MOLECULAR INSIGHTS TO CRUSTACEAN PHYLOGENY

DISSERTATION

zur Erlangung des Doktorgrades (Dr. rer. nat.)
an der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

BJOERN MARCUS VON REUMONT

Bonn – Oktober 2009

Angefertigt mit der Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn.

Diese Dissertation wurden am Zoologischen Forschungsmuseum Alexander Koenig in Bonn durchgeführt.

Tag der mündlichen Prüfung 29.01.2010
Erscheinungsjahr 2010

Betreuer

Prof. Dr. Johann-Wolfgang Waegele
Prof. Dr. Bernhard Y. Misof

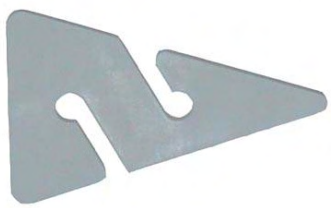
[…] Make up your mind, plan before and follow that plan.

[…] An unforgivenable sin is quitting. Never give up and keep on going. The only struggle should be to solve the problem or survive. […] Focus on the task and on the moment. […]

Excerpts of a cave diving manual

This basic philosophy for cave diving is not only useful in cave diving but also generally in live and was indeed helping not only once conducting this thesis. And while diving…

Dedicated to my family and all good friends – a permanent, safe mainline



CONTENTS

Above picture: *Derocheilocaris typicus*, a specimen of the Mystacocarida

Cover picture: Some of the rather small but beautiful crustaceans (from above left: Branchiura, Mystacocarida, Copepoda, Ostracoda, Cladocera, Branchiopoda and Remipedia.

ABSTRACT

A key role in arthropod phylogeny plays a group of organisms that was already in the focus of taxonomic research of Charles Darwin in the mid of the 19th century, namely the Crustacea. This extremely divers group comprises small species like the Mystacocarida (*Derocheilocaris typicus*) with only 0.3 mm body size or such big representatives like the Japanese giant crab (*Macrocheira kaempferi)* with a span width of almost 4 m. Generally accepted are six major crustacean taxa, the Malacostraca (Latreille, 1802), Branchiopoda (Latreille, 1817), Remipedia (Yager, 1981), Cephalocarida (Sanders, 1955), Maxillopoda (Dahl, 1956) and Ostracoda (Latreille 1802). The validity of the taxon Maxillopoda is to date still disputed. The monophyly of some crustacean groups like the Malacostraca and Branchiopoda is generally accepted, but for several other groups unclear. This thesis aims to resolve internal relationships of the major crustacean groups inferring phylogenies with molecular data. The crustaceans are in addition of eminent interest to enlight the question how land was successfully conquered by arthropod taxa. New molecular and neuroanatomical data support the scenario that the Hexapoda might have evolved from Crustacea. The thesis further seeks to address the possible close relationship of Crustacea and Hexapoda. That issue is closely linked to the partly still debated position of crustaceans within arthropods and the supposable sister-group of the Crustacea.

Most molecular studies of crustaceans relied on single gene or multigene analyses in which for most cases partly sequenced rRNA genes were used. However, intensive data quality and alignment assessments prior to phylogenetic reconstructions are not conducted in most studies. Additionally, a complex modeling and the implementation of compositional base heterogeneity along lineages are missing. One methodological aim in this thesis was to implement new tools to infer data quality, to improve alignment quality and to test the impact of complex modeling of the data. Two of the three phylogenetic analyses in this thesis are also based on rRNA genes.

In analysis (A) 16S rRNA, 18S rRNA and COI sequences were analyzed. RY coding of the COI fragment, an alignment procedure that considers the secondary structure of RNA molecules and the exclusion of alignment positions of ambiguous positional homology was performed to improve data quality. Anyhow, by extensive network reconstructions it was shown that the signal quality in the chosen and commonly used markers is not suitable to infer crustacean phylogeny, despite the extensive data processing and optimization. This result draws a new light on previous studies relying on these markers.

In analyses (B) completely sequenced 18S and 28S rRNA genes were used to reconstruct the phylogeny. Base compositional heterogeneity was taken into account based on the finding of analysis (A), additionally to secondary structure alignment optimization and alignment assessment. The complex modeling to compare time-heterogeneous versus time-homogenous processes in combination with mixed models for an implementation of secondary structures was only possible applying the Bayesian software package PHASE. The results clearly demonstrated that complex modeling counts and that ignoring time-

heterogeneous processes can mislead phylogenetic reconstructions. Some results enlight the phylogeny of Crustaceans, for the first time the Cephalocarida (*Hutchinsoniella macracantha*) were placed in a clade with the Branchiopoda, which morphologically is plausible. Unfortunately, the internal relationships of most crustacean groups were still poorly supported. Compared to the time-homogeneous tree the time-heterogeneous tree gives lower support values for some nodes. It can be suggested, that the incorporation of base compositional heterogeneity in phylogenetic analysis improves the reliability of the topology. The Pancrustacea are supported maximally in both approaches, but internal relations are not reliably reconstructed. One result of this analysis is that the phylogenetic signal in rRNA data might be eroded for crustaceans.

Recent publications presented analyses based on phylogenomic data, to reconstruct mainly metazoan phylogeny. Analyzing such a large number of sequences is possible with the "supertree" or "supermatrix" method. The supermatrix method seems to outperform the supertree approach. One main advantage is the possibility to apply modeling for each partition (each gene) separately. Within this thesis crustaceans were collected to conduct EST sequencing projects and to include the resulting sequences combined with public sequence data into a phylogenomic analysis (C). In this analysis the supermatrix approach was applied. New and innovative reduction heuristics were performed to condense the dataset. The strategy of the reduction heuristics relies on the potential relative information content of each gene of each taxon to use a more objective criterion to select taxa and genes. Again, the alignment evaluation and processing was a major aspect for the analysis design. The results showed that the matrix implementation of the reduced dataset ends in a more reliable topology in which most node values are highly supported. In analysis (C) the Branchiopoda were positioned as sister-group to Hexapoda, a differing result to analysis (A), but that is in line with other phylogenomic studies. Unfortunately, important crustacean taxa are still missing to conduct an extensive phylogenomic analysis. Some EST sequencing projects of the collected crustaceans for this thesis were delayed for technical reasons, e.g. the ESTs for *Sarsinebalia urgorrii* (non-derived malacostracan) and *Speleonectes tulumensis* (Remipedia) are still in progress. A preliminary result obtained with sequences isolated from remipede tissue is suggesting that remipedes and hexapods are closely related based on homologous hemocyanin subunits.

The conclusion of the analyses conducted in the framework of this thesis is that alignment evaluation and processing improves the resulting inference of the phylogeny. Assessing the quality of the signal or potential conflicts in the dataset is extremely important, also for further decisions on the selection of substitution models and final phylogenetic reconstructions. Complex models can improve the phylogeny reconstruction additionally. This was explicitly demonstrated in analysis B. The supermatrix approach relying on a more objective criterion to select genes and taxa compared to cut-off values is very promising for future studies. However, for the Crustacea it was also demonstrated that this group is problematic regarding the phylogenetic signal of the analyzed single gene data. The hope is, that phylogenomic data with similar complex models as applied in analysis B, in combination

with a denser taxon sampling can improve our knowledge about crustacean phylogeny in future studies. This thesis presents essential new methodological but also phylogenetic findings for this challenging task.

ZUSAMMENFASSUNG

Eine Schlüsselrolle in der Evolution der Arthropoda spielen die Krebse (Crustacea), einige Krebsgruppen waren bereits Studienobjekt seitens Charles Darwins in der Mitte des 19. Jahrhunderts. Die Crustacea sind eine extrem diverse Gruppe und umfassen so kleine Arten wie die Mystacocarida (*Derocheilocaris typicus*) von nur 0.3 mm Körpergröße und so große Arten wie die Japanische Riesenkrabbe mit fast 4 m Spannweite. Allgemein anerkannt sind sechs Großgruppen der Crustacea, die Malacostraca (Latreille, 1802), Branchiopoda (Latreille, 1817), Remipedia (Yager, 1981), Cephalocarida (Sanders, 1955), Maxillopoda (Dahl, 1956) und Ostracoda (Latreille 1802). Die Maxillopoda sind als valides Taxon recht umstritten. Die Monophylie einiger Gruppen der Crustacea, wie z.B. der Malacostraca und Branchiopoda ist allgemein akzeptiert, für die meisten Gruppen jedoch noch immer unklar. Diese Doktorarbeit soll unter anderem die größtenteils noch unklaren Verwandschaftsverhältnisse zwischen den Gruppen der Crustacea mit molekularen Methoden untersuchen. Die Crustacea sind zusätzlich von großem Interesse, um die Frage zu beantworten, wie die Arthropoda so erfolgreich das Land besiedeln konnten. Neuere Ergebnisse von molekularen und neuroanatomischen Studien unterstützen ein Szenario, in welchem die Hexapoda von den Crustacea abstammen. Die Frage, ob Hexapoda nah verwandt mit den Crustacea sind und eventuell aus diesen evolvierten, soll ebenfalls in der Arbeit untersucht werden. Eng verbunden mit dieser Frage ist die teilweise noch immer diskutierte Stellung der Crustacea innerhalb der Arthropoda.

Die Analysen der meisten Studien zur Phylogenie der Crustacea basieren auf Einzelgenen, oder "Multigenanalysen" mit nur wenigen Genen. Ribosomale RNA Gene wurden besonders häufig hierfür verwendet. Allerdings erfolgte nur in den wenigsten Studien auch eine Analyse bezogen auf Qualität der Daten und im Besonderen des Alignments. Hinzu kommt, dass eine komplexe Modellierung der Daten und vor allem die Implementierung der Inhomogenität der Basenkompositionen in den meisten Analysen fehlt. Ein methodisches Ziel in dieser Arbeit war neue Methoden und Werkzeuge zu verwenden um Daten und Alignmentqualität zu evaluieren und zu verbessern, bei gleichzeitiger Verwendung von komplexen Modellierungen.

In zwei von den drei hier vorgestellten Analysen, werden auch rRNA Gene verwendet. In Analyse (A) wurden 16S rRNA, 18S rRNA und COI Sequenzen analysiert. RY Kodierung für das COI Fragment, ein mittels Sekundärstrukturen optimiertes Alignment und der Ausschluss von zufällig gleich alinierten Positionen im Alignment dienten der Verbesserung der Qualität der Daten. Allerdings wurde durch die Verwendung von Netzwerk Rekonstruktionen gezeigt, dass die verwendeten molekularen Marker nur eingeschränkt geeignet sind, um die

Phylogenie der Crustacea zu beleuchten. Dieses Ergebnis lässt die bisherigen Studien, die zum großen Teil auf diesen Markern beruhen, in einem anderen Licht erscheinen.

In Analyse (B) wurden deshalb vollständig sequenzierte Sequenzen von 18 und 28S rRNA zur phylogenetischen Rekonstruktion verwendet. Inhomogene Basenkomposition wurde berücksichtigt und analog zu Analyse (A) erfolgte ein sekundärstruktur-optimiertes Alignment mit anschließender Evaluierung des Alignments, um zufällig ähnliche Positionen im Alignment zu exkludieren. Das Anwenden von komplexen Modellen zum Vergleich von zeithomo- und zeitheterogenen Prozessen war in Kombination mit der Implementierung von gemischten Modellen zur realistischen Darstellung von Sekundärstrukturen nur mit der Software PHASE möglich. Die Ergebnisse aus dieser Analyse machen deutlich, dass komplexes Modellieren extrem wichtig ist, und das stillschweigende Ignorieren von zeit-heterogene Prozessen die Baumrekonstruktion beeinflussen kann. Einige der Ergebnisse werfen neues Licht auf die Evolution der Crustacea, zum ersten Mal wurde *Hutchinsoniella macracantha* (Cephalocarida) zu den Branchiopoda gestellt, was morphologisch recht plausibel ist. Leider wurden die internen Knoten für die Krebse in beiden Ansätzen schlecht aufgelöst. Die Pancrustacea wurden aber in beiden Ansätzen maximal unterstützt. Es ist zu vermuten, dass im zeitheterogenen Ansatz eine realistischere, partiell nicht gut gestützte Topologie rekonstruiert wurde. Allerdings ist auch ein klares Ergebnis dieser Analyse, dass die rRNA Daten nur gering zur Aufklärung der Crustacea Phylogenie beitragen können, ihr Signal ist für diese Gruppe offensichtlich stark erodiert.

Neuere Studien (überwiegend zur Metazoa Evolution) basieren auf phylogenomischen Daten. Das Analysieren solch großer Datensätze ist mittels der "Supertree" oder "Supermatrix" Methode möglich. Momentan ist die Supermatrix Methode performanter, vor allem ist ein Vorteil, dass die Partitionen (Gene) getrennt durch verschiedene Modelle beschrieben werden können. Im Rahmen dieser Doktorarbeit wurden Crustacea gesammelt um EST Sequenzierungsprojekte durchzuführen und eine phylogenomische Analyse zu starten In Analyse (C) wurden die Sequenzdaten auch von publizierten EST und Genomprojekten mittels des Supermatrix Ansatzes untersucht. Hierfür wurde eine neue Strategie angewandt, die über den relativen Informationsgehalts der Gene, Gene und Taxa aussucht. Uninformative Gene werden exkludiert, Taxa die nur solche Gene aufweisen, ebenfalls. Ein weiterer wichtiger Punkt war wieder die Qualität der Daten und des Alignments, analog zu Analysen (A) und (B). Das Ergebnis zeigt, dass die Reduktion des Datensatzes mit der neuen Methode eine plausiblere und besser gestützte Topologie zur Folge hat. In dieser Analyse sind die Branchiopoda die Schwestergruppe zu den Hexapoda, was Analyse (B) widerspricht. Allerdings ist dieses Ergebnis kongruent zu anderen phylogenomischen Daten. Leider fehlen zu einer klaren Aussage zur Phylogenie der Crustacea noch einige Crustacea Arten in den phylogenomischen Analysen. Einige geplante EST Sequenzierungsprojekte, von im Rahmen dieser Arbeit gesammelten Arten, haben sich aus technischen Gründen verzögert. Dies sind z.B. *Sarsinebalia urgorrii (*Malacostraca) und *Speleonectes tulumensis* (Remipedia). Ein vorläufiges Ergebnis aus den EST Daten der

Remipedia war der Fund von Hämocyanin Untereinheiten, die auf eine nahe Verwandtschaft von Remipedia und Hexapoda schließen lassen.

Die Analysen, die im Rahmen dieser Arbeit durchgeführt wurden, zeigen deutlich, dass Alignment, Evaluierung und Optimierung die resultierenden Baumrekonstruktionen verbessern. Ein Überprüfen der Datenqualität und der Qualität des Signals für den jeweiligen Datensatz ist von enormer Wichtigkeit. Einen ähnlich großen Einfluss können komplexe Modelle haben, wenn Prozesse wie inhomogene Basenkomposition auftreten. Der Supermatrix Ansatz ist vielversprechend für weitere Studien. Gene und Arten nach einem objektiven Kriterium (relativer Informationsgehalt) auszuwählen, worauf die so "kondensierten" Datensets in eine finale Analyse einfließen, resultiert in robusteren Phylogenien, anstatt mit Schwellenwerten zu arbeiten. Es wurde jedoch auch gezeigt, dass die Crustacea eine recht problematische Gruppe sind. Eventuell durch ihr Alter scheint das phylogenetische Signal in den durchgeführten Einzelgenanalysen stark erodiert zu sein. Die phylogenomischen Daten sind zurzeit nicht ganz aussagekräftig, da noch immer viele Taxa der Crustacea fehlen. Zu hoffe ist, dass komplexes Modellieren in Kombination mit einer guten Artenauswahl in weiteren phylogenomischen Analysen unsere Einsichten in die Evolution der Crustacea verbessert. Mit dieser Arbeit wurde unter anderem für diese weiteren Analysen eine wichtige methodische aber auch phylogenetische Basis geschaffen.

1. INTRODUCTION

Evolutionary biology, in contrast with physics and chemistry, is a historical science, the evolutionist attempts to explain events and processes that have already taken place. Laws and experiments are inappropriate techniques for the explication of such events and processes. Instead one constructs a historical narrative, consisting of a tentative reconstruction of the particular scenario that led to the events one is trying to explain (ERNST MAYR).

In the "Darwin-year" 2009 evolution draws a special attention to the public audience. The evolutionary processes resulted in the diversity of species we find today on our planet. One of the most diverse invertebrate groups are the Arthropoda. A persistent challenge in phylogenetic systematics concerns the evolution and genealogical relationships of the Arthropoda. The evolution of this group is debated since the 19th century – a debate that started yet in the times of CHARLES DARWIN. Today it is still unclear how land and air were successfully conquered by arthropods. Their evolutionary success was accompanied by astounding transitions of body plan organisations, evolving more than three times as many species compared to other multicellular organisms. Currently hypotheses are conflicting to explain the successful evolution of the arthropods. Crustaceans may play a key role in the scenario how arthropods colonized land. Possibly freshwater crustaceans (Branchiopoda) constituted a link to early hexapods starting to crawl on land (GLENNER ET AL. 2006) or the enigmatic crustacean group Remipedia evolved from ancestors shared with the Tracheata (FANENBRUCK 2003).

The phylogenetic relationships of the five major traditional euarthropod groups, the Hexapoda, Myriapoda, Crustacea, Chelicerata, and the extinct Trilobitomorpha, are debated since the 19th century (e.g. LANKESTER 1904; LATREILLE 1817; POCOCK 1893A; POCOCK 1893B). Although arthropod phylogeny has long been debated based on morphological and developmental evidence (BÄCKER ET AL. 2008; BITSCH & BITSCH 2004; HARZSCH 2006; UNGERER & SCHOLTZ 2008) there is since several years additionally a strong focus on results from molecular data, derived from mitochondrial, nuclear and phylogenomic datasets (BOORE ET AL. 1995; HASSANIN 2006; MALLATT ET AL. 2004; MALLATT & GIRIBET 2006; DUNN ET AL. 2008; ROEDING ET AL. 2007).

Crustaceans are of eminent interest for the study of arthropod phylogeny since molecular analyses often reveal this clade paraphyletic in respect to hexapods (e.g. REGIER & SHULTZ 2001; BABBITT & PATEL 2005; MALLATT & GIRIBET 2006). The position of crustaceans within arthropods is controversially discussed in classical morphological concepts positioning them as sister-group to Tracheata (SNODGRASS 1935), Chelicerata (CISNE 1974) or Hexapoda

(PAULUS 1979; ZRZAVÝ & ŠTYS 1997; ZRZAVÝ & ŠTYS 1998A). Morphologically crustaceans are extremely divers and internal relationships are still unclear (e.g. MARTIN & DAVIS 2001). At the morphological front a "standoff" situation is created by conflicting hypotheses. One reason for this might be that crustaceans are a very old group, at least 500 million years old (e.g. WALOSSEK 1993, SIVETER ET AL. 2001, see contrary BUDD ET AL. 2001). New insights from developmental or neurobiological studies create an even more chaotic picture of crustacean phylogeny and their position within arthropods instead of enlighting this field. The hope that molecular studies could enhance the understanding of crustacean evolution in complementing and directing the contradicting morphological disputes was often disappointed. Instead, molecular studies contribute in many cases to contradicting scenarios of crustacean and arthropod evolution compared to other molecular but also morphological studies. Yet SPEARS AND ABELE (1998) mentioned the problems to select and interpret useful phylogenetic characters to infer crustacean phylogeny. They conclude that rapid radiation in younger crustacean lineages (see WÄGELE ET AL. 2003) but also the older arthropod nodes are problematic to reconstruct by molecular analysis. One has to keep in mind that due to the age of this group molecular analyses of crustaceans might operate at their limit because this long time span led to signal erosion in the sequences by several, multiple substitutions (WÄGELE & MAYER 2007; WÄGELE ET AL. 2009). A careful choice of molecular markers and molecular methods in combination with a rigorous quality assessment of the data and tree reconstruction methods is for this reason crucial. Especially the new field of phyolgenomics burgeoned the hope that stochastic errors of previous multi-gene analyses disappear and can enlight molecular phylogeny of crustaceans and arthropod.

1.1 Crustaceans and their controversial phylogeny – a short overview

The revelation of the internal phylogenetic relationships is not equal across the major extant arthropod taxa. A conspicuous relative lack of both attention and progress in understanding the phylogeny of the Crustacea exists compared to work on the higher-level phylogeny of hexapods, chelicerates and myriapods. Phylogenetic hypotheses about the evolution of body plan diversity of crustaceans are still chiefly based on morphological evidence (e.g., DAHL 1963; SCHRAM 1986; WILSON 1992; SCHRAM & HOF 1998; WILLS 1998; SCHRAM & KOENEMANN, 2004B), with little detailed consensus. Higher-level crustacean molecular phylogenetics started relatively late effectively in the late 1980s and 1990s by ABELE and SPEARS. There is cumulating morphological and molecular evidence that Crustacea *s. str.* may represent a paraphyletic assemblage of arthropods (MALLATT ET AL. 2004; MALLATT & GIRIBET 2006; REGIER ET AL 2005; RICHTER 2002; HARZSCH 2006). The concept of a hexapod-crustacean clade, Pancrustacea or Tetraconata, has been proposed independently in a number of studies (e.g. REGIER & SHULTZ 1997; SPEARS & ABELE 1998; ZRZAVÝ AND ŠTYS 1997;

GARCHIA-MACHADO ET AL. 1999; LAVROV ET AL. 2004; SCHRAM & KOENEMANN 2004B; COOK ET AL. 2005; REGIER ET AL. 2008) regardless if a para- or polyphyletic crustacean clade is inferred. With respect to extant taxa this means that hexapods are positioned within Crustacea, although it remains unclear to which extant the crustacean taxon Hexapoda would be closely related.

Six crustacean classes (figure 1.1) are recognized by MARTIN AND DAVIS (2001): Malacostraca, Branchiopoda, Remipedia, Cephalocarida, Maxillopoda and Ostracoda.

To date the most comprehensive higher-level phylogenetic analyses within Crustacea focus only on few groups, mainly Malacostraca (JENNER ET AL. 2009; MELAND & WILLASSEN 2007; SPEARS ET AL. 1992; 2005), Branchiopoda (BRABAND ET AL. 2002; DEWAARD ET AL. 2006; RICHTER ET AL. 2007; STENDERUP ET AL. 2006) and Thecostraca (PÉREZ-LOSADA ET AL. 2008; PÉREZ-LOSADA ET AL. 2004, see section Maxillopoda). These and larger-scale studies support the monophyly of Branchiopoda, Malacostraca and Thecostraca.

Remipedia and Cephalocarida are both considered monophyletic (MARTIN AND DAVIS 2001; KOENEMANN ET AL. 2007), but their phylogenetic positions remain unknown. In most studies (GIRIBET ET AL. 2001; REGIER ET AL. 2005) the clade of both taxa is reconstructed which might be influenced by the predominant long branches these taxa show in molecular analyses. Remipedes were described as "most primitive" crustaceans (SCHRAM 1983; 1986). However, results of neuroanatomical studies place remipedes as sister-group to Hexapoda (FANENBRUCK ET AL. 2004; FANENBRUCK & HARZSCH 2005) or Tracheata (FANENBRUCK 2003). The proposed clade Tracheata + Remipedia was named **Archilabiata-hypothesis** by FANENBRUCK (2003).

Maxillopoda are presented in MARTIN AND DAVIS (2001) as a "continuously terribly controversial assemblage". A similar taxon was earlier named Copepodoidea (BEKLEMISHEV 1952) with almost the same taxa included as in MARTIN & DAVIS (2001).

A monophyletic Maxillopoda in the composition first proposed by DAHL (1956) seems increasingly doubtful (overview in: MARTIN & DAVIS 2001), although maxillopodan monophyly is suggested on the basis of some morphological evidence (WALOSSEK & MÜLLER 1998B; WILLS 1998; AX 1999).

Excluding the Ostracoda, their positioning in Maxillopoda is doubted by molecular (ABELE 1992; SPEARS & ABELE 1998) and morphological data (WILSON 1992), this assemblage is constituted by the eight taxa: Copepoda, Mystacocarida, Branchiura, Thecostraca, Facetotecta, Ascothoracica, Cirripedia and Pentastomida. There are morphological studies that disagree (SCHRAM & KOENEMANN 2004B) and molecular evidence that contradicts maxillopodan monophyly (SPEARS & ABELE 1998; REGIER ET AL. 2005; 2008). Already BOXSHALL (1983) claimed that "Maxillopoda is not a valid taxon" which is underpinned by these recent analyses. However, although various studies include samples of maxillopod taxa, to date no broadly sampled molecular maxillopodan phylogeny is available.

Figure 1.1: Representatives of six major crustacean classes. A: *Sarsinebalia urgorrii* (MOREIRA, GESTOSO & TRONCOSO, 2003), a leptostracan as representative of the early Malacostraca. **B**: The branchiopods *Triops cancriformis* (BOSC, 1801) and *Daphnia magna* (STRAUS, 1820). **C**: *Hutchinsoniella macracantha* (SANDERSON, 1955), a cephalocarid. **D**: *Speleonectes tulumensis* (YAGER, 1987), Remipedia (picture kindly provided by KOENEMANN). **E**: *Derocheilocaris typica* (PENNAK & ZINN, 1943), Mystacocarida as representative of the copepodan lineage of the Maxillopoda. **F**: *Semibalanus balanoides* (LINNAEUS, 1758) and *Pollicipes pollicipes (GMELIN, 1789)* representing the thecostracan lineage of the Maxillopoda. **G**: Specimens of *Heterocypris incongruens* (RAMDOHR 1808), a freshwater ostracod.

Cirripedia were affirmed in extensive, recent studies (HØEG ET AL. 2009; PÉREZ-LOSADA ET AL. 2008) as monophyletic, within the clade Thecostraca (*Cirripedia* (Acrothoracica (Ascothoracica + Facetotecta))). Thecostraca were subject of previous studies (PÉREZ-LOSADA ET AL. 2008; PÉREZ-LOSADA ET AL. 2004) that support the monophyly of this clade by the same type of head-shield organ, the lattice organ (GRYGIER 1987; JENSEN ET AL. 1994; HOEG & KOLBASOV 2002).

Pentastomida are by now generally included into Maxillopoda, supported by molecular studies based on 18S and 28S rRNA (ABELE ET AL. 1989; MALLATT & GIRIBET 2006) and mitochondrial data (COOK ET AL. 2005; LAVROV ET AL. 2004). Also morphological (WINGSTRAND 1972) and combined morphological and molecular data (MØLLER ET AL. 2008) affirm these findings and position Pentastomida as sister-group to the Branchiura. However, there are morphological studies that contradict these hypotheses based on fossils of the Upper Cambrian 'Orsten' which place the Pentastomida outside the Euarthropoda (DE OLIVIERA ALMEIDA ET AL. 2008; WALOSZEK ET AL. 2005; WALOSZEK ET AL. 2006).

Ostracoda are traditionally considered monophyletic (MARTIN & DAVIS 2001), consistent with a recent morphological phylogenetic analysis (HORNE ET AL. 2005). Molecular evidence instead unites podocopid ostracodes more closely with branchiurans (and possibly pentastomids) than with myodocopids (SPEARS & ABELE 1998; REGIER ET AL. 2005; 2008).

Finally, it has to be stated again, that a consensus of the positions or even a generally accepted phylogeny of the crustacean groups is not yet in sight (see figure 1.2). There exist so many different hypotheses, that an obscure jungle of trees demands some thin out of this chaos. In this background the tested, main hypotheses discussed in this thesis are pictured in the next section.

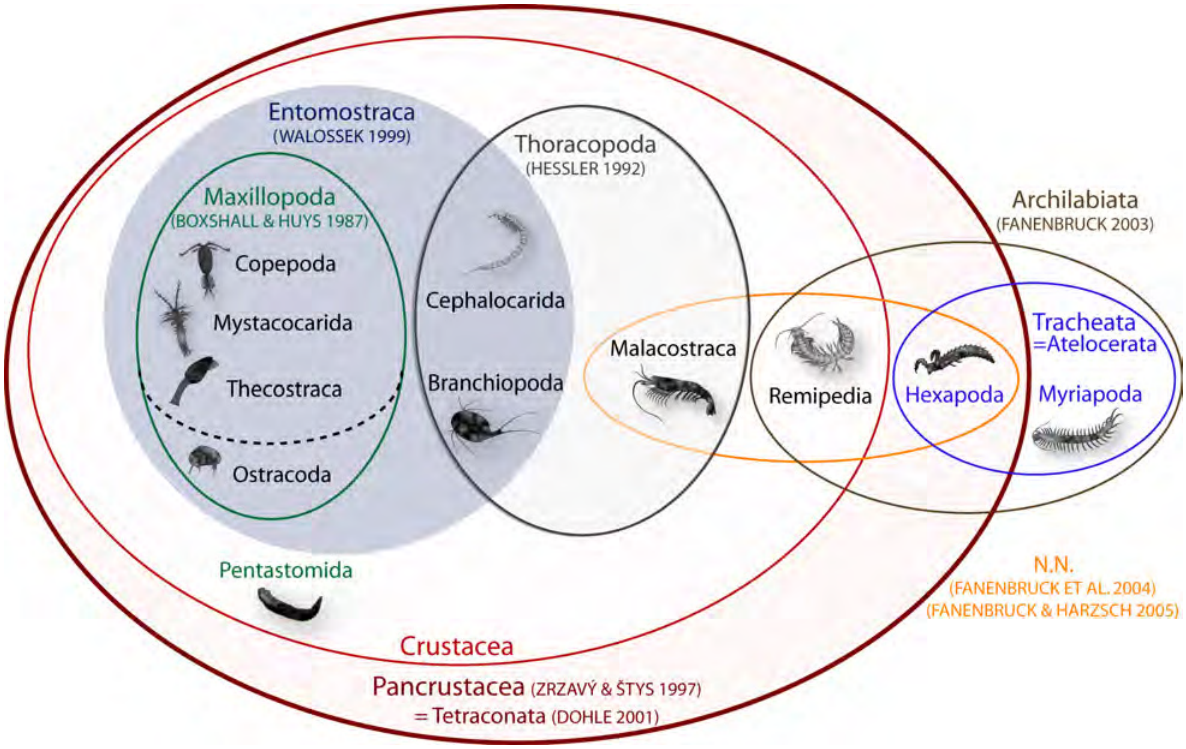


Figure 1.2: Conflicting hypotheses of crustacean phylogeny. The Van-Venn diagram shows the unclear internal relationships of the Crustacea. The Pentastomida are included to Maxillopoda in recent studies (SPEARS & ABELE 1998, MØLLER ET AL. 2008). Ostracoda are contrary excluded from Maxillopoda in some studies (ABELE ET AL. 1992; SPEARS & ABELE 1998).

1.2 Contradicting phylogeny hypotheses of major crustacean groups

Entomostraca: All non-malacostracan taxa except the Remipedia are often combined in the taxon Entomostraca. Most of these groups are represented by species of very small body size (e.g. Mystacocarida [~0.2 mm]; Copepoda [~0.7-2 mm]) and therefore harder to collect, sort and to study. WALOSZEK (WALOSSEK & MÜLLER 1998A+B; WALOSSEK 1999) proposes a set of morphological characters that are described as autapomorphies of Entomostraca. Malacostraca are placed as sister group to the Entomostraca.

Malacostraca are usually regarded to be a monophyletic taxon (SPEARS & ABELE 1998; GIRIBET & RIBERA 2000; JENNER ET AL. 2009; MELAND & WILLASSEN 2007; SPEARS ET AL. 1992; 2005). Several authors suggest that Malacostraca evolved later within Crustacea as a more derived crustacean group (FANENBRUCK 2003; FANENBRUCK ET AL. 2004; RICHTER & SCHOLTZ 2001). The malacostracan phylogeny is controversial for few internal groupings (RICHTER & SCHOLTZ 2001; JENNER ET AL. 2009).

Thoracopoda: HESSLER & NEWMAN (1975) proposed this clade including Cephalocarida, Branchiopoda and Malacostraca based on the existence on an epipodite (see also HESSLER 1982; 1992). This concept is clearly in conflict with the Entomostraca concept. EDGECOMBE ET AL. (2000) confirmed monophyletic Thoracopoda, using a dataset with morphological characters and two gene sequences. Later, ZHANG ET AL. (2007) questioned the existence of an epipodite as character for Thoracopoda. This study describes the existence of epipodites for Cambrian fossils (*Yicaris dianensis)*. The authors conclude a groundpattern of Eucrustacea including epipodites contradicting the Thoracopoda concept.

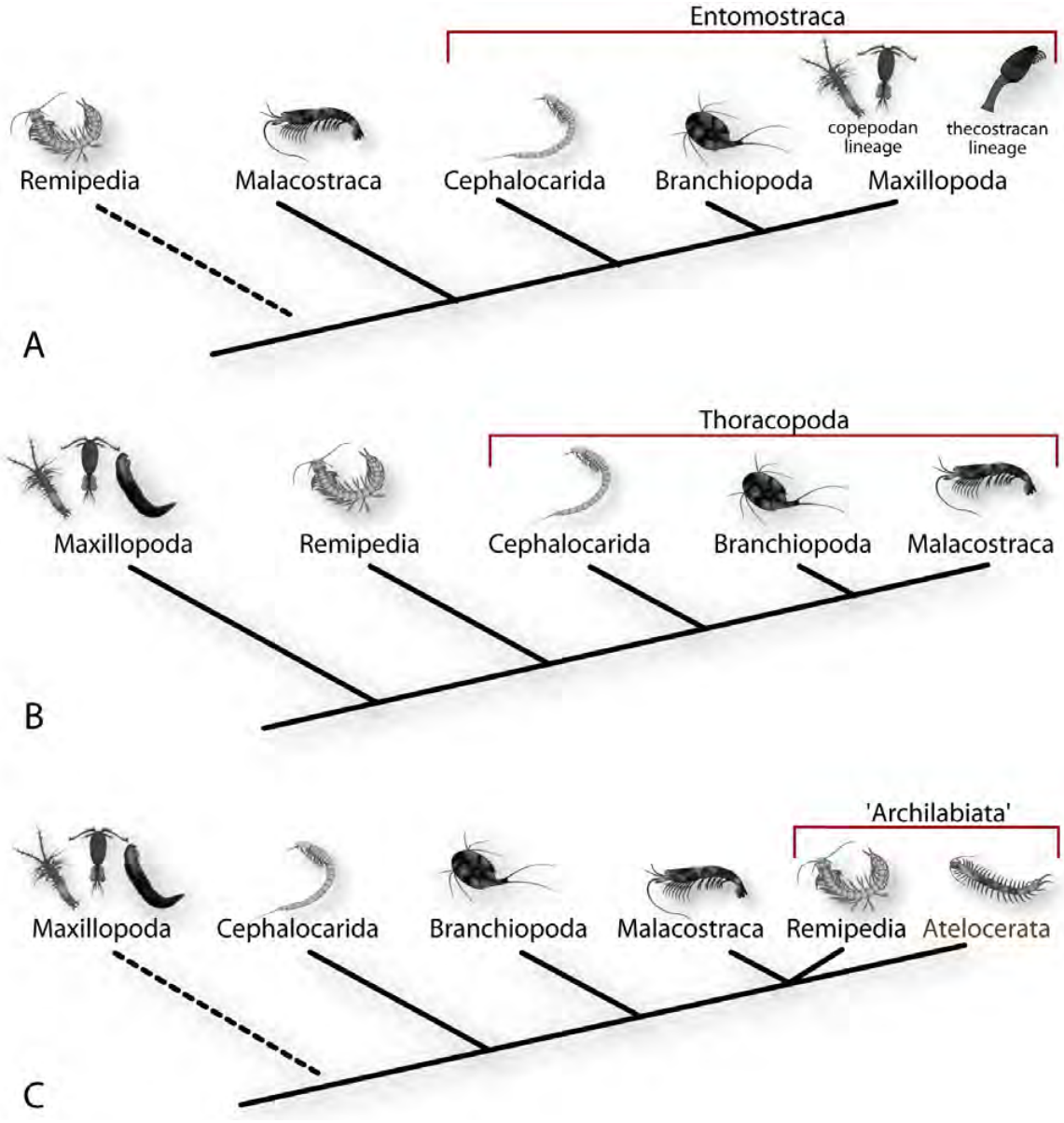


Figure 1.3: Cladograms representing the most commonly suggested, competing relationships of Crustacea.

