

# **Pancrustacean phylogeny in the light of new phylogenomic data: support for Remipedia as the possible sister group of Hexapoda**

## **Research Article**

Bjoern M von Reumont<sup>1§</sup>, Ronald A Jenner<sup>2</sup>, Matthew A Wills<sup>3</sup>, Emiliano Dell'Ampio<sup>4</sup>, Günther Pass<sup>4</sup>, Ingo Ebersberger<sup>5</sup>, Benjamin Meyer<sup>6</sup>, Stefan Koenemann<sup>7</sup>, Thomas M Iliffe<sup>8</sup>, Alexandros Stamatakis<sup>9</sup>, Oliver Niehuis<sup>1</sup>, Karen Meusemann<sup>1</sup>, Bernhard Misof<sup>1</sup>

<sup>1</sup> Zoologisches Forschungsmuseum Alexander Koenig, Adenauerallee 160, D-53113 Bonn, Germany

<sup>2</sup> Department of Zoology, The Natural History Museum, Cromwell Road, London SW7 5BD, United Kingdom

<sup>3</sup> Department of Biology and Biochemistry, University of Bath, The Avenue, Claverton Down, BA2 7AY, UK, Bath, United Kingdom

<sup>4</sup> Department of Evolutionary Biology, University of Vienna, Althanstraße 14, A-1090, Vienna, Austria

<sup>5</sup> Center for Integrative Bioinformatics Vienna (CIBIV), University of Vienna, Medical University of Vienna, University of Veterinary Medicine Vienna, Dr. Bohrgasse 9, A-1030 Vienna, Austria

<sup>6</sup> Biozentrum Grindel und Zoologisches Museum, Universität Hamburg, Martin-Luther-King Platz 3, D-20146 Hamburg, Germany

<sup>7</sup> Section Biology, Science and Technology, University of Siegen, Adolf-Reichwein-Straße 2, D-57068 Siegen, Germany

<sup>8</sup> Department of Marine Biology, Texas A&M University, 200 Seawolf Parkway, Galveston, Texas 77553, United States of America

<sup>9</sup> The Exelixis Lab, Scientific Computing Group, Heidelberg Institute for Theoretical Studies, Schloss-Wolfsbrunnenweg 35, D-69118, Heidelberg, Germany

§Corresponding author: Bjoern M von Reumont, Zoologisches Forschungsmuseum Alexander Koenig, Adenauerallee 160, 53113 Bonn,

E-mail: [bmvr@arcor.de](mailto:bmvr@arcor.de),

Phone: +49-228-9122-407

## Abstract

Remipedes are a small and enigmatic group of crustaceans, first described only 30 years ago. Analyses of both morphological and molecular data have recently suggested a close relationship between Remipedia and Hexapoda. If true, the remipedes occupy an important position in pancrustacean evolution and may be pivotal for understanding the evolutionary history of crustaceans and hexapods. However, it is important to test this hypothesis using new data and new types of analytical approaches. Here, we assembled a phylogenomic data set of 131 taxa, incorporating newly generated 454 EST data from six species of crustaceans, representing five lineages (Remipedia, Laevicaudata, Spinicaudata, Ostracoda, and Malacostraca). This data set includes all crustacean species for which EST data are available (46 species), and our largest alignment encompasses 866,479 amino acid positions and 1,886 genes. A series of phylogenomic analyses was performed to evaluate pancrustacean relationships. We significantly improved the quality of our

data for predicting putative orthologous genes and for generating data subsets by matrix reduction procedures, thereby improving the signal to noise ratio in the data.

Eight different data sets were constructed, representing various combinations of orthologous genes, data subsets, and taxa. Our results demonstrate that the different ways to compile an initial data set of core orthologs and the selection of data subsets by matrix reduction can have marked effects on the reconstructed phylogenetic trees. Nonetheless, all eight data sets strongly support Pancrustacea with Remipedia as the sister group to Hexapoda. This is the first time that a sister group relationship of Remipedia and Hexapoda has been inferred using a comprehensive phylogenomic data set that is based on EST data. We also show that selecting data subsets with increased overall signal can help to identify and prevent artifacts in phylogenetic analyses.

## **Keywords**

Phylogenomics, EST, matrix reduction, orthology prediction, Crustacea, Remipedia

## **Running head**

Pancrustacean phylogeny and the position of Remipedia

## **Abbreviations**

The following abbreviations will be used in this article: BLAST = Basic Local Alignment Search Tool, BS = bootstrap support, EST = Expressed Sequence Tag, HaMStR = Hidden Markov Model based Search for Orthologs using Reciprocity, ML = Maximum Likelihood, NCBI = National Center for Biotechnology Information, NGS = next generation sequencing, pHMM = profile Hidden Markov Model

## Background

A monophyletic taxon Pancrustacea is supported by phylogenies that are based on mitochondrial, single nuclear gene, multi-gene, and large phylogenomic analyses (Friedrich and Tautz 1995; Shultz and Regier 2000; Friedrich and Tautz 2001; Giribet et al. 2001; Hwang et al. 2001; Regier and Shultz 2001; Nardi et al. 2003; Carapelli et al. 2005; Carapelli et al. 2007). These results all support the hypothesis that hexapods are more closely related to crustaceans than to myriapods, and thus contradict the Atelocerata (a.k.a. Tracheata) hypothesis, which assumes a sister group relationship of hexapods and myriapods (Pocock 1893; Heymons 1901). If the Pancrustacea hypothesis (Zrzavý and Stys 1997) is accepted, it still remains unclear which among the major crustacean groups represents the sister group of Hexapoda. Many studies based on large molecular data sets have proposed Branchiopoda as the sister group of Hexapoda (Roeding et al. 2007; Dunn et al. 2008; Timmermans et al. 2008; Roeding et al. 2009; Meusemann et al. 2010; Rota-Stabelli et al. 2011). However, these studies are characterized by a relatively poor sampling of crustacean taxa. A recent, comprehensively sampled molecular phylogenetic analysis of arthropods instead suggests that hexapods are the sister group to a clade “Xenocarida”, which comprises Remipedia and Cephalocarida (Regier et al. 2010). A close relationship between hexapods and Remipedia was previously suggested by a phylogenetic analysis of hemocyanin sequences (Ertas et al. 2009) as well as by several morphological studies (Moura and Christoffersen 1996; Fanenbruck et al. 2004; Fanenbruck and Harzsch 2005; Bäcker et al. 2008). By contrast, other morphological analyses inferred Remipedia and Malacostraca as being sister taxa (Koenemann et al. 2007; Koenemann et al. 2009).

To shed light on the higher-level pancrustacean phylogeny and the closest

crustacean relatives of hexapods, we performed a series of phylogenetic analyses on the most exhaustive crustacean phylogenomic data set derived from ESTs compiled to date. This includes 454 EST data from six hitherto unsampled crustacean species, namely *Lynceus brachyurus* (Laevicaudata, Branchiopoda), *Spinicaudata* sp. (Branchiopoda), *Cypridininae* sp. (Ostracoda), *Sarsinebalia urgorrhii*, *Nebalia bipes* (Leptostraca, Malacostraca), and *Speleonectes* cf. *tulumensis* (Remipedia). Data from Cephalocarida, however, were not available for inclusion in this study, despite a tremendous effort and several field trips to collect sufficient specimens of this taxon.

Non-phylogenetic signal (Felsenstein 1988; Philippe et al. 2005; Philippe et al. 2011) can seriously mislead phylogenomic studies. The greatest challenges are therefore to optimize the quality of the data, to separate signal from noise, and to handle efficiently missing data (Driskell et al. 2004; Philippe et al. 2005; Dunn et al. 2008; Hartmann and Vision, 2008; Wiens and Moen 2008; Meusemann et al. 2010; Philippe et al. 2011). Here these issues are addressed by using the HaMStR approach (Hidden Markov Model based Search for Orthologs using Reciprocity) for orthology prediction (Ebersberger et al. 2009), automated alignment evaluation and masking (Misof and Misof 2009; Kück et al. 2010), and a recently developed approach to matrix reduction that selects optimal data subsets featuring increased signal (see Meusemann et al. 2010; Meyer and Misof 2010, <http://mare.zfmk.de>).

In summary, this study has three goals. 1) To address pancrustacean phylogeny with the largest phylogenomic data set derived from ESTs compiled to date, including data from hitherto unsampled key taxa such as Remipedia. 2) To assess the likely sister group of Hexapoda based on phylogenomic EST data. 3) To evaluate the impact of matrix reduction procedure on inferred trees by selecting optimal data subsets derived from two different orthologous gene sets.

## Methods

### Molecular techniques

454-pyrosequencing (ROCHE) was used to generate EST sequences from six crustacean species (see supplementary file 1). Fresh tissue was preserved in RNAlater and stored at -20°C or -80°C. Total RNA of Cypridininae sp. (Ostracoda), *Speleonectes* cf. *tulumensis* (Remipedia), and *Sarsinebalia urgorrhii* (Leptostraca) was extracted (Absolutely RNA kit, Stratagene) and its corresponding cDNA synthesized (Mint kit, Evrogen) at the Max Planck Institute for Molecular Genetics (MPIMG), Berlin, Germany. Subsequently, cDNA fragments were size-selected with the Chromaspin 1000 kit (Clontech), and the cDNA library was normalized with the Trimmer kit (Evrogen). cDNA was digested with the restriction enzyme SfiI (NEB). The digested cDNA was purified with the Qiagen PCR kit and subsequently ligated with 454 pyrosequencing adaptors (Roche). 1,000,000 reads per species were sequenced on a Titanium FLX sequencer (Roche). Total RNA of *Nebalia bipes* (Leptostraca), *Lynceus branchyurus* (Laevicaudata), and Spinicaudata sp. (a new species that is currently being described by Nicolas Rabet, Université Pierre et Marie Curie, Paris) was extracted with the Qiagen RNeasy Kit by RAJ at the University of Bath. Synthesis of cDNA, construction of non-normalized cDNA libraries, 454-pyrosequencing (100,000 to 140,000 reads per species), and sequence assembly were performed at the GenePool genomics facility, University of Edinburgh, United Kingdom.

### Sequence processing, orthology prediction, and alignment masking

Vector sequences of the obtained reads were removed with CrossMatch (Green 1994-1996, 0.990329) using UNIVeC (build 5.2, Dec. 2009;

<http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html>) after lowercase nucleotides were clipped with the aid of a custom made PERL script written by Sascha Strauss (CIBIV, Vienna, Austria). Additionally, vector sequences and poly-A tails were removed with SeqClean (<http://compbio.dfci.harvard.edu/tgi/software/>) using UNIVC (build 5.2, Dec. 2009). Subsequently sequences were masked with RepeatMasker (Smit et al. 1996-2010, open-3.1.6) and RepBase (20061006; <http://www.girinst.org/server/RepBase>). Clustering and assembly was performed using MIRA 3.0.3 (Chevreux et al. 2004). EST sequences of other taxa (see additional table 1) were retrieved from GenBank. All crustaceans, for which EST sequences are available (39 species), were added to our data set. The data set comprised a total of 46 crustaceans, 46 hexapods, 32 chelicerates, and three myriapods, as well as three onychophorans and one polychaete (*Capitella* sp.) (see supplementary file 1). Onychophorans and the polychaete were included as outgroup taxa. All EST sequences were quality checked and assembled in the processing pipeline described above. Assembled sequences of our own 454 projects were submitted to the Transcriptome Sequences Assembly database (TSA) at NCBI (accession numbers are summarized in supplementary file 1).

