies. Bootstrap support for both Xenocarida and Miracrustacea is strong for noLRall1+nt2 and degen1 analyses (93–100%), moderately strong for codon analysis (79–89%) and weak for amino-acid analysis (17–54%).

The two classes of Xenocarida were discovered in the second half of the twentieth century, and for several decades each was viewed as the morphological model of the ancestral 'urcrustacean'[21]. Apart from the absence of eyes and the possible compensatory enhancement of olfactory nerve centres[22–24], Remipedia and Cephalocarida share few obvious synapomorphies. On the other hand, neurobiological studies have rejected the hypothesis that xenocarid brains are 'primitive'[22,23], and have proposed either Malacostraca or—as in our study—Hexapoda as possible relatives. Remipedes are relatively large predators with a long series of biramous swimming legs and are the only extant pancrustaceans that lack significant postcephalic tagmosis (Fig. 2). Cephalocarids are tiny particle feeders with large cephalic shields, distinct pereons (thoraxes) with foliaceous swimming legs, and long, legless pleons (abdomens). Despite these differences, a close relationship between remipedes and cephalocarids has been suggested before[4,21]; Fig. 1 provides robust likelihood bootstrap support for this proposal.

Our results strongly support the monophyly of Hexapoda, in contrast to mitochondrial studies that place Collembola (springtails) among 'crustaceans' rather than other hexapods[25]. Within Hexapoda, our results agree with long-standing, morphology-based hypotheses of the basal lineages[26]. Specifically, we recover Entognatha (non-insectan hexapods) as the sister group to Insecta, Archaeognatha (jumping bristletails) as the sister group to all other insects and Zygentoma (silverfish, firebrats and so on) as the sister group to Pterygota (winged insects). Relationships among pterygotes recovered here are largely non-controversial, although relationships between the extant paleopterous orders Ephemeroptera (mayflies) and Odonata (dragonflies and damselflies) have been a subject of persistent debate. We recover them as a monophyletic group.

The sister group to Miracrustacea is another unanticipated group, Vericrustacea ('true crustaceans'), which joins Malacostraca (crabs, shrimp and so on), Branchiopoda (fairy shrimp, water fleas and so on) and some members of the polyphyletic 'Maxillopoda', namely Thecostraca (barnacles) and Copepoda. The Vericrustacea encompass the most familiar and diverse groups from the traditional 'crustaceans', including species of economic significance and model organisms. Within the Vericrustacea are two other groupings not anticipated by morphology: the Multicrustacea ('numerous crustaceans': Malacostraca plus Thecostraca plus Copepoda) and the Communostraca ('common shelled ones': Malacostraca plus Thecostraca).

Our results agree with a recent ribosomal study[6] supporting both the monophyly of Oligostraca and its position as the sister group to all other pancrustaceans (the same study also identifies Communostraca). As originally proposed[27], Oligostraca included Ostracoda (seed shrimp) and Ichthyostraca, which encompass the highly derived, endoparasitic Pentastomida (tongue worms) and ectoparasitic Branchiura (fish lice). Significantly, our analysis adds Mystacocarida to Oligostraca. The mystacocarids are small, enigmatic crustaceans that live between sand grains along marine shores. Oligostracans are a disparate, ancient clade, and there is little in their gross morphology other than reduction in the number of body segments that would suggest a close relationship among them.

The fully terrestrial Myriapoda is the sister group to Pancrustacea. The basic internal phylogenetic structure of Myriapoda recovered here is consistent with that favoured by morphology, including monophyly of its constituent classes, namely Diplopoda (millipedes), Symphyla and Chilopoda (centipedes), and Pauropoda (only one species sampled). It also recovers Progoneata (Diplopoda plus Pauropoda plus Symphyla), the members of which each have an anteriorly placed gonopore. Our results differ from morphology-inspired hypotheses in uniting Pauropoda with Symphyla rather than with Diplopoda, a result that is also seen in recent analyses of nuclear ribosomal sequences[6,28].

In conclusion, our phylogenomic study provides a strongly supported phylogenetic framework for the arthropods, but the problem of reconstructing and interpreting morphological evolution within this diverse group remains. Our phylogeny highlights the large gaps in the morphological spectrum of extant arthropods that have complicated the task of morphology-based systematists. Our result has significant implications, as it requires taxonomists to acknowledge crustaceans as a paraphyletic grade of primitively aquatic mandibulates and to classify hexapods as a terrestrial clade within Pancrustacea. In particular, the position of Xenocarida (Remipedia plus Cephalocarida) as the sister group to Hexapoda, and the relatively derived placement of supposedly 'primitive' groups such as Branchiopoda, promises to alter views on the evolution of morphology and morphogenesis in Arthropoda.