A: Entomostraca-concept as postulated by WALOSSEK (1999). Remipedes are not within

crustaceans according to Waloszek. He describes a copepodan and thecostracan lineage of the Maxillopoda. Pentastomida are considered as Arthropoda *sensu latu* and not Crustacea.

B: Thoracopoda-concept after HESSLER (1992) based on the character of an existing epipodite.

C: Archilabiata-concept presented by FANENBRUCK (2003), see 1.1. At the moment molecular and neuroanatomical evidence favors a polyphyletic clade of Malacostraca, Remipedia, Atelocerata instead of 'Archilabiata'.

Unclear positions of taxa in phylogeny hypothesis A-C are represented by dotted lines.

1.3 Early concepts of arthropod phylogeny and major arthropod clades in a modern background

Several of the familiar higher-level groupings, such as Atelocerata (= Tracheata), Uniramia, and Mandibulata have their origin as far back as the 19th century. Since then a lot of hypotheses of rejection and evaluation of these taxa marked this area of phylogeny. In the following section main hypotheses around these taxa are described in more detail.

Pancrustacea (=Tetraconata): In contrast to the Tracheata, the hypothesis of a clade, consisting of Crustacea and Hexapoda was championed by PAULUS (1979) and confirmed by molecular data (FRIEDRICH & TAUTZ 1995). ZRZAVÝ & ŠTYS (1997) named this clade "Pancrustacea", a term used in several molecular studies (FRIEDRICH & TAUTZ 2001; GIRIBET ET AL. 2001; HWANG ET AL. 2001; REGIER & SHULTZ 2001; SHULTZ & REGIER 2000). DOHLE (2001) advocated the Tetraconata concept based on the structure of the ommatidia and postulated monophyletic Crustacea and Hexapoda as sister groups. Today both terms are mostly used synonymously. Neuroanatomical studies delivered possible synapomorphies for the Pancrustacea in adult nervous systems (HARZSCH 2006; HARZSCH ET AL. 2005; HARZSCH ET AL. 2006) and in developmental pathways of neuroblasts (UNGERER & SCHOLTZ 2008). FANENBRUCK ET AL. (2004) also favored a derivation of Hexapoda from within Crustacea based on neuroanatomical data. In recent molecular studies, Branchiopoda (REGIER ET AL. 2005) or Copepoda (MALLATT & GIRIBET 2006) emerged as sister group of Hexapoda. Also non - monophyletic Hexapoda interspersed within Crustacea (CARAPELLI ET AL. 2007; CARAPELLI ET AL. 2005; NARDI ET AL. 2003) have been proposed.

Tracheata (=Atelocerata, Antennata): In 1866 HAECKEL erected the taxon Tracheata, to which he assigned all arthropods with tracheal breathing, the Arachnida, Myriapoda, and Hexapoda. The Tracheata were redefined by POCOCK (1893A+B), who subsequently excluded the arachnids. POCOCK furthermore considered the Myriapoda "an unnatural assemblage of beings", composed of (diplopods + pauropods) and (chilopods + hexapods) as the two most closely related groups, and symphylans in an unassigned position ("a question for future discussion"). However, based on a detailed comparison of metameric structures, HEYMONS (1901) continued to support myriapods and hexapods as sister groups, and proposed to

unite them under the new name Atelocerata. Today, both concepts, Tracheata and Atelocerata, are usually used as synonyms. Interestingly, in the phylogenetic analysis of combined molecular and morphological evidence of WHEELER ET AL. (2004), a monophyletic Atelocerata is supported whether or not selected fossils are included in the analysis.

Mandibulata: Another concept of a major arthropod clade goes back to SNODGRASS (1935), who erected the Mandibulata (Crustacea, Myriapoda, Hexapoda) as a taxon encompassing Crustacea + Atelocerata (figure 1.4, A), groups that both share, in particular, the possession of distinctly shaped mandibles and two pairs of maxillae. The monophyly of Mandibulata is generally supported by several morphological (VACCARI ET AL. 2004; WHEELER ET AL. 2004; GIRIBET ET AL. 2005), neuroanatomical and molecular studies (BOORE ET AL. 1995; GIRIBET ET AL. 2001; KUSCHE ET AL. 2003). However, a clade Mandibulata is often not supported in molecular analyses (figure 1.4, B) or contradicted by the Myriochelata concept (figure 1.4, C).

Myriochelata (=Paradoxopoda): The Mandibulata hypothesis has recently come under fire from molecular phylogenetic analyses that instead unite Myriapoda and Chelicerata as a clade "Paradoxopoda" (HASSANIN 2006; HASSANIN ET AL. 2005; MALLATT ET AL. 2004; ROTA-STABELLI & TELFORD 2008) or "Myriochelata", which is a synonymous term (PISANI 2004; ROTA-STABELLI & TELFORD 2008), see figure 1.4 (C). It has been discussed that support for Paradoxopoda based on mitochondrial evidence is an artifact of out-group choice (ROTA-STABELLI AND TELFORD 2008). Analyses based on nuclear sequence data (BOURLAT ET AL. 2008; DUNN ET AL. 2008; REGIER ET AL. 2008) support either Mandibulata or Paradoxopoda.

Schizoramia: The "Schizoramia" hypothesis ("TCC" = Trilobita Chelicerata Crustacea concept) that groups the chelicerates and crustaceans based on morphological characters (CISNE 1974) contradicts the Mandibulata concept, see figure 1.4 (D). Paleontologists had favored this concept.

Uniramia: Some early hypotheses about the evolutionary relationships of arthropods included other segmented animals, such as Onychophora, as basal arthropods, from which modern, extant forms were believed to have been derived (e.g., SNODGRAS 1935; 1938). MANTON (1973) went a step further and proposed the taxon Uniramia to embrace hexapods, myriapods, and onychophorans, three groups characterized by segmented trunks, single-branch limbs, one pair of (first) antennae, and reduced post-oral mouthparts (figure 1.4, D). According to this concept, Crustacea was considered the closest relative of the Uniramia and the arthropods and euarthropods are considered to be polyphyletic. The Uniramia hypothesis is now generally considered obsolete (see WÄGELE 1993). However, neuroanatomical data (STRAUSFELD ET AL. 2006A+B) phylogenomic (MARLÉTAZ ET AL. 2008) and single gene analyses (BALLARD ET AL. 1992) place Onychophora as sister group to Chelicerata within Euarthropoda.

Relevance of morphology and fossils: Molecular evidence has become a crucial source of data but comparative morphology retains an important role in systematizing both extant and fossil arthropods. The study of WHEELER ET AL. (2004) is emblematic for the importance of morphology, especially in showing the power of fossils to influence relationships among extant taxa. This study showed that the inclusion of just a small number of fossil taxa can

significantly change the relationships of the major arthropod taxa (alternatively supporting Atelocerata or Pancrustacea) based on morphological or combined molecular and morphological evidence. Our current understanding of the phylogenetic position and evolution of extinct lineages is of course highly dependent on the assignment and interpretation of morphological data (COBBETT ET AL. 2007; VACCARI ET AL. 2004). Finally, excellent morphological work on fossils allowed unique insights into the composition of stem-lineages that support the extant crown groups of arthropods (e.g., WALOßEK 1993; WALOßEK & MÜLLER 1998A+B; MAAS & WALOSZEK 2001; EDGECOMBE 2004). Anyhow, analyses of some questioned fossils (like representatives of the lobopodians) cannot enlight unambiguously either, e.g the position of Onychophora and Tardigrada, which is still unclear; see EDGECOMBE (2009) and BUDD & TELFORD (2009).

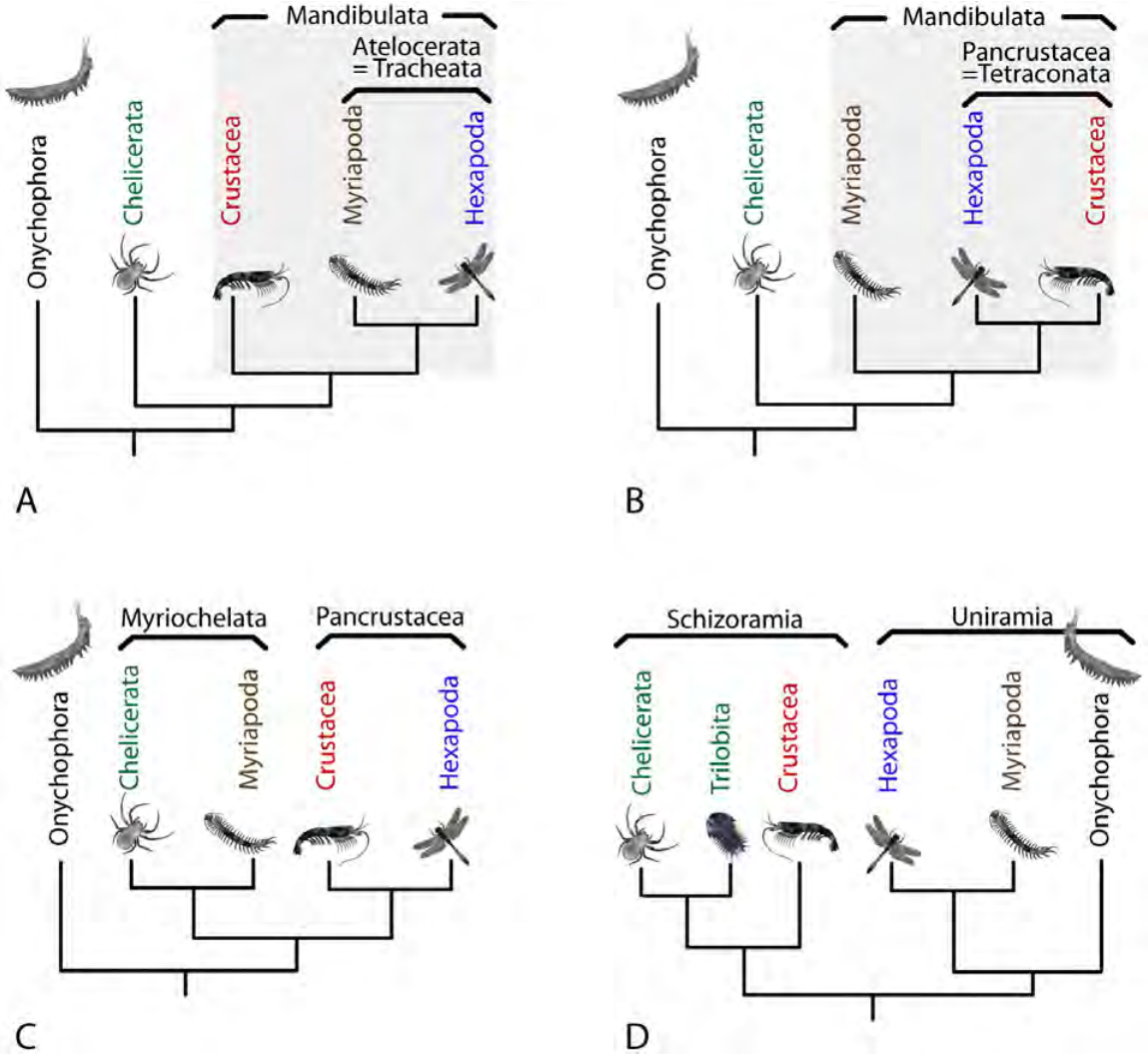


Figure 1.4: Conflicting hypothesis on the phylogeny within Arthropoda. Modified and complemented after RICHTER & WIRKNER (2004), FANENBRUCK (2003) and MÜLLER (2007).

1.4 Quintessence of recent arthropod studies

A large number of molecular phylogenetic analyses of major arthropod relationships (some also including morphological data) has been published, but despite some emerging consensus many unresolved issues remain (e.g., GIRIBET ET AL. 2004; GIRIBET ET AL. 1996; GIRIBET ET AL. 2001; GIRIBET ET AL. 2005; HWANG ET AL. 2001; GLENNER ET AL. 2006; REGIER ET AL. 2008; ZRZAVÝ ET AL. 1998A; FORTEY & THOMAS 1998; WHEELER 1998). As pointed out by REGIER ET AL. (2008), deep arthropod phylogeny shares many of the problems that plague deep metazoan phylogenetics. Original phylogenetic signal has saturated significantly over the hundreds of millions of years of independent evolution separating the major taxa, and as data density grows systematic errors become apparent. Thus results are very sensitive to choice of method and data treatment (see 1.5 methodological background). Recent studies (REGIER ET AL. 2008; ROTA-STABELLI & TELFORD 2008) provide clear illustrations of the difficulties involved. Studies also show some striking conflicts between mitochondrial (CARAPELLI ET AL. 2007; HASSANAIN 2006) and nuclear data (MALLATT ET AL. 2004, MALLATT & GIRIBET 2006; REGIER ET AL. 2005; 2008), for example with respect to the monophyly of Hexapoda and Crustacea.

Morphological studies have the obvious problem that authors interpret many characters or character transformations differently, thus conflicting evolutionary scenarios are created. An example might be the different morphological concepts for crustaceans (see 1.3, Tetraconata vs. Entomostraca vs. Schizoramia). Apart from these internal conflicts they provide in many cases a backbone or test case for molecular studies to detect artifacts of molecular phylogenies. An example might be the result in HASSANIN (2006) that *Vargula*, (an Ostracoda) groups as sister-group to Myriapoda + Chelicerata + remaining Hexapoda and Crustacea in HASSANIN (2006). Based on morphological data the author doubts this reconstructed topology.

The previously sketched problems can be enhanced by combined or total evidence studies. Molecular phylogenetic analyses in most cases still need improvement regarding data quality analyses and phylogenetic modeling. Often the molecular results do not represent the best results that are achievable with existing models. Combining this "half-baked" analyzed molecular data with morphological datasets can only enforce conflicts. It has additionally to be kept in mind in this context that mathematical models for the transformation of morphological character states do not exist like for molecular data. Consequently, the same recommendations made for future studies of metazoan phylogenetics can be made for higher-level arthropod phylogenetics (JENNER & LITTLEWOOD 2008), acknowledging that much still needs to be done.

A provisional consensus in this highly dynamic field (see figure 1.5) can nevertheless be drawn from the most recent comprehensive studies (GIRIBET ET AL. 2004, 2005; WHEELER ET

AL. 2004; BOURLAT ET AL. 2008; REGIER ET AL. 2005, 2008; TIMMERMANS ET AL. 2008; BUDD & TELFORD 2009).

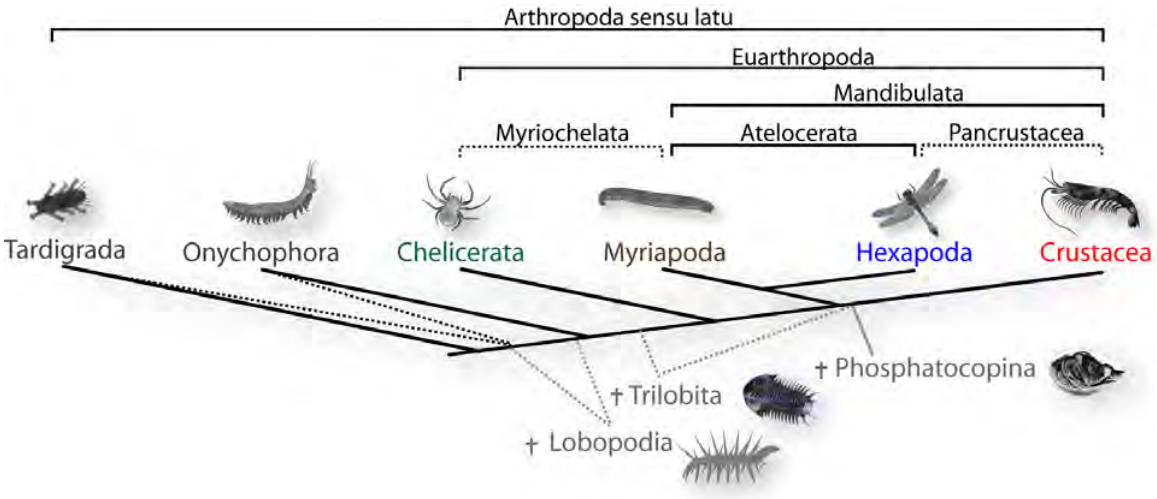


Figure 1.5: Summarized recent hypotheses of arthropod evolution. Contradicting hypotheses are represented by dotted lines, fossil taxa in grey. The Mandibulata concept is in concurrence to Myriochelata combining Chelicerata and Myriapoda to a clade. Note that if Pancrustacea is supposed, Mandibulata still exists with (Myriapoda (Hexapoda + Crustacea)). The stem lineage of Euarthropoda is supported by fossils but partly still discussed. Position of Phosphatocopina following MAAS AND WALOSZEK (2005). The Trilobita are generally placed as sistergroup to Chelicerata, but interpretation of new fossil evidence of ontogeny (larval states) suggests a closer relationship to Crustacea (HAUG 2009, PhD thesis).

Arthropoda is monophyletic and comprises at least four extant clades: Pycnogonida, Chelicerata, Pancrustacea (hexapods and crustaceans), and Myriapoda. The monophyly of Pycnogonida and Chelicerata is mainly accepted, whereas the monophyly of Pancrustacea is increasingly well supported on the basis of molecular evidence. In contrast, the monophyly of Hexapoda (NARDI ET AL. 2003; COOK ET AL. 2005) and Myriapoda (STRAUSFELD ET AL. 2006A) is less certain. Especially phylogenetic analyses based on mitochondrial sequences have questioned hexapod monophyly, suggesting that collembolans do not group with the remaining hexapods (NARDI ET AL. 2003; CARAPELLI ET AL. 2007). Nevertheless both hexapod and myriapod monophyly are generally supported in the most comprehensive analyses (REGIER ET AL. 2005; 2008), respectively evidence for a monophyletic Hexapoda is growing (DELSUC ET AL. 2003; TIMMERMANNS ET AL. 2008). Crustacea may be para- or even polyphyletic (SCHRAM & KOENEMANN 2004A+B; REGIER ET AL. 2008).

The position of Tardigrada and Onychophora to Euarthropoda in COLGAN ET AL. (2008) is sensitive to method of analysis, and remains also ambiguous in other molecular phylogenetic analyses (MALLATT & GIRIBET 2006; DUNN ET AL. 2008; PODSIADLOWSKI ET AL. 2008 for

onychophorans; PAPS ET AL. 2009 for tardigrades). The position of Onychophora as a sister group to Chelicerata in ROEDING ET AL. (2007) and MARLÉTAZ ET AL. (2008) may very well be influenced by the absence of Myriapoda in these analyses, and needs further testing. Recent publications using morphological, neuroanatomical and EST data (EDGECOMBE 2009; ZANTKE ET AL. 2008;) favor the Tardigrada as sister-group to Onychophora + Euarthropoda. In contrast other recent studies also support a position of tardigrades within nematodes (DUNN ET AL. 2008; HEJNOL ET AL. 2009). The fossil report regarding tardigrades and onychophorans gives no clear answer to solve their position. Lobopodia for example are discussed controversial in this respect. RAMSKÖLD & JUNYUAN (1998) conclude a closer relationship to Onychophorans and rank Tardigrada as first group of the Arthropoda *sensu latu*; which is in line with EDGECOMBE (2009).

1.5 Methodological background

Phylogenetic relationships of animals were classically considered based on morphological data. As stated in 1.2 and 1.3 (see figure 1.2) morphological analyses still leave questions open regarding the phylogeny of many internal relationships of arthropod and crustacean groups. Furthermore, conflicts are created by morphological results that contradict against each other. With the beginning era of molecular phylogenetics and the first phylogenies based on single molecular markers like 18S rRNA (AGUINALDO ET AL. 1997; SPEARS & ABELE 1998), the hope arose that these conflicts and open questions are reliably resolved by the molecular data (BRINKMANN & PHILIPPE 2008). However, the results of most single gene analyses partly strongly contradict each other. Subsequently, datasets of different genes were concatenated (KLUGE 1989) with the hope to reconstruct more reliable and robustly resolved trees compared to analyses based on single genes (e.g. SHULTZ & REGIER 2000). Anyhow, this technique increased the resolution only slightly in most cases as stated in BRINKMANN & PHILIPPE (2008). One promise attended by the new era of phylogenomics borne by the revolutionary progress in DNA sequencing methods (see section 2.2 for technical overview) was to boost the number of genes (or even genomes) implemented into analyses and to improve the robustness and reliability of reconstructed trees (BRINKMANN & PHILIPPE 2008). Several studies confirm that a stochastical or sampling error should vanish if the number of genes that are added to the analyses is as large as in phylogenomic data (ROKAS ET AL. 2003; MADSEN ET AL. 2001) and the taxon sampling is improved. However, the systematic error increases with more data, bringing the methodological aspect to the front.

In parallel to phylogenomic data accumulation we find a continuous progress in the development of phylogenetic models, the theory of DNA sequence evolution and application of models and software for maximum likelihood and Bayesian approaches. In particular, for often used single gene markers like rRNA genes sophisticated models were developed to incorporate existent background knowledge, e.g. on the secondary structure of these genes

(see section 1.5.2). In general, more and more of these findings prevail that both, sophisticated models and new methodological approaches are essential to draw a more realistic picture of metazoan phylogeny (PHILIPPE ET AL. 2005). Triggered by the phylogenomic data some rather old insights to phylogenetic reconstructions were revived. FELSENSTEIN (1988) noted already: "[…] molecular evolutionists who use methods for inferring phylogenies do not engage in much discussion of the properties of the methods they use since they focus on the difficult task of collecting the data […]". This quotation matches exactly the results of the recent phylogenomic studies (PHILIPPE ET AL. 2005; BAURAIN ET AL. 2006) which demonstrate that adding more and more data or taxa is not the way to solve the fundamental phylogenetic problems because the systematic errors will stay (PHILIPPE & TELFORD 2006).

1.5.1 The fundament for all molecular phylogenetic analyses – taxon choice and alignment reconstruction

A general remark to taxon choice – placing the fundament: The sampled taxa are the first stone in the fundament to infer a reliable phylogeny. One of the major problems in many molecular analyses is an unbalanced taxon sampling. On the one hand highly derived species are included and on the other hand the taxon set might be incomplete regarding all subgroups. This is in line with an argumentation promoted by AGUINALDO ET AL. (1997), BRINKMANN ET AL. (2005) and PHILIPPE ET AL. (2005), that a widespread taxon sampling can avoid Long Branch Artifacts (LBA). LBA is the phenomenon that was relative early described for Parsimony methods by FELSENSTEIN (1978). If two taxa have significantly longer branches compared to the other taxa it is very likely that the long branches will cluster together, in spite of no close phylogenetic relationships. The effect is increasing with sequence lengths and leads to inconsistency of the reconstruction method. WÄGELE AND MAYER (2007) coin this effect a class III LBA. It is basically caused by chance similarities or convergent positions in the sequences that outnumber apomorphic positions. Exclusion of long branch taxa is one possibility to avoid this effect. Another chance seems to be the inclusion of taxa that show no terminal long branches and to add these taxa for clades that partly show long branches.

The effort of the species collection for this thesis can be seen in the subsection collection work (chapter 2). The main reason was the intention to reduce from the outset a bias caused by long branches (as described above).

The crucial step of sequence alignment: The second stone in the fundament of phylogenetic reconstructions is the alignment. Multiple sequence alignments (MSA) are an essential prerequisite for alignment-based phylogenetic analyses, because they establish fundamental homology assessments of primary sequence characters (alignment positions of nucleotides or amino acids). Yet the multiple sequence alignment problem is NP-hard, which means, to find a solution is impossible with more than a few sequences (WALLACE ET AL. 2006). This is one reason why so many approaches have been developed (over 50 MSA programs) to approximate this problem. For recent reviews see NOTREDAME (2002) and

WALLACE ET AL. (2006). Commonly used methods are progressive and consistency based alignment procedures. In progressive alignments pairs of sequences are aligned first. A guide tree determines the order to align sequences or pairs of alignments; the most similar sequences are aligned first (WALLACE ET AL. 2005). A problem with this method is that alignment errors introduced in an early step cannot be corrected. Consistency based methods try to compensate this problem generating an alignment that is consistent with a set of pairwise alignments which is like a "library of alignment information" that guides the progressive alignment procedure (NOTREDAME ET AL. 2000; NOTREDAME 2002; WALLACE ET AL. 2006). In particular, if sequences are highly divergent, the introduction of gaps for example becomes more and more complicated and can currently not be fully governed by formal algorithms. The major problem is that finding the most accurate alignment parameters in progressive and consistency based alignment approaches is difficult due to the incomplete knowledge of the evolutionary history of sequences and/or heterogeneous processes along sequences (NUIN 2006). As a result problematic sequence alignments will contain sections of ambiguous positions with doubtful positional homology.

Alignment masking and processing: Recent studies underline that alignment errors can influence the correctness of tree reconstructions (DRESS ET AL. 2008; LÖYTYNOJA & GOLDMAN 2005; OGDEN & ROSENBERG 2006). To deal with this problem at the step of sequence alignment, different approaches and alignment software tools have been developed to assess the alignment quality. Despite major advances, alignment quality is still mostly dependent on arbitrary user-given parameters, e.g. gap costs (NOTREDAME 2002; MORRISON 2006). It has been shown that a selection of unambiguously aligned sections, or the exclusion of ambiguous positions (alignment masking: HARTMANN & VISION 2008), improves phylogenetic reconstructions in many cases (WÄGELE & MAYER 2007; DRESS ET AL. 2008; WONG ET AL. 2008; MISOF & MISOF 2009). However, an objective criterion of selecting unambiguous alignment sections or profiling multiple sequence alignments was still not available for a long time. Different automated heuristic profiling approaches to assess the quality of protein and nucleotide alignments have been developed to fill this gap. The recently developed software ALISCORE was used here for this task (chapter 2.3).

1.5.2 Single gene data - incorporating background knowledge to rRNA analyses

Biologically realistic modeling for rRNA genes: rRNA genes possess loop regions and stem regions. Loop regions show the same evolutionary pattern like standard DNA sequences, each nucleotide evolves independently. In stem regions nucleotides are paired forming secondary structures and thus the change of one nucleotide is covaring with the paired site due to the selection pressure to conserve the molecule's tertiary structure. It has been demonstrated that ignoring this correlated variance may mislead tree reconstructions they can be biased by an overemphasis of changes in paired sites (JOW ET AL. 2002; GALTIER 2004; MISOF ET AL. 2007). Evolutionary constraints on rRNA molecules are well known, for

example constraints resulting from secondary structure interactions as described above. The accuracy of rRNA comparative structure models (WOESE & FOX 1975; WUYTS ET AL. 2000; GUTELL ET AL. 2002) has been confirmed by crystallographic analyses (BAN ET AL. 2000; NOLLER ET AL 2005). Based on that background knowledge, rRNA sequences are an ideal test case to study the effect of biologically realistic substitution models on tree reconstructions. Recent studies of genome scale data revealed, that carefully chosen biologically realistic substitution models are of particular importance (LARTILLOT & PHILIPPE 2008; RODRIGUEZ-EZPELETA ET AL. 2007; PHILIPPE ET AL. 2000). However, the extent to which biological processes can/should be modeled in detail is still unclear. The analyses of rRNA sequences can still deliver new insights into this direction, because different aspects of the substitution processes can nicely be separated. In order to model covariation in stem regions of rRNA sequences, we estimated secondary structure interactions by applying a new approach implemented in RNAsalsa (STOCSITS ET AL. 2009) to avoid inadequate modeling of rRNA substitution processes in deep phylogenetic inference (e.g. MISOF ET AL. 2007; BROWN & LEMMON 2007). Essentially, this approach combines prior knowledge of conserved site interactions modeled in a canonical eukaryote secondary structure consensus model with the estimation of alternative and / or additional site interactions supported by the specific data.

Non-stationary substitution processes: Inhomogeneous base composition across taxa is a frequently observed phenomenon indicating non-stationary substitution processes (GALTIER & GOUY 1995; TARRIO ET AL. 2001; GOWRI-SHANKAR & RATTRAY 2007). Non-stationary processes if present clearly violate assumptions of stationary processes that are regularly assumed in phylogenetic analyses (BLANQUART & LARTILLOT 2006; GOWRI-SHANKAR & RATTRAY 2006; 2007). Thus, non-stationary processes were modeled combined with the application of mixed substitution models in a Bayesian approach using the *PHASE2.0* software package (GOWRI-SHANKAR & JOW 2006) to provide a better fit to our data than standard substitution models (TELFORD ET AL. 2005; GOWRI-SHANKAR & RATTAY 2007).

1.5.3 Phylogenomic data – a general overview

"Phylogenomics" is a newly coined term (EISEN 1998; O'BRIEN & STANYON 1999) that comprises several research fields in molecular biology and evolution (PHILIPPE ET AL. 2005A; BOUCK & VISION 2007) and can be summarized as an approach that is applied at a genome-scale level (combining genes in a large scale) to phylogenetic inference (JEFFROY ET AL. 2006). Recent and ongoing progress in DNA sequencing methods (MELDRUM 2000; HUDSON 2008; SCHUSTER 2008) deviating from the classical Sanger method of sequencing by di-desoxy chain termination (SANGER ET AL. 1977; SHENDURE ET AL. 2004) provide automated and faster capability to large-scale sequencing. Complete genome sequencing will be a standard technique also to infer molecular phylogeny (for a short "technical" overview see chapter 2.2.2) in a fast and cheap way.

Expressed Sequence Tags (ESTs) represent to date the largest part of phylogenomic data containing an increasingly large part of the transcriptome for many species (JONGENEEL

2000; RUDD 2003). The concept to produce single read sequences via cDNA from reverse transcribed cellular mRNA (the "transcriptome") was developed in the 1980s (PUTNEY ET AL. 1983) to detect previously unknown protein coding genes. In the course of the Human Genome Project (HGP) the sequencing of expressed genes came to the foreground as an effective sequencing method - omitting the "uninteresting", non coding DNA parts (BRENNER 1990). The term "EST" was first published by Mark Adams (ADAMS ET AL. 1991) describing this new method and its potential for automated large-scale sequencing.

The phylogenomic approach: The idea of the phylogenomic approach is to overcome conflicting results of single gene based analyses by a genome-scale approach (JEFFROY ET AL. 2006; PHILIPPE & TELFORD 2006). The impact of stochastic or sampling error is reduced by the phylogenomic approaches (PHILIPPE & TELFORD 2006; BRINKMANN & PHILIPPE 2008) if datasets contain more than 100 genes. Consequently, the statistical support and node values for reconstructed trees are greatly increased leading generally to a better resolution (DELSUC ET AL. 2005). However, with the dawn of the new phylogenomic era some important points and problematic issues are to address, some of them are rather old to phylogenetic analyses.

Systematic errors have a crucial influence, when using phylogenomic data (BRINKMANN & PHILIPPE 2008; PHILIPPE & TELFORD 2006). Systematic errors occur always in cases in which reconstruction methods will infer a wrong tree evoked by the data sensitivity of the reconstruction method. An increasing amount of data will also increase this inconsistency of the method. The best example of methodological inconsistency using maximum parsimony is the long branch attraction artifact (FELSENSTEIN 1978). If taxa evolve rather heterogeneous compared to each other maximum parsimony groups long branch taxa together independently from their evolutionary relation. Other probabilistic methods (maximum likelihood and Bayesian inference) are relatively robust against this phenomenon. Detection of systematic errors and the effect they may have on the resulting topology is one of the big challenges working with phylogenomic data (PHILIPPE ET AL. 2005; BRINKMANN & PHILIPPE 2008). Several strategies exist to evaluate systematic errors (reviewed in e.g. PHILIPPE ET AL. 2005; PHILIPPE & TELFORD 2006; BRINKMANN & PHILIPPE 2008) of which two important points are briefly addressed. A broad taxon sampling (see section above) to break down long branches and to identify different evolutionary rates between related taxa. Excluding the fast evolving ones could also be a solution (BRINKMANN & PHILIPPE 2008).

Models of protein evolution are the second important point to address systematic errors. They characterize the evolutionary substitution process in protein coding sequences, describing the probabilities of change from one amino acid to another (Thorne, 2000; Thorne and Goldman, 2003, Abascal 2005). Similar to nucleotide substitution models, protein models can be used for reconstructing phylogenetic trees with distance, maximum likelihood and Bayesian methods. One type of protein models relies on a matrix for single amino acid replacement (20x20), which estimates substitution rates from any amino acid to another empirically based on biological, chemical and physical properties of amino acids (QUANG ET AL. 2008). Properties of amino acids are for example positive or negative charge. Only proteins with e.g. similar charge are likely to be substituted. DAYHOFF ET AL. (1978)

introduced the first commonly used matrix, a 20-state time reversible homogeneous Markov model. Contrary to nucleotide models the extremely large number of parameters are not estimated in the phylogenetic reconstruction but calculated from data prior to analysis (ABASCAL 2005). Several of such matrices are used, like Dayhoff matrix (DAYHOFF ET AL. 1978), the JTT matrix (JONES ET AL. 1992), the mtREV matrix (ADACHI & HASEGAWA, 1996) or the WAG matrix (Whelan and Goldman, 2001). In contrast, the commonly used CAT model (LARTILLOT & PHILIPPE 2008) estimates for each site its substitution history from a number of classes. Each class is characterized by its own set of equilibrium frequencies. The model is implemented in a Markov Chain Monte Carlo Process to perform the complex estimation.