Our strategies for orthology prediction and for alignment masking followed the procedures described in Meusemann et al. (2010). Two sets of orthologous genes were constructed using the InParanoid transitive closure (TC) approach described by Ebersberger et al. (2009). Selection of orthologs in these two sets was guided by protein sequences available in proteome data sets of the so called ‘primer taxa’. Sequences of primer taxa were aligned within each set of orthologs and used to generate profile hidden Markov models (pHMMs). Subsequently, the pHMMs were used to search for putative orthologous sequences among the translated ESTs from all the taxa in our data set. Ortholog set 1 included the amino acid sequences of those

genes for which the algorithm 4.1s from InParanoid (Berglund et al. 2008; Ostlund et al. 2010) inferred orthologous sequences based on the following five primer taxa: *Ixodes scapularis* (Chelicerata), *Daphnia pulex* (Crustacea), *Apis mellifera*, *Aedes aegypti* (Hexapoda), and *Capitella* sp. (Polychaeta). Ortholog set 2 included genes for which InParanoid 7 inferred orthologous sequences based on the following six primer taxa: *Ixodes scapularis* (Chelicerata), *Daphnia pulex* (Crustacea), *Apis mellifera*, *Tribolium castaneum*, *Bombyx mori* (Hexapoda), and *Capitella* sp. (Polychaeta). HaMStR then assigned ESTs to the core ortholog groups (Ebersberger et al. 2009) (options `-representative`, `-strict`, and `-eval_limit=0.01`). Each group of orthologous amino acid sequences was aligned separately with MAFFT L-INS-I (Kato and Toh 2008). Randomly similar aligned positions were identified with ALISCORE. We applied the default sliding window size, the maximal number of pairwise comparisons (`-r`), and a special scoring for gappy amino acid data (`-e`) (Misof and Misof 2009; Kück et al. 2010). Randomly aligned positions were subsequently removed with ALICUT (Kück 2009, <http://www.utilities.zfmk.de>). All masked gene alignments were finally concatenated with FASconCAT (Kück and Meusemann 2010).

Orthology prediction resulted in two data sets: ortholog set 1 (hereafter, set 1A<sub>unred</sub>) encompasses 1,886 genes and ortholog set 2 (set 2A<sub>unred</sub>) contains 1,579 genes (see supplementary files 2, 3, and 4). Each set consists of 131 taxa. The reference species for the reciprocal BLAST procedure are given in supplementary file 1. To generate additional data sets, the number of hexapod and chelicerate species was reduced. This *a priori* exclusion of taxa allowed the quartet mapping and subsequent gene selection procedures (see next paragraph) to preferentially retain genes that are proportionally more represented in crustaceans and eventually more informative for resolving relationships among crustaceans. In addition, we removed several hexapod and chelicerate taxa with long terminal branches in the trees inferred from set 1 and 2



(e.g., *Glycophagus domesticus*) in order to reduce the impact of possible long-branch attraction artifacts (see supplementary file 1). This yielded the additional data sets 1B<sub>unred</sub> and 2B<sub>unred</sub>, each with 105 species.

We assessed the overlap of our putatively orthologous genes with those presented in Meusemann et al. (2010) and with the sequences analyzed by Regier et al. (2010). Of the data presented in Regier et al (2010), all mRNA sequences for nine representatives of the major crustacean taxa present in and complementary to our data set (including Remipedia and Cephalocarida, supplementary file 5a) were downloaded from NCBI (September 2010). Sequences of these nine crustaceans were also analyzed with HaMStR (same settings as before) to search for orthologous genes that correspond to those in our data sets (supplementary file 5a).

### **Matrix reduction and selection of data subsets**

There are various strategies to handle highly incomplete matrices (i.e., data sets with a large proportion of missing entries or gaps). Most often, concatenated ‘supermatrices’ are filtered using predefined thresholds of data availability (Dunn et al. 2008; Philippe et al. 2009). We utilized an alternative approach to data reduction here, selecting a subset of genes and taxa from each supermatrix based on information content in addition to data availability (MARE v 0.1.2-alpha; Meyer and Misof 2010, <http://mare.zfmk.de>). The approach yields a condensed and more informative data set by maximizing the ratio of signal to noise, and by reducing the number of uninformative genes and poorly sampled taxa. MARE first evaluates the ‘tree likeness’ of each single gene. Tree likeness reflects the fraction of all possible (but not more than 20,000, due to computational limitations) quartets that are resolved for a given sequence alignment. The process is based on geometry-weighted quartet mapping (Nieselt-Struwe and von Haeseler 2001), extended to amino acid data. For

further details on the procedure and the algorithm, see Meyer and Misof (2010; <http://mare.zfmk.de>).

Matrix reduction was performed on the four data sets ( $1A_{\text{unred}}$ ,  $1B_{\text{unred}}$ ,  $2A_{\text{unred}}$ ,  $2B_{\text{unred}}$ ) defined above (see supplementary file 4 for an overview), using *Acerentomon franzi* (Protura, Hexapoda) and *Balanus amphritite* (Cirripedia, Crustacea) as constraint taxa and applying a taxon weighting parameter ( $-t$ ) of 1.5 to keep more taxa. The constraints aim to maximize the retention of entognathous hexapods and cirripede crustacean taxa, respectively. The matrix reduction resulted in the reduced data sets  $1A_{\text{red}}$ ,  $1B_{\text{red}}$ ,  $2A_{\text{red}}$ , and  $2B_{\text{red}}$ .

### Phylogenetic analyses

Phylogenetic relationships were inferred by analyzing data sets  $1A_{\text{red}}$ ,  $1B_{\text{red}}$ ,  $2A_{\text{red}}$ , and  $2B_{\text{red}}$  under the Maximum Likelihood (ML) optimality criterion in RAxML v7.2.6 (Stamatakis 2006; Ott et al. 2007) (see table 1). Tree searching and bootstrapping were conducted simultaneously (PROTCATWAG,  $-f a$ , 1,000 bootstrap replicates). In all analyses, the “bootstopping” criterion (Pattengale et al. 2010) was used (default settings) *a posteriori* to assess whether or not a sufficient number of bootstrap replicates had been computed for evaluating tree robustness. Additionally, the unreduced data sets (i.e.  $1A_{\text{unred}}$ ,  $1B_{\text{unred}}$ ,  $2A_{\text{unred}}$ ,  $2B_{\text{unred}}$ ) were analyzed using the same procedures, except that we used the “on-the-fly” bootstopping criterion (to avoid unnecessary computations and save computational resources) with the SSE-3-vectorized Pthreads-parallelized version 7.2.8 of RAxML. All analyses were done on the High Performance Computing (HPC) clusters at the ZFMK Bonn (Zoologisches Forschungsmuseum Alexander Koenig, Bonn), the RRZK in Cologne (Regionales Rechenzentrum Köln: SUGI - Sustainable Grid Infrastructure project - and CHEOPS - Cologne High Efficient Operating Platform for Science), and the HITS gGmbH in

Heidelberg (Heidelberg Institute for Theoretical Studies). Leaf stability indices were computed with Phyutility (Smith and Dunn 2008) on the respective sets of bootstrap trees from each data set. The indices are a measure for the consistency of the position of each terminal taxon (leaf) relative to remaining taxa across replicates. Potentially unstable positions of taxa can be detected in the reconstructed topologies using this method. The lower the leaf stability index for a given terminal taxon is, the less stable is its phylogenetic position.

## Results

### Sets of orthologous genes

Set  $1A_{\text{unred}}$  of orthologous genes comprises sequences of 131 taxa, 1,886 genes, and 831,013 aligned amino acid positions (supplementary files 2 and 6). Set  $2A_{\text{unred}}$  includes sequences of 131 taxa, 1,579 genes, and 711,430 aligned amino acid positions (see supplementary files 3 and 7). The two sets have 1,410 genes in common (see supplementary files 2, 3, 5). After applying MARE, the information content in each data subset was approximately doubled (see table 1). MARE removed nearly the same species from each data set such that the two a priori reduced data sets ( $1A_{\text{red}}$  and  $2A_{\text{red}}$ ) had very similar taxon samples (supplementary file 1). 496 of these genes are present in the unreduced data set analyzed by Meusemann et al. (2010). Of the 129 genes present in the reduced data set (selected optimal data subset) of Meusemann et al. (2010), 75 were found in the reduced data sets  $1A_{\text{red}}$  and  $2A_{\text{red}}$ , and 74 genes in the reduced data sets  $1B_{\text{red}}$  and  $2B_{\text{red}}$  (see additional file 5).

Of the sequences of Regier et al. (2010), 42 sequences were assigned to our groups of orthologous sequences in data set  $1A_{\text{unred}}$  and 37 to our groups of orthologous sequences in data set  $2A_{\text{unred}}$ . However, only 19 sequences of Remipedia and

Cephalocarida overlap with set 1A<sub>unred</sub>, and 18 overlap with set 2A<sub>unred</sub>. Only four remipede and cephalocarid genes used in Regier et al. 2010 were present in our data sets 1A<sub>red</sub> and 2A<sub>red</sub>; five genes were shared with our data sets 1B<sub>red</sub> and 2B<sub>red</sub> (see additional file 5).

### **Pancrustacean relationships in the trees inferred from reduced data sets**

The monophyly of Pancrustacea received 99%-100% bootstrap support in all of our trees. Likewise, many major crustacean groups (i.e. Malacostraca, Branchiopoda, Copepoda, Cirripedia) have high (BS= 99%) or maximal support in all trees (see table 1).