METHODS SUMMARY

A flow chart of activities leading to the phylogenetic trees shown in Figs 1 and 2 can be found in Supplementary Fig. 1. Laboratory procedures are described in Methods. Supplementary Tables 1–3 respectively include a list of taxa, tests of nucleotide homogeneity and bootstrap support for taxonomic groups recovered in single-gene analyses. Supplementary Tables 4 and 5 include GenBank accession numbers. Likelihood analyses of DNA and amino acids for the 80 taxa can be found in Figs 1 and 2 and Supplementary Fig. 2, Bayesian analyses can be found in Supplementary Figs 3–5 and parsimony analysis can be found in Supplementary Fig. 6. Likelihood analysis of 85 taxa with expanded outgroup sampling can be found in Supplementary Fig. 7. Separate supplementary files include the 80-taxon nucleotide data matrix, the 80-taxon degen1 data matrix, and the 85-taxon amino-acid data matrix. Also available as supplementary files are the PERL script and instructions to produce degenerated nucleotide matrices (Degen1_v1_2.pl, Degen1_README.txt; http://www.phylotools.com).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 1 September; accepted 10 December 2009.
Published online 10 February 2010.

1. Abele, L. G., Kim, W. & Felgenhauer, B. E. Molecular evidence for inclusion of the phylum Pentastomida in the Crustacea. *Mol. Biol. Evol.* **6,** 685–691 (1989).
2. Friedrich, M. & Tautz, D. Ribosomal DNA phylogeny of the major extant arthropod classes and the evolution of myriapods. *Nature* **376,** 165–167 (1995).
3. Regier, J. C. & Shultz, J. W. Molecular phylogeny of the major arthropod groups indicates polyphyly of crustaceans and a new hypothesis for the origin of hexapods. *Mol. Biol. Evol.* **14,** 902–913 (1997).
4. Giribet, G., Edgecombe, G. D. & Wheeler, W. C. Arthropod phylogeny based on eight molecular loci and morphology. *Nature* **413,** 157–161 (2001).
5. Hwang, U. W., Friedrich, M., Tautz, D., Park, C. J. & Kim, W. Mitochondrial protein phylogeny joins myriapods with chelicerates. *Nature* **413,** 154–157 (2001).
6. Mallatt, J. & Giribet, G. Further use of nearly complete 28S and 18S rRNA genes to classify Ecdysozoa: 37 more arthropods and a kinorhynch. *Mol. Phylogenet. Evol.* **40,** 772–794 (2006).
7. Budd, G. E. & Telford, M. J. The origin and evolution of arthropods. *Nature* **457,** 812–817 (2009).
8. Boore, J. L., Lavrov, D. V. & Brown, W. M. Gene translocation links insects and crustaceans. *Nature* **392,** 667–668 (1998).
9. Phillips, M. J., Delsuc, F. & Penny, D. Genome-scale phylogeny and the detection of systematic biases. *Mol. Biol. Evol.* **21,** 1455–1458 (2004).
10. Rota-Stabelli, O. & Telford, M. J. A multi criterion approach for the selection of optimal outgroups in phylogeny: recovering some support for Mandibulata over Myriochelata using mitogenomics. *Mol. Phylogenet. Evol.* **48,** 103–111 (2008).
11. Snodgrass, R. E. *Evolution of the Annelida, Onychophora and Arthropoda* (Smithsonian Inst. Press, 1938).
12. Mallatt, J. M., Garey, J. R. & Shultz, J. W. Ecdysozoan phylogeny and Bayesian inference: first use of nearly complete 28S and 18S rRNA gene sequences to classify the arthropods and their kin. *Mol. Phylogenet. Evol.* **31,** 178–191 (2004).
13. Dunn, C. W. *et al.* Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* **452,** 745–749 (2008).
14. Timmermans, M. J. T. N., Roelofs, D., Mariën, J. & Van Straalen, N. M. Revealing pancrustacean relationships: phylogenetic analysis of ribosomal protein genes places Collembola (springtails) in a monophyletic Hexapoda and reinforces the discrepancy between mitochondrial and nuclear DNA markers. *BMC Evol. Biol.* **8,** 83 (2008).
15. Philippe, H. *et al.* Phylogenomics revives traditional views on deep animal relationships. *Curr. Biol.* **19,** 706–712 (2009).