Missing data is often observed for phylogenomic and EST data, respectively. Its importance or influence on resulting topologies is not yet to evaluate. Some studies demonstrate that missing data influence an unstable placement of taxa with incomplete or missing protein sequences (WIENS 1998). Contrary, several studies show that taxa with missing data have a minimal effect if the total number of positions is large (WIENS 2003; WIENS 2005; PHILIPPE ET AL. 2004; WIENS & MOEN 2008) or can even improve the results by breaking up long branches (WIENS 2006). Thus, at the moment is seems that the total number of existent positions and the signal quality in this positions is important and the absent or missing data has less impact on tree reconstructions.

Orthology prediction: The distinction between orthologous and paralogous genes was first made by FITCH (1970) and is an essential fundament for analyzing both, single gene and phylogenomic data. Orthologs are genes that descend from a shared ancestral gene that was existent in the last shared ancestor. Their lineages are split by speciation events. They are more likely to keep their original functionality (KOONIN 2005). Paralogs are genes that descend from a shared ancestor gene. Their lineages were created by gene duplication (FITCH 2000) and always reflect the evolution of genes instead of species trees. But paralogs can also appear if the gene duplication event occurred after speciation. These "inparalogs" can form gene groups that are ortholog to genes in different species, and thus infer a species tree. The paralogs evoked by gene duplication before a speciation event are also called "outparalogs" and are never orthologs and infer exclusively gene trees (see O'BRIEN, ET AL. 2005). The differentiation between species and gene trees is eminently important for phylogenetic inference, which should be based on species trees (TATENO ET AL. 1982; RANNALA & YANG 2008).

Data supermatrix: After identification of the orthologous sequences among taxa the resulting data is the starting point for following analyses and has to be prepared for phylogenetic tree reconstruction. Working with EST data one should consider that many orthologous genes are missing or the identification may not have been successful, thus missing data is in most cases existing and affects the reconstruction. Generally, two possible methods were proposed to handle this data amount with partly missing data: the "supertree" and the "supermatrix" approach (SANDERSON 1998; DE QUEIROZ & GATESY 2006). A supertree represents the single, joint phylogeny estimation that results from separately analyzed, subdivided datasets (SANDERSON 1998; BININDA-EMONDS 2004). The alternative supermatrix

instead combines all characters to one matrix. All characters are considered simultaneously the basic idea was already described by KLUGE (1989).

1.6 Aims of the thesis

Many internal relationships of crustaceans are still unclear. Two specific questions of this thesis concern: [1] how the major lineages of crustaceans are related, and [2] what is the position of the clade Crustacea within arthropods. To address those questions single gene sequences [analysis A-B, see 1.7] but also phylogenomic data [analysis C, see 1.7] is analyzed with new methods to enlight molecular crustacean phylogeny.

Following crustacean phylogeny concepts should be tested in particular:

1) Entomostraca hypothesis
2) Thoracopoda hypothesis
3) Maxillopoda hypothesis

Within crustaceans in particular the positions of the single crustacean clades Malacostraca, Remipedia, Mystacocarida, Cephalocarida and Pentastomida should be tested.

Arthropod phylogeny hypotheses that were tested are:

1) Pancrustacea [Hexapoda + Crustacea] versus Atelocerata [Tracheata (=Antennata) + Crustacea (=Diantennata)]
2) Archilabiata hypothesis
3) Mandibulata versus Myriochelata
4) Arthropoda= Tardigrada + Onychophora + Euarthropoda

That the relationships of some arthropod taxa are highly interwoven with crustaceans is to recognize for arthropod phylogeny hypothesis 1) and 2) which test of course also internal relationships of crustaceans, e.g. if they might be paraphyletic in respect to Hexapoda or Tracheata.

Following methodologically questions should be addressed in this thesis:

1) Can an (automated) alignment evaluation and masking process improve the phylogenetic reconstruction?
2) Does the implementation of secondary structure based mixed models result in more reliable and improved phylogenetic trees based on rRNA data?
3) What is the impact of inhomogeneous base frequencies to phylogenetic reconstructions and can incorporation of time- heterogeneity improve the phylogenetic reconstruction compared to the standard time- homogeneous approaches?

4) Enlightens phylogenomic data based on a supermatrix approach the crustacean and arthropod phylogeny?

Resulting trees finally always upend in a bifurcation and separated clades, wether or not the separation is based on distinct signal. Handling only trees makes one sneaky falling into the trap to believe that there is a distinct signal. A general aim was to demonstrate that information content in the data can be evaluated by network reconstruction to gain a first impression of eventually conflicting signal. To assess the result of alignment processing this network reconstruction might be beneficial before and after alignment processing.

1.7 Short introduction and overview of analyses [A-C]

Analysis [A]: deals with the inner relationships of Crustacea. The potential phylogenetic signal is analyzed in a combined dataset of three commonly used loci (18S rRNA, 16S rRNA, COI) with the software MrBayes 3.0. This analysis goes beyond previous efforts by 1) an increased sampling of taxa within Crustacea, 2) the use of newly developed software to improve the quality of multiple sequence alignments (MAFFT and MUSCLE). The aim of analyses [A] and [B] was to improve the analyses by fitting biologically realistic mixed DNA/RNA substitution models to the rRNA data. Both analyses include for the first time representatives of Mystacocarida, Pentastomida, Branchiura, Remipedia, Cephalocarida and Leptostraca.

Analysis [B]: evolution of crustacean and higher arthropod taxa is addressed using a more extensive taxon sampling (compared to analysis [A]) of 148 arthropod taxa representing all major euarthropod clades. Thus onychophorans and tardigrades were chosen as out-group taxa. To focus on the relationships of high-ranking arthropod taxa the number of representatives of each taxon was trimmed to a concerted quantity of species representing the four major euarthropod groups. For a detailed comparison with analysis A/C 27 crustacean taxa (mostly two species per group) were finally included. This is a tribute to the inclusion of completely sequenced 18S and 28S rRNA genes, which do not yet exist for too many taxa. The more sophisticated phylogenetic reconstruction compared to analysis [A] implemented a time-heterogeneous substitution process to compare the results with the standard time- homogeneous approach.

Analysis [C]: The dataset of analysis [C] is the to date largest phylogenomic dataset for arthropods. The crustacean taxon sampling including species sequenced in this thesis comprises several species of the Maxillopoda and potential crustacean sister group taxa (e.g. Branchiopoda, Copepoda) to a hexapod clade, to test the putative para- or polyphyly of Crustacea. The main strength is the implemented reduction heuristics (see 2.6), which is to that extent absolutely novel in phylogenomic data.

2. Material and Methods

The fairest thing we can experience is the mysterious. It is the fundamental emotion, which stands at the cradle of true art and true science. He who know it not and can no longer wonder, no longer feel amazement, is as good as dead, a snuffed-out can. (ALBERT EINSTEIN)

2.1 Species choice, collection and fieldwork

„What is a scientist after all? It is a curious man looking through a keyhole, the keyhole of nature, trying to know what's going on.“ (JACQUES YVES COUSTEAU)

The taxon sampling was designed to cover all major crustacean groups, following MARTIN & DAVIS (2001), including representatives of early malacostracan lineages with at least one representative species for each group. Generally, it was tried to collect species (table S1) that do not differ too widely from the hypothetical morphological ground pattern of the represented group, whenever possible (LARTILLOT & PHILIPPE 2008; PHILIPPE ET AL. 2005). For all analyses presented in this thesis at least two less derived representatives of each major grouping were included (if possible) in the dataset. In published sequences of crustaceans, we find an overhead of malacostracan groups (e.g. Decapoda) while lower crustaceans are represented with only small sequence numbers.

To conduct multi-gene analyses with housekeeping genes and EST data for all crustaceans, the focus for fieldwork and collection laid respectively in achieving sequencing of missing lower crustaceans sequences complementary to public databases. Species collection consequently focused on those lower crustacean groups, which are underrepresented in public databases. For EST and 454 sequences the picture of existing public database entries is even worse. Sequencing projects of most lower crustacean groups are not yet conducted (STILLMAN ET AL. 2008, see supplementary table S14). Especially within lower crustaceans we find groups that are extremely hard to collect, which is one reason for the lack of sequences for these taxa.

In the field, crustaceans were collected with several methods. Vessel-based were dredges, Van Veen grab sampler and Multicorer used as sampling equipment (HIGGINS & THIEL 1988). By scuba diving I collected more specific, single target species operating bait-traps and hand-corer. Filtering and decanting samples from different zones from the seafloor up to the littoral was an important method either. Sampled specimens were always preserved in 94-98 % Ethanol for DNA sequencing and morphological determination. Specimens for EST or 454 sequencing projects were preserved in RNA later (Qiagen) or liquid nitrogen and additionally some voucher specimens in Ethanol. Samples were stored at -20°C, tissue for EST

sequencing at – 80°C. For collection localities and details see supplementary table S2. The collection effort to finish successfully this thesis is demonstrated by highlighting some of the collecting methods (figure 2.1) and the more detailed description of the collection of two important key taxa, the Pentastomida and Remipedia.

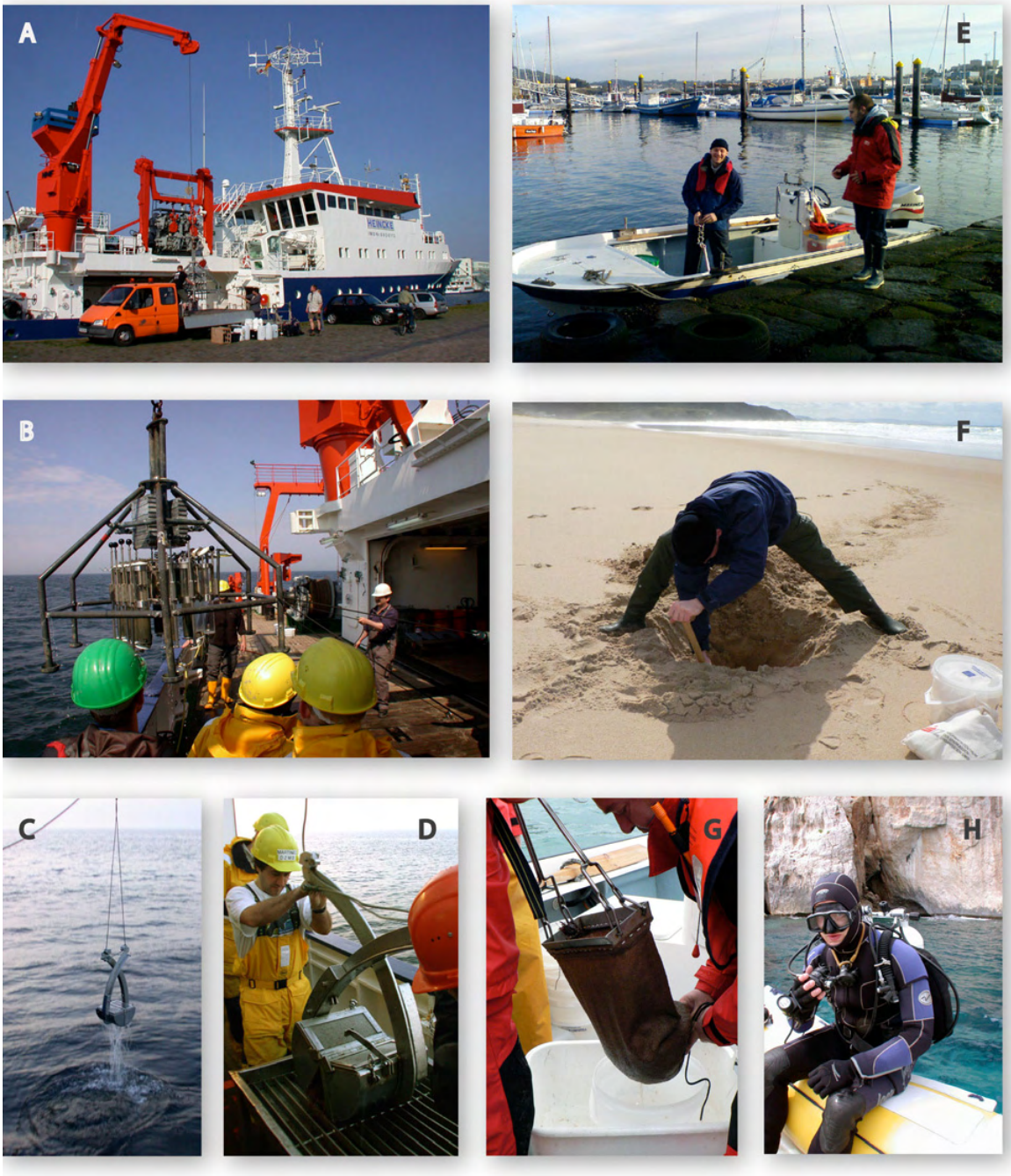


Figure 2.1: Some impressions of the field and species collection work. **A** shows the FS Heincke (Alfred Wegener Institut, Bremerhaven), a German research vessel before joining an offshore trip of several days on the North Sea. **B**: the Multicorer is lifted on deck. **C – D:** the Van Veen grab sampler coming up (**C**) and its opening to sort the samples (**D**).

E: Preparing the equipment for a dredging trip on the Ria de Ferrol in Galicia, Spain. **G:** One of the hand-dredges that were used. **F:** Digging a hole for filtering interstitial water to collect the tiny Mystacocarida. **H:** Last checks before descending with double tanks into the depth of the Mediterranean Sea (Sardinia) to lay baited traps for Ostracoda and Leptostraca and to take samples of Cephalocarida (the latter issue was unfortunately not successful at all.

To find parasitic tongue worms (Pentastomida) over 120 reptile host specimens were screened. Squamates are typical end-hosts of the genus *Raillietiella* (BOSCH 1987) and finally five pentastomids of this genus were identified (determined by H. BOSCH) in two *Hemidactylus frenatus* (house gecko) hosts. The specimens parasitized the trachea system of *Hemidactylus;* the lobes of the lung are the typical location (figure 2.2) of pentastomid larvae IV (the last stage) and adults in the reptile organism. For further details see BOSCH (1986; 1987).

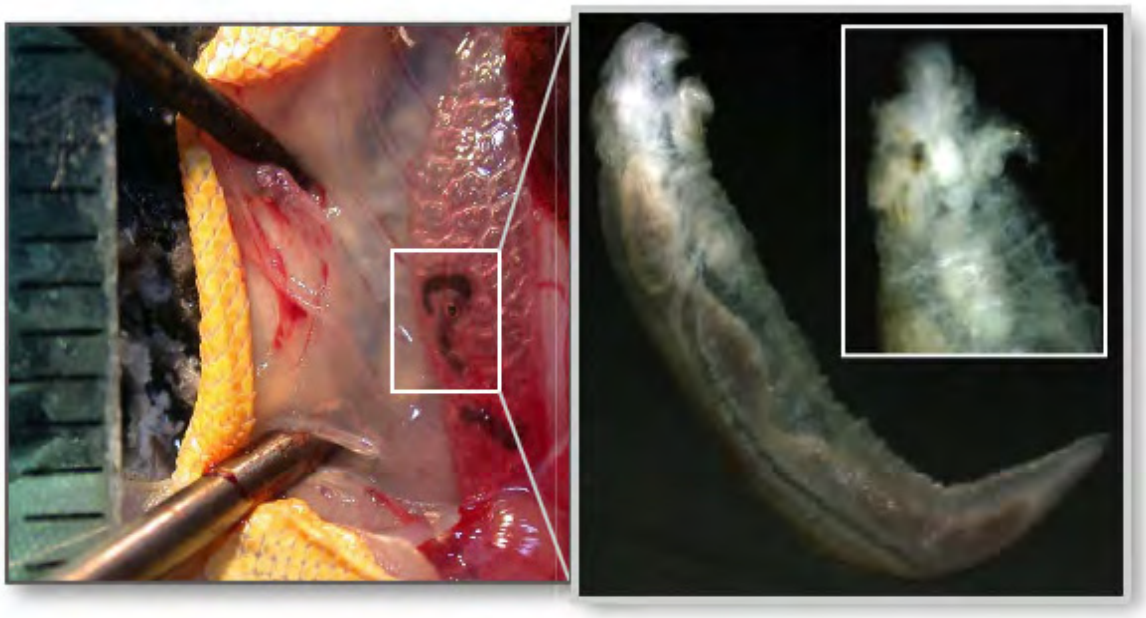


Figure 2.2: Pentastomid parasitizing *Hemidactylus* host. The left picture shows *Raillietiella* parasitizing in the right lung lobe of *Hemidactylus frenatus* (smaller white frame). The black scale bars on the left edge are in millimeters. On the right a specimen is pictured under the microscope in total and with details of the typical hooks (smaller white frame).

Remipedia are relatively late discovered (YAGER 1981), enigmatic crustaceans of subterranean anchialine environments. A center of distribution of species and high specimen abundance for species like *Speleonectes tulumensis* is the region of the Caribbean Sea and the Gulf of Mexico (KOENEMANN ET AL. 2003). Yet by characterizing their habitats it becomes clear that the collection of remipedes is an adventure. They live restricted to anchialine (HOLTHUIS 1973) cave systems that feature an underground connection with salt water to the sea, while the entrance and surface part of the water column is composed of freshwater (ILIFFE 1992A+B; ILIFFE & SARBU 1990). Fresh and saltwater form generally two layers of water separated by a cline-area, the halocline, in which both layers mix (figure 2.3). The vertical spreading and dimension of this interface layer is conditioned to the topology in the cave and varies also dependent on characteristics of existing currents. The remipedes are observed at the present only below the halocline (KOENEMANN ET AL. 2007) making the collection dangerous and only possible by cave diving. After being cave diving trained, it was possible to collect specimens of this mysterious group in the years 2007 and 2008 together with T. ILIFFE in Mexican cave systems (figure 2.4) on the Yucatan Peninsula.

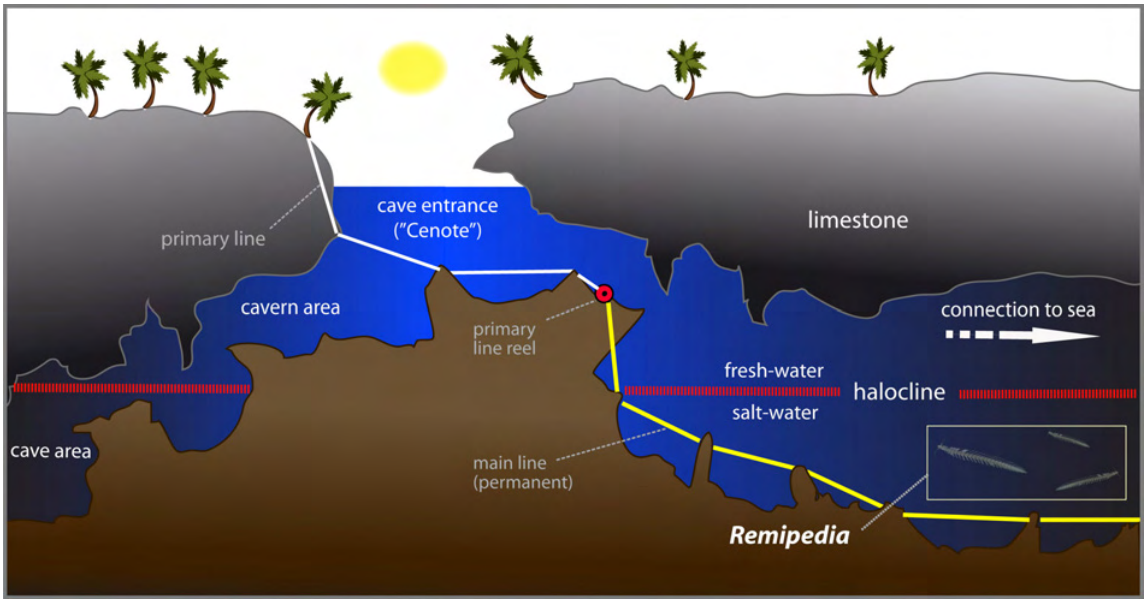


Figure 2.3: Longitudinal section of an anchialine cave system. The thicker, dashed red line represents the "halocline" in the cave, the interface layer between salt and freshwater. Remipedia occur only below the halocline. The difference between cave diving to normal scuba diving is the overhead environment. In case of an emergency resurfacing vertically is not possible. In addition to the more extensive equipment and training one main difference is also the complex use of "guide lines" that are permanently or temporarily laid in the cave. Always, a primary line is connected from the open water into the cave and connected to the main cave line. This is crucial for surviving in case of "silt outs", incidents evoked by disturbed sediment that create a sudden, zero sight situation.

Figure 2.4: Cenote Crustacea on the Eastern Yucatan, Mexico. At the left the entrance of the anchialine cave system (called "Cenote" in Mexican) is pictured, on the right the author is diving within the cave system with standard double tanks. The diving equipment matched the NACD (National Association for Cave Diving) and IANTD (International Association for Nitrox and Technical Diving) standards for cave diving.

2.2 Laboratory methods

"I am among those who think that science has great beauty. A scientist in his laboratory is not only a technician: he is also a child placed before natural phenomena which impress him like a fairy tale". (MARIE CURIE)

The molecular laboratory methods used in the framework of this thesis and their results were the base for several separate analyses. Three complex analyses are included, two using standard gene sequencing [A+B], the third [C] is based on EST sequencing. In this section are described [1] the combined molecular methods conducted in the molecular laboratory to achieve the sequence amplification and [2] in general the different sequencing methods used to conduct DNA- and EST analyses.

DNA extraction from complete specimens or tissue samples (e.g. from *Squilla mantis*, Stomatopoda) followed a standard protocol. The column DNA extraction kits DNeasy Blood & Tissue Kit (QUIAGEN) and NucleoSpin Tissue Kit (MACHERY-NAGEL) were used following the manuals. Only single specimens were macerated, samples were not pooled. Manufacturer protocols were slightly modified incubating the samples over night and adding 8 µl RNAse [10 mg/ml] after lysis. For small (*Derocheilocaris typica*, Mystacocarida) or very rare (*Lightiella incisa*, Cephalocarida) crustacean specimens the extracted DNA was amplified with the Illustra GenomiPhi V2 Amplification Kit (GE HEALTHCARE).

PCR was conducted for all nuclear and mitochondrial gene fragments using published and modified primers (supplementary table S3) that were ordered from METABION. The amplification and sequencing of complete 18S and 28S rRNA genes was a main focus, requiring highly complex primer combinations and PCR settings (see supplementary table S5, supplementary figure S1). The amplification of the complete 28S rRNA gene was performed using nine overlapping fragments (supplementary figure S1) starting approximately in the middle of the rRNA 5.8S and ending in the final part of the D12 of 28S rRNA. The 18S rRNA was completely amplified in one PCR product and sequenced using four primer combinations (supplementary figure S1). Of the mitochondrial genes 16S rRNA and COI only standard fragments were amplified using different primer combinations (supplementary table S3). PCR products were purified with the NucleoSpin ExtractionII (MACHERY-NAGEL) and QIAquick PCR purification kit (QUIAGEN). In case of multiple bands fragments with the expected size were cut from 1.5% agarose gel and purified according to the manufacturer protocol.

Cycle Sequencing reactions and electrophoreses were performed on different thermocyclers from ABI (GenAmp 2300, 2320 & 9600) and BIOMETRA (T-gradient) and on BECKMAN COULTER 8000 & 8800 capillary array sequencers. Some problematic PCR products were sent to MACROGEN (Inc.) in Korea.

Please refer to the supplement for more detailed information on chemicals (supplementary table S4), PCR profiles (supplementary table S5) and primer combinations (supplementary

figure S1). The in molecular sense (sequence amplification and sequencing performance) extremely heterogeneous crustaceans made an intensive testing for each gene essential to establish the settings for laboratory work. For an overview of sequenced genes for all taxa see supplementary table S6.

Single gene sequencing*:* Over three decades have passed since DNA sequencing based on electrophoretic methods has been established (SHENDURE ET AL. 2004). Since that an enormous development of sequencing technologies regarding automation (MELDRUM 2000), parallelization and cost reducing refinements occurred, mostly based on the principle of the Sanger sequencing method (SANGER ET AL. 1977). Capillary sequencers that sequence micro-plates of at least 96 wells in few hours are now standard for the majority of laboratories. Especially for standard DNA sequencing a new trend is that samples are shipped to sequencing companies, like MACROGEN or AGOWA, instead of doing "home-sequencing". While finishing the laboratory work for this thesis few last samples were also sent to MACROGEN (KOREA). Anyhow, most samples were sequenced on capillary sequencers (BECKMAN COULTER 8000 & 8800) in the laboratory of the ZFMK, Bonn.

EST sequencing: EST sequencing was one of the first high-throughput or large-scale technologies (ADAMS ET AL. 1991; BOGUSKI 1995; GERHOLD & CASKEY 1996) established in the framework of the HGP (Human Genome Project). "EST" is an abbreviation for Expressed Sequence Tags, referring to the mRNA fragments in the cell that are "tagged" by poly-A tails. Fishing those tagged fragments, only coded and expressed parts (Exons) of the DNA are sequenced (figure 2.5). This transcriptome sequencing via ESTs became today a standard method in molecular phylogeny (BOUCK & VISION 2007; HUGHES ET AL. 2006; SANDERSON & MCMAHON 2007) and molecular ecology (BOUCK & VISION 2007), see also methodological background in 1.5. Three crustaceans were sequenced in EST projects to find new marker genes and to conduct phylogenomic analyses to address questions of crustacean phylogeny and their position within arthropods. The EST project for the Remipedia (*Speleonectes tulumensis*) was not successfully finished; the only "by-product" is presented in 2.7. It is to state here, that the specimens for EST sequencing were only preserved and isolated in the field and/or laboratory in Bonn. The remaining laboratory procedure including rRNA amplification and final sequencing (figure 2.4) was conducted at the Max Planck Institute for Molecular Genetics (MPI Berlin). The described procedure [1-7] relies on the dideoxy, stop-nucleotide Sanger method.

The final EST sequences are processed and analyzed in data pipelines to remove vector sequences etc. and are finally assembled to EST contigs (fig. 2.8). The EST contigs are then the base for phylogenetic analyses (fig. 2.9).

EST – next generation sequencing: The tremendous pace for the faster, more efficient and cheaper high-throughput sequencing method can be recognized by the fact that EST sequencing based on cDNA clones is by now old fashioned. New high throughput sequencing technologies based on alternative sequencing methods were developed recently (HUDSON 2008). They encompass the limitations of the Sanger based methods used in the older EST technology, namely comparatively low throughput, long runtimes and high costs (RONAGHI

2001). For Remipedia, *Sarsinebalia* (Leptostraca) and Ostracoda we used consequently during the last period of this thesis one of the next generation, non-Sanger-based methods (HUDSON 2008; SCHUSTER 2008), the pyrosequencing technology (RONAGHI ET AL 1996; RONAGHI ET AL. 1998) from ROCHE, as implemented in the FLX 454 machine. The aim to include these species in the phylogenomic data failed unfortunately, the 454 sequencing was delayed for several, technically reasons.

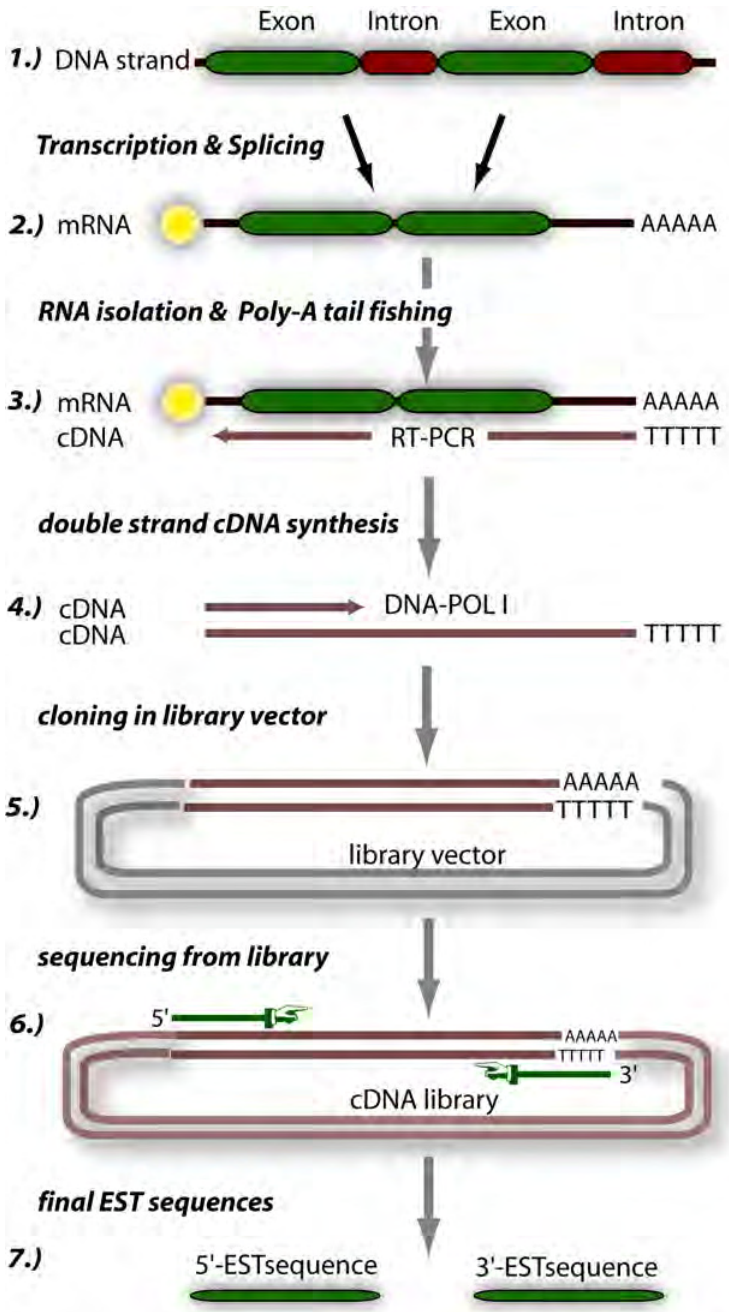


Figure 2.5: EST cloning and sequencing procedure, modified after BOUCK & VISION (2007). Non-coding DNA regions (exons) of a DNA strand [1] are spliced and transcribed

from DNA to finally mRNA [2]. The mRNA is characterized by a poly-A tail and a 5'-end cap structure with a modified guanine nucleotide (yellow circle). After isolating total RNA the poly-A tail is targeted to perform a Reverse Transcriptase PCR and to synthesize a complementary cDNA strand from the mRNA template [3].
Then follows digestion of the mRNA adding RNase H and afterwards DNA polymerase 1 synthesizes a second strand [4], resulting in a double stranded cDNA. This double stranded cDNA is inserted into cloning vectors [5]. The vector subsequently is inserted into competent cells and [6] a cDNA library is the final result. EST sequences can be generated of this library sequencing in 5' or 3' end direction. [7].

2.3 Data analyses methods prior to phylogenetic tree reconstruction

„Whoever, in the pursuit of science, seeks after immediate practical utility, may generally rest assured that he will seek in vain.“ (HERMANN VON HELMHOLTZ)

After finishing the molecular work including sequencing follow the procedures of sequence processing, data quality assessment and multiple sequence alignment reconstruction and evaluation. The resulting alignments are the basis for the phylogenetic tree reconstruction.
A methodological emphasis in this thesis was the use of new software and the design of process flows (or "pipelines") to assess data quality and to improve the crucial step of alignment reconstruction (figure 2.6).

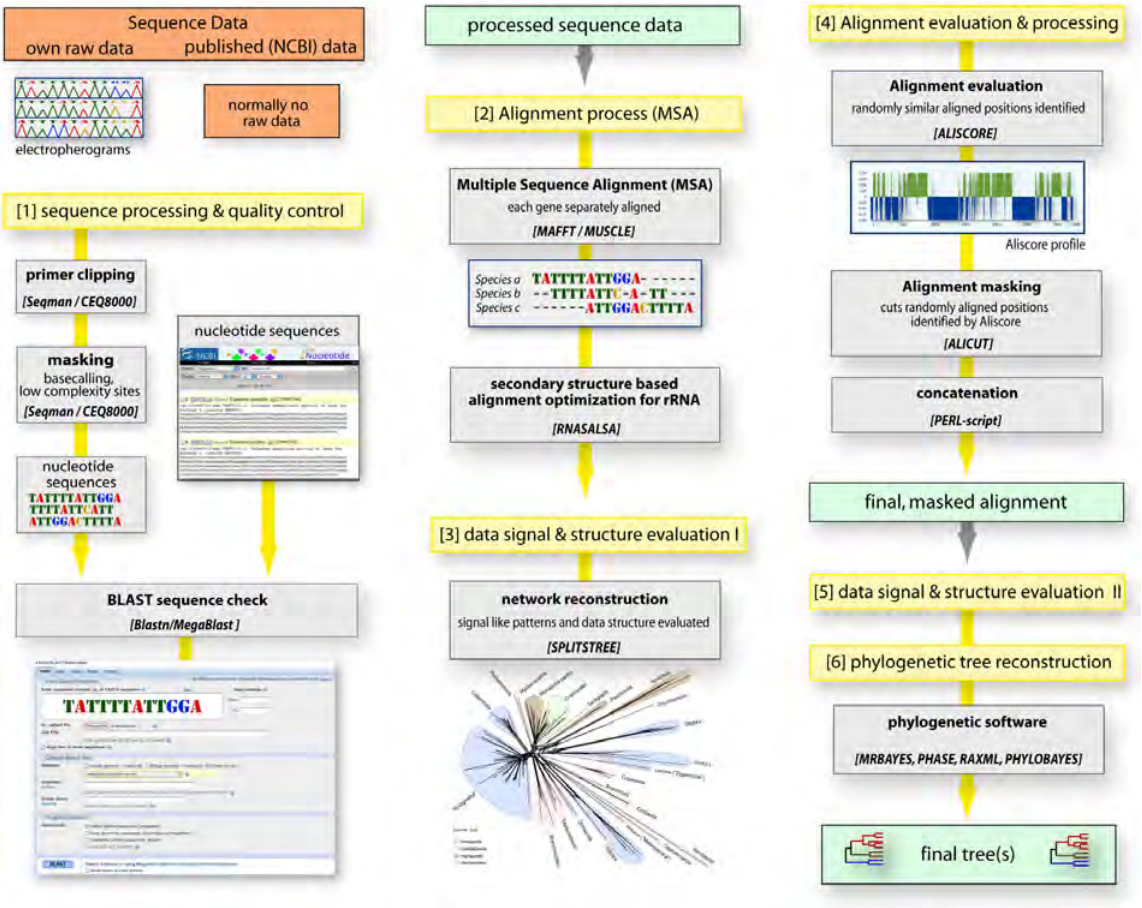


Figure 2.6: Phylogenetic analyses process prior to tree reconstruction. Sequence data (orange) is processed and quality controlled [1]. Electropherograms of raw sequences are quality checked and subsequently controlled own and published nucleotide sequences are blasted to identify eventual contamination. It is problematic that normally electropherograms selected from public databases are not available for published

sequences. Therefore sequence errors cannot be discovered in these data. Processed sequence data is prealigned [2] applying multiple sequence alignment programs. In case of rRNA genes a secondary structure-based alignment optimization follows. To gain a first impression of the information in the data, its structure is evaluated by phylogenetic network reconstruction. It follows the final alignment evaluation & processing [4]. For each gene ALISCORE (MISOF & MISOF 2009) is performed to identify randomly similar aligned positions and ALICUT excludes those positions found by ALISCORE (=masking process). Single, masked alignments are concatenated by a PERL-script to the final alignment. For most analyses it is useful to compare data structure before and after the alignment process in a network reconstruction [5]. After this the last step is the phylogenetic tree reconstruction [6].