Two large clades are found in the trees inferred from the reduced data sets: a clade composed of Malacostraca, Cirripedia, and Copepoda, and another comprising Branchiopoda, Remipedia, and Hexapoda. Support for the first clade is much higher in the trees that we derived from the submatrices of set 2 (figure 2: data set 2A<sub>red</sub> and figure 4: data set 2B<sub>red</sub>; BS = 75% and 100% respectively) than in the trees derived from the submatrices of set 1. However, the relationships of the major lineages within this clade (i.e., Malacostraca, Cirripedia, and Copepoda) differ between the trees inferred from submatrices of sets 1 and 2. The reduced data sets of set 1 suggest a sister group relationship of cirripedes and malacostracans (figure 1: data set 1A<sub>red</sub> and figure 3: 1B<sub>red</sub>). In contrast, the reduced data sets of set 2 imply that cirripedes and copepods are sister groups (figure 2: data sets 2A<sub>red</sub> and figure 4: 2B<sub>red</sub>). Similarly, the clade comprising branchiopods, remipedes, and hexapods receives stronger support in the trees (83% and 100%) that were inferred from submatrices of set 2. Trees based upon set 2 also show higher average leaf stability indices (see figures 1-4) than those based upon set 1. Importantly, all trees inferred from the reduced data sets support the relationship (Branchiopoda (Remipedia, Hexapoda)). Data sets 1A<sub>red</sub>, 1B<sub>red</sub>, and 2B<sub>red</sub>

maximally support a clade (Ostracoda, remaining pancrustaceans), while data set 2A<sub>red</sub> suggests a clade (Ostracoda ((Malacostraca (Copepoda, Cirripedia)) (BS=79%).

Inferred phylogenetic relationships within the monophyletic higher-level crustacean taxa are consistent between our data sets. Within Malacostraca, both the unreduced and the reduced data sets suggest a sister group relationship of Leptostraca and Eumalacostraca. Eucarida (Euphausiacea, Decapoda) are supported in three of the four reduced trees (figures 1, 2, and 4). In the fourth tree (figure 3), *Euphausia superba* (Euphausiacea) was not present because this taxon had been excluded from the data set during matrix reduction. In all inferred trees, Eucarida and Peracarida (represented by Amphipoda) are sister taxa. All trees, except one of the unreduced trees (supplementary file 6, data set 1A<sub>unred</sub>), support the same phylogenetic relationships within Decapoda. Decapoda is divided into two sister clades. The first unites Caridea and Dendrobranchiata as sister taxa. The second clade supports the relationships ((Anomura, Brachyura) (Astacidea, Achelata)). Within branchiopods, all our analyses suggest the same topology: (Anostraca (Notostraca (Laevicaudata (Spinicaudata, Cladocera)))). Finally, in Hexapoda, a split between Ectognatha and Entognatha (Insecta) is recovered consistently. Within Entognatha, Collembola is inferred invariably as the sister group to Protura (together constituting the clade Ellipura).

### **Comparison of trees inferred from unreduced and reduced data sets**

All of the trees inferred from the unreduced data sets suggest consistently the monophyly of Mandibulata (BS=83%-99%). In all of these, Myriapoda is the sister group of Pancrustacea. In contrast, Mandibulata is not supported by any of the trees derived from the reduced data sets. Rather, a clade (Chelicerata, Pancrustacea) is recovered with weak to maximal support (BS=51%-100%). Within Pancrustacea, the

trees based upon the unreduced data sets strongly support a clade of cirripedes and malacostracans (BS=99%-100%). The same relationships are obtained when analyzing the reduced subsets of set 1 (i.e., data sets 1A<sub>red</sub>, and 1B<sub>red</sub>). Three out of four phylogenetic trees inferred from the reduced data sets show a sister group relationship of Ostracoda plus the remaining pancrustaceans (figures 1, 3 and 4) although the precise phylogenetic position of Ostracoda remains uncertain.

## Discussion

### *Pancrustaceans*

The monophyly of Pancrustacea (Zrzavý and Stys 1997) has been suggested by several studies that investigated nuclear and/or mitochondrial sequences (Friedrich and Tautz 1995; Shultz and Regier 2000; Friedrich and Tautz 2001; Giribet et al. 2001; Hwang et al. 2001; Regier and Shultz 2001; Nardi et al. 2003; Carapelli et al. 2005; Carapelli et al. 2007). This clade, sometimes also referred to as Tetraconata (Dohle 2001), has also been advocated because of conspicuous similarities in the ommatidia of the compound eyes shared between hexapods and crustaceans (but see also Harzsch and Hafner 2006), and because of similarities in their neuroanatomy and neuroembryology (Harzsch et al. 2005; Harzsch 2006; Ungerer and Scholtz 2008). Recent phylogenomic analyses (Roeding et al. 2009; Meusemann et al. 2010; Rota-Stabelli et al. 2011) also strongly support the monophyly of Pancrustacea. Our results corroborate strongly a clade Pancrustacea, which is maximally or highly supported in all trees inferred from our data sets.

### *Malacostraca*

Malacostraca was consistently recovered as a clade in our analyses. Nonetheless, the phylogenetic relationships of the major lineages within Malacostraca as well as the

phylogenetic position of the Malacostraca within the Pancrustacea are still unclear (Jenner 2010). Our data support a split of the Malacostraca into the lineages Leptostraca and Eumalacostraca. This is consistent with morphological data (Wills et al. 1995; Wills 1998; Richter and Scholtz 2001; Jenner et al. 2009; Wills et al. 2009). Although our phylogenomic data are unable to completely resolve the relationships within Eumalacostraca, they do suggest a common origin of Anomura, Brachyura, Astacidea, and Achelata by exclusion of Dendrobranchiata and Caridea. This last result is largely consistent with recently published molecular phylogenetic investigations including these taxa (Bracken et al. 2009; Toon et al. 2009; Bracken et al. 2010). Most contentious of all is the position of the Malacostraca within crustaceans (Jenner 2010). Even if we only focus on recently published molecular phylogenetic (von Reumont et al. 2009; Koenemann et al. 2010; Regier et al. 2010) and phylogenomic studies with reasonable sampling of crustacean taxa (Meusemann et al. 2010; Andrew 2011), no consistent pattern emerges. Our current results support two alternative sister groups for Malacostraca: Cirripedia (representing Thecostraca) or (Cirripedia, Copepoda). In Meusemann et al. (2010), these two alternatives were inferred from the same data set using Bayesian and likelihood methods, respectively. In our study, the results of six out of eight analyses support (Malacostraca, Cirripedia), with only the reduced data sets based on ortholog set 2 supporting (Malacostraca (Cirripedia, Copepoda)). Since matrix reduction is shown to increase the signal to noise ratio (Table 1), we speculate that the clade (Malacostraca, Cirripedia), which was also found by Regier et al. (2010) and Andrew (2011), might be an artifact, a hypothesis at least consistent with the slight drop in support value for this clade in the reduced data sets based on ortholog set 1. More importantly perhaps, support for this clade was also significantly reduced in the analysis of Regier et al. (2011) that was corrected for heterogeneity in base composition. A closer affinity of

copepods and cirripedes would also be more congruent with some analyses of morphological data (e.g., Wills 1998a, b; Martin and Davis 2001).

### *Branchiopoda*

Our results strongly support monophyly of Branchiopoda, in line with earlier molecular and morphological studies (Wills 1998a, b; Stenderup et al. 2006; Olesen 2007; Richter et al. 2007). Furthermore, we found the conchostracans to be paraphyletic, in agreement with recent studies (Braband et al. 2002; Olesen 2007; Richter et al. 2007). Unfortunately, there is still no agreement on the position of Branchiopoda within the crustaceans. In terms of the number of recently proposed alternative hypotheses, the placement of Branchiopoda remains one of the most intriguing challenges in higher-level pancrustacean phylogenetics (Jenner 2010). One recent, well-supported hypothesis that has attracted considerable interest is the possible sister group relationship of branchiopods and hexapods (Glenner et al. 2006; Roeding et al. 2009; Meusemann et al. 2010; Andrew 2011; Rota-Stabelli et al. 2011). Indeed, this hypothesis underpins a seductive scenario, in which hexapods are conjectured to have evolved from marine ancestors via some Late Silurian, freshwater, branchiopod-like intermediate (Glenner et al. 2006). However, if the marine fossil *Rehbachella kinnekullensis* (Walossek 1993) represents a stem group branchiopod (Schram and Koenemann 2001), then branchiopods themselves are also likely to be ancestrally marine (Olesen 2007), contrary to Glenner et al. (2006): see also figure 5.

Importantly, no previous phylogenomic analyses of EST data have included the enigmatic remipedes. Our new EST data strongly suggest that Branchiopoda is the sister group of Remipedia plus Hexapoda (with the single exception of our unreduced set 2A<sub>unred</sub>, supplementary file 7). Our data thus challenge the monophyly of



Vericrustacea (= (Branchiopoda (Copepoda (Malacostraca, Thecostraca)))) found by Regier et al. (2010).

The conflict between molecular and morphological data regarding the evolutionary history of Branchiopoda, Malacostraca, and Remipedia is illustrated in figure 5. Our data, in common with most molecular studies (Regier et al. 2005; Mallat and Giribet 2006; Regier et al. 2008; Roeding et al. 2009; von Reumont et al. 2009; Meusemann et al. 2010; Andrew 2011; Rota-Stabelli et al. 2011), imply that Branchiopoda is more closely related to Hexapoda and Remipedia than is Malacostraca. In conflict with these molecular results are morphological and neuroanatomical studies that support a clade of Malacostraca, Remipedia and Hexapoda (Fanenbruck et al. 2004; Fanenbruck and Harzsch 2005).

#### *Is Remipedia the sister group to Hexapoda?*

Remipedes have been considered crucial for understanding the origin of crustaceans ever since they were first described in the 1980s (Yager 1981; Yager and Schram 1986). Their homonomous trunks and the presence of a pair of biramous appendages on each segment have usually been interpreted as crustacean plesiomorphies (Schram 1986; Emerson and Schram 1991; Schram and Hof 1998; Ax 1999; Wills 1999; Wheeler et al. 2004). However, new and substantially more comprehensive molecular, morphological, neuroanatomical, and developmental data have started to challenge the idea that remipedes diverged early during crustacean evolution. Similarities in neuroanatomy suggest a close relationship of remipedes, malacostracans, cephalocarids, and hexapods, which has been used to argue for a less basal position of remipedes. These taxa possess highly complex brains with a markedly different construction from those of other crustaceans (Fanenbruck et al. 2004; Fanenbruck and Harzsch 2005). Intriguingly, remipede larvae show many similarities with those of

some malacostracans (Koenemann et al. 2007; Koenemann et al. 2009).

Until recently, molecular phylogenetic analyses provided evidence for conflicting hypotheses with respect to the position of remipedes within pancrustaceans (see Jenner 2010). For example, mitochondrial and nuclear ribosomal RNA sequences suggested a sister group relationship of remipedes to cirripedes (Carapelli 2000; Lavrov et al. 2004; Hassanin 2006; Lim and Hwang, 2006), to ostracods (Cook et al. 2005), to collembolans (Cook et al. 2005; Hassanin 2006), to diplurans (Carapelli et al. 2007), and to various ‘maxillopodan’ taxa (Lavrov et al. 2004; von Reumont et al. 2009). The set of possible crustacean sister groups proposed for hexapods has been equally diverse, including branchiopods (Babbitt and Patel 2005; Glenner et al. 2006; Roeding et al. 2009; Meusemann et al. 2010; Andrews 2011), malacostracans (Lim and Hwang 2006; Strausfeld et al. 2009), and copepods (Mallat and Giribet 2006; Dell’Ampio et al. 2009; von Reumont et al. 2009). However, the taxonomic coverage in these studies was often sparse and usually did not include remipedes.