16. Regier, J. C. *et al.* Resolving arthropod phylogeny: exploring phylogenetic signal within 41 kb of protein-coding nuclear gene sequence. *Syst. Biol.* **57,** 920–938 (2008).
17. Holder, M. T., Zwickl, D. J. & Dessimoz, C. Evaluating the robustness of phylogenetic methods to among-site variability in substitution processes. *Phil. Trans. R. Soc. B* **363,** 4013–4021 (2008).
18. Seo, T. & Kishino, H. Statistical comparison of nucleotide, amino acid, and codon substitution models for evolutionary analysis of protein-coding sequences. *Syst. Biol.* **58,** 199–210 (2009).
19. Podsiadlowski, L., Braband, A. & Mayer, G. 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 (2008).
20. Shultz, J. W. A phylogenetic analysis of the arachnid orders based on morphological characters. *Zool. J. Linn. Soc.* **150,** 221–265 (2007).
21. Schram, F. R. (ed.) *Crustacean Phylogeny* (Balkema, 1983).
22. Fanenbruck, M., Harzsch, S. & Wagele, J. The brain of the Remipedia (Crustacea) and an alternative hypothesis on their phylogenetic relationships. *Proc. Natl Acad. Sci. USA* **101,** 3868–3873 (2004).
23. Harzsch, S. Neurophylogeny: architecture of the nervous system and a fresh view on arthropod phyologeny. *Integr. Comp. Biol.* **46,** 162–194 (2006).
24. Boxshall, G. A. Crustacean classification: on-going controversies and unresolved problems. *Zootaxa* **1668,** 313–325 (2007).
25. Carapelli, A., Liò, P., Nardi, F., van der Wath, E. & Frati, F. Phylogenetic analysis of mitochondrial protein coding genes confirms the reciprocal paraphyly of Hexapoda and Crustacea. *BMC Evol. Biol.* **7** (suppl. 2), S8 (2007).
26. Hennig, W. *Insect Phylogeny* (Wiley, 1981).
27. Zrzavy, J., Hypsa, V. & Vlaskova, M. in *Arthropod Relationships* (eds Fortey, R. A. & Thomas, R. H.) 97–107 (Chapman and Hall, 1997).
28. Gai, Y.-H., Song, D., Sun, H. & Zhou, K. Myriapod monophyly and relationships among myriapod classes based on nearly complete 28S and 18S rDNA sequences. *Zool. Sci.* **23,** 1101–1108 (2006).
29. Zwickl, D. J. *Genetic Algorithm Approaches for the Phylogenetic Analysis of Large Biological Sequence Datasets under the Maximum Likelihood Criterion*. PhD thesis, Univ. Texas Austin (2006).
30. Goldman, N., Thorne, J. L. & Jones, D. T. Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *J. Mol. Biol.* **263,** 196–208 (1996).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements C.W.C. thanks W. Hartman for early insight into questions of arthropod phylogeny and D. Percy for sequencing. J.W.M. and R.W. thank N. Tait, G. Hampson and R. Hessler for help collecting samples. J.C.R. and A.Z. thank M. Cummings and A. Bazinet for making available grid computing, and the DNA Sequencing Facility at the Center for Biosystems Research, University of Maryland Biotechnology Institute. J.W.S. was supported by the Maryland Agricultural Experiment Station. C.W.C. was supported by the Whiteley Center. This work was funded by two programmes at the US National Science Foundation, namely Biocomplexity in the Environment: Genome-Enabled Environmental Science and Engineering, and Assembling the Tree of Life.

Author Contributions C.W.C., J.C.R., J.W.S., A.Z. and J.W.M. designed the project. J.W.S., J.W.M., R.W. and J.C.R. designed and carried out taxon sampling and collection. J.C.R. and C.W.C. supervised DNA sequencing and editing, with PCR templates generated by J.C.R., B.B. and others. J.C.R., A.Z., C.W.C. and J.W.S. decided on the strategy for data analysis and its implementation, with the degen1 coding method developed and implemented by J.C.R., A.H. and A.Z. J.C.R. and A.Z. assembled the Supplementary Information and submitted sequences to GenBank. J.W.S. and J.W.M. proposed the names for the new, strongly supported clades in the Pancrustacea. C.W.C. wrote the first draft of the manuscript, with major additions by J.C.R. and J.W.S. and additional contributions by J.W.M. and A.Z. All authors commented on the manuscript.