Data analyses were conducted in close cooperation with several co-workers of the molecular laboratory at the ZFMK in Bonn (namely P. KÜCK, R. STOCSITS and H. LETSCH) and at the University of Hamburg (B. MISOF), who programmed the software. For this reason these newly developed software tools could be used and further improved by direct collaboration and discussing the results of the analyses (table 2.1).

Table 2.1: Main analyses included. Used marker genes and main focus is shown.

Analyses	Marker genes	Main focus in analysis
[A]	16S, 18S, COI	Standard vs. secondary structure guided alignments
[B]	18S, 28S	Time-homo vs. heterogeneous processes, secondary structure
[C]	EST sequences	Orthologous gene selection, relative information content of genes

Because this thesis was conducted in the framework priority program "deep metazoan phylogeny" of the DFG (Deutsche Forschungsgemeinschaft) a close cooperation existed with collaborators of the other three arthropod groups involved in this program. This was in general the case with K. MEUSEMANN.
The applied methods and programs to reconstruct trees for this thesis will be explained in the specific analyses sections. It is to mention that the EST analysis procedure followed in general the same principles as shown in figure 2.6 but differed slightly according to the differences between phylogenomic and single gene data. A detailed flowchart for the EST analysis is given in chapter 2.6. For that phylogenomic analysis some bioinformatics and computational issues had to be outsourced to the bioinformatics group (V. HÄSELER, Vienna, Austria) of the priority program, but see also the flowchart in chapter 2.6

2.3.1 Sequence processing and quality control

All resulting sequence electropherograms were analyzed and assembled using the software programs SeqMan (DNASTAR, Lasergene), CEQ 8000 (BECKMAN COULTER) or Bioedit 7.0 (HALL 1999). Unfortunately most published sequences are not linked to their trace files and consequently the quality of electropherograms cannot be determined. All final sequences and composed fragments were blasted in NCBI using BLASTN, MEGABLAST and BLAST2SEQUENCES to exclude contaminations. This is the terminal but eminent important (and very often ignored) step to finish the laboratory work. Ambiguous own or published sequences were always excluded from analyses.

2.3.2 Multiple sequence alignment

Sequence pre-alignments were performed for each gene separately with the commonly applied alignment programs MUSCLE (EDGAR 2004A; 2004B) and MAFFT (KATOH ET AL. 2002). For a comparison of MUSCLE and MAFFT alignments see (2.4). Tests of MAFFT have indicated that its LINSI-algorithm is more reliable for rRNA genes. These often inhabit expansion segments and ambiguous regions with variable length polymorphisms, which require a different estimation and judging for introducing gaps (KATOH & TOH 2008). Generally several different MSA programs were tested in parallel for this study. In addition to the above cited software for example the new version of CLUSTALX (LARKIN ET AL. 2007; THOMPSON ET AL. 1997; THOMSON ET AL. 1994) and T-COFFEE (NOTREDAME ET AL. 2000) were applied, but MUSCLE and MAFFT outperformed these and other programs regarding time and efficiency. Finally, the standard settings were used for all alignment programs.

2.3.3 Alignment optimization based on secondary structure information

The first step in alignment algorithms relies on identification of similar sequence regions which are subsequently arranged to sets of strings with maximized character identity in alignment positions, underlying homology hypotheses, see section 1.4.

The software RNAsalsa (STOCSITS ET AL. 2009) is a new approach to align structural rRNA sequences based on existing knowledge about structure patterns, using constraint directed thermodynamic folding algorithms and comparative evidence methods. This makes alignment reconstruction more objective. For each molecule in addition to sequence similarity a second trait, the structure is considered. RNAsalsa automatically and simultaneously generates both individual secondary structure predictions within a set of homologous RNA genes and a consensus structure for the dataset. Successively sequence and structure information is taken into account as part of the alignment's scoring function. Thus, functional properties of the investigated molecule are incorporated to corroborate homology hypotheses for individual sequence positions. The program employs a progressive multiple alignment method, which includes dynamic programming and affine gap penalties. Inferred site covariation patterns are used then to guide the application of mixed nucleotide / doublet substitution models in subsequent phylogenetic analyses. RNAsalsa needs a

prealignment as input. For a description of the exact algorithm and parameters of RNAsalsa, see STOCSITS ET AL. (2009), manual and software download see the homepage at: http: //rnasalsa.zfmk.de.

Secondary structure constraints for analyses [A] are based on the 16S (L20934) and 18S (78065) sequences of *Anopheles gambiae and albimanus.* For analysis [B] the 28SS+5.8S (U53879) and 18S (V01335) of *Saccharomyces cerevisae* were used. Corresponding secondary structures for the sequences were extracted from the European Ribosomal Database (DE RIJK ET AL. 2000; VAN DE PEER ET AL. 2000; WUYTS ET AL. 2000; WUYTS ET AL. 2004; WUYTS ET AL. 2002). Structure strings were converted into dot-bracket-format using Perl-scripts. Folding interactions between 28S and 5.8S (GILLESPIE 2005; GILLESPIE ET AL. 2006; MICHOT ET AL. 1983) required the inclusion of the 5.8S gene in the constraint to avoid artificial stems. Alignment sections presumably involved in the formation of pseudoknots were locked from folding to avoid artifacts. Pseudoknots in *Saccharomyces cerevisae* are known (WUYTS ET AL. 2000) for the 18S (stem 1 and stem 20, V4-region: stem E23\9, E23\10, E23\11 and E23\13) while they are lacking in the 28S secondary structure. Prealignments and constraints served as input, RNAsalsa ran with default settings.

2.3.4 Evaluating structure and signal by network reconstruction

Phylogenetic networks (HUSON & BRYANT 2006) were reconstructed to evaluate the general structure, potential conflicts and signal-like patterns in the alignments. Without constraining the results of a phylogenetic analysis in form of a bifurcated tree, these phylogenetic networks can be used to visualize the presence of conflicting signals in the data (HUSON ET AL. 2005). Conflicts are indicated by non-parallel edges that represent conflicting splits between taxa, and show the relative support for splits in the data by the length of parallel edges supporting a certain split (as an indicator for the weight of the split, analogous to branch lengths in a tree). For a detailed description of phylogenetic networks see (HUSON & BRYANT 2006; WÄGELE & MAYER 2007). With the software Splitstree 4.10 (HUSON 1998; HUSON & BRYANT 2006) was the neighbor-joining algorithm applied for network reconstruction in analyses [A and B] and additionally the LogDet transformation in analyses [B] to analyze the alignment of the complete 18S & 28S rRNA genes. LogDet is a distance transformation that corrects for biases in base composition (PENNY ET AL. 1994; STEEL ET AL. 2000).

2.3.5 Alignment evaluation and processing

Alignments were assessed with the software ALISCORE (MISOF & MISOF 2009) to identify ambiguous or randomly similar aligned sections. ALISCORE uses for this purpose a parametric approach, relying on defined models of sequence evolution. This results in a more objective and reliable procedure to infer topologies but not necessarily better performance regarding resulting topologies. In contrast, GBLOCKS (CASTRESANA 2000) the currently most frequently used software does not make explicit use of models of sequence evolution, which makes its usage in a specific way "subjective".

ALISCORE generates profiles of randomness using a sliding window approach. Sequence positions within this window are assumed to have random-like nucleotide patterns when the observed score does not exceed 95% of scores of random sequences of similar window size and character composition generated by a Monte Carlo resampling process. ALISCORE generates a listfile of all putative randomly similar sections. No distinction is made between random similarity caused by mutational saturation and alignment ambiguity. The default settings were used, the window size was (w=6), gaps were treated as ambiguities (- N option) and the maximum number of possible random pairwise comparisons (- r option) was analyzed.

The alignment masking process was conducted with the program ALICUT (KÜCK, http://utilities.zfmk.de). This perl-script masks the alignment by excluding the positions identified in an ALISCORE analysis to be randomly similar.

The consensus secondary structure of rRNA genes given in RNAsalsa was included into the alignment in analyses [A and B]. Consequently, both the aligned sequences and the consensus sequence are masked. In this way, the user can consider secondary structure information for phylogenetic analysis, for example, by implementing mixed models for RNA molecules. By default, ALICUT excludes also stem positions if identified as "randomly similar aligned" and converts the corresponding stem nucleotide into a dot ignoring covariation. However, it is plausible that evolution of stem positions is constrained by secondary structure and covariation patterns. Therefore, the -s function in ALICUT was used to keep all stem positions in the alignment.

2.4 Analyses [A]: Can 16S, 18S and COI marker genes improve inference of crustacean phylogeny? Comparing "usual" standard vs. secondary structure based approaches

"Science... never solves a problem without creating ten more" (GEORGE BERNHARD SHAW)

2.4.1 Objectives

To evaluate alignment methods was one central objective for this analysis, in particular for ribosomal RNA genes.

Previously published and new sequences were included for three marker gene loci: 18S rRNA, 16S rRNA, and cytochrome *c* oxidase I (COI). However, the intention to include as many representatives of all major arthropod and respectively crustacean groups as possible entailed a trade-off regarding the choice of genetic markers. The genes of the preferred choice were not available for all of the selected taxa. Therefore, incomplete gene sequences and even missing markers were tolerated for some taxa, see supplementary table S7.

2.4.2 Taxon sampling

The selected published and new sequences (supplementary table S7) represent species of all major extant groups of crustaceans, insects, myriapods, and chelicerates to (1) evaluate sister group relationships within the Arthropoda, and (2) appraise which groups constitute possible higher monophyla such as the Pancrustacea.

38 new crustacean sequences were contributed: 11 for the 16S rRNA, 16 for the 18S rRNA and 11 for the COI, respectively. The three sequences of *Pleomotra apletocheles* were kindly provided by the group KOENEMANN (TiHo Hannover). The taxon sample includes 88 terminal taxa representing all major extant groups of the Crustacea (57 taxa), Hexapoda (13 taxa), Myriapoda (5 taxa), Chelicerata (11 taxa), and two out-group phyla, the Onychophora and Tardigrada. In view of recent suggestions based on neuroanatomical (HOMBERG 2008) and phylogenomic evidence (DUNN ET AL. 2008; ROEDING ET AL. 2007) that onychophorans may either be a sistergroup to euarthropods or positioned within euarthropods, only tardigrades were designated as out-group. This allows to test the phylogenetic position of the onychophorans.

2.4.3 Analysis design

In order to maximize data density per taxon, composite (chimerical) higher-level terminal units were constructed in several cases by combining gene sequences of closely related taxa. It is argued that this strategy should not distort phylogenetic analyses, provided the composite taxa are monophyletic with respect to other, closely related terminal units (SPRINGER ET AL. 2004). Given the relatively distant relationships between the included

terminals this assumption appears justified. In the phylogenetic trees, chimerical taxa are named after the next available or an unambiguous higher rank, for example, *Hypochilus thorelli* + *H. pococki* = *Hypochilus*. The only exceptions are the two out-group taxa that were named Onychophora and Tardigrada for convenience.

Additionally it was verified that the COI data did not contain any nuclear copies of mitochondrial-derived genes (numts; see BUHAY 2009). For two terminal taxa, there were multiple 18S sequences available that differed conspicuously. Since it was not possible to unambiguously identify the "correct" sequence, both 18S sequences were included for these two taxa, the mystacocarid *Derocheilocaris typica* (one own and one published sequence from NCBI) and the symphylan *Scutigerella causeyae (both sequences published in NCBI)*; both species are represented as doubled terminal taxa.

The influence of multiple sequence alignment methods on phylogenetic analysis was one focus of this study. Consequently, a series of analyses was conducted to determine the effects of different combinations of these variables on the dataset (see table 2.2 for an overview). These included:

(1) Alternative methods of multiple sequence alignment using either MUSCLE or MAFFT,

(2) Alignment methods based on secondary structure information,

(3) Identification and removal of ambiguously aligned regions (alignment masking),

(4) RY-coding for the mitochondrial marker COI and the loop regions of 16S rRNA to encompass saturation effects and compositional biases.

An alternative, manual alignment processing was carried out in addition to the alignment processing described in 2.3 (using ALISCORE and RNAsalsa). Two manually aligned datasets (16S and 18S rRNA sequences) and the following description of this process (which was only slightly modified by the author) were kindly provided by the collaborating working group of S. KOENEMANN (TiHO Hannover). This allows a comparison of results of hand vs. automized alignment processing (optimization):

"Both terminal regions were clipped by hand from the pre-aligned dataset since these regions appeared to contain erroneous or doubtful sequence fragments for a number of taxa. Subsequently, the pre-alignments of 18S and 16S were realigned manually based on secondary structure information. For methodical suggestions see KJER ET AL. (2007) and KJER (1995). Reconstructions of the secondary structure that were available on the Comparative RNA Web site (CRW) (CANNONE ET AL. 2002) and the European ribosomal RNA database (WUYTS ET AL. 2004) for some of the taxa included in this analysis were used. After a general identification of homologous structures, it was possible to reallocate entire sections of the sequence as well as smaller motifs within these sections.

For example, large sections of 18S sequences were found in the alignments that were likely misaligned for five taxa (*Derocheilocaris*, Tanaidacea, *Lightiella*, *Allopaurus* and *Scutigerella*). These large misaligned sections contained several hundred bps, and were entirely moved a distance of 700 up to 1,500 positions within the 18S alignment. In their new positions, the sections could be unambiguously allocated and realigned according to highly conserved

structures. In addition, numerous smaller sections were realigned based on recognizable, unique motifs, so that an estimated 40% of the positions in the standard alignment were resolved and rearranged. The structural optimization also revealed that the two 18S sequences of *Scutigerella*, which differed markedly in the pre-alignment, were highly compatible after rearrangements. Therefore, the shorter one of the two sequences (AF007106) was excluded from further analysis. Similar misaligned sections were found within the 16S partition. After structural optimization, two datasets were prepared for phylogenetic analysis. The manually adapted "hand alignment" is composed of 4902 characters (18S = 3569, 16S = 673, and COI = 658 characters) without alignment masking. For the dataset masked and adapted by hand single uninformative sites were deleted (sites containing nucleotides for only one taxon) and highly variable sections that could not be resolved according to secondary structure information. The smaller dataset has an 18S partition with 2184 characters, the partition of the 16S has 444 positions".

2.4.4 Phylogenetic tree reconstruction

Compositional heterogeneity of the dataset was tested using the program PAUP 4.0b10 (SWOFFORD 2002), and additionally the RY-coding was chosen to accommodate inhomogeneous base compositions.

RY-coding was applied in pretests (supplementary table S8) for the COI sequences and for the 16S rRNA loop regions to counteract the effects of saturation and inhomogeneous base composition. RY-coding was originally used to assign third codon positions of protein-coding mitochondrial genes to one of two categories, purines (R) or pyrimidines (Y) (PHILLIPS & PENNY 2003). In this analysis all alignment positions of chosen regions were RY-coded, generating an improved model likelihood (LnL). RY-coding was used for Runs 1-3 (table 2.2).

In a series of extensive pretests (supplementary table S8) of the partitioned data matrix, Bayes Factor Test A and B were carried out to identify the best model and settings for the final runs following the criteria of KASS & RAFTERY (1995). For detailed descriptions of the Bayes Factor Test (BFT) see NYLANDER ET AL. (2004) and KASS & RAFTERY (1993; 1995). Generally, the convergence of each parameter was checked for each run both "by hand" and using the software Tracer 1.4 (DRUMMOND & RAMBAUT 2007).

The Bayes Factor Test A showed significant convergence problems with the settings nst=6. Therefore, the second-best model (nst=2) was chosen for the three partitions of final runs 3 and 6-9 (table 2.2). Also the unlinking of partitions improved the model likelihood in the pretests (BFT A+B); however, parameter convergence of preliminary test runs with unlinked partitions was more problematic than for runs without unlinked. The unlinking is a more realistic scenario (delivering also a better model likelihood in the pretests), because it implies that all partitions can show different rates along the genes. For the final runs (table 2.2) the dataset with unlinked partitions was additionally tested (against the more conservative model setting without unlinking). The consideration was that longer iterations (doubled from 20 to 40 million) could enable a convergence stability for the nst=2 model.

To compare alignment methods and their impact on the resulting topologies, differing settings were additionally performed for the final runs next to the best model (see **run 2**, table 2.2) identified in the pretests. A second BFT and parameter convergence was conducted to check again the resulting final runs (supplementary table S9).

Table 2.2: Final runs for the dataset. Each run was performed with 2 x 4 chains on a Linux cluster running the parallel (MPI) MRBAYES version (HUELSENBECK & RONQUIST 2001). The cluster is a HP Blade system at the ZFMK (Bonn) with HP ProLiant DL380 G5 blades (Xeon dual quadcore E5345 2.33 GHz, 2x 4MB L2-cache, 32 GB RAM, 8 HDD, 64bit system). The RAXML run was computed with RAXML 7.04 P-THREADS (STAMATAKIS 2006).

Final runs	Prealignment	Processing	Model setting	Iteration	Burnin
run 1	MAFFT	RNASALSA	nst=2+1	40 mio	10 mio.
mixed	2528 bp	ALISCORE (-r)	5 partitions		
models		ALICUT	*gamma*		
			RY-coded		
run 2	MAFFT	RNASALSA	nst=2+1	40 mio	10 mio.
mixed	2528 bp	ALISCORE (-r)	5 partitions		
models		ALICUT	*gamma*		
part. unlinked			RY-coded		
run 3	MAFFT	RNASALSA	nst=2+1	40 mio	10 mio.
3 partitions	2528 bp	ALISCORE (-r)	*gamma*		
		ALICUT	unlinked		
			RY-coded		
run 4	MUSCLE	RNASALSA	nst=2+1	40 mio	10 mio.
mixed	2528 bp	ALISCORE (-r)	5 partitions		
models		ALICUT	*gamma*		
			RY-coded		
run 5	MUSCLE	RNASALSA	nst=2+1	40 mio	10 mio.
mixed	2528 bp	ALISCORE (-r)	5 partitions		
models		ALICUT	*gamma*		
part. unlinked			RY-coded		
run 6	MUSCLE	Manual aligment	nst=2	40 mio	10 mio.
3 partitions	4902 bp	(no masking)	*gamma*		
run 7	MUSCLE	Manual aligment	nst=2	40 mio	10 mio.
3 partitions	3288 bp	masked by hand	*gamma*		
run 8	MUSCLE	Manual aligment	nst=2	40 mio	10 mio.
3 partitions	2449 bp	ALISCORE (-r)	*gamma*		
		ALICUT	unlinked		
run 9	MUSCLE	Manual aligment	nst=2+1	40 mio	10 mio.
3 partitions	2449 bp	ALISCORE (-r)	3 partitions		
part. unlinked		ALICUT	*gamma*		
			unlinked		
			RY-coded		
run 10	MAFFT	RNASALSA	unpartitioned	(-f, a, GTR+CAT)	
RAXML	2528 bp	ALISCORE (-r)		10000 bootstrap-	

	ALICUT	replicates

For Runs 1-2 and 4-5, a mixed-model setting was used with five partitions as follows: 1 = 18S loop regions, 2 = 18S stems regions, 3 = 16S loop regions, 3 = 16S stem regions, 5 = COI (see Table 2.3).

Table 2.3: Number of alignment positions for the partitioned datasets based on alignments using the programs MAFFT and MUSCLE.

Gene (region)	18S loop	18S stem	16S loop	16S stem	COI	Total
Run 1-2 (MAFFT)	1011	688	208	94	546	2547
Run 4-5 (MUSCLE)	982	698	218	86	544	2528

For partitions 1, 3 and 5, the 4by4 (Standard DNA) model was chosen, for partitions 2 and 4, we applied the doublet model (nst=2) to account for secondary structures and covariation of paired stem positions. For the 18S loop (partition 1), nst=2 was chosen, while the loop region of 16S and the COI sequences were RY-coded to compensate for saturation effects. In this case, the setting nst=1 was chosen to consider that transitions and transversions are equally likely.

Run 3 was conducted with 3 partitions only to compare the influence of mixed models on the analyses.

The unmasked manual alignment (**run 6**) was optimized manually (**run 7**), processed with ALISCORE and ALICUT (**run 8**) and additionally RY coded for COI and 16S loop regions (**run 9**) to see the influence of an "objective" procedure on this data.

The Bayesian analyses were performed on a parallel version (MPI) of MRBAYES 3.0 (HUELSENBECK & RONQUIST 2001; RONQUIST & HUELSENBECK 2003) with 40 million generations for each final run with standard settings and 4 chains for each of the two parallel runs in MRBAYES.

A likelihood analysis was computed using RAXML 7.0.4 P-THREADS (OTT ET AL. 2007; STAMATAKIS 2006). **Run 10** was performed with the unpartitioned MAFFT dataset.

2.5 Analyses [B]: Is implementation of secondary structure based on alignment optimization and time-heterogeneity a solution to improve inference of crustacean phylogeny?

"Science is wonderfully equipped to answer the question –How? - But it gets terribly confused when you ask the question -Why? - (ERWIN CHARGAFF)

2.5.1 Objectives

In this analysis the methods of analysis [A| were massively expanded. Automated alignment processing and optimization using RNASALSA were applied "by default". The central objective for this analysis was to perform a time-heterogeneous analysis and to compare its results with a "simple" time-homogeneous analysis. This main focus on time-heterogeneity should compensate inhomogeneous base frequencies found for the dataset [B] and yet for the dataset in analysis [A].

The assumption was that completely sequenced SSU & LSU rRNA genes are useful to resolve the deeper splits of crustacean phylogeny and the position of crustaceans within arthropods. Some results are discussed and published in BMC Evolutionary Biology, see the attached manuscript I in the supplement.

2.5.2 Taxon sampling

In total 148 concatenated 18S and 28S rRNA sequences were included in the analysis (supplementary table S10).

27 new sequences for crustaceans were gained: 13 for the 18S and 14 for the 28S rRNA gene, respectively. Sequences of Pterygota and basal Hexapoda were kindly provided from collaborating working groups of B. MISOF, G. PASS and H. HADRYS, a close cooperation for this analysis existed in general with K. MEUSEMANN.

Only sequences which span at least 1500 bp for the 18S gene and 3000 bp for the 28S gene were finally included. For 29 taxa concatenated, "composite" sequences of 18S and 28S rRNA sequences were reconstructed. These are marked with an asterisk (see supplementary table S11). GenBank species were chosen as closely related as possible. Composite 18S sequences were constructed manually of *Speleonectes tulumensis* (EU370431, present study and L81936) and 28S sequences of *Raillietiella* sp. (EU370448, present study and AY744894). Concerning the 18S of *Speleonectes tulumensis* we combined positions 1-1644 of L81936 and positions 1645-3436 of sequence EU370043. Regarding the 28S of *Raillietiella* we combined positions 1-3331 of AY744894 with positions 3332-7838 of sequence EU370448. Position numbers refer to aligned positions. The out-group included the concatenated 18S and 28S rRNA sequences of *Milnesium* sp. (Tardigrada).

2.5.3 Analysis design

Secondary structures of rRNA genes were considered as advocated by BUCKLEY ET AL. (2000), HICKSON ET AL. (2000), KJER (1995) AND MISOF ET AL. (2006) to improve sequence alignment. Structural features are the targets of natural selection, thus the primary sequence may vary, while the functional domains are structurally retained.

Alignments and their preparation for analyses were executed with the multiple sequences for each gene separately. Sequences were prealigned using MUSCLE v3.6 (EDGAR 2004A). Sequences of 24 taxa of Pterygota were additionally added applying a profile-profile alignment (EDGAR 2004B). The 28S sequences of *Hutchinsoniella macracantha* (Cephalocarida), *Speleonectes tulumensis* (Remipedia), *Raillietiella* sp. (Pentastomida), *Eosentomon* sp. (Protura) and *Lepisma saccharina* (Zygentoma) were incomplete. Apart from *L. saccharina,* prealignments of these taxa had to be corrected manually. The "BLAST 2 SEQUENCES" tool was used to identify the correct position of sequence fragments in the multiple sequence alignment (MSA) for these incomplete sequences.

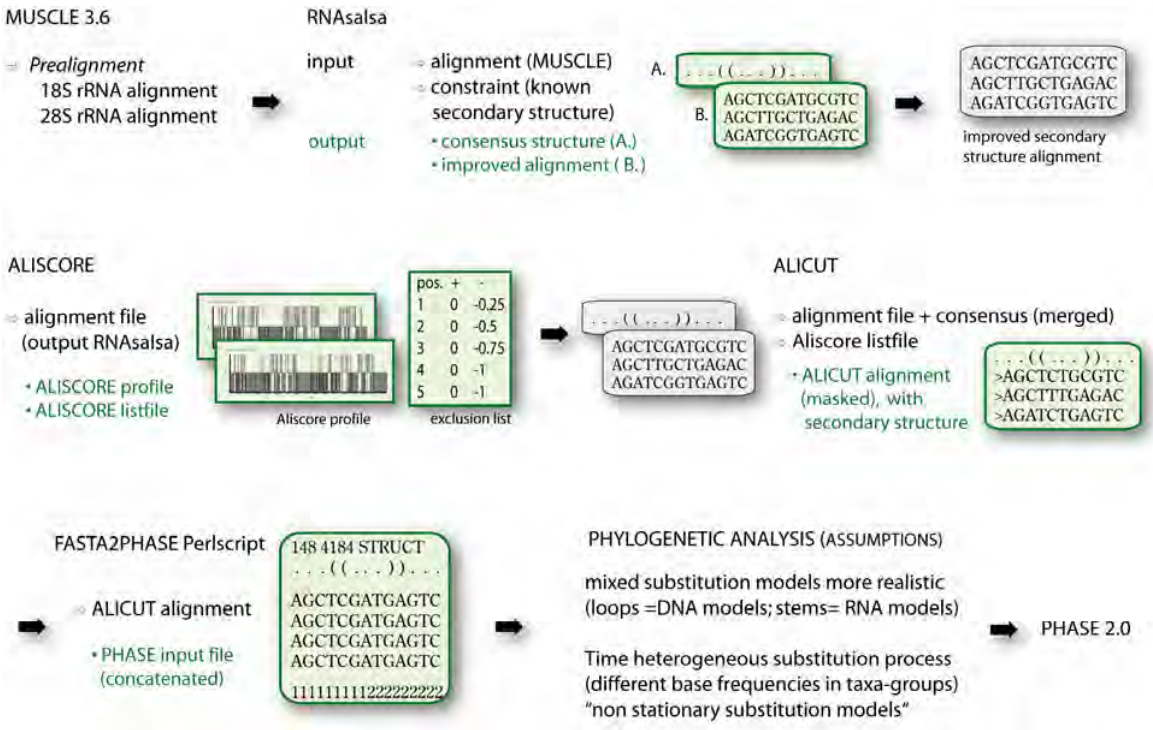


Figure 2.7: Design of analysis [B]. Input files of analyses are green, output files grey colored. The process flow shows the structure of the files for the phylogenetic software. The analysis design was conducted in close cooperation with K. MEUSEMANN.

2.5.4 Phylogenetic tree reconstruction

Mixed DNA/RNA substitution models were chosen, in which sequence partitions corresponding to loop regions were governed by DNA models and partitions corresponding to stem regions by RNA models that consider co-variation. Among-site- rate variation (YANG 1996) was implemented in both types of substitution models. Base frequency tests indicated that base composition was inhomogeneous among taxa (see results), suggesting non-stationary processes of sequence evolution. To take such processes into account the analyses were performed in PHASE-2.0 (GOWRI-SHANKAR & RATTRAY 2006) to accommodate this compositional heterogeneity to minimize bias in tree reconstruction. Base compositional heterogeneity is implemented in PHASE-2.0 according to the ideas developed by FOSTER (2004).

The number of candidate models was limited to the HKY85+Gamma (A), TN93+Gamma (B) and REV+Gamma (C) models for loop regions and the corresponding RNA16K+Gamma, RNA16J+Gamma and RNA16I+Gamma models for stem regions. Site heterogeneity was modeled by a discrete gamma distribution (YANG 1994) with six categories. The extent of invariant characters was not estimated since it was shown to correlate strongly with the estimation of the shape parameter of the gamma distribution (WADELL ET AL. 1997; SULLIVAN & SWOFFORD 2001; KELCHNER & THOMAS 2007; YANG 1996). The data was partitioned into four units representing loop and stem regions of 18S rRNA and loop and stem regions of 28S rRNA. DNA and RNA substitution model parameters were independently estimated for each partition. Substitution models were selected based on results of time-homogeneous setups. Three different combinations of substitution models were tested, REV +Gamma & RNA16I +Gamma, TN93 +Gamma & RNA16J +Gamma and HKY85 +Gamma & RNA16K +Gamma. Dirichlet distribution was used for priors, proposal distribution and Dirichlet priors and proposals for a set of exchangeability parameters (supplementary table S12) described in GOWRI-SHANKAR & RATTRAY (2007).

Appropriate visiting of the parameter space according to the posterior density function (ZWICKL & HOLDER 2004) was checked by plotting values of each parameter and monitoring their convergence. This was calculated for all combinations after 500,000 generations (sampling period: 150 generations). Models in which values of several parameters did not converge were discarded. For models that displayed convergence of nearly all parameter values, re-runs of the MCMC processes were performed with 3,000,000 generations and a sampling period of 150 generations. Prior to comparison of the harmonic means of lnL values, 299,999 generations were discarded as burn-in. After a second check for convergence the model with the best fitness was selected applying a Bayes Factor Test (BFT) to the positive values of the harmonic means calculated from lnL values (KASS & RAFTERY 1995; NYLANDER ET AL. 2004). The favored model ($2\ln B_{10}>10$) was used for final phylogenetic reconstructions.

To compare time-homogeneous vs. time-heterogeneous models in Bayesian analyses, 14 independent chains of 7,000,000 generations and two chains of 10 million generations were run for both setups on a Linux cluster with HP ProLiant DL380 G5 blades

(ZFMK Bonn). For each chain the first two million generations were discarded as burn-in (sampling period of 1000). The setup for the time-homogeneous approach was identical to the pre-run except for number of generations, sampling period and burn-in. The setting for the time-heterogeneous approach differed (figure 2.8). It followed the method of FOSTER (2004) and GOWRI-SHANKAR & RATTRAY (2007) in the non-homogeneous setup whereby only a limited number of composition vectors can be shared by different branches in the tree. Exchangeability parameters (average substitution rate ratio values, rate ratios and alpha shape parameter) were fixed as input values. Values for these parameters were computed from results of the preliminary time-homogeneous pre-run (3,000,000 generations). A consensus tree was inferred in the PHASE software with the option (like a sub-program) *Mcmcsummarize* using the output of the pre-run (see figure 2.8). This consensus tree topology and the model file of this run served as input for a ML estimation of parameters in the PHASE software option *Optimizer* (see figure 2.8).

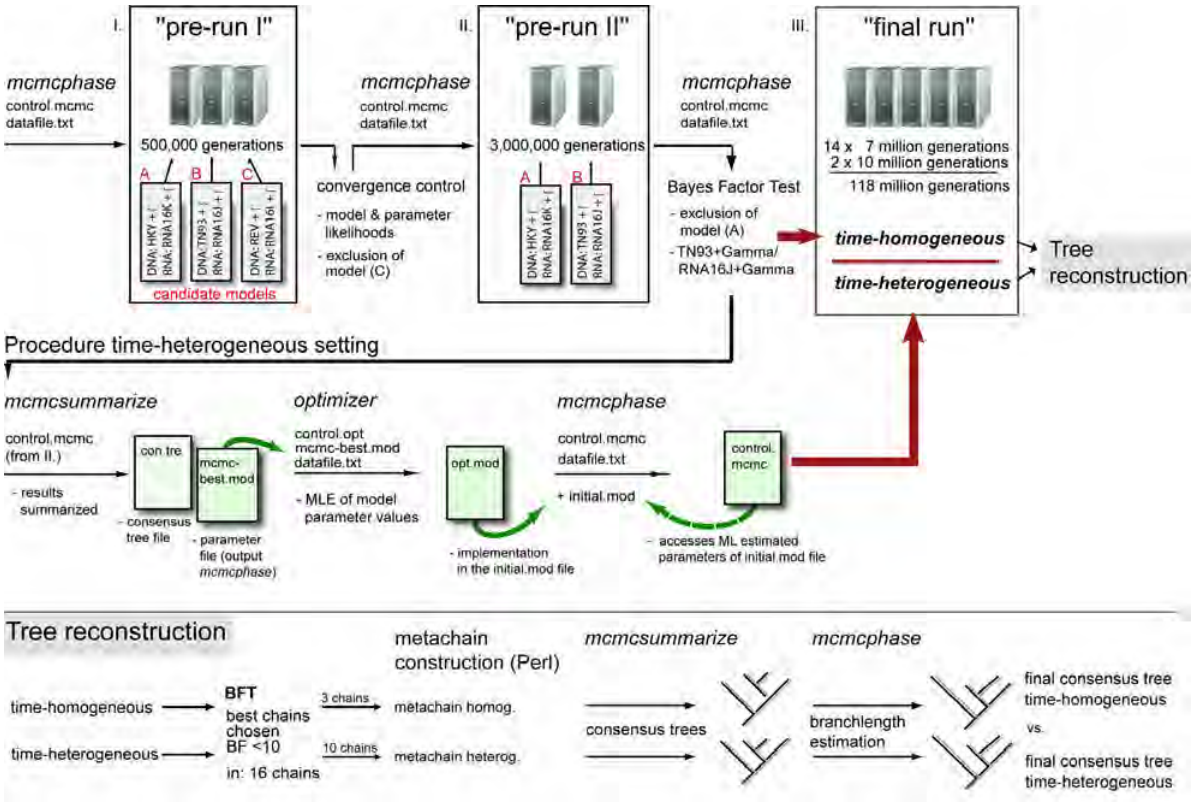


Figure 2.8: Detailed flow of the procedure of analyses [B] using the software package PHASE-2.0. Options (or software sub-program) used in PHASE-2.0 are italicized above the arrows and are followed by input files. Black arrows represent general flows of the analysis procedure, green arrows show that results or parameter values after single steps were inserted or accessed in a further process. Red block-arrows mark the final run of the time-heterogeneous and time-homogeneous approach with 16 chains each (2 x 118,000,000 generations).