In rDNA-based phylogenies, Remipedia and Cephalocarida show long branches, and at least the cephalocarids are affected by non-stationary substitution processes (von Reumont et al. 2009). Spears and Abele (1998) interpreted a sister group relationship of Cephalocarida and Remipedia inferred from 18S rDNA sequence data with caution, and suggested the possibility of pseudogenes in addition to non-stationary substitution processes. The putative sister group relationship of these two taxa must therefore be regarded with caution (von Reumont et al. 2009; Koenemann et al. 2010). Nonetheless, remipedes and cephalocarids have also emerged as close relatives from analyses of nuclear coding genes (Shultz and Regier 2000; Regier et al. 2005, 2008), but mostly without strong support. However, in the study by Regier et al. (2010), support for this clade was higher when models were applied that exclude the degenerated third codon positions on nucleotide level. Yet, support was again weak

when studying the phylogenetic relationships at the amino acid level. Testing this hypothesis by means of analyzing exhaustive phylogenomic data must await the generation of EST data for cephalocarids.

Ertas et al. (2009) provided the first molecular phylogenetic support for a close relationship of remipedes and hexapods. This result was soon corroborated by a multi-gene analysis at the nucleotide level by Regier et al. (2010), which recovered the clade Xenocarida = (Remipedia, Cephalocarida) as a sister group of Hexapoda. Our results provide strong support for a close relationship of remipedes and hexapods, and on the basis of significantly more nuclear protein coding genes than analyzed before. Given the minimal overlap between our data and those by Regier et al., our results offer a largely independent test of this hypothesis. We therefore propose an evolutionary scenario, in which the last common ancestor of remipedes and hexapods lived in a shallow marine environment, from which crown group remipedes and hexapods colonized their respective anchialine and terrestrial habitats (Figure 5).

### **Impact of ortholog sets and matrix reduction**

This study shows that the size and precise composition of phylogenomic data sets can have marked effects on the results of phylogenetic inference. Large size alone clearly does not make a data set reliable (Philippe et al. 2011). However, understanding the relative contributions of the size and composition of data sets on the results requires more studies in the future. Using the HaMStR approach, the set of orthologous genes selected for analysis (both the total number and identity) is strongly dependent upon the choice of primer taxa (supplementary files 2, 3 and 5). Of course one expects a smaller set of orthologs when using more primer taxa. The exclusion of the dipteran *Aedes* and inclusion of *Tribolium* and *Bombyx* consequently results in a smaller number of orthologous genes in data sets derived from ortholog set 2 (see

supplementary file 4 and 5). The percentage of present genes that overlap between the two ortholog sets is significantly higher in data sets derived from ortholog set 2 (90%  $2A_{\text{unred}}$ , 89%  $2A_{\text{red}}$ , 92%  $2B_{\text{red}}$ ) compared to data sets from ortholog set 1 (75%  $1A_{\text{unred}}$ , 77%  $1A_{\text{red}}$ , 73%  $1B_{\text{red}}$ ): see table 2. Overlapping genes between the unreduced and reduced data sets within ortholog set 1 and 2 is nearly identical, see table 2 and supplementary file 5b.

Nonetheless, it remains difficult to determine which ortholog set should be considered as the most “reliable”. Not only the contribution of each gene to the inferred relationships is unknown, the interactions of signals present in all genes also remain wholly unexplored. The software MARE attempts to address the first of these issues by excluding genes with low tree-likeness in order to reduce noise. However, more studies are needed to fully explore the efficiency and performance of this approach. For example, the clade Mandibulata is replaced with a clade Chelicerata + Pancrustacea in the topologies of the reduced data sets. This could conceivably be an artifact of matrix reduction. During the random substitution process one expects that older phylogenetic signal is more likely to be substituted by multiple hits (noise) than younger phylogenetic signal. Since MARE excludes genes that have lower tree-likeness scores, it could be that it disproportionally removes genes that contain older and distorted phylogenetic signal. This could lead to a loss of support for deeper nodes in the tree. However, because MARE does not distinguish between such secondarily noisy genes, and pure noise, the potential loss of some phylogenetic signal is an inescapable side effect of trying to increase the overall signal to noise ratio of the data.

An important methodological issue may be illustrated by considering the variable placement of Cirripedia. Data sets based on set 1 support a clade Cirripedia and Malacostraca, independent of matrix reduction (albeit with decreased support in the

reduced data sets). In contrast, when data sets based on set 2 are reduced with MARE (sets 2A<sub>red</sub> and 2B<sub>red</sub>), Cirripedia are inferred as the sister group to Copepoda (figures 2 and 4). The latter hypothesis is in line with results from morphological and several molecular analyses (see Martin and Davis 2001; Jenner 2010). This indicates that some genes that are found exclusively in both reduced matrices of set 1 (supplementary file 5) apparently obscure the signal for a clade (Cirripedia, Copepoda). Interestingly, the clade (Cirripedia, Malacostraca) collapses in the study by Regier et al. (2010) when these authors tried to reduce the effects of sequence saturation corroborating the suggestion that conflicting signal is present in some genes.

## Conclusions

- 1) This is the first phylogenomic analysis (including new EST data) which supports a sister group relationship of Remipedia and Hexapoda (Ertas et al. 2009; Fanenbruck et al. 2004; Fanenbruck and Harzsch 2005). This particular conclusion is unaffected by the precise procedures used for identifying orthologous genes, or for reducing the data sets.
- 2) Our results suggest that Pancrustacea is divided into two clades: i) Malacostraca, Copepoda, and Cirripedia, and ii) Branchiopoda, Remipedia, and Hexapoda.
- 3) The methods used for selection of putative orthologous genes, namely the primer taxa choice for the HaMStR approach and matrix reduction by selecting optimal data subsets can markedly influence the inferred relationships. For example, matrix reduction indicates that the clade Communostraca (Malacostraca, Thecostraca), with Cirripedia representing Thecostraca in our study that was supported by Regier et al. (2010) and by the phylogenomic analysis of Andrew (2011) might be artificial. This underlines the importance of implementing the most appropriate methods for

compiling data sets and for controlling their quality

4) By increasing the information content of data sets via matrix reduction, some conflicts in the data become visible and can be removed like (Malacostraca, Cirripedia). However, this study serves in parallel as a test case to explore the idea that MARE might introduce potential artifacts such as a collapse of Mandibulata in the reduced topologies.

5) High-level pancrustacean phylogeny remains a challenging area of research. The recent study by Regier et al. (2010) sampled significantly more genes and taxa than its forebears and represented a major advance. In view of the limited overlap between the genes included in that study and ours, our results allow an ostensibly independent test of some of the more surprising relationships reported by Regier et al. (2010). Future work should aim to incorporate hitherto unsampled taxa in phylogenomic data sets, most notably Cephalocarida.

6) An alternative approach to the one employed here, is to assemble genomic data for more pancrustacean taxa to infer more pancrustacean-typical putative orthologous genes that might carry a less noisy signal. Critically, the prediction of orthologous genes could then be based on a larger sample of completely sequenced genomes. HaMStR could represent one possible strategy to identify the ortholog genes. In an additional second step after the HaMStR approach, gene subsets could be selected with MARE targeting in general only those genes that show a high tree-likeness and chance to be informative. Subsequently, the sequences of the identified genes can be used to reconstruct primer toolboxes to amplify genes for specific taxa groups. This method will allow us additionally to include species that can be collected for DNA-based work, but which are difficult to collect fresh, and in sufficient quantity for mRNA-based EST sequencing.

## Authors' contributions

BMvR and BM designed the study and analyses. EST processing, design of orthologous sequence groups and orthology search was conducted by IE. RAJ and MAW generated the sequence data for *Nebalia bipes*, *Lynceus brachyurus*, and the undescribed species of spinicaudatan. BMvR generated the sequence data for *Speleonectes* cf. *tulumensis*, *Sarsinebalia urgorrhii*, GP and EDA for the Cypridininae sp. (Ostracoda). The manuscript was written by BMvR, RAJ, MAW, EDA, GP, IE and BM, with helpful comments from KM, ON, SK, TMI, AS. KM helped to handle the sequence data of Regier et al. (2010) and provided automated scripts for ALISCORE and ran some RAxML analyses. AS ran the RAxML analyses for the unreduced data sets. BMvR collected *Sarsinebalia urgorrhii* and the Ostracoda. TMI and BMvR collected *Speleonectes* cf. *tulumensis*.

## Acknowledgements

BMvR is supported by the DFG grant WA530/34 and thanks J-W Wägele for the encouragement for this study. BM and KM are supported by the DFG grant MI 649/6-3. Both DFG projects collaborate together in the “Arthropod Network” within the Deutsche Forschungsgemeinschaft (DFG) Priority Program 1174 “Deep Metazoan Phylogeny” (<http://www.deep-phylogeny.org>). MAW thanks the BBSRC (Grant BB/C006682/1) and the Leverhulme Trust (Grant F/00 351/Z) for funding. The contributions of EdA and GP were financially supported by the Austrian Science Fund (FWF) Project 20497-B17. RAJ and MAW thank GenePool staff for their sequencing expertise, in particular Stephen Bridgett for his work assembling the sequences, and Dr. Nicolas Rabet for collecting the specimens of *Nebalia bipes*, *Lynceus brachyurus*, and the undescribed spinicaudatan species. Guillermo Díaz-

Agras and Juan Moreira Da Rocha from the Marine Biological Station A Grana (University of Santiago de Compostela), Ferrol, Spain helped to collect the *Sarsinebalia* specimens. B. Gonzalez assisted with the collection of *Speleonectes* cf. *tulumensis*. We thank B. Sames, Institute of Palaentology, University of Vienna for determing the Ostracoda species. We are grateful to Viktor Achter, Volker Winkelmann and Sebastian Breuers for their help regarding the setup, scripts and likelihood analyses on SUGI (cluster system of the RRZK, University of Cologne; part of the BMBF granted project “Sustainable Grid Infrastructures as part of the D-grind initiative to support the scientific life science community”) and CHEOPS (Cologne High Efficiency Operating Platform for Sciences, a DFG granted high performance cluster at the RRZK of the University of Cologne). IE acknowledges support by a grant of the Wiener Wissenschafts-, Forschungs- und Technologie Fonds (WWTF) to Arndt von Haeseler, and from the DFG priority program SPP 1174 Deep Metazoan Phylogeny (Grant HA 1628/9). We thank Sascha Strauss for processing and assembling of the EST data. Christoph Mayer and Gerrit Hartig, ZFMK, Bonn, provided scripts for the submission of the EST data. We also thank J. Regier for his fast and direct answers regarding our questions about the matrices published in Regier et al. (2010). Janus Borner, Biozentrum Grindel und Zoologisches Museum, University of hamburg, helped essentially to sort the data from Regier and colleagues. All Data sets and checked EST raw data are available at:

<http://www.zfmk.de/web/Forschung/Molekularlabor/Datenstze/index.en.html>

## References

Andrew DR. 2011. A new view of insect-crustacean relationships II. Inferences from expressed sequence tags and comparisons with neural cladistics. *Arthropod Struct Dev*. Advanced access: doi 10.1016/j.asd.2011.02.001



- Ax P. 1999. Das System der Metazoa II. Ein Lehrbuch der phylogenetischen Systematik. Stuttgart: Gustav Fischer Verlag.
- Babbitt CC, Patel NH. 2005. Relationships within the Pancrustacea: Examining the influence of additional malacostracan 18S and 28S rDNA. In: S Koenemann, RA Jenner, editors. Crustacea and arthropod relationships. London: Taylor and Francis.
- Bäcker H, Fanenbruck M, Wägele JW. 2008. A forgotten homology supporting the monophyly of Tracheata: The subcoxa of insects and myriapods re-visited. *Zool Anz.* 247:185-207.
- Braband A, Richter S, Hiesel R, Scholtz G. 2002. Phylogenetic relationships within the Phyllopoda (Crustacea, Branchiopoda) based on mitochondrial and nuclear markers. *Mol Phylogen Evol.* 25:229-244.
- Bracken HD, De Grave S, Toon A, Felder DL, Crandall KA. 2010. Phylogenetic position, systematic status, and divergence time of the Procarididea (Crustacea: Decapoda). *Zool Scripta.* 39:198-212.
- Bracken HD, Toon A, Felder DL, Martin JW, Finley M, Rasmussen J, Palero F, Crandall KA. 2009. The decapod tree of life: compiling the data and moving toward a consensus of decapod evolution. *Arthropod Struct Dev.* 67:99-116.
- Carapelli A. 2000. Molecular phylogeny of the apterygotan insects based on nuclear and mitochondrial genes. *Pedobiologia.* 44:361-373.
- Carapelli A, Liò P, Nardi F, Van Der Wath E, Frati F. 2007. Phylogenetic analysis of mitochondrial protein coding genes confirms the reciprocal paraphyly of Hexapoda and Crustacea. *BMC Evol Biol.* 7:1-13.
- Carapelli A, Nardi F, Dallai R, Boore J, Liò P, Frati F. 2005. Relationships between hexapods and crustaceans based on four mitochondrial genes. In: S Koenemann, RA Jenner, editors. Crustacea and arthropod relationships. Boca

- Raton: Taylor and Francis, CRC press. P. 295-306.
- Chevreur B, Pfisterer T, Drescher B, Driesel AJ, Muller WE, Wetter T, Suhai S. 2004. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res.* 14:1147-1159.
- Chou H, Holmes MH. 2001. DNA sequence quality trimming and vector removal. *Bioinformatics.* 17:1093.
- Cook C, Yue Q, Akam M. 2005. Mitochondrial genomes suggest that hexapods and crustaceans are mutually paraphyletic. *Proc R Soc Lond B.* 272:1295-1304.
- Dell'Ampio E, Szucsich NU, Carapelli A, Frati F, Steiner G, Steinacher A, Pass G. 2009. Testing for misleading effects in the phylogenetic reconstruction of ancient lineages of hexapods: influence of character dependence and character choice in analyses of 28S rRNA sequences. *Zool Scripta.* 38:155-170
- Dohle W. 2001. Are the insects terrestrial crustaceans? A discussion of some new facts and arguments and the proposal of the proper name 'Tetraconata' for the monophyletic unit Crustacea + Hexapoda. *Ann Soc Entomol Fr (New Series).* 37:85-103.
- Driskell AC, Ané C, Burleigh JG, McMahon MM, O'Meara BC, Sanderson MJ. 2004. Prospects for building the tree of life from large sequence databases. *Science.* 306:1172-1174.
- Dunn CW, Hejnol A, Matus DQ, et al. (18 co-authors). 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature.* 452:745-749.
- Ebersberger I, Strauss S, Von Haeseler A. 2009. HaMStR: Profile Hidden markov Model Based Search for Orthologs in ESTs. *BMC Evol Biol.* 9:1-9.
- Emerson MJ, Schram F. 1991. Remipedia; part II, Palaeontology. *Proc. San Diego Soc. Nat. Hist.* 7:1-52.
- Ertas B, von Reumont BM, Wägele J, Misof B, Burmester T. 2009. Hemocyanin

- suggests a close relationship of Remipedia and Hexapoda. *Mol Biol Evol.* 26:2711-2718.
- Fanenbruck M, Harzsch S. 2005. A brain atlas of *Godzilloognomus frondosus* Yager, 1989 (Remipedia, Godzilliidae) and comparison with the brain of *Speleonectes tulumensis* Yager, 1987 (Remipedia, Speleonectidae): implications for arthropod relationships. *Arthropod Struct Dev.* 34:343-378.
- Fanenbruck M, Harzsch S, Wägele JW. 2004. The brain of the Remipedia (Crustacea) and an alternative hypothesis on their phylogenetic relationships. *Proc Natl Acad Sci USA.* 101:3868-3873.
- Friedrich M, Tautz D. 1995. Ribosomal DNA phylogeny of the major extant arthropod classes and the evolution of myriapods. *Nature.* 376:165-167.
- Friedrich M, Tautz D. 2001. Arthropod rDNA phylogeny revisited: A consistency analysis using Monte Carlo simulation. *Ann Soc Entomol Fr (New Series).* 37:21-40.
- Giribet G, Edgecombe GD, Wheeler WC. 2001. Arthropod phylogeny based on eight molecular loci and morphology. *Nature.* 413:157-161.
- Glenner H, Thomsen P, Hebsgaard M, Sorensen M, Willerslev E. 2006. Evolution: The Origin of Insects. *Science.* 314:1883-1884.
- Green, P. Crossmatch. (1993-1996). Online-access:  
[http://www.incogen.com/public\\_documents/vibe/details/crossmatch.html](http://www.incogen.com/public_documents/vibe/details/crossmatch.html).
- Hartmann S, Vision TJ. 2008. Using ESTs for phylogenomics: can one accurately infer a phylogenetic tree from a gappy alignment? *BMC Evol Biol.* 8:95.
- Harzsch S. 2006. Neurophylogeny: Architecture of the nervous system and a fresh view on arthropod phylogeny. *Integr Comp Biol.* 46:162-194.
- Harzsch S, Hafner G. 2006. Evolution of eye development in arthropods: phylogenetic aspects. *Arthropod Struct Dev.* 35:319-340.

- Harzsch S, Müller C, Wolf H. 2005. From variable to constant cell numbers: cellular characteristics of the arthropod nervous system argue against a sister-group relationship of Chelicerata and 'Myriapoda' but favour the Mandibulata concept. *Dev Genes Evol.* 215:53-68.
- Hassanin A. 2006. Phylogeny of Arthropoda inferred from mitochondrial sequences: strategies for limiting the misleading effects of multiple changes in pattern and rates of substitution. *Mol Phylogen Evol.* 38:100-116.
- Heymons R. 1901. Die Entwicklungsgeschichte der Scolopender. *Zoologica* 33:1-244
- Hwang UW, Friedrich M, Tautz D, Park CJ, Kim W. 2001. Mitochondrial protein phylogeny joins myriapods with chelicerates. *Nature.* 413:154-157.
- Jenner RA. 2010. Higher-level crustacean phylogeny: Consensus and conflicting hypotheses. *Arthropod Struct Dev.* 39:143-153.
- Jenner RA, Ní Dhubhghaill C, Ferla MP, Wills MA. 2009. Eumalacostracan phylogeny and total evidence: limitations of the usual suspects. *BMC Evol Biol.* 9:21.
- Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform.* 9:286-298.
- Koenemann S, Jenner RA, Hoenemann M, Stemme T, von Reumont BM. 2010. Arthropod phylogeny revisited, with a focus on crustacean relationships. *Arthropod Struct Dev.* 39:88-110.
- Koenemann S, Olesen J, Alwes F, Iliffe TM, Hoenemann M, Ungerer P, Wolff C, Scholtz G. 2009. The post-embryonic development of Remipedia (Crustacea)-additional results and new insights. *Dev Genes Evol.* 219:131-145.
- Koenemann S, Schram FR, Bloechl A, Iliffe TM, Hoenemann M, Held C. 2007. Post-embryonic development of remipede crustaceans. *Evol Devel.* 9:117-121.
- Kück P (2009) ALICUT: a Perlscript which cuts ALISCORE identified RSS.

Department of Bioinformatics, Zoologisches Forschungsmuseum A. Koenig  
(ZFMK), Bonn, Germany, version 2.0 edition. Available  
<http://www.zfmk.utilities.de>

Kück P, Meusemann K. 2010. FASconCAT: Convenient handling of data matrices.  
*Mol Phylogenet Evol.* 56:1115-1118.

Kück P, Meusemann K, Dambach J, Thormann B, von Reumont BM, Waagele JW,  
Misof B. 2010. Parametric and non-parametric masking of randomness in  
sequence alignments can be improved and leads to better resolved trees. *Front  
Zool.* 7:10.

Lavrov D, Brown W, Boore J. 2004. Phylogenetic position of the Pentastomida and  
(pan)crustacean relationships. *Proc R Soc Lond B.* 271:537-544.

Lim JT, Hwang UW. 2006. The complete mitochondrial genome of *Pollicipes mitella*  
(Crustacea, Maxillopoda, Cirripedia): non-monophylies of Maxillopoda and  
Crustacea. *Mol Cells.* 22:314-322.

Mallatt J, Giribet G. 2006. 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.

Martin J, Davis GE. 2001. An update classification of the recent Crustacea. *Nat Hist  
Mus Los Angel Cty Sci Ser.* 39:1-124.

Meusemann K, von Reumont BM, Simon S, et al. (16 co-authors). 2010. A  
phylogenomic approach to resolve the arthropod tree of life. *Mol Biol Evol.*  
27:2451-2464.