Author Information All sequences generated for this publication have been deposited in GenBank under the accession numbers given in Supplementary Tables 4 and 5. Full data matrices are available in Supplementary Information. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.W.C. (cliff@duke.edu).

METHODS

Specimen storage. Typically, live specimens were placed in at least 15 volume equivalents of 100% ethanol and stored at −85 °C. Sometimes, reaching a temperature of −85 °C took several weeks, in which case temporary storage at −20 °C or 0 °C proved satisfactory. For smaller specimens, storage in 100% ethanol at room temperature (~22 °C) was satisfactory, even for as long as one year.

Nucleic-acid extraction. We extracted total RNA from specimens using the SV Total RNA Isolation System (catalogue no. Z3100; Promega) with the DNase-digestion steps omitted and with intentional vortexing to fragment the genomic DNA.

PCR primers and amplification strategies. This study differs from most animal phylogenomic studies because the gene regions were screened a priori for orthology and for an intermediate rate of substitution[16]. Instead of screening EST libraries for genes that are most consistently expressed, our study amplified genes using reverse-transcription PCR (RT–PCR) of mRNA extractions. All PCR primers and amplification strategies have been published (see Appendix 1 of ref. 16, available at http://systbio.org/?q=node/295). To summarize, primers were designed from single-copy, orthologous protein-coding nuclear genes as described in ref. 16. We made pairwise alignments for 5,274 putative orthologues from *D. melanogaster* and *H. sapiens* identified using the InParanoid method[31], and further reduced them to 595 genes with >55% sequence identity between *D. melanogaster* and *H. sapiens* (sequences from *C. elegans* were added to the alignments when suitable[16]). Our primers amplified messenger RNAs encoded by 62 distinct genes from most of the 80 taxa in this study (Supplementary Table 1) and many Bilateria (C.W.C., unpublished results). The primers were completely degenerate with respect to the amino-acid code. All forward and reverse primers had 18-nucleotide-long M13REV or M13(-21) sequences appended to facilitate amplification[32] and automated sequencing.

Preparation of DNA fragments for automated sequencing. Laboratory methods have been described in detail (see Appendix 2 of ref. 16; http://systbio.org/?q=node/295). To summarize, we reverse transcribed total RNA and amplified the resulting cDNA using PCR (RT–PCR). This approach avoided problems with introns. These problems include lack of phylogenetic informativeness at the taxonomic level of this study and the difficulty of identifying the correctly sized amplicon due to variation in intron length and location. This strategy also kept our fragments within the size range for convenient sequencing. Without introns, expected amplicon lengths could be accurately inferred from orthologues of other species (for example *Drosophila* and other test taxa) and gel isolated. A problem with the RT–PCR approach is that amplification is probably limited to genes that are expressed in moderate to abundant amounts and in a relatively non-cell-specific manner, but we were still able to amplify 62 distinct genes.

Following gel isolation, the RT–PCR amplicon was typically re-amplified using one nested primer and one of primers used for RT–PCR (hemi-nested re-amplification), again followed by gel isolation. If the template concentration was insufficient for sequencing, we re-amplified the amplicon a second time using only M13REV and M13(-21) primers, and again followed this with gel isolation. Templates were sequenced using a 3730 DNA Analyser (Applied Biosystems).

Construction of data matrices. Sequencer chromatograms were edited and individual-gene, individual-taxon sequences were assembled using the PREGAP4 and GAP4 programs in the Staden package[33] (version staden_solaris-1-5-3) or SEQUENCHER (version 4.5; GeneCodes). Multi-sequence alignments were performed manually using the sequence editor Genetic Data Environment[34] (version 2.2) and with MAFFT[35] (version 6.716b). Sequences from individual gene regions were concatenated into the final nt123 nucleotide data matrix (39,261 sites with 6.5% unalignable sites removed using criteria described in ref. 16), and character sets were defined.