First row: I.) 3 control files (control.mcmc) were prepared for *Mcmcphase* using three

different mixed models. This "pre-run'' was used for a first model selection (500,000 generations for each setting). Model (C) was excluded from the candidate models (colored in red, see "pre-run I") based on non-convergence of parameter values. **II.)** Step one (I.) was repeated with 3,000,000 generations using similar control files (different number of generations and random seeds) of the two remaining model settings. Calculated ln likelihoods values of both chains were compared in a BFT resulting in the exclusion of mixed model (A). Parameter values of the remaining model (B) were implemented in the time-heterogeneous setting. **III.)** We started the final analysis (final run) using sixteen chains for both the time-homogeneous and the time-heterogeneous approach. In the final time-homogeneous approach, the control files were similar to step II.) except for a different number of generations and random seeds.

Second row: Additional steps were necessary prior to the computation of the final time-heterogeneous chains. *Mcmcsummarize* was applied for the selected mixed model (B) to calculate a consensus tree. *Optimizer* was executed to conduct a ML estimation for each parameter value (given in the opt.mod file) based on the inferred consensus tree and optimized parameter-values (mcmc-best.mod), a data file delivered by *Mcmcphase*. Estimated values were implemented in an initial.mod file. The initial.mod file and its parameter values was accessed by the control files of the final time-heterogeneous chains (only topology and base frequencies estimated).

Third row: Trees were reconstructed separately for the time-homogeneous and time-heterogeneous setting. All chains of each approach were tested in a BFT against the chain with the best lnL. We only included chains with a 2ln B_{10}-value >10. From these chains a metachain was constructed for each setting using Perl and applied *Mcmcsummarize* to infer the consensus topology. To estimate branch lengths properly *Mcmcphase* was conducted, resulting branch lengths were implemented in the consensus trees.

Estimated values of exchangeability parameters from the resulting *Optimizer* output file and estimated start values for base frequencies were fed into *Mcmcphase* for the time-heterogeneous analysis. Values of exchangeability parameters remained fixed during the analysis. The number of allowed base frequency categories (models) along the tree was also fixed. The number of base frequency groups was set to three "submodels", reflecting base frequency heterogeneity. Harmonic means of ln Likelihood values of these 16 independent chains were again compared with a BFT to identify possible local optima in which a single chain might have been trapped. Only sample data of chains with a 2ln B_{10}-value <10 (KASS & RAFTERY 1995) was merged using a Perl-script to construct a "metachain'' (BEIKO ET AL. 2006).

Finally ten time-heterogeneous chains and three time-homogeneous chains were included. The assembled meta-chains included 56 million generations for the non-stationary approach (table 2.4) and 18 million generations for the time-homogeneous approach (table 2.4), burn-ins were discarded.

Table 2.4: Chains included to infer the time-heterogeneous and homogeneous consensus trees.

Chains included for time-heterogeneous models	generations (burnin excluded)	harmonic mean ln-Likelihood	2ln (B10)
1	5 million	78999.3699	-
2	5 million	78999.6928	0.6458
3	5 million	78999.7010	0.6800
4	5 million	79000.0580	1.3770
5	5 million	79001.9540	5.1689
6	5 million	79002.5669	6.3941
7	8 million	79002.8227	6.9057
8	8 million	79003.4490	8.1575
9	5 million	79003.5540	8.3680
10	5 million	79004.0280	9.3156
Chains included for time homogeneous models	**generations (burnin excluded)**	**harmonic mean ln-Likelihood**	**2ln (B10)**
1	8 million	79680.9820	-
2	5 million	79683.9097	5.8554
3	5 million	79685.0871	8.2102

Consensus trees and posterior probability values were inferred using *Mcmcsummarize*. Branch lengths of the time-homogeneous and time-heterogeneous consensus tree were estimated using three *Mcmcphase* chains (4 million generations, sampling period 500, topology changes turned off, starting tree = consensus tree, burn-in: 1 million generations) from different initial states with a GOWRI-SHANKAR modified PHASE version. To infer mean branch lengths data was combined with the described branch lengths and *Mcmcsummarize*. These mean branch lengths were used to redraw the consensus tree (figure 2.8).

2.6 Analyses [C]: Do phylogenomic data enlight crustacean phylogeny within arthropods - or do old problems stick to the analysis of this new large-scale data?

"Facts do not 'speak for themselves', they are read in the light of theory" (STEPHEN JAY GOULD)

2.6.1 Objectives

Extensive sequence data from genome sequencing and expressed sequence tags (ESTs) projects were recently used to infer deep metazoan phylogeny (DUNN ET AL. 2008; PHILIPPE ET AL. 2009; ROEDING ET AL. 2007). These studies report robust results concerning crustacean and arthropod relationships, but are however deficient in taxon sampling of arthropod groups, e.g. large groups like crustaceans are covered by only few taxa and important groups of chelicerates, myriapods or primary wingless hexapods are completely missing. Until recently, arthropod data with broad taxon sampling were predominantly restricted to single gene analyses, e.g. rRNA genes (MALLAT & GIRIBET 2006; MALLAT ET AL. 2004).

The largest phylogenomic dataset of crustaceans and arthropods including 233 taxa was compiled for this analysis in order to alleviate the restrictions of current phylogenomic data. Previous phylogenomic analyses have shown that massive accumulation of data is not sufficient to guarantee reliable tree reconstruction, but instead selection of orthologous loci, consideration of data quality and missing data and model fitting must be part of the analysis pipeline (DUNN ET AL. 2008; PHILIPPE ET AL. 2009; ROEDING ET AL. 2007). Therefore, new tools were used for orthologous gene prediction (HaMStR, EBERSBERGER ET AL. 2009, see figure 2.9), alignment masking (MISOF & MISOF 2009) and new heuristics for selection of genes and taxa were applied.

2.6.2 Taxon sampling

Three new crustacean EST taxa of three different underrepresented major crustacean groups (table 2.5) were added to other published sequences (supplementary table S13).

Table 2.5: EST sequencing projects for crustacean taxa. Number of sequences gained during the processing from raw ESTs to final orthologous gene sequences (see analysis design).

Crustacean group	Species	Raw ESTs	ESTs processed	EST contigs	Orthologous Sequences
Branchiopoda	*Triops cancriformis*	3981	3932	2542	115
Cirripedia	*Pollicipes pollicipes*	4224	4193	1721	107
Copepoda	*Tigriopus californicus*	5024	5007	2598	65

Until recently data of only one copepod representative was published with only few genes. Therefore *Tigriopus californicus* was additionally sequenced. Branchiopoda are to date only represented by extremely derived species of the anostracan group in phylogenomic data, namely several *Artemia* species. Genome projects of *Daphnia magna/pulex* exist additionally. That was the reason to choose *Triops cancriformis*, a representative of the Notostraca. *Pollicipes pollicipes* is to date the only cirripede for which EST data exists.

For an overview of published crustacean EST data (27 projects in total) see supplementary table S14. Especially for phylogenomic data the over-representation of malacostracan crustaceans is conspicuous. Some to date unpublished sequences which were generously provided by the arthropod working groups of T. BURMESTER (Myriapods & Chelicerates), B. MISOF (basal Hexapods), and H. HADRYS (Hexapods) could be included in this analysis. The analysis was conducted in cooperation with K. MEUSEMANN.

233 taxa were implemented for final analysis (214 taxa of Euarthropoda plus representatives of 3 onychophorans, 2 tardigrades and out-group taxa) with a total of 775 putative orthologous gene loci (supplementary table S15).

2.6.3 Analysis design

The general idea of this approach is to reduce effects of under sampling of taxa and genes and filter instable taxa before tree reconstruction. A strategy is chosen in which a subset of the concatenated super matrix is selected to condense it to a maximally informative set of taxa and genes. Informativeness of genes and taxa was assessed by integrating established methods (see figures 2.10 – 2.13). This preprocessing of the total data helped to considerably reduce the effort spent in tree reconstructions. It also opens a route to assess whether it is worthwhile to include new taxa or genes in a pre-existing supermatrix based on their contribution to the total informativeness of the supermatrix without using tree reconstructions.

Sequence processing and orthology assignment (figure 2.9). New EST data were preprocessed with the software tool LUCY (CHOU & HOLMES 2001) for DNA sequence quality trimming and vector removal. EST data available for arthropods (myriapods, chelicerates, pancrustaceans, onychophorans) plus tardigrades and selected species of nematodes, annelids and molluscs were extracted from dbEST (NCBI), the Gene Index Project or the NCBI Trace Archive. EST sequences of 244 taxa (222 euarthropods) were screened for contamination, trace files and low-quality ends of sequences quality checked. To obtain contigs, the ESTs were clustered using the TGICL, a software system for fast clustering of large EST datasets (PERTEA ET AL. 2003) and the contigs were translated into amino acid level for orthology prediciton by the HaMStR approach (EBERSBERGER ET AL. 2009). Therefore, a reference gene set was used with 13 mainly arthropod species plus three vertebrates (figure 2.9 and supplementary table S13). The pipeline for this procedure was developed in the bioinformatics group of A. V. HAESELER (CIBIV, Vienna) within the "deep metazoan phylogeny

program", the settings for the reference gene set was suggested by the present author in cooperation with others.

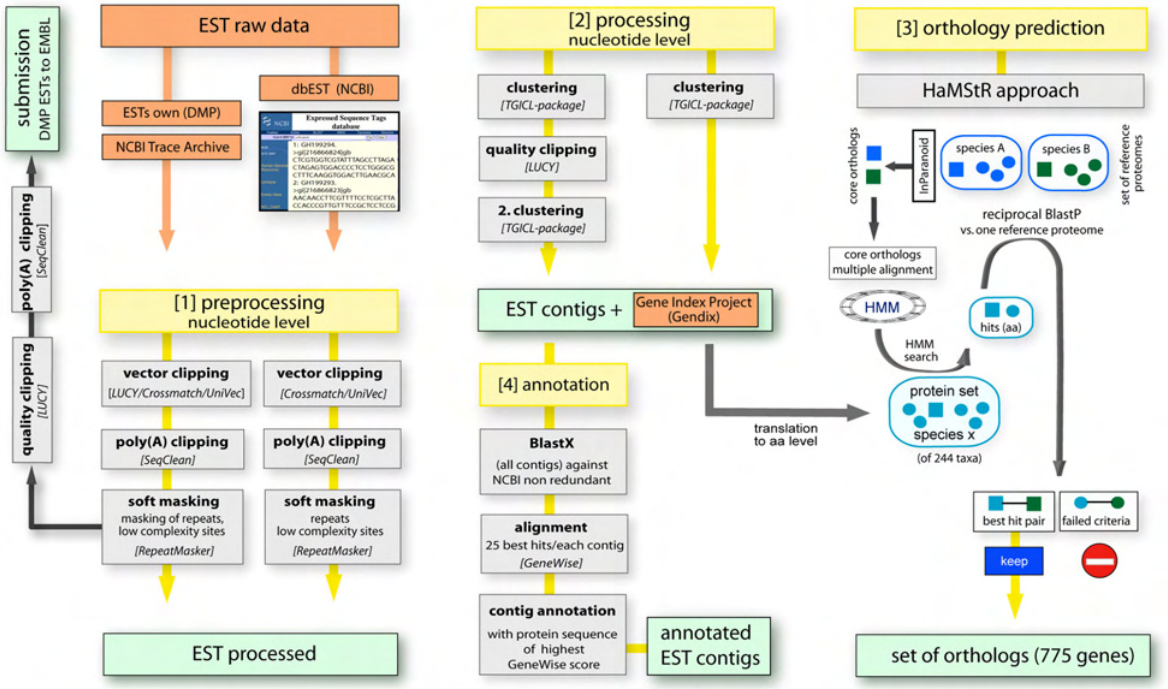


Figure 2.9: Processing of EST data and orthology assignment.

EST raw data (orange) of own and public EST projects were mined and processed in four major steps (yellow): preprocessing, processing, orthology prediction and annotation. [1] In preprocessing, EST sequences were screened for vectors and poly(A) tails using LUCY (CHOU & HOLMES 2001). Then all sequences including published ESTs were screened for contamination by comparison against UNIVEC (www.ncbi.nlm.nih.gov/VecScreen/UniVec.html) with CROSSMATCH (www.incogen.com/public_documents/vibe/details/crossmatch.html) and SEQCLEAN (www.tigr.org/tdb/tgi/software/). SEQCLEAN screened for poly(A) tails as well. Sequences < 100 nucleotides were discarded. Afterwards repetitive elements in the remaining ESTs were soft masked with REPEATMASKER (SMIT 1996) using REPBASE (JURKA ET AL. 2005). [2] ESTs were clustered using the TGICL package (PERTEA ET AL. 2003). ESTs generated in own projects were quality clipped with LUCY and clustered a second time to obtain and keep longer sequences for the EST contigs. The contigs were translated into amino acid level and [3] integrated in the orthology prediction (HaMStR). A set of reference proteomes (INPARANOID, http://inparanoid6.sbc.su.se) was compiled with *D. pulex, T. castaneum, B. mori, A. aegypti, A. melifera, D. melanogaster, C. elegans, C. briggsae, Capitella sp., L. gigantea, H. sapiens, T. nigroviridis* and *X. tropicalis* as 'primer' taxa. Multiple alignments of 'core' orthologs for primer taxa were used to train Hidden Markov Models (HMMs) to search in each protein set of the 244 taxa for hits. A reciprocal BLASTP decides about the surviving of a hit. For the re-blast step always the proteome of the presumably evolutionary closest primer taxon for each considered species was chosen. Finally a set of 775 putative orthologous gene loci is constituted. [4] In a side step EST

contigs were annotated using a BLASTX search against NCBI's non-redundant protein database. The protein sequences of the 25 best hits for each contig were aligned with GENEWISE (BIRNEY ET AL. 2004). The contig is annotated according to the protein sequence with the highest GeneWise score. Single EST reads were submitted to EMBL.

Alignments and alignment masking. All 775 putative orthologous gene loci were aligned (figure 2.10) with MAFFT L-INSI (KATOH & TOH 2008). The complete dataset comprised 222 euarthropods, 3 onychophorans, 2 water bears, 3 vertebrates, 8 nematodes, 3 annelids and 3 molluscs as out-group. All vertebrates and eight Drosophila species were excluded from further processing to avoid overrepresentation of this genus.

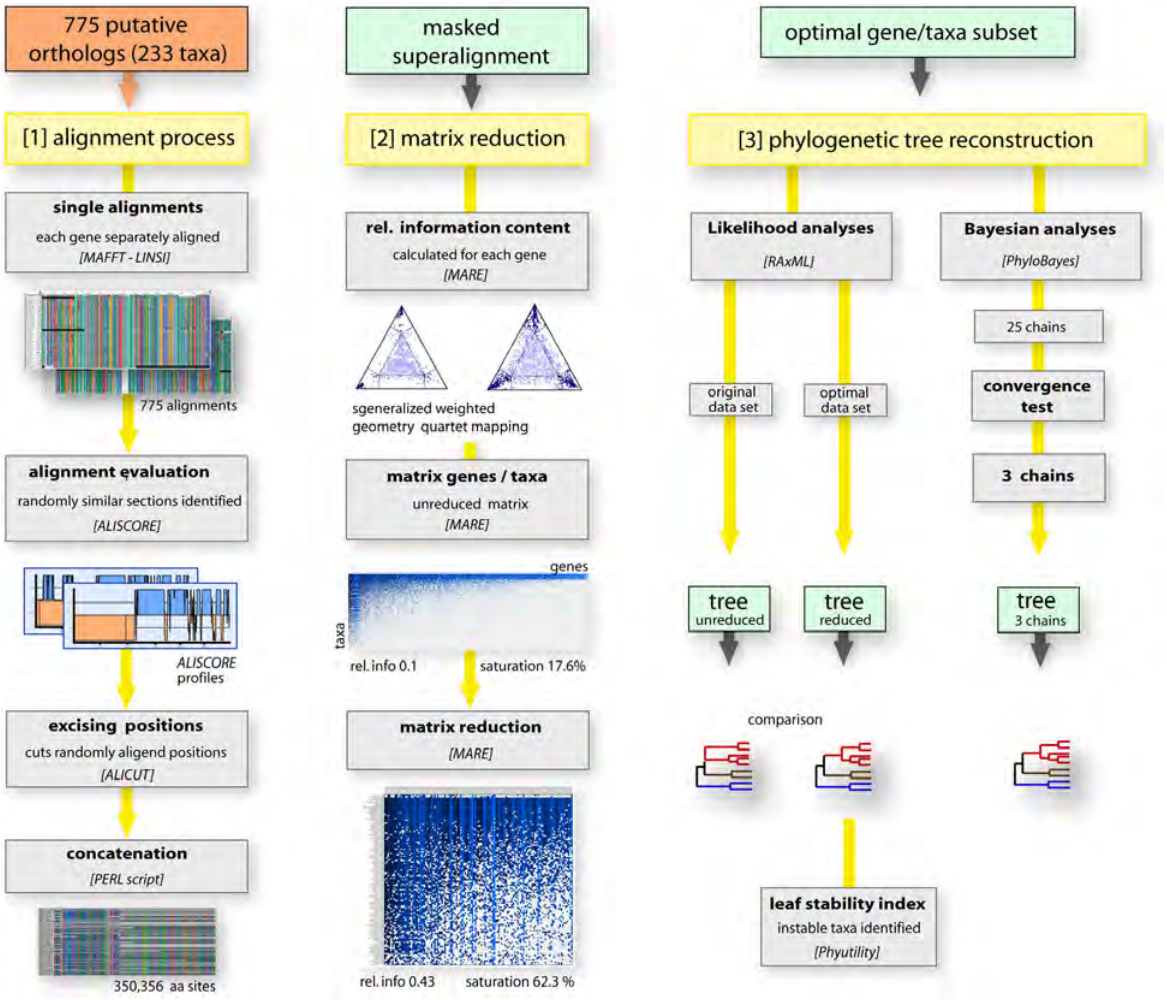


Figure 2.10: Alignment masking, selecting an optimal data-subset and phylogenetic analyses. Based on the 775 putative orthologous genes (orange) the workflow consists of three major steps (yellow): alignment process, matrix reduction and phylogenetic reconstruction. [1] The alignment process starts aligning multiple sequence alignments for each single gene separately using MAFFT (KATOH & TOH 2008). ALISCORE

(MISOF & MISOF 2009) identifies randomly similar sections in each alignment. ALICUT (http://utilities.zfmk.de) excises all positions scored negatively by ALISCORE. The genes are concatenated to a masked "super alignment" (green). [2] The step of reduction heuristics starts with the calculation of the relative information content of each gene in our masked superalignment. The generated matrix (taxa vs. genes) is provided with a value for relative information content of each gene. A subset (green) is selected by excluding genes and taxa showing low relative information content [3] Phylogenetic trees were reconstructed using RAXML and PHYLOBAYES. The two resulting ML trees (green) are derived from the data subset and the original dataset. PHYUTILITY was used to identify 'instable' taxa. 25 chains were set (PHYLOBAYES) after testing for topological incongruencies (see "Phylogenetic tree reconstructions") resulting in a 'triple' consensus tree (3 chains).

Being aware that excluding randomly similar aligned sections can make phylogenetic analyses more reliable prior to tree reconstruction (CASTRESANA 2000) every single gene alignment was masked with ALISCORE (MISOF & MISOF 2009) on amino acid levels with default window size and maximal number of pairwise comparisons. For each gene, only sequences comprising more than one half of the average sequence information (sequence lengths) were included in the ALISCORE analyses. Sections scored as randomly similar were discarded with ALICUT and all alignments were concatenated to a "super alignment" comprising 233 taxa and 350,356 amino acid positions.

Selecting a data subset using new matrix reduction heuristics. With the software MARE (MAtrix REduction) (MISOF ET AL., in prep.) potential relative phylogenetic information content of each single partition (gene) within a "super alignment" was calculated based on weighted geometry quartet mapping (NIESELT-STRUWE & HAESELER 2001) which was extended to amino acid data (figure 2.10 and 2.11).

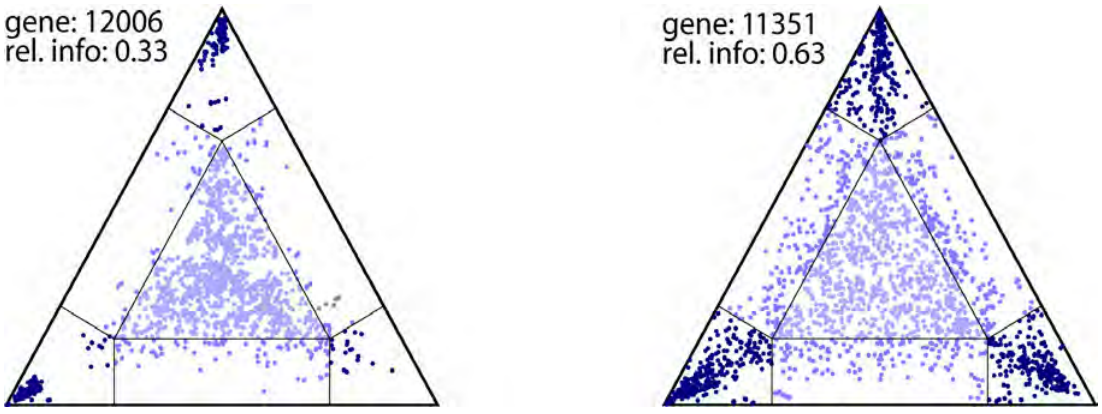


Figure 2.11: Potential relative information content of genes visualized by 2D simplex bipartite graphs. Potential information content (rel. info) of a single partition (gene) is defined as the relative tree-likeness of the data using geometry mapping (NIESELT-STRUWE & HAESELER 2001), extended to amino acids incorporating the BLOSUM62

substitution matrix. Relative tree-likeness corresponds to the relative frequency of simplex points each representing a quartet of taxa within the outer areas of at least partially resolved trees compared to the total number of simplex points. Genes containing less than four sequences and taxa containing less than 1/3 of the single gene sequence are considered as absent.

Each gene received a value of informativeness between 0.0 and 1.0, reflecting the relative number of resolved quartet trees. A data availability matrix was then transformed into a matrix of potential information content of each taxon and gene by multiplying availability (0 or 1) with scores of informativeness. Relative information content of each gene was calculated as the average value over all taxa including missing data. The total average information content of a super matrix was calculated as the sum over all genes (figure 2.12).

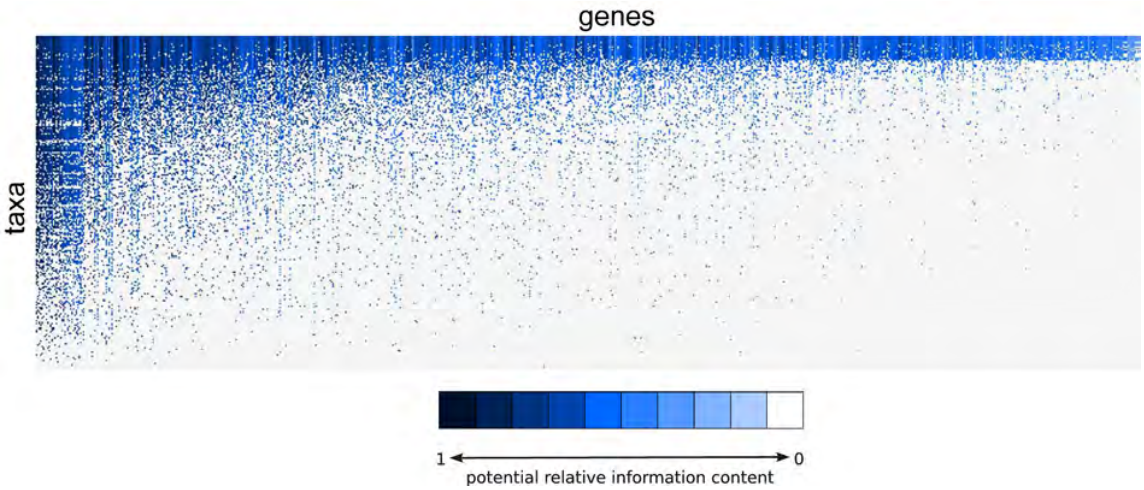


Figure 2.12: Original data matrix with potential relative information content of each gene and taxon. The matrix comprises 233 taxa (rows) and 775 genes (columns). Potential relative information content ranges from 0.0 – 1.0 (10 units). Potential relative information content is color coded (see color scale) from dark blue (> 0.9 – 1.0) to white (relative information content of ≥ 0 - 0.1 or missing data). Genes with a relative information content < 0.04 were considered as absent. Overall relative information content of the matrix: 0.1, overall saturation: 17.6%.

To select an optimal subset of taxa and genes with high total average information content a simple hill climbing procedure was used. Reduction starts with dropping either a taxon (row) or a gene (column) with the lowest average information content, generating a new matrix. Consequently taxa or genes with lowest average information content will be discarded from the matrix, and a data subset with increased relative information content (figure 2.13) is obtained.

The copepod *Tigriopus* and the chilopod *Scutigera* were defined as constrained taxa, thus they were not dropped from the sub-matrix. These taxa were thought to be important to shorten long branches and therefore retained. In order to reach an optimum of matrix

reduction, an optimality function was defined, which takes into account that size reduction and low average information content are penalized. The connectivity between taxa was monitored. Details of the new reduction algorithm will be published elsewhere (MISOF ET AL., in prep.). Finally, the original "super-alignment" was rewritten based on the optimal subset of the data

Although this final matrix of the data subset is partly a result of this analysis it will be presented in this section for a better understanding of the entire reduction procedure.

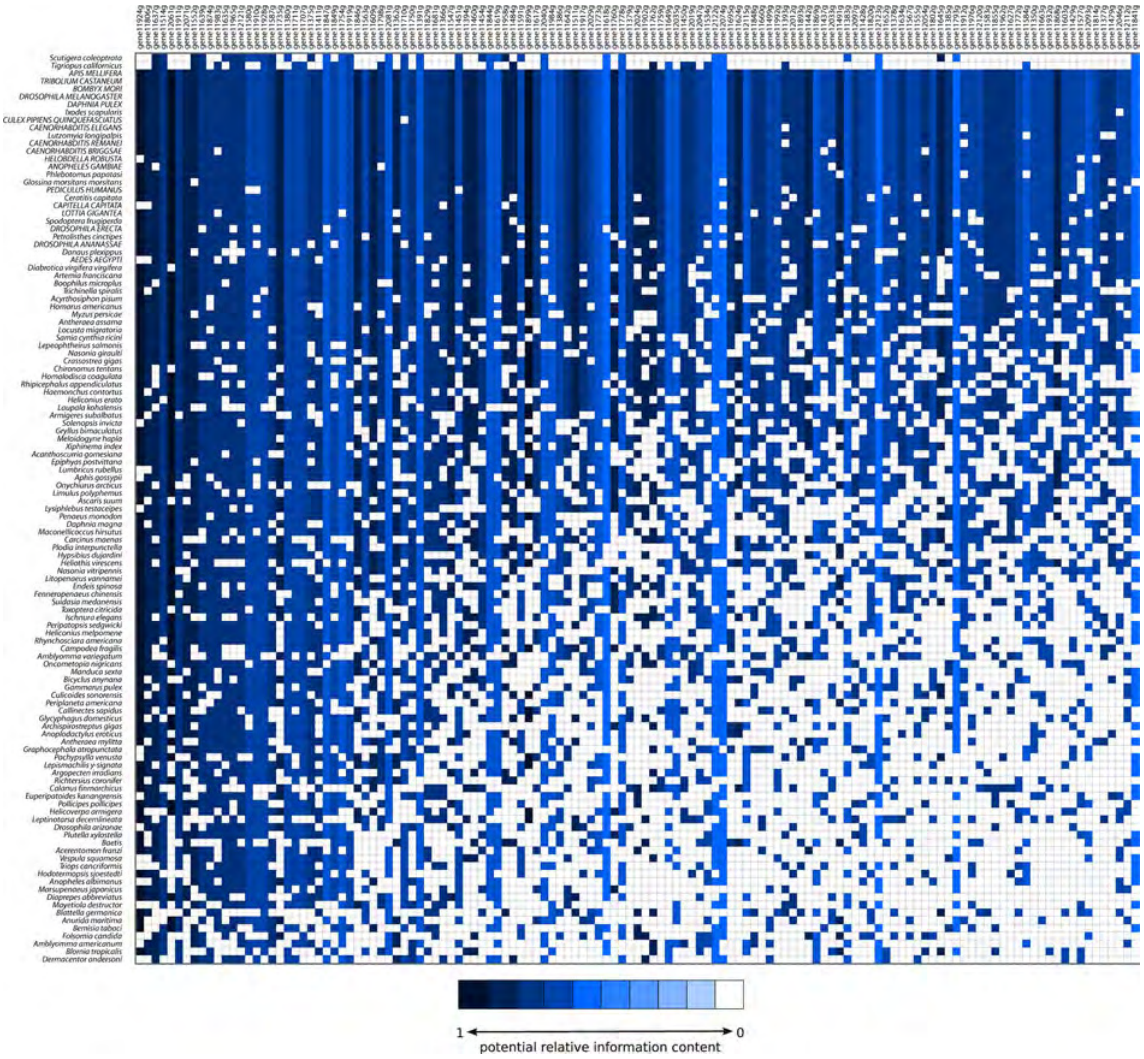


Figure 2.13: Optimal data subset. The matrix comprises 117 taxa (rows) and 129 genes (columns). Potential relative information content ranges from 0.0 – 1.0 (10 units). The color code is specified in Supplementary Fig. 4. *Tigriopus* (Copepoda, Crustacea) and *Scutigera* (Diplopoda, Myriapoda) are defined as constraints (first and second row), thus remain in the matrix although they show few gene hits. Overall relative information content of the matrix: 0.43, overall saturation: 62.3%.

2.6.4 Phylogenetic tree reconstruction

ML analyses were performed using RAXML PTHREADS 7.0.0 (OTT ET AL. 2007; STAMATAKIS 2006). The optimal data subset comprised 129 genes, 117 taxa (101 arthropods including onychophorans, two tardigrades, nematodes, annelids and three molluscs) and 37,476 aa positions. ML tree search and rapid bootstrapping were applied within one step (-f a, 1,000 bootstrap replicates). To examine possible biases, ten single ML tree searches and separate bootstrapping (100 replicates) for the original concatenated supermatrix (350,356 aa positions) were conducted. Due to restricted computational power, calculation of branch lengths was not possible. The ML tree with the best likelihood value was chosen to plot bootstrap values (results, figure 3.13). All ML searches were calculated with the PROTMIX substitution model and the WAG matrix (WHELAN & GOLDMAN 2001). The WAG matrix is commonly used for protein coding genes, which is based on empirical data for globular proteins. It allows the prediction of convergences and reversions among amino acids with biochemical structure. A major disadvantage is that highly saturated positions will be underestimated with the WAG matrix implementing substitution processes only in a small level reaching equilibrium rather early (LARTILLOT ET AL. 2007A). This means after reaching this equilibrium very fast for highly saturated positions the WAG model equals a site-homogeneous model.

Bayesian analyses for the data subset were inferred using PHYLOBAYES version 2.3c (LARTILLOT ET AL. 2007B) running the CAT mixture model (LARTILLOT & PHILIPPE 2004). In contrast to the WAG matrix a mixture model as the CAT model (LARTILLOT & PHILIPPE 2004) reflects more realistic a multiple substitution process by sorting the positions into groups with several profiles of amino acid frequencies.

25 MCMC chains ran for 20,000 cycles each, sampling every cycle. All parameter values were checked for convergence to define the burn-in (5,000 cycles). To infer a majority rule consensus (mrc) tree the discrepancy observed across all bipartitions (maxdiff) was checked of all chains by pairwise comparison and comparing 'triple' chain-combinations with the bpcomp tool. Harmonic means of the likelihood values of each chain (burn-in excluded) were calculated. To infer the Bayesian mrc tree three chains were included showing the lowest maxdiff value (0.186) while featuring the best likelihood values (harmonic means) of all 'triple-chain combinations' (table 2.6).

Table 2.6: Likelihood values and chain combinations of the 25 PHYLOBAYES runs. Chain ID-identifier for each PHYLOBAYES-chain from the 25 runs; log LH (harmean) harmonic mean of all log likelihood values, 15,000 cycles per chain while the burnin was excluded; chain combination – chain combinations consisting of three chains each per combinations (triple) for which the maxdiff values (PHYLOBAYES) were below 0.3; maxdiff – discrepancy value observed across all for the given triple-chain.