Meyer B, Misof B. 2010. MARE: Matrix Reduction – A tool to select optimized data  
subsets from supermatrices for phylogenetic inference. Zentrum für molekulare  
Biodiversitätsforschung (zmb) am ZFMK, Adenauerallee 160, 53113 Bonn,  
Germany. Version 01-alpha, August 2010, available upon request, current

version: <http://mare.zfmk.de>

- Misof B, Misof K. 2009. A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. *Syst Biol.* 58:21-34.
- Moura G, Christoffersen ML. 1996. The system of the mandibulate arthropods: Tracheata and Remipedia as sistergroups, crustacea non-monophyletic. *J Comp Biol.*1.
- Nardi F, Spinsanti G, Boore J, Carapelli A, Dallai R, Frati F. 2003. Hexapod Origins: Monophyletic or Paraphyletic? *Science.* 299:1887-1889.
- Nieselt-Struwe K, Von Haeseler A. 2001. Quartet-mapping, a generalization of the likelihood-mapping procedure. *Mol Biol Evol.* 18:1204-1219.
- Olesen J. 2007. Monophyly and phylogeny of Branchiopoda, with focus on morphology and homologies of branchiopod phyllopodus limbs. *J Crust Biol.* 27:165-183.
- Ostlund G, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S, Frings O, Sonnhammer EL. 2010. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.* 38:196-203.
- Ott M, Zola J, Stamatakis A, Aluru S. 2007. Large-scale maximum likelihood-based phylogenetic analysis on the IBM BlueGene/L. Proceedings of the 2007 ACM/IEEE conference on Supercomputing. Reno, Nevada: ACM.
- Pattengale ND, Alipour M, Bininda-Emonds OR, Moret BM, Stamatakis A. 2010. How many bootstrap replicates are necessary? *J Comput Biol.* 17:337-354.
- Philippe H, Brinkmann H, Lavrov DV, Littlewood DT, Manuel M, Worheide G, Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. *Plos Biol.* 9:e1000602.
- Philippe H, Delsuc F, Brinkmann H, Lartillot N. 2005. Phylogenomics. *Annu. Rev.*

*Ecol Evol Syst.* 36:541-562.

Philippe H, Derelle R, Lopez P, et al. 2009. Phylogenomics revives traditional views on deep animal relationships. *Curr Biol.* 19:706-712.

Pocock RI. 1893. On the classification of the tracheate Arthropoda. *Zoolog Anz.* 16:271-275.

Regier J, Shultz J, Kambic R. 2005. Pancrustacean phylogeny: hexapods are terrestrial crustaceans and maxillopods are not monophyletic. *Proc R Soc Lond B.* 272:395-401.

Regier J, Shultz J, Zwick A, Hussey A, Ball B, Wetzer R, Martin JW, Cunningham CW. 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature.* 463:1079-1083.

Regier JC, Shultz JW. 2001. Elongation factor-2: a useful gene for arthropod phylogenetics. *Mol Phylogenet Evol.* 20:136-148.

Regier JC, Shultz JW, Ganley AR, et al. (12 co-authors). 2008. Resolving arthropod phylogeny: exploring phylogenetic signal within 41 kb of protein-coding nuclear gene sequence. *Syst Biol.* 57:920-938.

Richter S, Olesen J, Wheeler WC. 2007. Phylogeny of Branchiopoda (Crustacea) based on a combined analysis of morphological data and six molecular loci. *Cladistics.* 23:301-336.

Richter S, Scholtz G. 2001. Phylogenetic analysis of the Malacostraca (Crustacea). *J. Zool Syst Evol Res.* 39:113-136.

Roeding F, Borner J, Kube M, Klages S, Reinhardt R, Burmester T. 2009. A 454 sequencing approach for large scale phylogenomic analysis of the common emperor scorpion (*Pandinus imperator*). *Mol Phylogenet Evol.* 53:826-834.

Roeding F, Hagner-Holler S, Ruhberg H, Ebersberger I, Von Haeseler A, Kube M, Reinhardt R, Burmester T. 2007. EST sequencing of Onychophora and

- phylogenomic analysis of Metazoa. *Mol Phylogenet Evol.* 45:942-951.
- Rota-Stabelli O, Campbell L, Brinkmann H, Edgecombe GD, Longhorn SJ, Peterson KJ, Pisani D, Philippe H, Telford MJ. 2011. A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. *Proc R Soc Lond B.* 278:298-306.
- Schram FR. 1986. Crustacea. Oxford: Oxford University Press.
- Schram FR, Hof CHJ. 1998. Fossils and interrelationships of major crustacean groups. In: GD Edgecomb, editor. Arthropod Fossils and Phylogeny. New York: Columbia University Press. P. 233-302.
- Schram FR, Koenemann S. 2001. Developmental genetics and arthropod evolution: part I, on legs. *Evol Dev.* 3 (5):343-354.
- Shultz JW, Regier JC. 2000. Phylogenetic analysis of arthropods using two nuclear protein-encoding genes supports a crustacean + hexapod clade. *Proc Biol Sci.* 267:1011-1019.
- Smit, AFA, Hubley, R, Green, P. RepeatMasker Open-3.0. 1996-2010. Open online access: <http://www.repeatmasker.org>
- Smith S, Dunn C. 2008. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics.* 24:715-716.
- Stamatakis A. 2006. RaxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics.* 22:2688-2690.
- Stenderup J, Olesen J, Glenner H. 2006. Molecular phylogeny of the Branchiopoda (Crustacea)—Multiple approaches suggest a ‘diplostracan’ ancestry of the Notostraca. *Mol Phyl Evol.* 41:182-194.
- Timmermans M, Roelofs D, Marien J, van Straalen NM. 2008. Revealing pancrustacean relationships: Phylogenetic analysis of ribosomal protein genes



- places Collembola (springtails) in a monophyletic Hexapoda and reinforces the discrepancy between mitochondrial and nuclear markers. *BMC Evol Biol.* 8:1-10.
- Toon A, Finley M, Staples J, Crandall KA. 2009. Decapod phylogenetics and molecular evolution. In: JW Martin, KA Crandall, DL Felder, editors. Decapod crustacean phylogenetics. Boca Raton: CRC Publishing. P. 15-29.
- Ungerer P, Scholtz G. 2008. Filling the gap between identified neuroblasts and neurons in crustaceans adds new support for Tetraconata. *Proc R Soc Lond B.* 275:369-376.
- von Reumont BM, Meusemann K, Szucsich N, et al. (14 co-authors). 2009. Can comprehensive background knowledge be incorporated into substitution models to improve phylogenetic analyses? A case study on major arthropod relationships. *BMC Evol Biol.* 9:119.
- Walossek D. 1993. The Upper Cambrian *Rehbachella* and the phylogeny of Branchiopoda and Crustacea. *Fossils and Strata.* 12.
- Wheeler WC, Giribet G, Edgecombe GD. 2004. Arthropod systematics. The comparative study of genomic, anatomical, and paleontological information. In: J Cracraft, MJ Donoghue, editors. Assembling the tree of life: Oxford University Press, New York. P. 281-295.
- Wiens JJ, Moen DS. 2008. Missing data and the accuracy of Bayesian phylogenetics. *J Syst Evol.* 46:307-314.
- Wills MA, Briggs DEG, Fortey RA, Wilkinson M. 1995. The significance of fossils in understanding arthropod evolution. *Verhandlungen der Deutschen Zoologischen Gesellschaft.* 88: 203-215.
- Wills MA. 1998a. A phylogeny of recent and fossil Crustacea derived from morphological characters. In: RA Fortey, RH Thomas, editors. Arthropod

- Relationships. London: Chapman and Hall.
- Wills MA. 1998b. Crustacean disparity through the Phanerozoic: comparing morphological and stratigraphic data. *Biol J Linn Soc.* 65:455-500.
- Wills MA, Jenner RA, Dhubhghaill CA. 2009. Eumalacostracan evolution: conflict between three sources of data. *Arth Syst Phyl.* 67:71-90.
- Yager J. 1981. Remipedia, a new class of Crustacea from a marine cave in the Bahamas. *J Crust Biol.* 1:328-333.
- Yager J, Schram FR. 1986. *Lasionectes-Entrichoma*, New Genus, New Species, (Crustacea, Remipedia) from Anchialine Caves in the Turks-and-Caicos, British West-Indies. *Proc Biol Soc Wash.* 99:65-70.
- Zrzavý J, Stys P. 1997. The basic body plan of arthropods: insights from evolutionary morphology and developmental biology. *J Evol Biol.* 10:353-367.

## Figures

### Figure 1 – Phylogram derived from data matrix 1A<sub>red</sub> (91 taxa, 316 genes) in RaxML

Topology inferred from set 1A<sub>red</sub> in RAxML (PROTCATWAG, 1,000 BS replicates, -f a). Bootstrap values are given only for nodes that lack maximum support. Ellipses on the branches (as explained in the bottom left corner) summarize the leaf stability values calculated with Phyutility (Smith and Dunn, 2008), the value for the highly unstable Ostracoda is shown in italic for this branch. 1,000 sampled bootstrap trees converged after 50 replicates applying the *a posteriori* bootstop function (Pattengale et al. 2010). Color code: crustaceans red and orange; hexapods blue; chelicerates green; myriapods brown; outgroup taxa black. Species that are marked by an asterisk

(\*) are newly sequenced in this study, species marked by an (#) are only present in the respective data set, species written in CAPITALS represent proteome taxa, a (P) indicates the used primer taxa.

**Figure 2 – RAxML topology derived from data matrix 2A<sub>red</sub> (92 taxa, 272 genes).**

Topology inferred from set 2A<sub>red</sub> in RAxML (PROTCATWAG, 1,000 BS replicates, - f a). Taxa are represented with the same colors as described in figure 1. Bootstrap values are given only for nodes that lack maximal support. 1,000 sampled bootstrap trees converged after 50 replicates. For color codes and Phyutility usage see figure 1.

**Figure 3 – RAxML topology derived from data matrix 1B<sub>red</sub> with a priori taxa exclusion (62 taxa, 351 genes).**

Topology inferred from set 1B<sub>red</sub> in RAxML (PROTCATWAG, 1,000 BS replicates, - f a). Taxa are colored as described in figure 1. Bootstrap values are given only for nodes that lack maximal support. 1,000 sampled bootstrap trees converged after 50 replicates. For color codes and Phyutility usage see figure 1.

**Figure 4 – RAxML topology derived from data matrix 2B<sub>red</sub> (67 taxa, 280 genes).**

Topology inferred from set 2B<sub>red</sub> in RAxML (PROTCATWAG, 1,000 BS replicates, - f a). Taxa are colored as seen in figure 1. Bootstrap values are given only for nodes that lack maximal support. 1,000 sampled bootstrap trees converged after 100 replicates. For color codes and Phyutility usage see figure 1.

**Figure 5 – Schematic illustrating the proposed evolutionary scenario highlighting conflicts between morphological and molecular data of pancrustaceans.**

Brown arrows and lines represent evolutionary lineages. The impact of predatory fishes as a possible evolutionary driver is illustrated by the grey waves. Circles

represent nodes that are strongly supported by morphological and molecular data. Dashed lines indicate more weakly supported relationships. Red question marks indicate branches whose position is uncertain: variously because of ambiguity in the molecular data, conflict with morphological data, or a large gap in the fossil record. Molecular and morphological evidence suggest conflicting positions for Branchiopoda and Malacostraca. Molecular analyses generally place Branchiopoda closer to Hexapoda, while selected morphological, neuroanatomical, and larval-development data suggest a closer relationship of Malacostraca to Remipedia and Hexapoda. The figure illustrates the close relationship of Remipedia and Hexapoda, which is strongly supported by the present study.