Phylogenetic analysis. Phylogenetic analyses were performed using PAUP* 4.0b10 (parsimony; ref. 36), GARLI 0.96b8 (likelihood; ref. 29) and MRBAYES 3.2 (Bayesian; refs 37–39) with the best-fit models as determined in MRMODELTEST 2.3 (ref. 40). The nt123 data matrix was analysed under a codon model (13,087 sites). The codon model incorporated four discrete d*N*/d*S* categories, where d*N*/d*S* refers to the ratio of nonsynonymous to synonymous change, and a GTR–gamma model to describe underlying nucleotide change. From the nt123 matrix, three other matrices were derived and analysed, as follows. In the matrix used for degen1 analyses (39,261 sites), codons for each amino acid were fully degenerated for the first and third codon positions using ambiguity coding, an extension of RY coding[9]. The resulting partially polymorphic nucleotide data matrix was analysed under the best-fit GTR–gamma–invariant model. In the matrix used for noLRall1+nt2 analyses (21,823 sites), nucleotide characters were excluded at all third-codon positions and at LRall1 sites (that is, first-codon positions that encoded one or more leucine or arginine codons[16,41]). Removing LRall1 eliminated significant heterogeneity in base composition (Supplementary Table 2). The resulting nucleotide data matrix was analysed under the best-fit GTR–gamma–invariant model. In the matrix used for amino acid analyses (13,087 sites), amino acids were first automatically generated using GARLI 0.96b8 and then analysed under the best-fit model of amino-acid change (JTT–gamma–F). For each strategy, multiple search replicates were performed to find the maximum-likelihood topology, and involved 11, 600, 675 and 574 search replicates for the strategies codon model, degen1, noLRall1+nt2 and amino acid, respectively. The non-parametric bootstrap analyses in Fig. 1 involved one search replicate for each of 105, 1065, 1005, and 1024 pseudoreplicates for the respective strategies. Scripts written in PERL for identifying LRall1 sites (LeuArg1_v1_2.pl) are available from files supplementary to ref. 16 (http://systbio.org/?q=node/295) and for the degen1 method in Supplementary Information (Methods Summary). The latest versions of both scripts will be available at http://www.phylotools.com.

Computational resources. To make thorough maximum-likelihood and bootstrap searches (involving, for example, hundreds of maximum-likelihood search replicates per data matrix) possible under a GTR model, we performed analyses in parallel using grid computing[38,42], through The Lattice Project[43]. Additionally, a dedicated cluster of Linux servers with 40 cores at 2.5 GHz each and 80 GB RAM was set up to allow the computationally very demanding codon-model analyses. Additional analyses were carried out using the CIPRES cluster at the San Diego Supercomputing Center, and on the Duke Shared Resource Cluster.

31. Remm, M., Storm, C. E. & Sonnhammer, E. L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **314,** 1041–1052 (2001).
32. Regier, J. C. & Shi, D. Increased yield of PCR product from degenerate primers with nondegenerate, nonhomologous 5′ tails. *Biotechniques* **38,** 34–38 (2005).
33. Staden, R., Judge, D. & Bonfield, J. Sequence assembly and finishing methods. *Methods Biochem. Anal.* **43,** 303–322 (2001).
34. Smith, S. W., Overbeck, R., Woese, C. R., Gilbert, W. & Gillevet, P. M. The genetic data environment and expandable GUI for multiple sequence analysis. *Comp. Appl. Biosci.* **10,** 671–675 (1994).
35. Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33,** 511–518 (2005).
36. PAUP*. v.4.0 (Sinauer Associates, Sunderland, Massachusetts, 2002).
37. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* **17,** 754–755 (2001).
38. Ronquist, F. & Huelsenbeck, J. P. MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19,** 1572–1574 (2003).
39. Altekar, G., Dwarkadas, S., Huelsenbeck, J. P. & Ronquist, F. Parallel Metropolis-coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics* **20,** 407–415 (2004).
40. MRMODELTEST. v.2 (Evolutionary Biology Centre, Uppsala University, 2004).
41. Regier, J. C. & Shultz, J. W. Elongation factor-2: a useful gene for arthropod phylogenetics. *Mol. Phylogenet. Evol.* **20,** 136–148 (2001).
42. Cummings, M. & Huskamp, J. Grid computing. *EDUCAUSE Rev.* **40,** 116–117 (2005).
43. Bazinet, A. L. & Cummings, M. C. in *Distributed & Grid Computing–Science Made Transparent for Everyone. Principles, Applications and Supporting Communities* (ed. Weber, M. H. W.) 2–13 (Tectum, 2009).