Chain ID	Log likelihood (harmean, burnin excluded)	Chain combination	maxdiff
c18	948174.861	c04 - c18 - c20	0.186
c04	948217.994	c23 - c01 - c06	0.203
c20	948376.710	c21 - c23 - c08	0.208
c16	948469.642	c21 - c23 - c01	0.188
c05	948525.741	c01 - c23 - c08	0.208
c22	948678.822	c21 - c08 - c01	0.208
c23	948708.712	c22 - c05 - c14	0.236
c21	948752.990	c22 - c05 - c16	0.187
c08	948757.764	c22 - c14 - c16	0.236
c14	948779.210	c05 - c14 - c16	0.162
c01	948865.845	all 25 chains	1

For identification of 'instable' taxa leaf stability indices (THORLEY & WILKINSON 1999) were calculated from the collected bootstrap trees of the ML analysis using PHYUTILITY (SMITH & DUNN 2008). Instable taxa are identified by taxa triplets in PHYUTILITY by calculating a stability index of taxa combination for these triplets. A threshold of < 95% was defined as 'instable'. All analyses ran several months on Linux Clusters of the ZFMK and of the SuGI (Sustainable Grid Intrastructure) project, VIKTOR ACHTER, University of Cologne.

Consensus network and statistics of Bayesian topologies. Due to differences between single topologies of the 25 PHYLOBAYES (LARTILLOT & PHILIPPE 2004) chains a consensus network (HOLLAND & MOULTON 2003) was computed with SplitsTree 4.8 (HUSON & BRYANT 2006). To visualize conflicts a threshold of 0.01 was chosen and averaged edge weights incorporated.

2.7 Analysis of hemocyanin structure in Remipedia

It is a good morning exercise for a research scientist to discard a pet hypothesis every day before breakfast. It keeps him young (KONRAD LORENZ).

2.7.1 Objectives

Oxygen transport in the hemolymph of various arthropod taxa is facilitated by copper-proteins referred to as hemocyanins (BURMESTER 2002; MARKL & DECKER 1992; VAN HOLDE & MILLER 1995). Hemocyanins of Arthropoda are large hexameric or oligohexameric proteins composed of similar or identical subunits in the size range of 75 kDa. Each subunit can bind to an O_2 molecule by the virtue of two copper ions that are coordinated by six histidine residues. Hemocyanins have been thoroughly studied in Chelicerata and Crustacea, occur in Myriapoda (KUSCHE & BURMESTER 2001), and have recently been identified in Onychophora (KUSCHE ET AL. 2002) and Hexapoda (HAGNER-HOLLER ET AL. 2004; PICK ET AL. 2008; PICK ET AL. 2009). Within Crustacea, hemocyanins have been thought to be confined to Malacostraca (MANGUM 1985; MARKL & DECKER 1992).

Hemocyanin sequences have been shown to be informative for the inference of phylogenies within Arthropoda (BURMESTER 2001; KUSCHE & BURMESTER 2001). Thus, lineage-specific presence and phylogenetic analyses could be useful in assessing the phylogenetic position of Remipedia. Screening the preliminary EST data of *Speleonectes tulumensis* hemocyanin, three subunits were found and analyzed.

2.7.2 Analysis design

This analysis was conducted in close cooperation with B. ERTAS and T.BURMESTER of the University Hamburg, it was published in the Journal "Molecular Biology and Evolution", see attached manuscript II in the supplement.

Screening the preliminary *Speleonectes tulumensis* EST database for hemocyanin three subunits of this respiratory protein were found. From fresh tissue sequences were specifically amplified and sequenced. The laboratory work (cloning of identified sequences and western blotting for protein amplification) was accomplished by B. ERTAS in the lab of T. BURMESTER. The present author collected material and co-worked on the manuscript.

2.7.3 Phylogenetic tree reconstruction

Bayesian phylogenetic analysis was performed with 8,000,000 generations using MrBayes 3.1.2 (HUELSENBECK & RONQUIST 2001), assuming the WAG model with a gamma distribution of substitution rates. The program Tracer 1.4 (http://tree.bio.ed.ac.uk/software/tracer/) was used to examine log-likelihood plots and MCMC summaries for all parameters. Posterior probabilities were estimated on the final 60,000 trees (burnin = 20,000).

A maximum likelihood analyses was performed in RAXML 7.0.4 (WAG model) and the resulting tree was tested by bootstrapping with 1000 replicates. TREE-PUZZLE 5.2 (SCHMIDT ET AL. 2002) was used to test alternative tree topologies.

The phylogenetic analysis was performed at the University of Hamburg using Linux cluster systems, for more details of the laboratory work, molecular analysis and tree reconstruction see attached manuscript II in the supplement.

3. RESULTS

There are many hypotheses in science, which are wrong. That's perfectly all right; they're the aperture to finding out what's right. Science is a self-correcting process. To be accepted, new ideas must survive the most rigorous standards of evidence and scrutiny. (CARL SAGAN)

3.1 Analyses [A]: Can 16S, 18S and COI marker genes improve inference of crustacean phylogeny? Comparing "usual" standard vs. secondary structure based approaches

3.1.1 Data signal and split supporting patterns

Neighbor-joining networks visualize the presence and nature of potentially conflicting signals in the data (figures 3.1-3.2). The networks clearly show that conflicting signals prevail, indicated by the preponderance of non-parallel edges that represent conflicting splits of groups of taxa. This lack of a strong tree-like signal is additionally reflected by the presence of many unresolved areas, and low clade support values in the phylogenetic trees. Certain crustacean clades are apparent in the networks, such as Cirripedia, Copepoda, and Branchiopoda. Accordingly, we recovered these clades in all our phylogenetic analyses. Each of the networks evidences that the data is obviously biased with long branch problems. A conspicuous grouping of long branch taxa is observed in all reconstructed networks.

New alignment processing methods: The networks show that data processing and optimization of alignments (figure 3.1, A and B) as described in chapter 2 can to an extent grade improve the structure of the data by removing conflict. The RY coding (figure 3.1, B) improved the dataset by eliminating more conflicts. This is illustrated by the pycnogonids. Both networks are based on the same MAFFT alignment, but only in the RY coded network the three included pycnogonids, an expected clade, group together.

Manually aligned dataset: The dataset aligned (figure 3.2, A) and optimized by hand (figure 3.2, B) at the working group of S. KOENEMANN do not differ significantly from each other. Optimization and masking by hand did not greatly improve the general data structure. Contrary, one can see that there is obviously a problem in this procedure: optimizing data by hand drives one eventually in the problematic situation that homology assumptions and random sequence similarity are not clearly to distinguish. The best example for this problem is the clade Mystacocarida. The two mystacocarids cluster together in the manually optimized alignments, which is more than doubtful and suspicious regarding the fact that the published mystacocarid 18S sequence is a contamination (see 3.3 problematic data).

A final result of the network reconstructions is that the dataset based on the marker genes 18S, 16S and COI shows a problematic data structure prior to tree reconstruction.

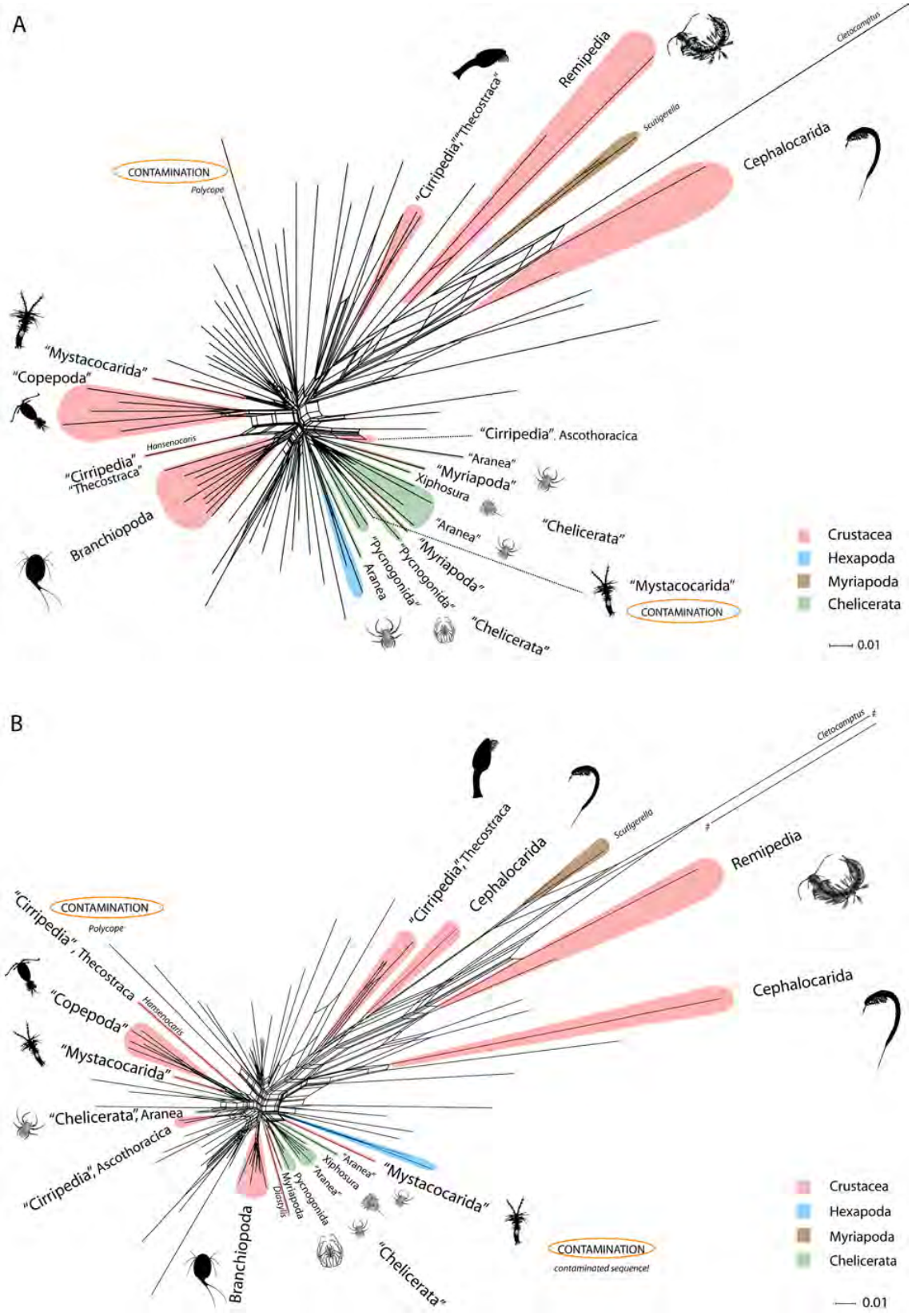


Figure 3.1: Networks of the processed, MAFFT-aligned (A) and additionally RY-coded (B) dataset. Sequences identified as contaminated are marked with an orange circle. For color code, see graphics. Crustaceans are marked red and some groups are highlighted with species pictures (only some taxa are exemplarily colored).

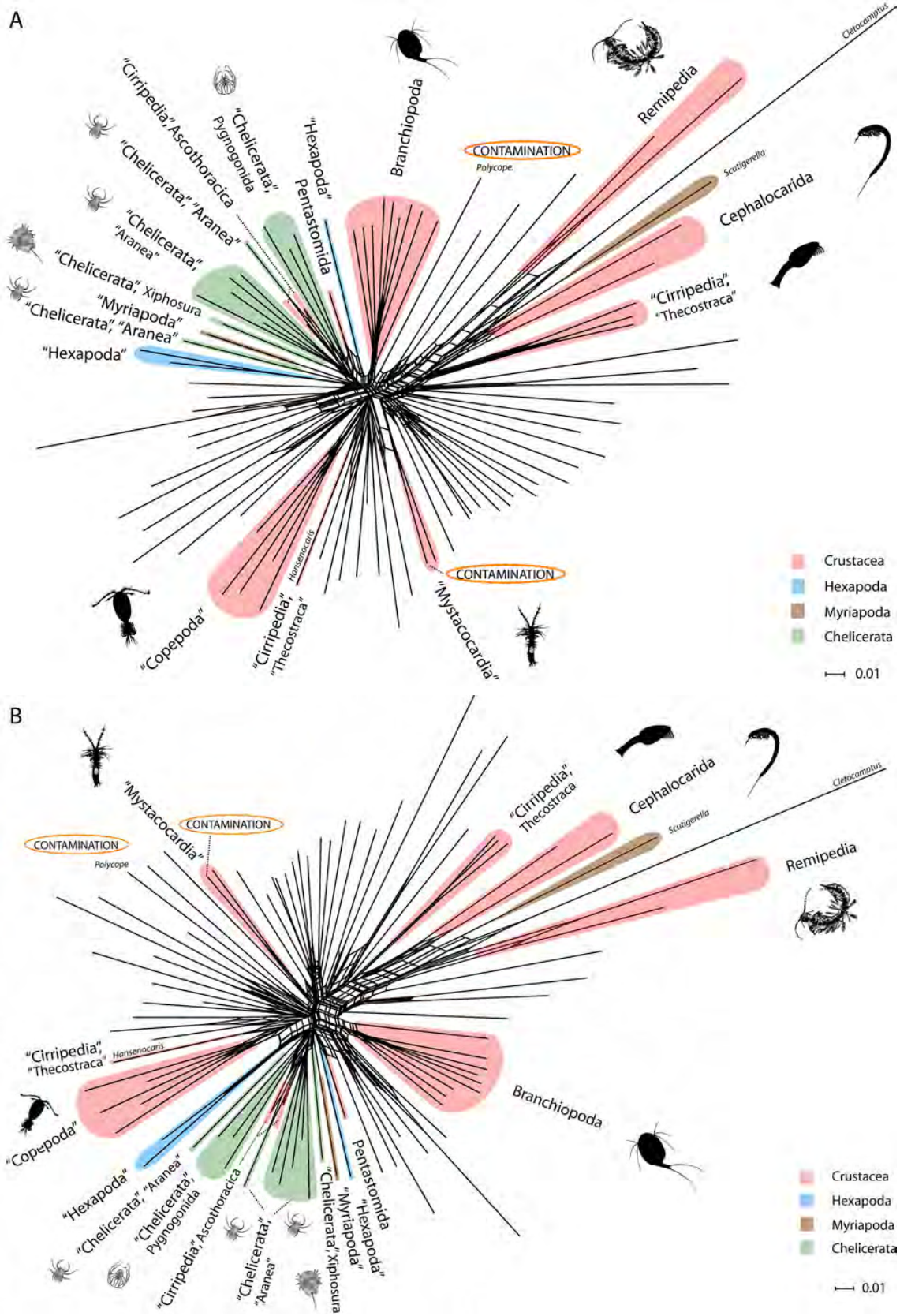


Figure 3.2: Networks of the manually aligned (A) and additionally manually optimized (B) dataset.

3.1.2 Base compositions

Heterogeneous base composition was checked for the MAFFT-aligned, data matrix of Run 1-4 (table 2.2). Compositional base homogeneity was rejected for the total dataset (including all taxa) and also for the set of all crustaceans (P-value of P=0.000000). In contrast, base composition homogeneity could not be rejected for more restricted branchiopod and maxillopodan groups. RY coding was implemented to handle this inhomogeneous base composition.

Table 3.1: Base frequency test for the dataset, MAFFT-aligned without RY coding. P-values above 0.05 represent base homogeneity. A (-) indicates the taxa, which were excluded from the dataset (or group) before the base frequency test was applied. Name of corresponding most inclusive dataset in bold.

Groups	Taxa	P -value
All taxa	88	P = 0.00000000
(-) Myriapoda	83	P = 0.00000000
(-) Myriapoda, Arachnida	81	P = 0.00000000
(-) Myriapoda, Chelicerata	72	P = 0.00000000
(-) Myriapoda, Chelicerata, Tardigrada	71	P = 0.00000000
(-) Myriapoda, Chelicerata, Tardigrada, Onychophora	70	P = 0.00000000
(-) Myriapoda, Chelicerata, Tardigrada, Onychophora, Pterygota	63	P = 0.00000000
Crustacea	57	P = 0.00000000
(-) Cephalocarida	55	P = 0.00000000
(-) Remipedia	56	P = 0.00000169
(-) Remipedia, *Cletocamptus*	55	P = 0.00006150
(-) Remipedia, Cephalocarida, Pentastomida, Mystacocarida	52	P = 0.00018036
(-) Remipedia, Cephalocarida, Pentastomida,	53	P = 0.00021264
(-) Remipedia, Cephalocarida	54	P = 0.00035780
(-) Remipedia, Cephalocarida, Pentastomida, Mystacocarida, *Cletocamptus*	51	P = 0.00356132
(-) Remipedia, Cephalocarida, Pentastomida, *Cletocamptus*	52	P = 0.00398729
(-) Remipedia, Cephalocarida*, Cletocamptus*	53	P = 0.00601048
(-) Remipedia, Cephalocarida, *Cletocamptus, Thetissplecaris*	52	P = 0.00708631
(-) Remipedia, Cephalocarida, *Cletocamptus, Thetissplecaris*, Tanaidacea,	51	P = 0.01087916
(-) Remipedia, Cephalocarida, *Cletocamptus*, *Thetissplecaris,* Tanaidacea,, *Thetysbaena*	50	P = 0.02002612
(-) Remipedia, Cephalocarida, *Cletocamptus*, T*hetissplecaris*, Tanaidacea, *Thetysbaena*, *Speleogripphus*	49	P = 0.05957794
Branchiopoda	9	P = 0.99999856
Branchiopoda + Mystacocardia	10	P = 0.99826169
Branchiopoda + Mystacocardia + Copepoda	16	P = 0.30029120
Branchiopoda + Mystacocardia + Copepoda, (-) *Clethocamptus*	15	P = 0.82277081
Branchiopoda + Ostracoda (-) *Heterocypris* + Copepoda (-) *Clethocamptus*	17	P = 0.35031847
Branchiopoda + Ostracoda (-) *Heterocypris* + Copepoda (-) *Clethocamptus* + Cirripedia	24	P = 0.37823571

Copepoda	6	P = 0.16156874
Copepoda (-) *Clethocamptus*	5	P = 0.79434022
Ostracoda	4	P = 0.17542550
Ostracoda (-) *Heterocypris*	3	P = 0.28077139
Cirripedia	6	P = 0.23684645

3.1.3 Phylogenetic reconstruction

Hypothesis testing in a second BFT of the final runs favored again the MAFFT aligned, unlinked partitioned dataset (run 2) against the linked MAFFT dataset (2ln B_{10}= 628.52, harmonic mean lnL_0= 41813.19; lnL_1=42127.45) and the two MUSCLE aligned datasets (see supplementary table S9). The MUSCLE aligned datasets (final run 4 and 5) showed a better model likelihood compared to final run2 but the resulting topology was less resolved and obviously biased by the MUSCLE alignment (see resulting topologies of supplementary figures S1-S9).

Some general results are striking. First, as seen in the network reconstructions, the question of marker choice can be answered, which is not optimal to address crustacean phylogeny within arthropods. Only major results of the topologies will be given here in respect to that. Second, the topologies reconstructed from the manually aligned dataset are more resolved compared to the "processed" data. This is especially the case for deeper splits, e.g. Chelicerata or Myriapoda. Third, it can be stated that for the processed data MAFFT produces obviously more "reliable" topologies compared to the MUSCLE results. In the MUSCLE aligned dataset suspicious clustering of long branch taxa is found in the resulting topologies (final run 4 and 5, supplementary figures S4 and S5) uniting unrelated taxa with the longest branches: the symphylan *Scutigerella,* the malacostracan *Spelaeogriphus*, the remipedes, and the cephalocarids. The last point concerns the reliability of the topologies. In the better-resolved trees of the manually aligned data suspicious clustering is found of e.g. the two Mystacocarida, of which one is a contamination (see 3.1.4).

Resulting topologies in general (figure 3.5-3.6 and supplementary figures S2-S9) show certain higher-level crustacean clades, including Branchiopoda, Copepoda, Thecostraca, Cladoceromorpha and non crustacean clades as Insecta (=true insects, following NCBI taxonomy), Protura + Diplura, Arachnida and Pycnogonida (figure 3.3).

A monophyletic Crustacea that excludes hexapods was not recovered in any of the trees. However, as shown in the results summary of figure 3.3, the analyses differ substantially in their ability to recover other higher-level groupings, such as Pancrustacea, Hexapoda, Chelicerata, Myriapoda and Malacostraca. Moreover, none of the trees supports or resolves hypothesized high-level clades such as Myriochelata or alternatively Mandibulata. However, the focus was primarily to investigate crustacean phylogeny and to test the ability of standard gene fragments (to date still commonly used) to reconstruct crustacean phylogeny within arthropods.

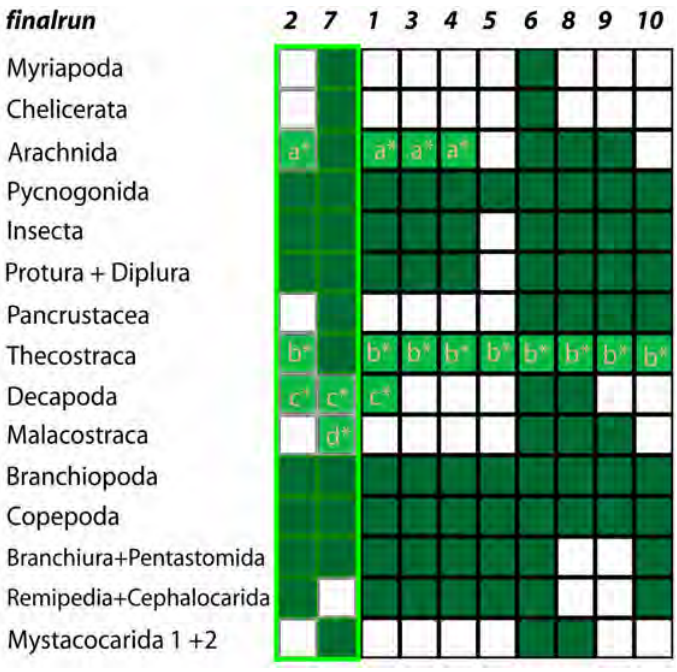


Figure 3.3: Navajo rugs showing the distribution of clades for all analyses. Dark-green squares indicate monophyly, while lighter green squares indicate monophyly with exeptions clarified by the aterisked* letters. White squares represent non monophyly for clades. Run 2 and 7 (resulting in favored topologies) are highlighted with a light green border.
a*= excluding *Amblyomma;* b*= *Heterosacchus*+ *Pollicipes* + *Semibalanus;* c*= excluding *Atoyida* d*= excluding *Speleogriphus*, clustering with Remipedia + Cephalocarida

Comparison of the two trees: In this section only the two favored trees of the manually aligned (figure 3.4) and "automatically processed" data (figure 3.5) will be presented. The resulting topologies of all other final runs are given in the supplement (supplementary figures S2-S9).

Only in the topologies of the manually aligned dataset (figure 3.4 and supplementary figures S6-S8) a monophyletic Arthropoda that excludes Onychophora is reconstructed. The trees show further well-supported monophyletic Myriapoda, Chelicerata, and Pancrustacea. They suggest two basic pancrustacean clades. The first clade includes the malacostracans, remipedes, cephalocarids and hexapods, while the other one contains the branchiopods and maxillopodans (copepods, thecostracans, mystacocarids, branchiurans, pentastomids and ostracodes, excluding the myodocopan ostracode *Polycope* (but see *Problematics of the data* below). In these trees, Maxillopoda is paraphyletic with respect to Branchiopoda. The trees agree consistent with dividing the maxillopodans across three clades, (1) Copepoda, (2) Cirripedia, and (3) Mystacocarida + Branchiura + Pentastomida + Ostracoda (except *Polycope*). Thecostraca is only a clade in figure 3.4. The other major clade includes hexapods, remipedes, cephalocarids, and malacostracans. However, although Insecta is monophyletic in these trees, Hexapoda is not (see *Problematics of the data*).

Though the deeper splits are not resolved in the tree for the automatically processed dataset, most crustacean groupings (see figure 3.3) are found. The tree differs from figure 3.4 in that the clade Thecostraca is not monophyletic, only regarding the taxa (*Heterosaccus*+(*Pollicipes*+*Semibalanus*)). This clade forms the sistergroup to the

Cephalocarida + Remipedia clade. A clade Maxillopoda is not supported and neither a monophyletic clade Malacostraca. The contaminated *Derocheilocaris* sequence is grouping with Myriapods.

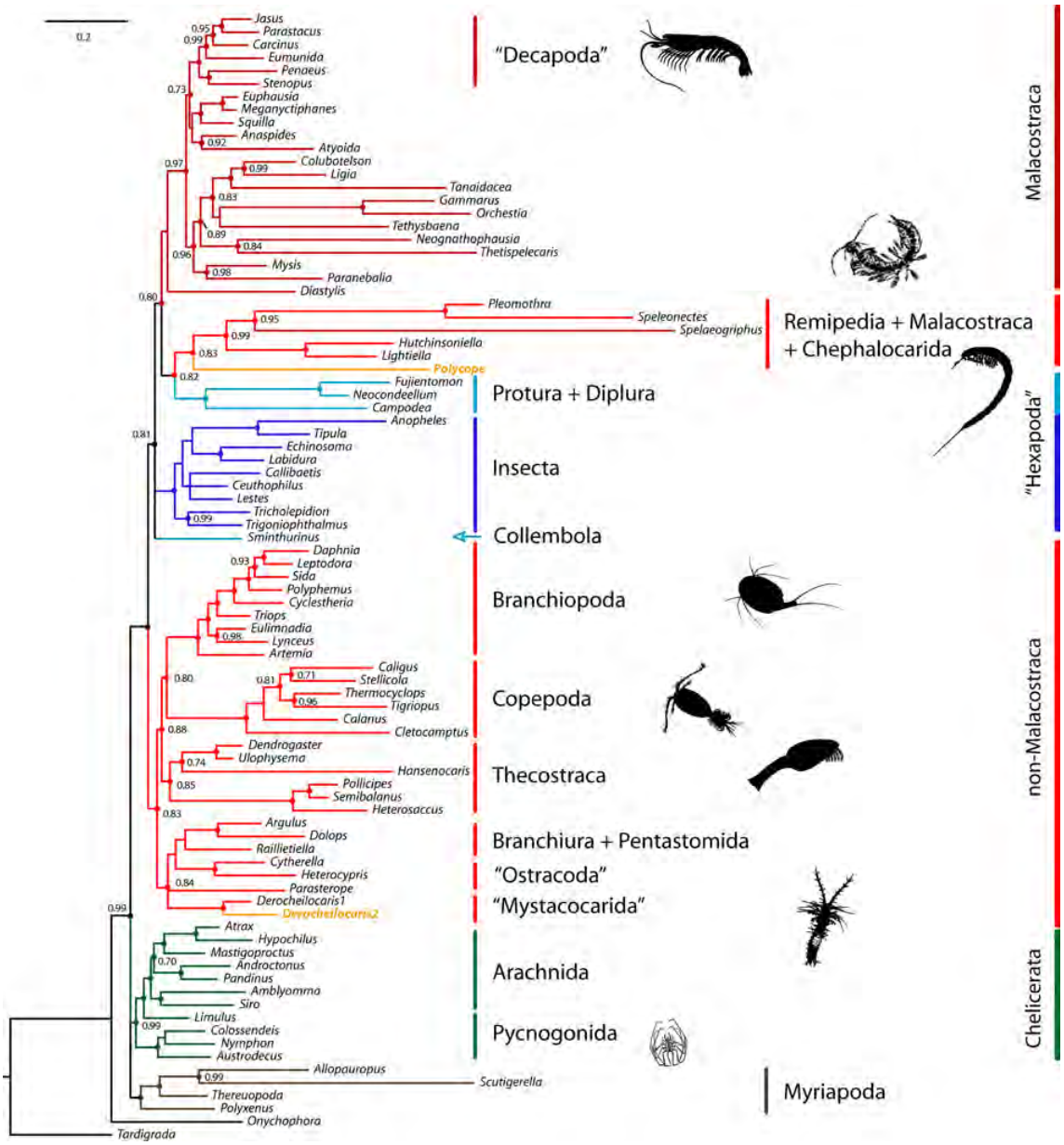


Figure 3.4: Resulting topology of the by hand aligned and optimized dataset (run 7). Bayesian Majority rule consensus tree with 40,000,000 Generations. Taxa groups are color coded: red= Malacostraca and non-Malacostraca, dark blue= Insecta, blue= basal Hexapods, green= Chelicerata, brown= Myriapoda. Out-group (tardigrades) and onychophorans are colored in black. Orange indicates contaminated sequences.

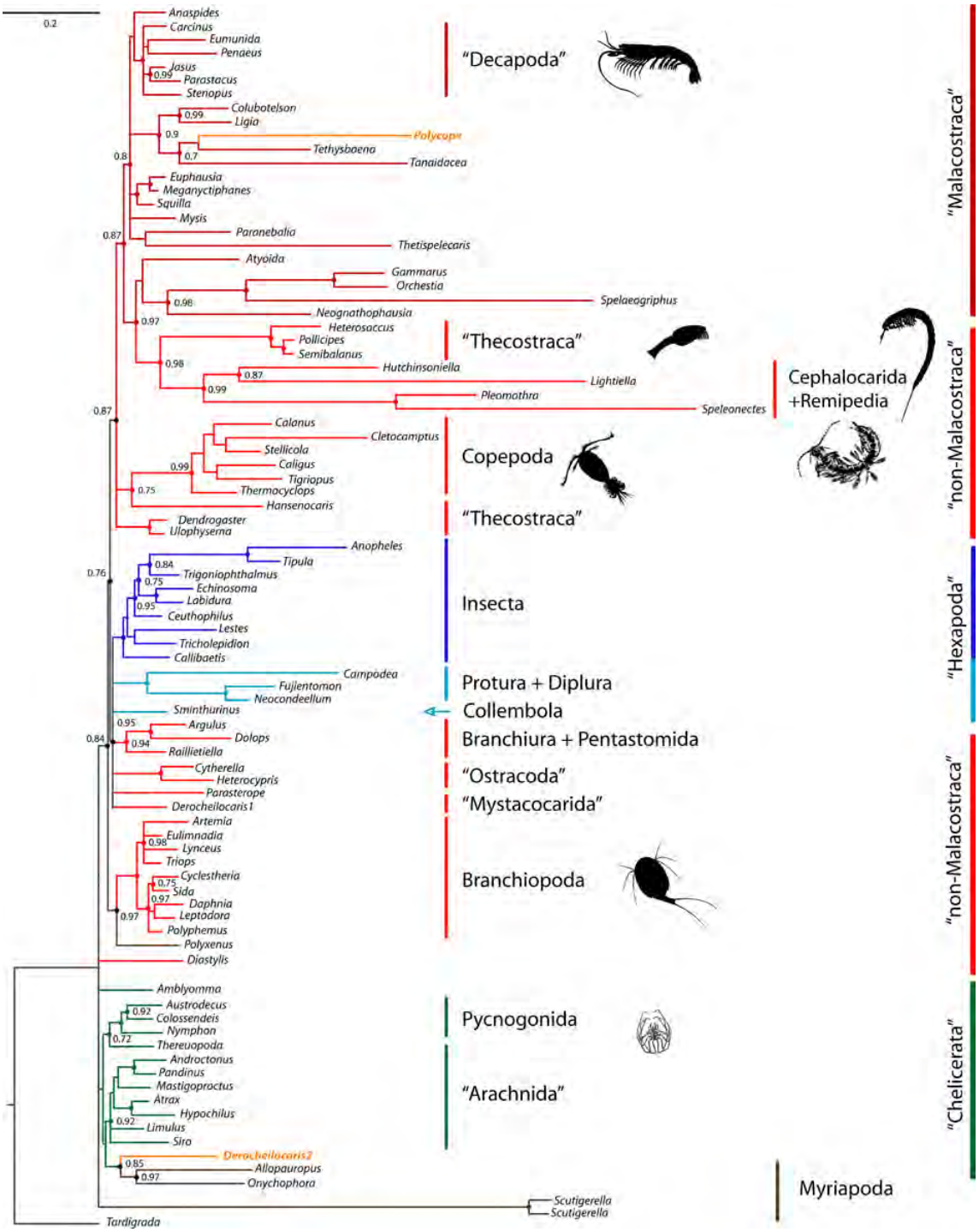


Figure 3.5: Resulting topology of the "processed" MAFFT-based dataset (run 2).

3.1.4 Problematics of the data

The placement of three taxa differs significantly between analyses. These taxa include the myodocopan ostracod *Polycope*, the cumacean *Diastylis*, and the mystacocarid *Derocheilocaris (sequence no. 2)*.

The ostracod *Polycope* is nested within a clade composed of (Remipedia + Cephalocarida) and (Diplura + Protura, figure 3.4). A BLAST search with the 18S sequence of *Polycope* indicates a high similarity with collembolans (springtails) and other non-ostracod, crustacean sequences, indicating possible contamination. YAMAGUCHI AND ENDO (2003), who included the 18S sequence of *Polycope* in a molecular analysis of Ostracoda, noted that the unusual length of their alignment was probably "owing to numerous inferred insertion and/or deletion events, especially in the sequences of *Polycope japonica*". They supposed that the position of *Polycope* in their tree was the result of long-branch attraction.

Similarly, all the highest BLAST hits for the 18S sequence of the *Diastylis* sequence are echinoderms, which would explain its basal position (figure 3.5) far apart from other Malacostraca. In the manually aligned dataset its position within the Malacostraca is problematic suggesting effects of subjective alignment arrangements.

The highest BLAST hits for the mystacocarid *Derocheilocaris (sequence no. 2)* 18S sequence are mites, and this is consistent with the finding of the mite *Acarus* being the sister taxon of *Derocheilocaris* in the 18S phylogeny of WHEELER ET AL. (2004). Intriguingly, however, this old mystacocarid sequence with high support is the sister taxon to the newly sequenced mystacocarid in the results of the manually aligned dataset (figure 3.4), and this clade groups within one of the maxillopodan clades. A few years ago a BLAST result for this sequence was relating to fungi, blasting this sequence to date gives closer hits to mites (see also more detailed in supplementary figure S10). As for *Diastylis,* its position close to *Derocheilocaris* (sequence no. 1) in the tree derived from the manually aligned dataset is suspicious.

The phylogenetic positions of some taxa in the trees may have been affected by long-branch attraction, including, possibly among others, the remipedes, the peracaridan *Spelaeogriphus*, the cephalocarids and the myriapod *Scutigerella*.

3.2 Analyses [B]: Is secondary structure based alignment optimization and implementation of time-heterogeneity a solution to improve inference of crustacean phylogeny?

3.2.1 Final dataset and split supporting patterns

Alignment filtering and concatenation of data: After the exclusion of randomly similar sections identified with ALISCORE, 1630 positions (originally 3503) of the 18S rRNA and 2472 (originally 8184) positions of the 28S rRNA remained. Filtered alignments were concatenated and comprised 4102 positions.