## Tables

**Table 1 – Comparison of the unreduced and reduced data sets and resulting support values for major taxa in both approaches.**

The numbers of taxa and genes, the alignment length and information content of all constructed matrices for both ortholog sets are given. Selected major taxa in all resulting topologies are listed with statistical support (bootstrap values). Dashes indicate low clade support (under 50%). Leaf stability values above 95% represent highly stable taxa. HS denotes high stability, IS an instable position, see figures.

Data set (matrix)	Set 1				Set 2			
	set 1A <sub>unred</sub>	reduced set 1A <sub>red</sub>	set 1B <sub>unred</sub>	reduced set 1B <sub>red</sub>	set 2A <sub>unred</sub>	reduced set 2A <sub>red</sub>	set 2B <sub>unred</sub>	reduced set 2B <sub>red</sub>
Number of included taxa	131	91	105	62	131	92	105	67
Number of included genes	1886	316	1886	351	1579	272	1579	280
Alignment lengths	831,013 aa	62,638 aa	866,479 aa	74,941 aa	711,430 aa	54,209 aa	736,307 aa	56,481 aa
Information content	0.276	0.617	0.26	0.62	0.276	0.62	0.26	0.614
Clade support								

Malacostraca	100	100	100	100	100	100	100	100
(Leptostraca, Eumalacostraca)	100	100	99	100	100	100	100	77
Decapoda	99	100	99	100	99	100	100	100
(Eucarida, Decapoda)	65	99	61	-	62	100	58	100
Cirripedia	100	100	100	100	100	100	100	100
(Cirripedia, Malacostraca)	100	88	99	94	99	-	100	-
(Cirripedia, Copepoda)	-	-	-	-	-	96	-	94
Copepoda	100	100	100	100	100	100	100	100
(Ostracoda, Copepoda)			22		-	-	69	-
(Ostracoda, (remaining Pancrustacea))	-	100 (IS)	-	100 (IS)	-	-	-	100 (IS)
(Ostracoda, (Malacostraca, (Cirripedia, Copepoda)))	-	-	-	-	-	79 (HS)	-	-
(Ostracoda, (Malacostraca, Cirripedia))	3	-	-	-	6	-	-	-
Branchiopoda	100	100	100	100	100	100	100	100
(Branchiopoda, (Remipedia, Hexapoda))	100	78	100	43	-	83	100	100
(Remipedia, Hexapoda)	100	98	94	100	100	96	100	100
Hexapoda	100	99	100	100	100	96	100	100
Pancrustacea	100	100	100	100	100	100	99	100
Mandibulata	91	-	96	-	99	-	83	-

**Table 2 – Comparison of gene overlap and exclusive gene occurrence in the data sets.**

The total numbers and the percentage of genes that are found in each data set derived from the two ortholog sets are given. Overlapping genes and exclusively represented genes for each data set are highlighted. Additionally the overlap with the reduced data set (SOS) from Meusemann et al. 2010 with each of our data sets is included. The sum-column shows the percentages of genes unique to each particular data set, and those shared with the corresponding data set derived from the other ortholog set (For a graphical comparison see supplementary file 5b).

Gene overlap of data sets	Numbers of genes	Percentage of genes	Sum
<b>Ortholog set 1: set 1A<sub>unred</sub> Total:1886</b>			
set 1 specific only	442	23%	Σ 25%
set 1 and SOS	34	2%	
set 1, set 2 and SOS	496	26%	Σ 75%
set 1 and set 2 only	914	49%	
<b>Ortholog set 2: set 2A<sub>unred</sub> Total: 1579</b>			
set 2 specific only	131	8%	Σ 10%
set 2 and SOS	38	2%	
set 2, set 1 and SOS	496	32%	Σ 90%
set 2 and set 1 only	914	58%	
<b>Ortholog set 1: set 1A<sub>red</sub> Total: 316</b>			
set 1 specific only	68	21%	Σ 23%
set 1 and SOS	6	2%	
set 1, set 2 and SOS	65	21%	Σ 77%
set 1 and set 2 only	177	56%	
<b>Ortholog set 2: set2A<sub>red</sub> Total: 272</b>			
set 2 specific only	26	10%	Σ 11%
set 2 and SOS	4	1%	
set 2, set 1 and SOS	65	24%	Σ 89%
set 2 and set 1 only	177	65%	
<b>Ortholog set 1: set 1B<sub>red</sub> Total: 351</b>			
set 1 specific only	90	26%	Σ 27%
set 1 and SOS	4	1%	
set 1, set 2 and SOS	66	19%	Σ 73%
set 1 and set 2 only	191	54%	
<b>Ortholog set 2: set 2B<sub>red</sub> Total: 280</b>			
set 2 specific only	19	7%	Σ 8%
set 2 and SOS	4	1%	
set 2, set 1 and SOS	66	24%	Σ 92%
set 2 and set 1 only	191	68%	