The neighbor-net graph, which results from a split decomposition based on uncorrected p-distances (figure 3.6) and LogDet correction plus invariant sites model (see figure 3.7) pictures a dense network, which hardly resembles a tree-like topology. This indicates the presence of some problems typical in studies of deep phylogeny: a) Some taxa like Diptera (which do not cluster with ectognathous insects), Diplura, Protura and Collembola each appear in a different part of the network, with Diplura and Protura separated from other hexapods. *Lepisma saccharina* is clearly separated from the second zygentoman *Ctenolepisma* that is nested within Ectognatha. Symphyla, Pauropoda, as well as Remipedia and Cephalocarida have very long branches.

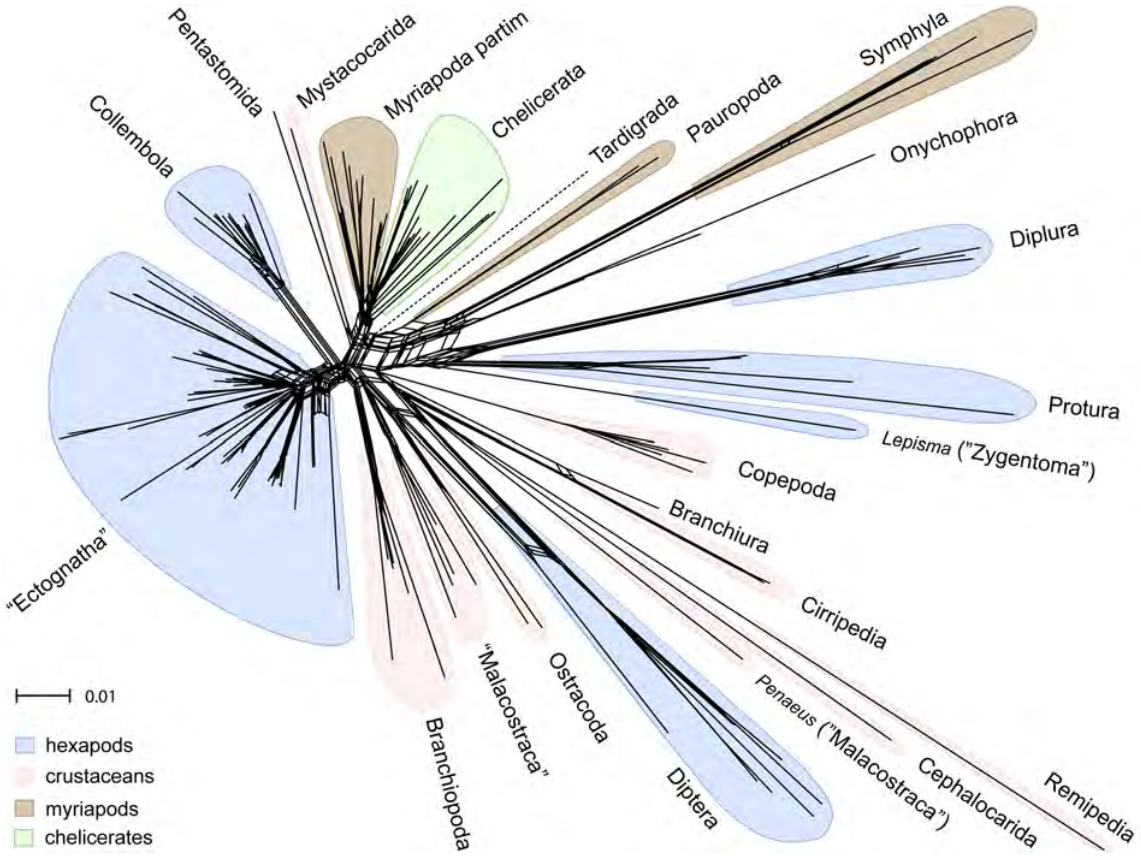


Figure 3.6: Neighbor-net graph of the concatenated 18S & 28S rRNA. Based on

uncorrected p-distances constructed in SPLITSTREE 4 after alignment filtering. Quotation marks indicate that monophyly is not supported in the given network graph. For the color code of groupings see graph.

The taxa may be misplaced in tree topologies due to signal erosion or occurrence of homoplasies, and their placement in trees must be discussed critically (WÄGELE & MAYER 2007). The usage of the LogDet distance adjusts the length of some branches but does not decrease the amount of conflicts in deep divergence splits. The inner part of the network shows little treeness, which indicates a high degree of conflicting signal.

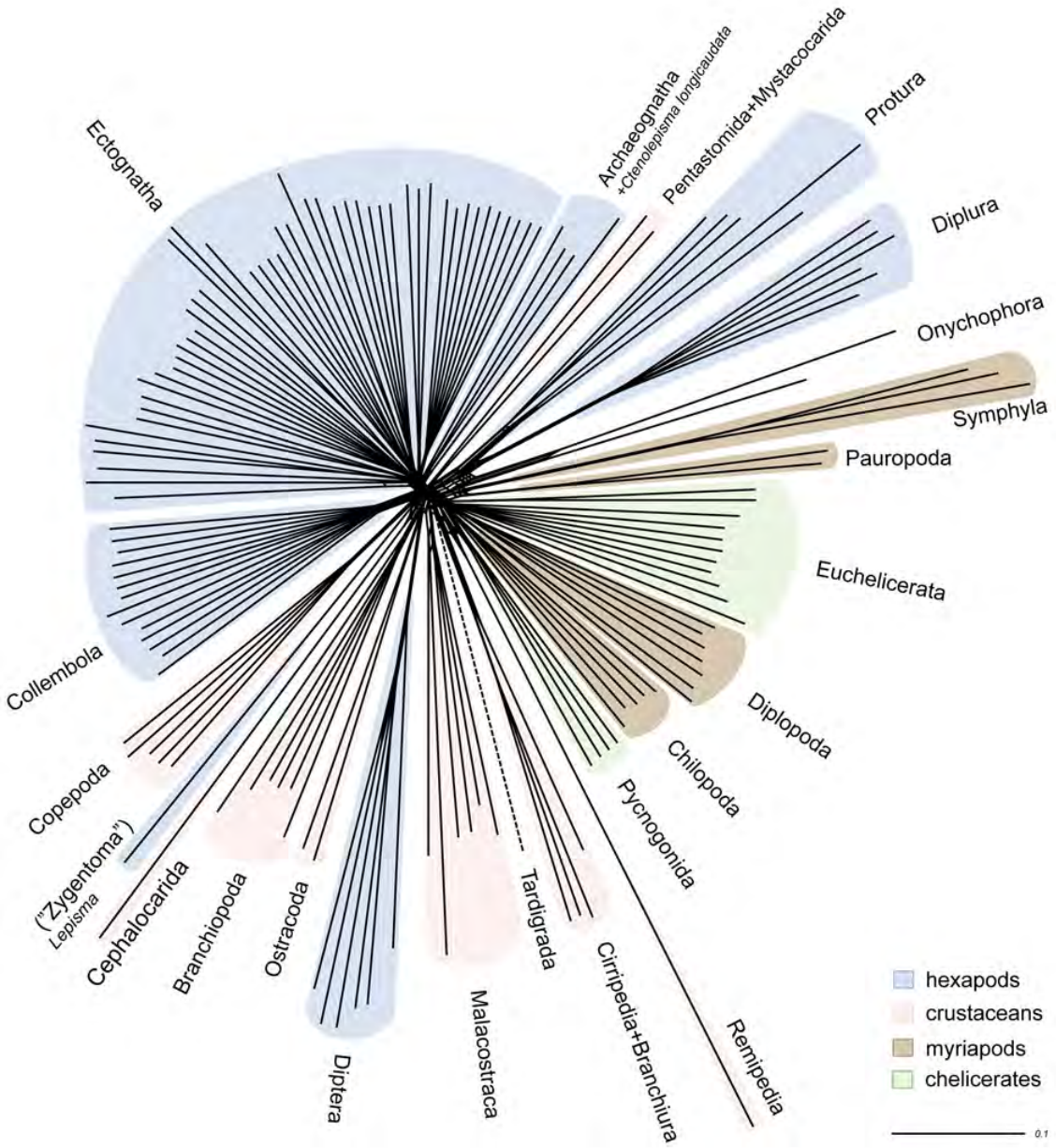


Figure 3.7: Neighbor-net graph based on LogDet correction reconstructed in Splitstree 4. Quotation marks indicate that monophyly is not supported in the given

network. For color code see graphic.

A remarkable observation seen in both phylogenetic networks is that some taxa have long stem-lineages, which means that the species share distinct nucleotide patterns not present in other taxa. Such well separated groups are Copepoda, Branchiopoda, Cirripedia, Symphyla, Collembola, Diplura, Protura and Diptera, while e.g. Myriapoda *partim*, Chelicerata and the Ectognatha (bristletails, silverfish/firebrats and pterygote insects) excluding Diptera share weaker patterns. This result is in line with the resulting patterns of the previous analysis [A]. It can be stated here that the strategy to change to completely sequenced 18S and 28S rRNA sequences results in a more suitable signal in the dataset compared to the previous one of analysis [A].

3.2.2 Compositional heterogeneity of base frequency

For the base compositional heterogeneity test parsimony uninformative positions were explicitly excluded with PAUP 4.0b10 (SWOFFORD 2002). Randomly similar alignment blocks identified with ALISCORE were excluded for both, the base compositional heterogeneity test and phylogenetic reconstructions. 901 characters of the 18S rRNA and 1152 characters of the 28S rRNA were separately checked for inhomogeneous base frequencies. Results led to a rejection of the null hypothesis H_0, which assumes homogeneous base composition among taxa (18S: chi^2=1168.94, df=441, P=0.00; 28S: chi^2=1279.98, df=441, P=0.00). Thus, base frequencies significantly differed across taxa in both 18S and 28S datasets.

A data partition into stems and loops revealed 477 unpaired positions and 424 paired positions in the 18S alignment, and 515 unpaired and 637 paired positions in the 28S alignment. Separate analyses of all four partitions confirmed heterogeneity of base frequencies across taxa in all sets (P=0.00 in all four partitions).

The homogeneity test was repeated for partitions as used in tree reconstruction. If base pairs were disrupted by the identification of the corresponding partner as "randomly similar " by ALISCORE, the remaining (formerly paired) positions were treated as unpaired. Hence, 1848 characters of the concatenated alignment (18S: 706; 28S: 1142) were treated as paired in all analyses. Again, the test revealed heterogeneity in unpaired characters of both the 18S and 28S alignments (P=0.00 for both genes; 18S: 506 characters; 28S: 567 characters). Examination at paired positions also rejected the null hypothesis H_0 (18S, 395 characters included: P<0.0003, 28S, 585 characters included: P=0.00). Since non-stationary processes were strongly indicated in all tests, we chose to apply time-heterogeneous models to account for lineage-specific substitution patterns. To fix the number of "free base frequency sub-models" in time-heterogeneous analyses, the minimal exclusive set of sequence groups was identified. Based on chi^2-tests the dataset could be divided into three groups for both rRNA genes. In both genes Diptera are characterized by a high A/T content and Diplura by a low A/T content. Exclusion of only one of the groups was not sufficient to retain a homogeneous dataset (18S: excluding Diptera: chi^2=972.91, df=423, P=0.00, excluding Diplura: chi^2=532.13, df=423, P<0.0003; 28S: excluding Diptera: chi^2=986.72,

df=423, P=0.00, excluding Diplura: chi^2=813.8, df=423, P=0.00). Simultaneous exclusion of both groups led to acceptance of H_0 for 18S sequences (chi^2=342.22, df=405, P=0.99). For the 28S set, after exclusion of both groups, H_0 was still rejected (chi^2=524.98, df=405, P<0.0001). After sorting taxa according to base frequencies in ascending order, additional exclusion of *Peripatus* sp. and *Sinentomon erythranum* resulted in a homogeneous base composition for the 28S gene (H_0: chi^2=434.99, df=399, P=0.1), likewise indicating that three sub-models are sufficient to cover the taxon set. The homogeneity-test was repeated for stem and loop regions of each gene separately. The exclusion of Diplura was sufficient to obtain homogeneity in the loop regions for both genes (18S: 474 characters, P=0.9757; 28S: 541 characters, P=0.0684). For stem regions in the 18S set it likewise was sufficient to exclude either Diptera (378 characters, P=0.6635) or Diplura (385 characters, P=0.99). These partitions would make two sub-models sufficient to cover the dataset. However, in the stem regions of the 28S homogeneity was obtained only after the exclusion of both Diptera and Diplura (547 characters, P=0.99). Since PHASE-2.0 does not allow to vary the number of chosen sub-models among partitions, three sub-models were applied and fitted to each data partition.

3.2.3 Phylogenetic model testing & reconstructions

Three combinations of mixed DNA/RNA models (REV +Gamma & RNA16I +Gamma, TN93 +Gamma & RNA16J +Gamma and HKY85 +Gamma & RNA16K +Gamma) were compared to select the best model set. Overall model ln likelihoods converged for all tested mixed models after a burn-in of 250,499 generations in an initial pre-run of 500,000 generations. However, most parameters did not converge for the combined REV +Gamma & RNA16I +Gamma models, consequently, this set up was excluded from further analyses. For each of the remaining two sets a chain was initiated for 3 million generations, with a burn-in set to 299,999 generations. The applied Bayes Factor Test (KASS & RAFTERY 1995; NYLANDER ET AL. 2004), favored the TN93 +Gamma & RNA16J +Gamma model combination (2ln B_{10}=425.39, harmonic mean ln L_0 (TN93 +Gamma & RNA16J +Gamma)=79791.08; harmonic mean ln L_1 (HKY85 +Gamma & RNA16K +Gamma) =80003.78. For each approach (table 2.4) all chains that passed a threshold value in a BFT were assembled to a metachain. Harmonic means of the ln likelihoods of included time-heterogeneous chains were compared against all ln likelihoods of included time-homogeneous chains (burn-in discarded) in a final BFT: the time-heterogeneous model was strongly favored (2ln B_{10}=1362.13). Each resulting "extended majority rule consensus tree" was rooted with *Milnesium*. Node support values for clades were deduced from 56,000 sampled trees for the time-heterogeneous set (figure 3.8) and from 18,000 sampled trees for the time-homogeneous set (figure 3.9). Detailed support values are shown in table 3.2).

Table 3.2: Bayesian support values for selected clades. Values for the time-homogeneous and the time-heterogeneous tree are given. pP= posterior probability value.

Clade	pP: time-hetero-geneous	pP: time-homo-geneous
Symphyla	1.0	1.0
(Pauropoda,Onychophora)	0.97	1.0
Pauropoda	1.0	1.0
Onychophora	1.0	1.0
Chelicerata	0.91	1.0
Pycnogonida	1.0	1.0
Euchelicerata (without Pycnogonida)	0.89	1.0
Myriapoda partim (excl. Symphyla & Pauropoda): (Diplopoda,Chilopoda)	0.97	0.98
Diplopoda	0.99	1.0
Chilopoda	1.0	1.0
Myriochelata partim: ((Diplopoda,Chilopoda)(Euchelicerata,Pycnogonida))	0.97	1.0
(Myriochelata partim,Pancrustacea)	0.95	0.98
Pancrustacea	1.0	1.0
((*Derocheilocaris*,Ostracoda)(*Speleonectes*(*Argulus*,Cirripedia))	0.33	-
(*Derocheilocaris*,Ostracoda)	0.62	-
((*Derocheilocaris,Raillietiella*)(*Speleonectes*(*Argulus*,Cirripedia))	-	0.59
(*Derocheilocaris,Raillietiella*)	-	0.75
(*Speleonectes*(*Argulus*,Cirripedia))	0.65	0.73
(*Argulus*,Cirripedia)	1.0	1.0
(Ostracoda,Malacostraca)	-	0.99
((Ostracoda,Malacostraca)(((*Derocheilocaris,Raillietiella*) (*Speleonectes*(*Argulus*,Cirripedia))))	-	0.61
(Malacostraca(*Raillietiella*((*Hutchinsoniella,*Branchiopoda) (Copepoda,Hexapoda))))	0.44	-
Malacostraca	1.0	1.0
(Raillietiella((*Hutchinsoniella,*Branchiopoda)(Copepoda,Hexapoda)))	0.60	-
((*Hutchinsoniella,*Branchiopoda)(Copepoda,Hexapoda))	0.65	-
(*Hutchinsoniella,*Branchiopoda)	0.59	-
Branchiopoda	1.0	1.0
(Copepoda,Hexapoda)	0.67	-
((Copepoda((*Lepisma,Hutchinsoniella*)(remaining hexapod taxa)))	-	0.70
((*Lepisma,Hutchinsoniella*)(remaining hexapod taxa))	-	0.58
Hexapoda	0.96	-
Entognatha: ((Protura,Diplura)(Collembola))	0.98	-
Nonoculata: (Protura,Diplura)	0.98	1.0
((*Lepisma,Hutchinsoniella*)(Protura,Diplura))	-	0.72
(*Lepisma,Hutchinsoniella*)	-	0.72
Protura	1.0	1.0
Diplura	1.0	1.0
Collembola	1.0	1.0
Ectognatha: (Archaeognatha(Zygentoma,Pterygota)	1.0	-
(Archaeognatha(*Ctenolepisma*,Pterygota))	-	1.0
Archaeognatha	1.0	1.0
Zygentoma	0.98	-

Dicondylia: (Zygentoma,Pterygota)	0.99	-
(*Ctenolepisma*,Pterygota)	-	0.99
Pterygota	0.97	0.94

3.2.4 Resulting topologies

In the following section, the focus lays on major arthropod clades and further on crustaceans and their position within arthropods. First similarities [1] and then differences [2] between the time-heterogeneous tree (figure 3.8) and time-homogeneous tree (figure 3.9) are pointed out. Hexapod clades are given in the trees (for a complete overview) but are only shortly mentioned in the text, except basal hexapods that are important for crustacean phylogeny as an eventual sister-group to crustaceans. In some cases the results of this far more sophisticated analysis are compared with the previous analysis [A].

Representatives of Symphyla and Pauropoda, already identified in the neighbor-net graph as taxa with conspicuously long branches (figure 3.6 and 3.7), assumed unorthodox positions in both trees which are clearly incongruent with morphological evidence and results obtained from other genes. Symphyla formed the sister group of all remaining arthropod clades, and Pauropoda clustered with Onychophora. Consequently, myriapods always appeared polyphyletic in both analyses. These results are considered as highly unlikely, since they contradict all independent evidence from morphology, development, and partly from other genes.

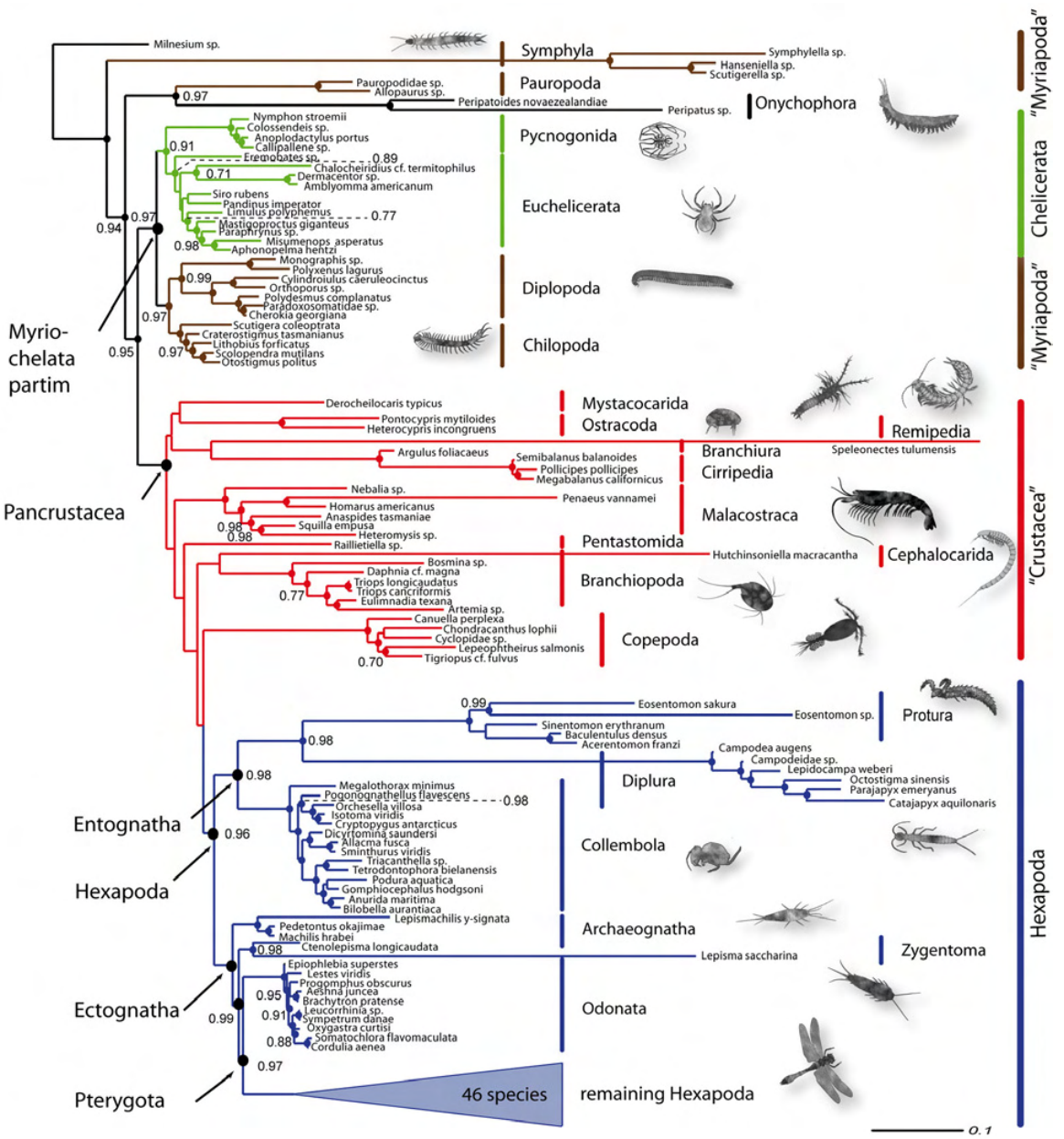


Figure 3.8: Time-heterogeneous consensus tree of the 18S and 28S dataset optimized considering secondary structure information. Consensus tree from 56.000 sampled trees of the time heterogeneous substitution process inferred with PHASE 2.0. Support values below 0.70 are not shown (nodes without dot) nodes with a maximum posterior probability (pP) of 1.0 are represented by dots only. Quotation marks indicate that monophyly is not supported in the given tree. Crustaceans are colored red, hexapods blue, chelicerates green and myriapods brown. Out-groups are black.

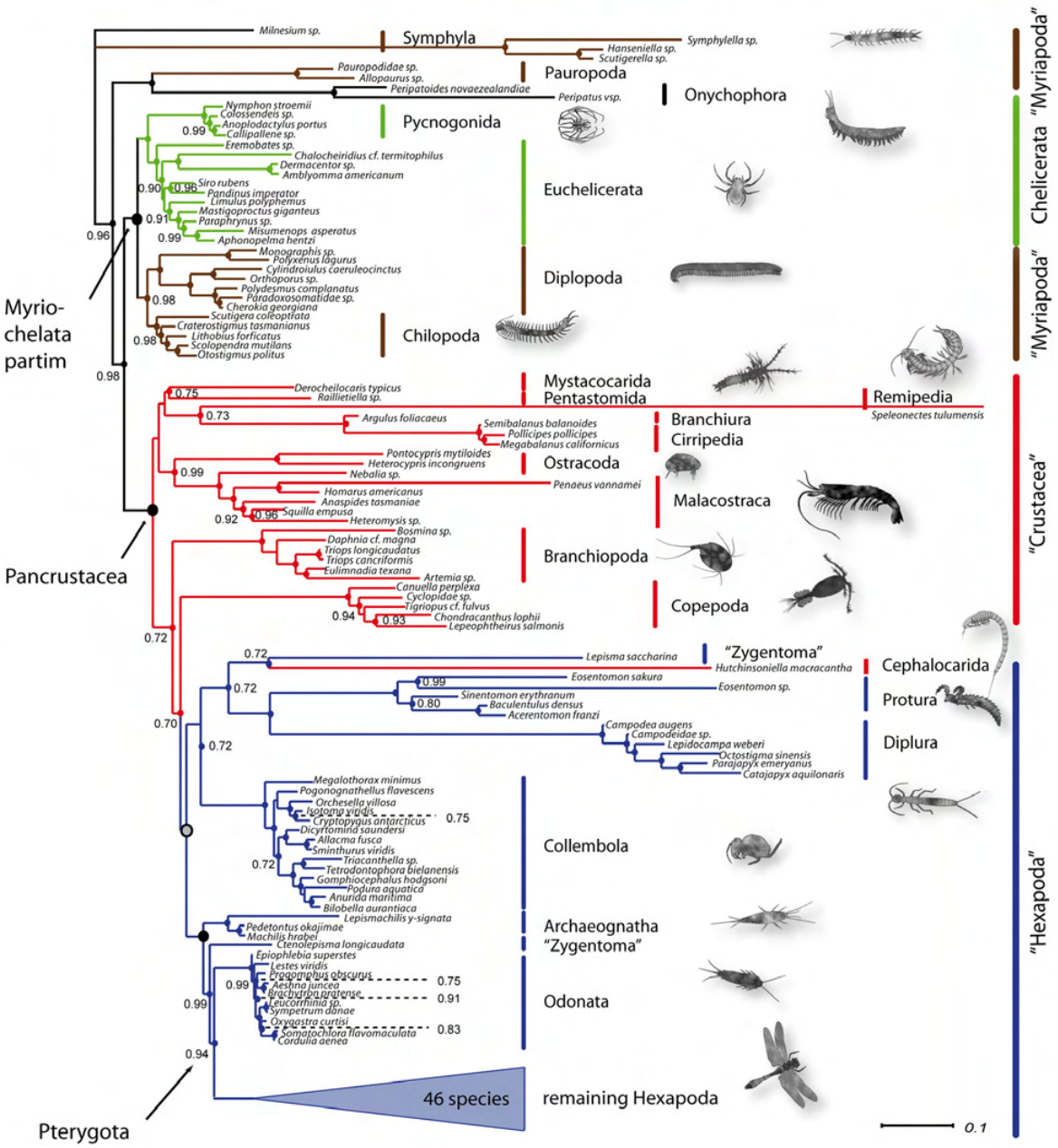


Fig 3.9: Time-homogeneous consensus tree of the 18S and 28S dataset optimized considering secondary structure information. Same setting and color-coding as given in figure 3.9**.** The grey dot indicates the clade containing all hexapod taxa including Hutchinsoniella (Crustacea) + Lepisma (Zygentoma); its node value is pP 0.58.

Congruent results [1]: Monophyletic Chelicerata are supported in the time-heterogeneous tree (pP 0.91) and with maximal support in the time-homogeneous tree with Pycnogonida (sea spiders) as sister group to remaining chelicerates. **Pycnogonida** received maximal support in both analyses. **Eucheliceratа** received highest support in the time-homogeneous approach while this clade in the time-heterogeneous approach received a

support of only pP 0.89. *Limulus polyphemus* (horseshoe crab) clustered within arachnids, but some internal relationships within Euchelicerata received only low support.

Chilopoda always formed the sister group of a monophyletic Diplopoda in both analyses with high support. Within the latter the most ancient split lied between Penicillata and Helminthomorpha. This myriapod assemblage "Myriapoda partim" formed the sister group of Chelicerata, thus giving support to the Myriochelata hypothesis, respectively Myriochelata partim, when the long-branch clades Symphyla and Pauropoda are disregarded.

Pancrustacea showed always maximal support. The monophyly of Malacostraca and of Branchiopoda received highest support in both approaches while their position varied. Branchiopoda was the sister group of a clade consisting of Copepoda + Hexapoda in the homogeneous tree (figure 3.9), however the cephalocarid *Hutchinsoniella* nested within hexapods.

Among hexapods, monophyly was unambiguously supported for Protura, Diplura, Collembola, Archaeognatha and Odonata (other Hexapoda are not discussed here). Diplura clustered with Protura, and gave support to a monophyletic Nonoculata.

Differing results [2]: While the time-heterogeneous and time-homogeneous trees corresponded in overall topologies, they differ in a number of remarkable details.

1) The cephalocarid *Hutchinsoniella* clustered among crustaceans as sister group to the Branchiopoda only in the heterogeneous approach. This clade formed the sister group to (Copepoda +Hexapoda), although with low support.

2) The time-homogeneous runs revealed highly supported (Malacostraca + Ostracoda) as the sister group to a clade ((Mystacocarida + Pentastomida) + (Branchiura + Cirripedia)). In contrast, in the time-heterogeneous analysis more terminal positioned Malacostraca are the sister group of a clade (Pentastomida ((Cephalocarida + Branchiopoda) + (Copepoda + Hexapoda))). The altered position of Pentastomida was only weakly supported in this tree.

3) In the homogeneous tree *Hutchinsoniella* emerged as sister taxon to *Lepisma* with low support (pP 0.72), and this cluster was positioned within the remaining hexapods (figure 3.9). Well-supported Hexapoda were monophyletic only in the time-heterogeneous approach, (pP 0.96,), with Copepoda as sister group, the latter with low support (pP 0.69).

4) In the time-homogeneous tree, Copepoda emerged as sister group, again with a low support value (pP 0.70) of *Lepisma* + *Hutchinsoniella* + "Hexapoda"').

5) Hexapoda (pP 0.96), Entognatha (pP 0.98) and Ectognatha (pP 1.0) were monophyletic only in the time-heterogeneous tree.

6) The time-heterogeneous tree showed the expected paraphyly of primarily wing-less insects with Archaeognatha as sister group to Zygentoma + Pterygota.

3.3 Analyses [C]: Do phylogenomic data enlight crustacean phylogeny within arthropods - or do old problems stick to the analysis of this new large-scale data?

3.3.1 Resulting topology of the unreduced dataset

Maximum likelihood tree reconstructions (STAMATAKIS 2006) based on the total unreduced supermatrix (see figure 2.11) showed some clades, e.g. Euarthropoda + Onychophora, Euchelicerata and Malacostraca with high support. However, several groups received only moderate support, like pancrustaceans (89% bootstrap support) or weak support, e.g. chelicerates or endopterygote insects, and many groups appeared para- or polyphyletic (Myriapoda, Hexapoda, see figure 3.10). The crustacean group Peracarida is not recovered as a monophyletic clade.

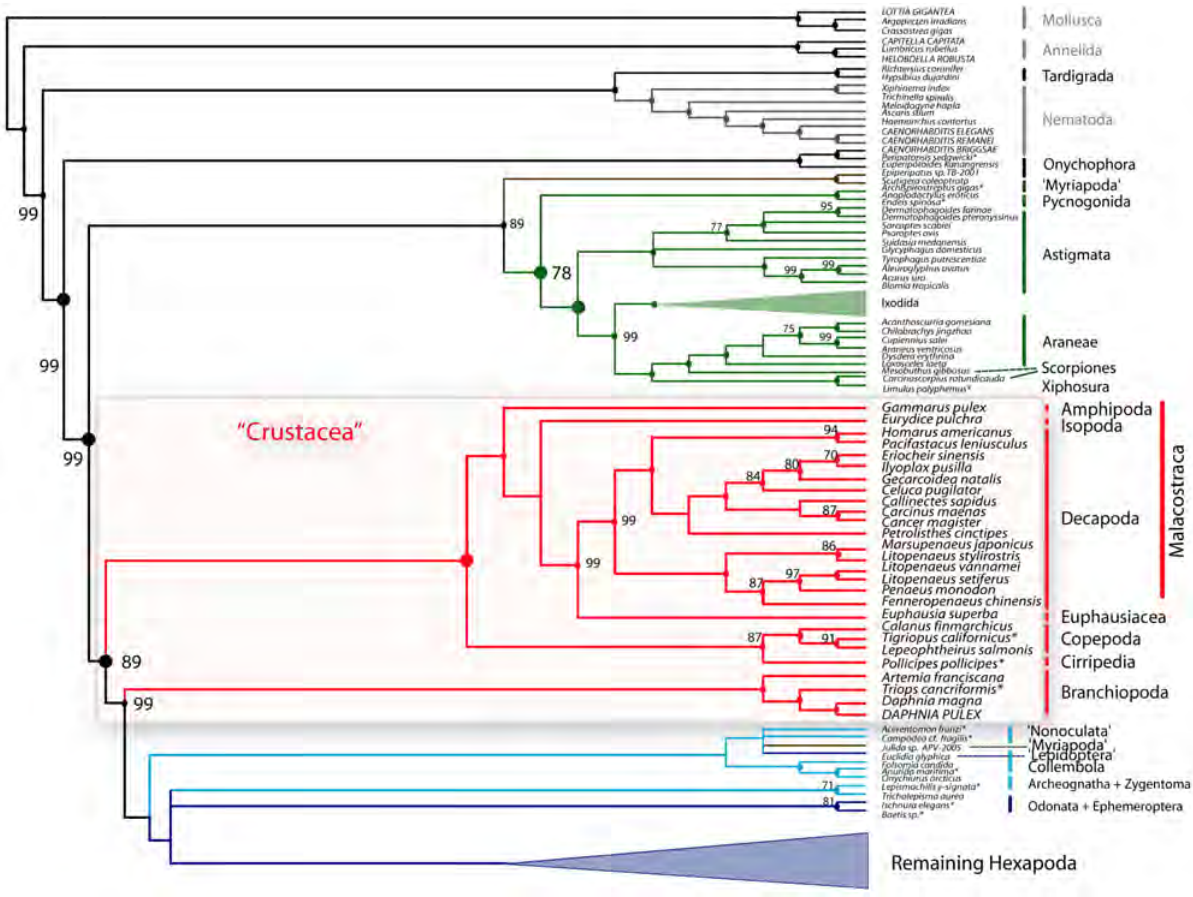


Figure 3.10: Cladogram of the 233-taxon RAXML analysis of the unreduced original dataset. Crustaceans are accentuated. ML tree search and bootstrapping (PROTMIX substitution model and WAG matrix) was conducted separately (see methods). Support values are derived from 100 bootstrap replicates. Support values (majority rule consensus tree) < 70: not shown, Support values = 100: represented by a dot only. Quotation marks indicate non-monophyly. Color code: molluscs, annelids and nematodes: lighter grey; tardigrades,

onychophorans: black; myriapods: brown; chelicerates: green; crustaceans: red; basal hexapods: light blue; pterygote insects: dark blue.