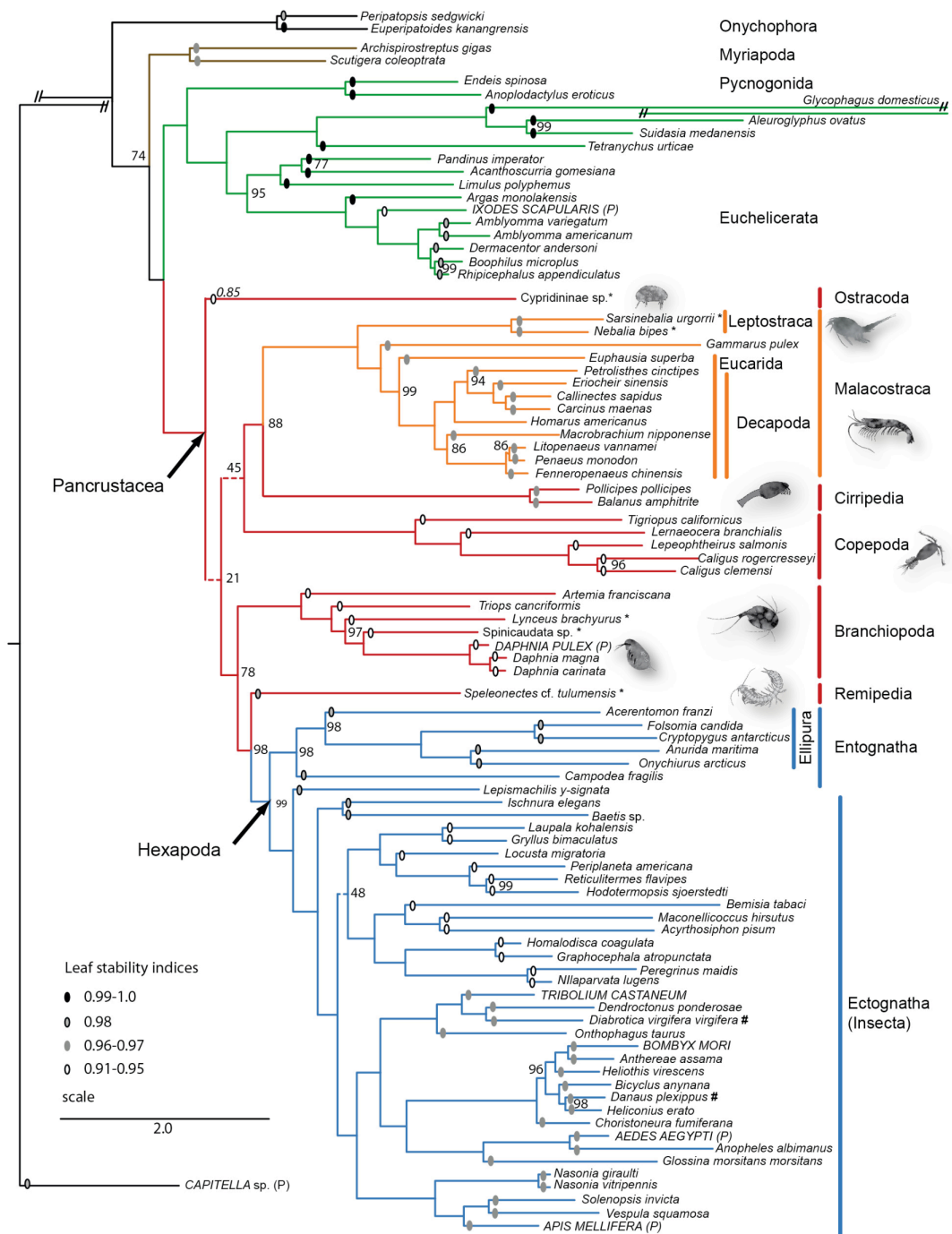


Figure 1

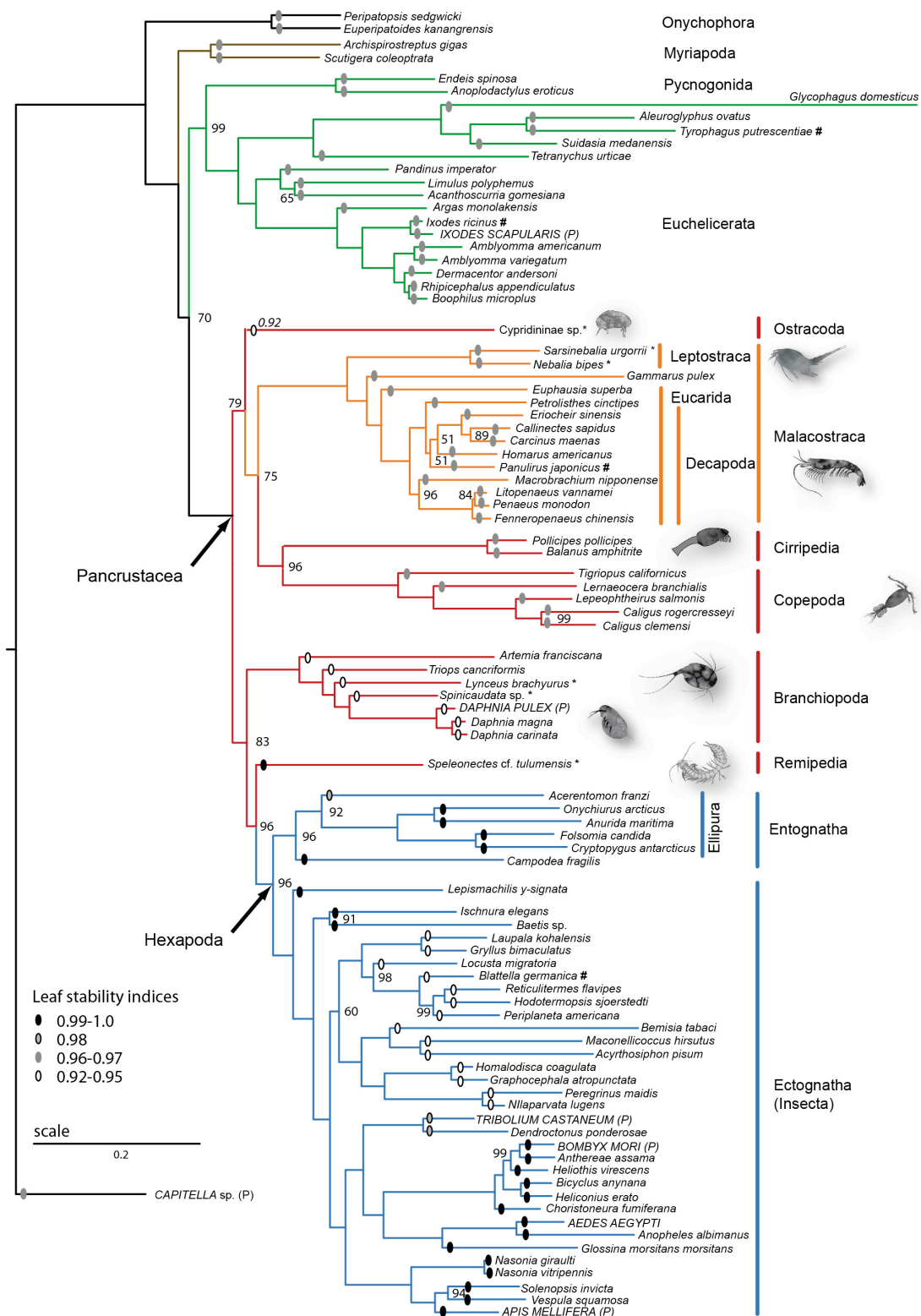


Figure 2



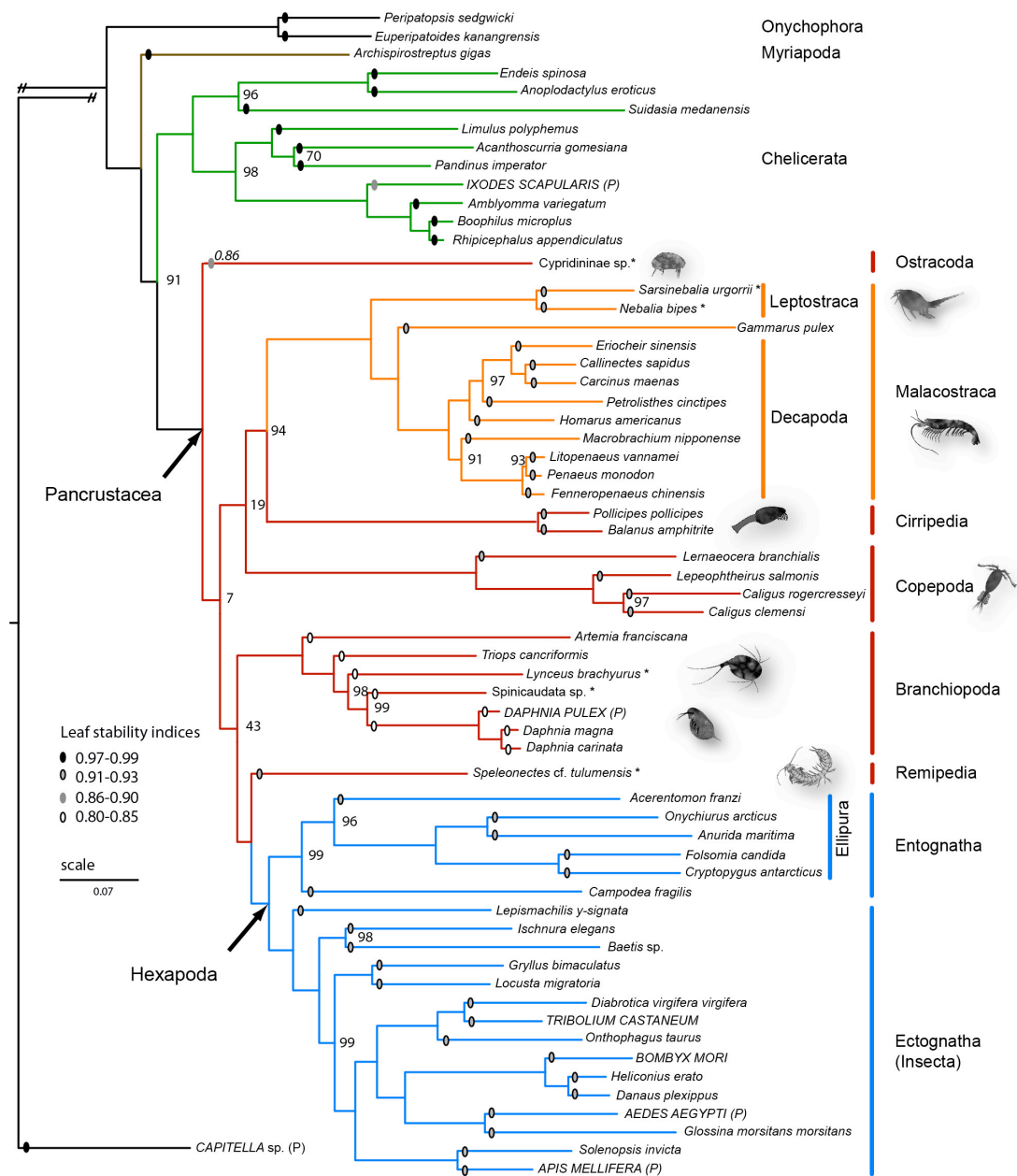


Figure 3

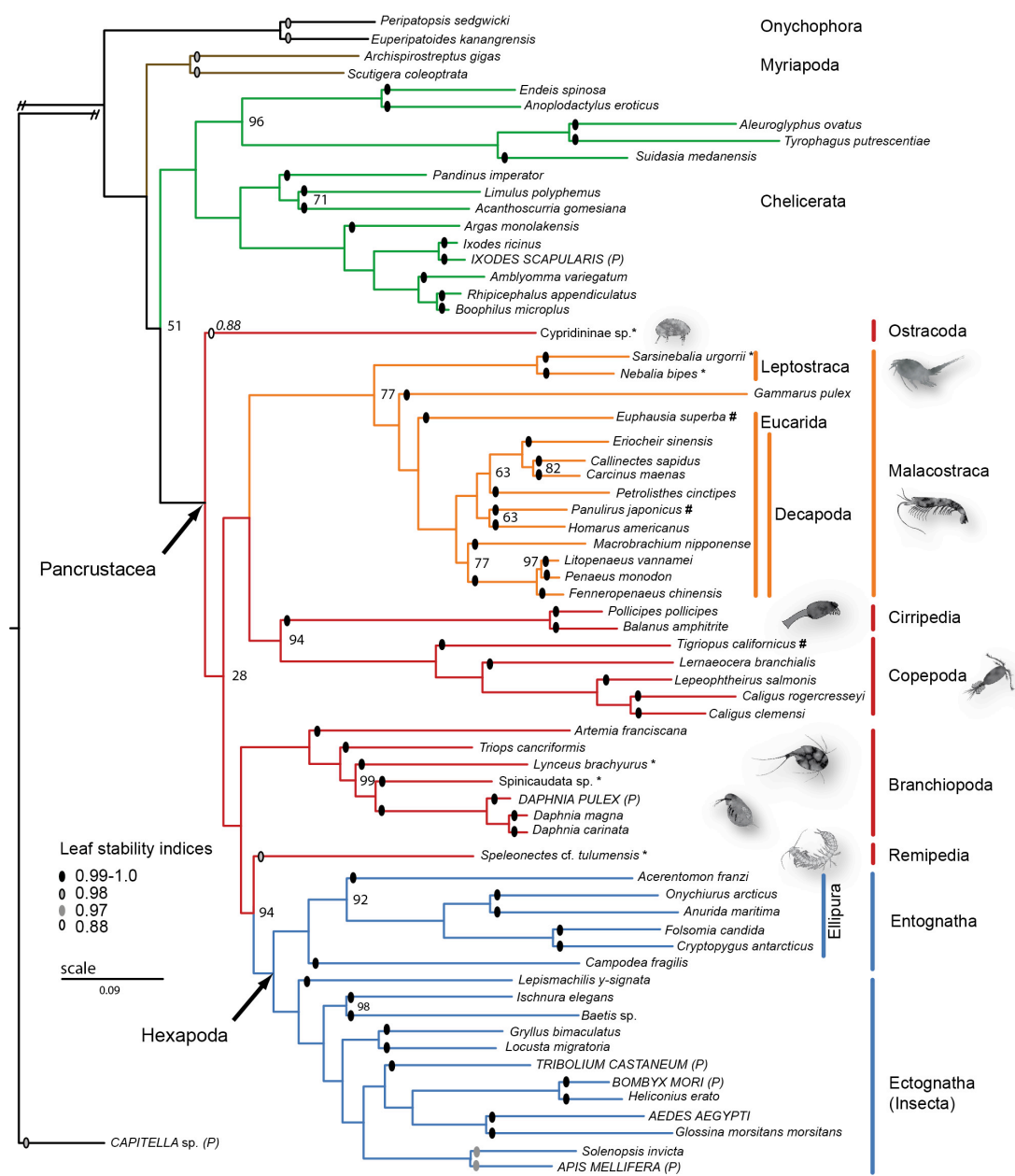


Figure 4

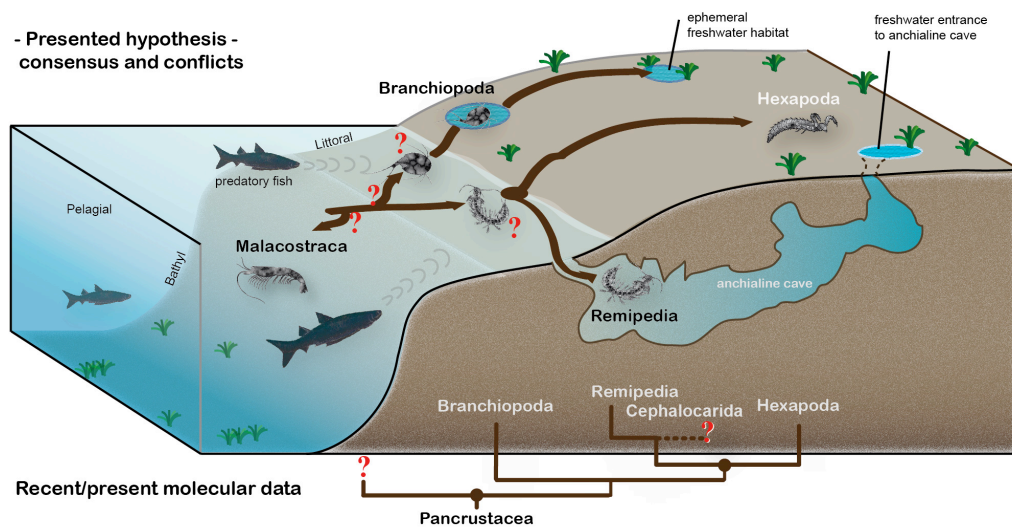


Figure 5