In contrast, the selected optimal subset of taxa and genes (see 3.3.2) yielded improved resolution of and strong support for nearly all groups.

3.3.2 Resulting topologies of the reduced, optimal data subset

The selected optimal subset included 117 taxa with 101 euarthropods (chelicerates, myriapods, crustaceans and hexapods), two onychophorans, two tardigrades and outgroups. The dataset comprised 129 genes of which 32 were ribosomal proteins (supplementary table S15). The concatenated masked alignment span more than 37 Mb amino acid positions (figure 2.12). Selecting a subset of the data (see 2.6.3 figures 2.11 - 2.13) raised the overall relative information content fourfold to 0.43 from originally 0.10. The overall matrix saturation (data availability) increased threefold to 62.3% from originally 17.6%. Taxa in our optimal dataset presented on average 84 genes (minimum 35, maximum 129) that were present on average in 76 taxa (minimum 46, maximum 109 taxa per gene). The relative information content of single genes varied from 0.42 - 0.92 with an average of 0.7 (supplementary table S15).

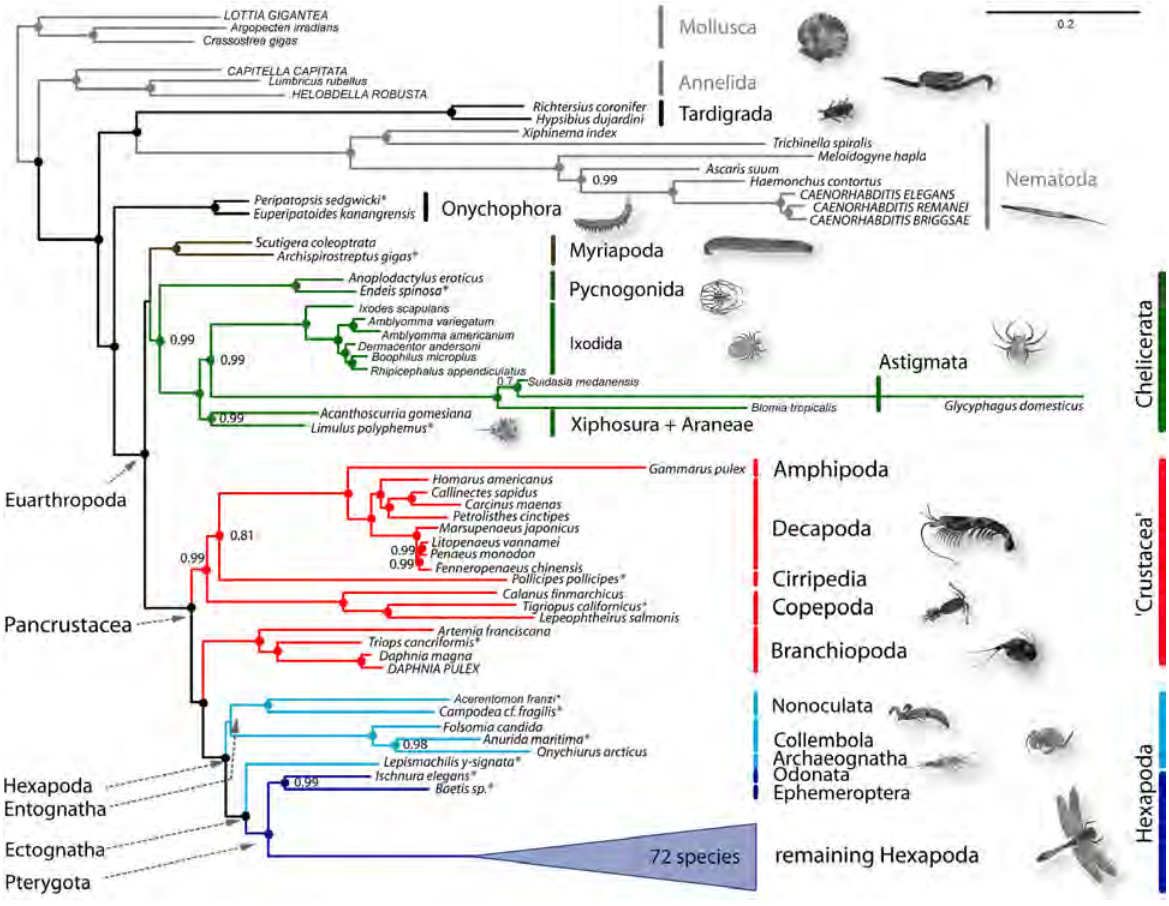


Figure 3.11: Phylogram of the 117-taxon Bayesian analysis. Bayesian majority rule

consensus tree (optimal data subset, 3 chains out of 25 chains, 20,000 cycles each, burn-in: 5,000 cycles). Posterior probabilities (pP) are estimated under the CAT mixture model (LARTILLOT & PHILIPPE 2004). The majority rule consensus tree is based on the 'triple' set (three chains) showing lowest maxdiff value (0.186) while each of these chains had the best harmonic mean of the likelihood values (burn-in excluded) of all possible 'triple'-chain combinations. pP-values < 0.7: not shown, pP-values = 1.0: represented by a dot only. Quotation marks indicate non-monophyly.

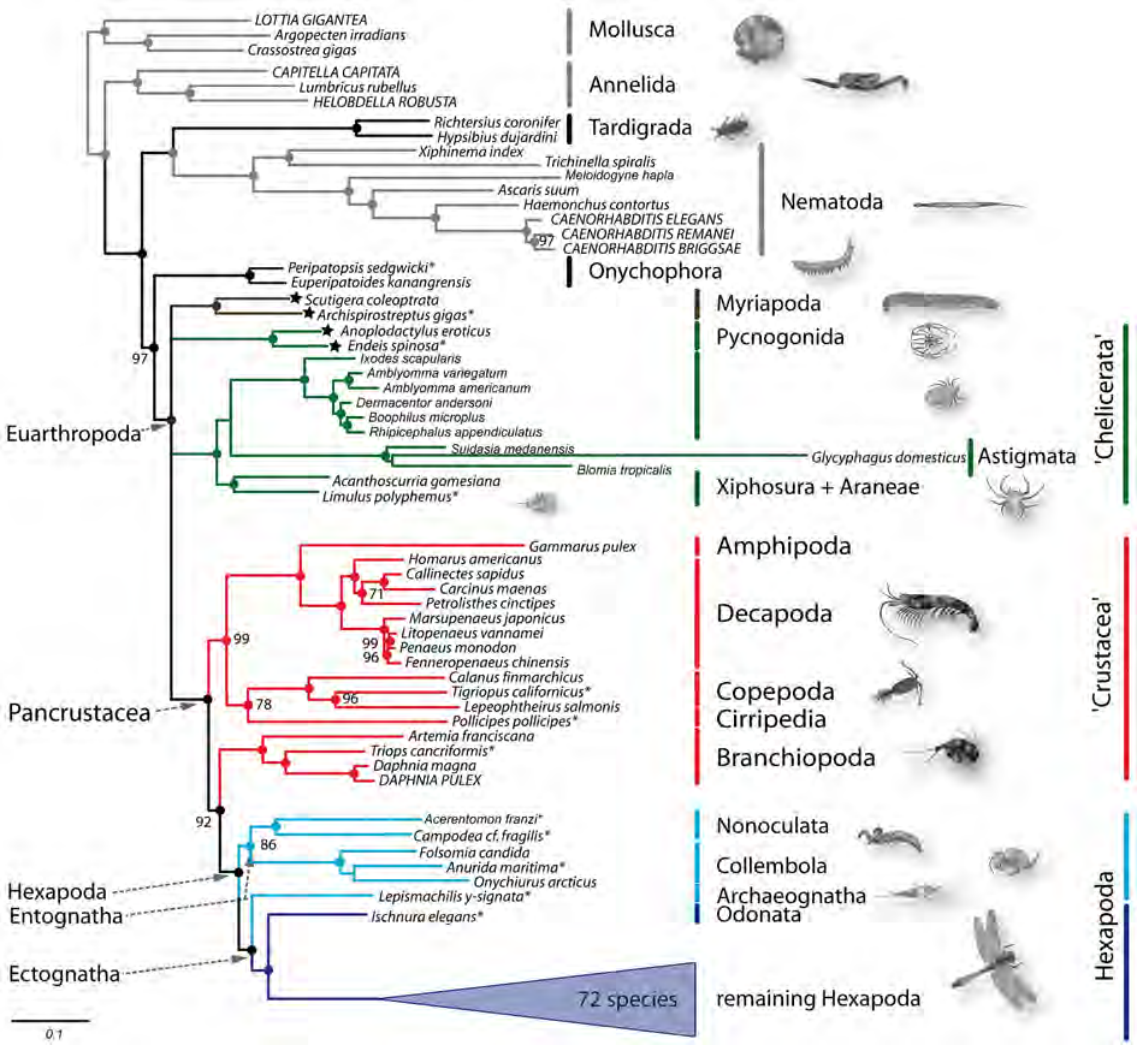


Figure 3.12: Phylogram of 117-taxon ML analysis. ML tree (majority rule consensus) of the optimal data subset (PROTMIX substitution model and WAG matrix). Support values are derived from 1,000 bootstrap replicates. Support values < 70: not shown, support values = 100: represented by a dot only. Quotation marks indicate non-monophyly. stars (*) after taxon names indicate EST taxa contributed by the authors. 'Instable' taxa (leaf stability index < 0.95) are marked by a star in front of the taxon name.

Euarthropoda: In general, maximum likelihood (WAG model including rate heterogeneity, STAMATAKIS 2006, figure 3.11) and Bayesian (CAT mixture model, LARTILLOT & PHILIPPE 2004, figure 3.12) analyses of the data subset showed strong support for Onychophora + Euarthropoda. Monophyletic myriapods, sea spiders and eucheliceratesw were recovered in ML and Bayesian trees while the resolution among these three groups differed between both approaches. Correspondingly, sea spiders (*Anoplodactylus, Endeis*) and millipeds (*Archispirostreptus, Scutigera*) showed relative low leaf stability indices (0.91; 0.91; 0.86; 0.86).

Pancrustacea were maximally supported in both approaches with paraphyletic crustaceans in relation to monophyletic hexapods.

Most extant arthropod orders, for example Ixodida (ticks) and Astigmata (mites) belonging to chelicerates, Decapoda, Copepoda (crustaceans) were strongly supported (Figure 3.11 and 3.12).

3.3.3 Differences in ML and Bayesian topologies of the reduced data subset

Some deep splits within arthropods, in previous studies thought to be resolved (DUNN ET AL. 2008) show remarkable sensitivity to available data and reconstruction methods.

Pycnogonida: While the position of sea spiders was not resolved in the ML tree (fig. 3.12), the Bayesian tree (fig. 3.11) showed monophyletic chelicerates with high support (posterior probability, pP 0.99), including sea spiders.

Limulus polyphemus (Xiphosura) is not separated from the Arachnida and groups in a clade with *Acanthoscurria* (Aranea).

Myriapoda emerged in the Bayesian tree (fig. 3.11) as the sister group to chelicerates with weak support. The relationships of myriapods, sea spiders, euchelicerates and pancrustaceans were not resolved in the ML tree reconstruction (figure 3.12) reconstructing a polytomic clade of these taxa.

Within Crustacea, cirripedes clustered with low support either with copepods (ML) or malacostracans (Bayesian). Branchiopoda were placed as sister group to Hexapoda with maximal support in the Bayesian approach and with moderate support (BS 92%) in ML analyses.

Entognatha (Protura, Diplura and Collembola) were recovered in ML and Bayesian approaches, albeit weakly supported. Within Entognatha, a strongly supported sistergroup relationship of Protura and Diplura was recovered.

3.3.4 Problematics in resulting topologies of the Bayesian chains

Topological incongruencies among the 25 Bayesian chains (topology of fig. 3.11) were visualized in a consensus network (HOLLAND & MOULTON 2003), see figure 3.13. Incongruent clustering of taxa within trees, e.g. placement of Archaeognatha, might result from different local optima of different chains. Incongruencies were caused by variable positions of few taxa.

[1] The barnacle *Pollicipes* (Cirripedia, Crustacea) emerged just in one chain as sister group to copepods with weak support (pP 0.51). However, also the alternative clade (*Pollicipes* + Malacostraca) (see fig. 3.13) showed a wide range from 0.56 – 0.96 (posterior probability) in different chains.

[2] The bristle tail *Lepismachilis* (Archaeognatha) was inferred as sister-goup to Blattaria + Isoptera in several trees, showing moderate or low support (posterior probability 0.52 - 0.82), additionally in these trees *Pediculus* (Phthiraptera) emerged as sister-group to this clade with maximal support. However, likelihoods (harmonic mean) of related chains were lower compared to chains used for the consensus tree and rejected after a BFT (NYLANDER ET AL. 2004).

[3] Mandibulata (Myriapoda + Pancrustacea) were found in trees of two chains maximally supported while Mandibulata + Chelicerata received negligible support (pP 0.52, pP 0.55) in chains of highest likelihoods.

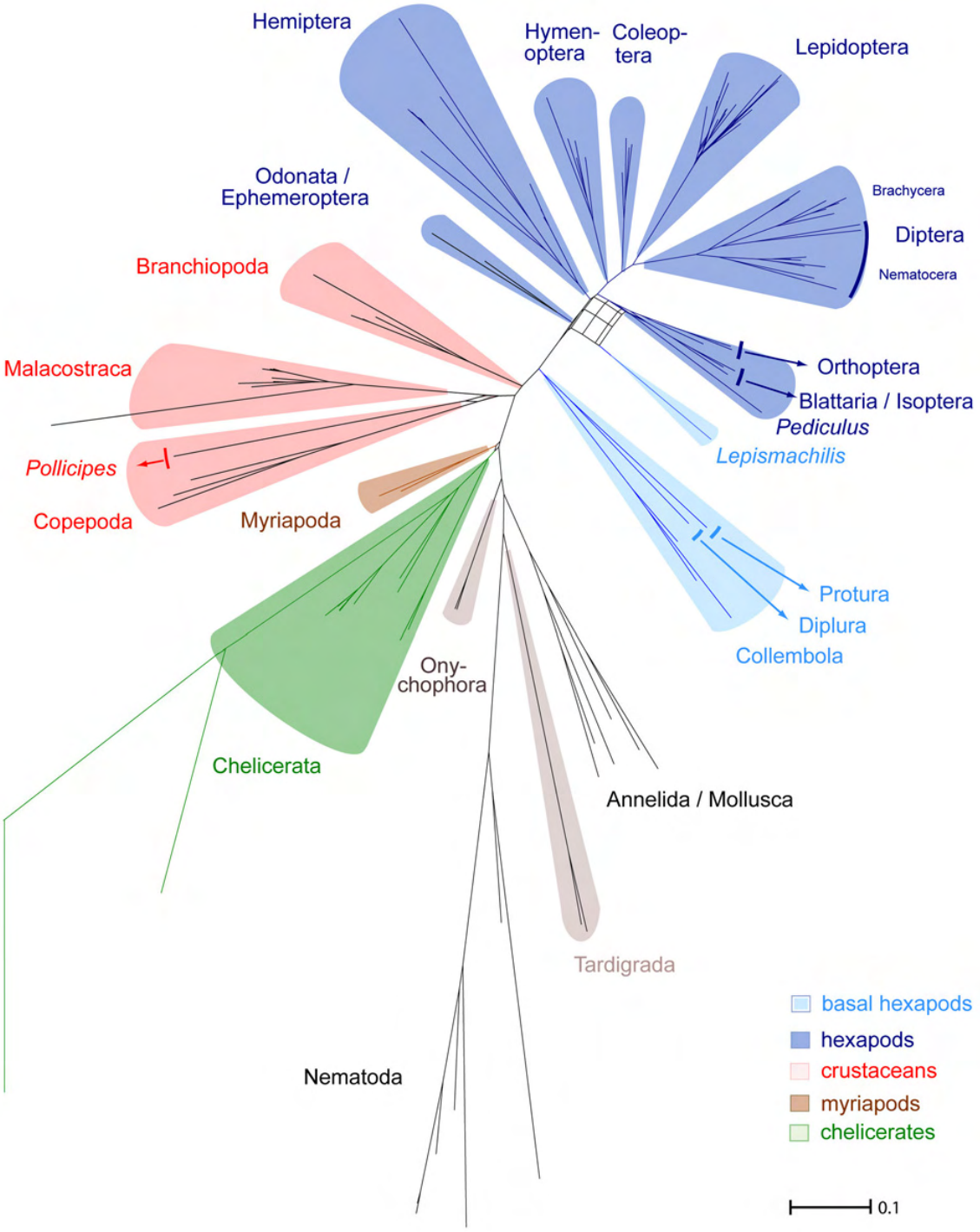


Figure 3.13: Consensus network of all PHYLOBAYES trees. The consensus network of all 25 PHYLOBAYES chains (optimal data subset) was calculated with SPLITSTREE 4.8 (HUSON & BRYANT 2006) in cooperation with K. MEUSEMANN. It visualizes incongruencies between 25 chains (threshold = 0.01, averaged edge weights). The color code is specified in the figure.

3.4 Analysis of hemocyanin structure in Remipedia

One problem was the failure of the planned EST project for Remipedia and consequently the absence of this challenging and still enigmatic group in the phylogenomic analyses presented in analysis [C]. Despite the extensive effort, and the availability of freshly collected tissue (see paragraph 2.1) and successfully isolated total RNA, the expression pattern of screened test clone sequences was so bad regarding numbers of genes that complete sequencing of 4000 sequences had to be stopped. A 454 pyrosequencing started in November 2008 with fresh material and is unfortunately still in progress - no sequences are available yet. At least one result of the EST project can be presented based on a screening of the few EST test sequences and the specifically cloned hemocyanin subunits from Remipedia. This was conducted in the lab and working group of T. BURMESTER, see also paragraph 2.7.

In summary, Remipedia do have functional hemocyanin sequences that occur in three subunits and are expressed in the adult organism. These are more similar to the insect hemocyanin types than to the hemocyanin types of other malacostracan crustaceans.

3.4.1 Phylogenetic reconstruction and resulting tree

Phylogenetic analyses employing maximum likelihood and Bayesian methods resulted in essentially identical trees (figure 3.14). *S. tulumensis* hemocyanins and hexapod proteins form a monophyletic clade not including other crustacean (malacostracan) sequences (ML bootstrap support: 48%; Bayesian posterior probability: 0.96). The other crustacean hemocyanins and pseudohemocyanins were monophyletic (100%; 1.0), but in none of the trees the hemocyanins of the Remipedia joined this clade.

The relative position of the myriapod hemocyanins remains uncertain, receiving poor support. These proteins were either associated with chelicerate hemocyanins (thereby supporting the "Paradoxopoda" hypothesis; MALLAT ET AL. 2004) or are in sister group position to the crustacean plus hexapod proteins (supporting the traditional "Mandibulata").

In none of the trees the myriapod hemocyanins joined hexapod hemocyanins and hexamerins. Hence, there is substantial molecular phylogenetic and structural evidence that the remipede hemocyanin subunits are orthologs of hexapod subunits. Thus, the lineage leading to the remipede and hexapod hemocyanins split into two distinct subunit types before these taxa diverged. This suggests a close relationship of hexapods with remipedes.

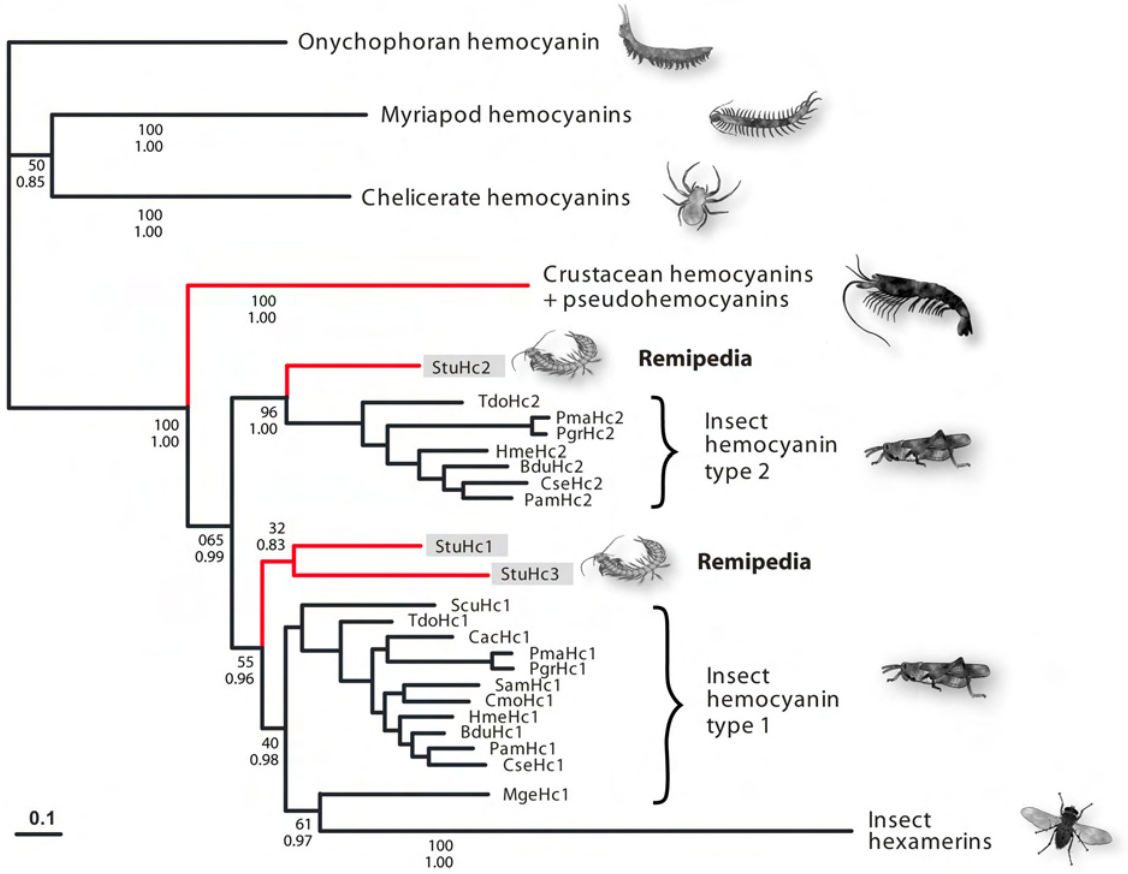


Figure 3.14: Simplified Bayesian phylogenetic tree of arthropod hemocyanins and hexamerins. The numbers at the nodes show maximum likelihood parametric bootstrap support (above) and Bayesian posterior probabilities (below). Branches representing crustaceans are colored in red. The complete trees are shown in the manuscript in the supplement.

4. DISCUSSION

A central lesson of science is that to understand complex issues (or even simple ones), we must try to free our minds of dogma and to guarantee the freedom to publish, to contradict, and to experiment. Arguments from authority are unacceptable (CARL SAGAN).

Methods and datasets are discussed first to facilitate the following phylogeny discussion. Then the inferred crustacean and arthropod phylogenies are discussed in the background of existing studies. A general methodological discussion and final conclusion are completing this chapter.

4.1 Separate methodological discussion of analysis [A-C]

4.1.1 Analyses [A]: Can 16S, 18S and COI marker genes improve inference of crustacean phylogeny? Comparing "usual" standard vs. secondary structure based approaches

This study is also a test case if data with smaller fragments of ribosomal RNA (16S and 18S rRNA genes) are suitable for phylogenetic reconstruction of deeper nodes like crustaceans and less ambitioned for arthropod relationships. Although certain relationships are robust in the trees, notably the monophyly of generally accepted clades such as Branchiopoda, Hexapoda, and Euchelicerata, many higher-level relationships remain unresolved. Results of the best-resolved trees (figures 3.4 and 3.5) suggest basic phylogenetic splits within Pancrustacea into Malacostraca, non-Malacostraca and Hexapoda but monophyly of these groups and the relationships within them generally lack statistically significant support (posterior probabilities < 0.95). In view of the methodological variations encompassed by this analysis, this clearly suggests the need for more and/or different data. These could consist of complete ribosomal sequences, a denser taxon sampling, incorporation of new loci, and the exploration of alternative out-groups that are separated from the in-group taxa by a shorter branch.

Manual versus automized alignment processing (optimization)

The results clearly reveal the crucial importance of alignments (KUMAR & FILIPSKI 2007) and the impact of different alignment strategies. MAFFT seems to perform better than MUSCLE regarding resulting topologies, which is in line with KATOH & TOH (2008) who find that the EINSI algorithm of MAFFT is more accurate and precise in aligning large expansion segments

and gap sections that occur normally in rRNA genes. Nevertheless misaligned sections are still apparent. But these misaligned sections can be identified much better by eye and realigned based on secondary structure information with the manual approach. A commonly held opinion is that manual alignments are subjective and (thus) not repeatable, implying that automated, computerized alignments are more objective. However, the automatically processed alignments of ribosomal genes contain obvious errors, such as gaps (ranging from one to several hundred nucleotides) that are correctly aligned for most taxa, but which are obviously misaligned for some other individual taxa. The resulting misalignment of conserved regions can easily be corrected manually. Therefore it can be agreed with KJER AT AL. (2007), who argue that ignoring apparently falsely aligned, non-homologous positions is in fact also a subjective decision that is likely to affect the resulting phylogenies. Some authors disagree about the relative merit of different strategies, notably manual versus automated sequence alignment (e.g. KJER 2004 vs. OGDEN ET AL. 2005). Although a thorough discussion of the theoretical pros and cons of the different approaches that were adopted is beyond the scope of this study, arguments can be made to prefer one or the other sets of results. The automated alignment processing contains obvious shortcomings regarding the resolution of deeper nodes like Myriapoda, Chelicerata and Malacostraca. Contrary, the results based on manual alignment are better resolved in deeper nodes (Chelicerata, Myriapoda, Malacostraca, see figures 3.5 and 3.6 and supplementary figures S2-S9) than the other analyses, but only latter analyses show expected clades such as a monophyletic Arthropoda and Pancrustacea. However, these analyses also show a clade of the two mystacocarids, one of which likely is contaminated with a mite sequence. Therefore, because this may be a spurious clade that is not produced in the results based on automated alignment processing and masking, one may prefer these less resolved but more reliable results. These are more objective and artifacts by randomly similar aligned sequence positions are minimized.

Importance of the secondary structure constraint and complete sequences

It might be discussed if the selected constraint played a veritable role for the results. For the automated analyses in this study constraints of *Anopheles* are used and alignments are not optimized by hand at all. The hand alignment is adapted with a specially adapted consensus secondary structure estimated from different sequences. An explanation for the better resolution in the topologies resulting from the manually adapted alignment is that by this procedure in total more positions and more stem positions are finally included into analysis compared to the automated alignment processing. In this analysis it becomes clear that the sequence alignment and masking programs can have difficulties with shorter sequence fragments, so that manual improvements become tempting. In such a situation of alternative, but non-perfect methods, one can perform various analyses to compare the results with an open mind. Using a manually adapted secondary structure constraint in combination with automated alignment processing like advocated might be the best objective compromise. Due to computational and time limitations it has been unfortunately impossible to perform a comparison of the results using the hand-adapted secondary

structure constraint for an automatically processed alignment and subsequently analyses. This direct comparison of manually versus automated alignment processing based on a manually optimized constraint is still to test for a final conclusion.

Additionally, it has to be stated that completely sequenced ribosomal RNA genes should be included into analyses to obtain more informative positions. Please note that comparison between different runs is complicated due to their requirement of different evolutionary models, so that variation in the results cannot unambiguously be ascribed to a single analysis variable. A caveat of the current study is that time limitations have prevented the author from doing additional analyses that may have improved the results. For example a comparative analysis based on time-homogeneous and time-heterogeneous models (see analyses B) including a more complex mixed model setting (e.g. as possible in the software package PHASE) would have been interesting. Yet, proceeding the analysis it becomes clear that the dataset by itself is insufficient to resolve this challenging phylogenetic problem. The signal in the data is too ambiguous (see figures 3.1 and 3.2) and probably eroded. Also a RY coding and several tests to come by this lack of resolution are not successful. Future high-level phylogenetic studies will have to explore additional markers, principally nuclear protein-coding genes, which show great promise (REGIER ET AL. 2005; 2008). In analysis B this exploration is made for completely sequenced 18S and 28S rRNA genes.

4.1.2 Analyses [B]: Is secondary structure based alignment optimization and implementation of time-heterogeneity a solution to solve crustacean phylogeny within arthropods?

It is one goal of this analysis to encompass limitations of analysis [A] and of many previously published studies. Apart of some exceptions most studies on phylogenetic relationships, at least partly, rely on nuclear rRNA data. Often, however, only one of the 18S and 28S genes is used, sometimes even just fragments of a gene (DELL AMPIO ET AL. 2009; KJER 2004; MISOF ET AL. 2007; D'HEASE 2002; GIRIBET ET AL. 2004; LUAN ET AL. 2005; EDGECOMBE & GIRIBET 2002; KJER ET AL. 2006; YAMAGUCHI & ENDO 2003), while only few studies use nearly complete 18S and 28S rRNA sequences (MALLATT & GIRIBET 2006; MALLATT ET AL. 2004; GAI ET AL. 2006). Despite this wide usage, the reliability of reconstructions based on rRNA markers is still debated (for contradicting views see MISOF ET AL. 2007; GILLESPIE ET AL. 2005 and JORDAL ET AL. 2008). A major cause of concern is the pronounced site heterogeneity of evolutionary rates, the non-stationarity of base composition among taxa and rate variation in time. The comparison of the time-homogeneous approach to the time-heterogeneous one is not only intended to show improvements in the application of more realistic models, but also to indicate which incongruencies may be causally explained by non-stationary processes during the evolution of these genes. All three phenomena quickly lead to the erosion of

phylogenetic signal (SIMON ET AL. 2006), which is also a problem in analysis [A]. Analysis [A] indicates that especially crustaceans are at least biased in phylogenetic reconstructions by inhomogeneous base compositions.

On the one hand, our understanding of the molecular structure of other markers and about taxon-dependent processes of molecular evolution remains poor. On the other hand, our vast background knowledge regarding the structure and variability of rRNA molecules offers a unique opportunity to study the effects of selection and application of substitution models in greater detail.

Automated alignment processing and time heterogeneous models

Phylogenetic signal in sequence data can get noisy due to (i) multiple substitution processes (saturation) and (ii) erroneous homology hypotheses caused by ambiguous sequence alignment. Both effects correspond in that they result in random similarity of alignment regions. Such noisy sections potentially bias tree reconstruction methods in several ways which have been appreciated for years but only recently been applied, that allow to account for these problems (WÄGELE & MAYER 2007; RODRÍGUEZ-EZPELATA ET AL. 2007; SUSKO ET AL. 2005; PHILIPPE ET AL. 2005). Exclusion of these ambiguously aligned or saturated regions can help to reduce noise, see e.g. MISOF & MISOF (2009). If this topic is addressed at all, the majority of studies include a manual alignment check for untrustworthy regions (MALLATT & GIRIBET 2006; FRIEDRICH & TAUTZ 1995; CARAPELLI ET AL. 2007; KJER 2004; MISOF ET AL. 2007; CARAPELLI ET AL. 2005; LUAN ET AL. 2005; KJER ET AL. 2006; YAMAGUCHI & ENDO 2003; GAI ET AL. 2006). Only some recent publications addressing arthropod relationships use automated tools (e.g. DUNN ET AL. 2008; ROEDING ET AL. 2007; PODSIADLOWSKI ET AL. 2007). To identify alignment sections of random similarity prior to tree reconstructions the alignment processing tested in analysis [A] was used. To improve the signal-to-noise ratio the character choice is restricted to alignment sections, which contain nucleotide patterns of more reliable positional homology.

Arthropod phylogenies have been inferred in the past with reconstruction methods like Maximum Parsimony, Maximum Likelihood and Bayesian approaches. Expanding massively the methods of analysis [A], knowledge about the evolution of rRNA is implemented in two ways: (i) the use of mixed DNA/RNA models is meant to account for known instances of character dependence due to compensatory mutations in stem regions, (ii) the application of time-heterogeneous models accounts for non-stationary processes that occurr in arthropod lineages. The consensus secondary structure of the dataset, generated with RNAsalsa can be understood as a model parameter that defines site interactions and thus character dependence due to compensatory mutations (MISOF ET AL. 2007; HANCOCK ET AL. 1988; STEPHAN 1996). Neglect of character dependence surely results in unrealistic support values. In single low supported nodes, where the signal-to-noise ratio is at the edge of resolution, such neglect theoretically can even turn the balance between two competing hypotheses. Additionally a consensus secondary

structure is necessary to apply a mixed model approach, since it determines whether the evolution of a given site is modeled by the DNA-model, or as part of a base pair by the RNA-model. The mixed model approach is optimized for DNA-corresponding 16–state RNA models (GOWRI-SHANKAR & JOW 2006). It can certainly be argued that the choice of 16–state models is problematic because it is difficult to fit these models to real data due to their parameter richness and heavy computational costs. However, even the best choice of a consensus secondary structure can only capture the predominantly conserved structural features among the sequences. This implies that the applied RNA models must be able to cope with mismatches in base-pairing. Less complex RNA models like those of the 6 and 7–state families either ignore mismatches completely or pool these mismatches into a single character state that produces artificial synapomorphies. Additionally, according to SCHÖNINGER & V. HAESELER (1994), it is more likely that co-variation is a multiple step process, which allows for the intermediate existence of instable (non Watson-Crick) pairs. These intermediate states are only described in 16–state RNA models. Concerning rRNA-genes of arthropods, shifts in base composition are mentioned for Diptera, Diplura, Protura and Symphyla (MALLATT & GIRIBET 2006; DELL' AMPO ET AL. 2009; MISOF ET AL. 2007; LUAN ET AL. 2005; GAI ET AL. 2006; FRIEDRICH & TAUTZ 1997).

Since base compositional heterogeneity within a dataset can mislead phylogenetic reconstruction (BLANQUART & LARTILLOT 2006; JERMIIN ET AL. 2004; FOSTER 2004) and (GOWRI-SHANKAR & RATTRAY 2007), some of these studies discuss observed but not incorporated non-stationary processes as possible explanations for misplacements of some taxa (MALLATT & GIRIBET 2006; MALLATT ET AL. 2004; DELL' AMPIO ET AL. 2009; HASSANIN ET AL. 2005; LUAN ET AL. 2005; GAI ET AL. 2006). The selective exclusion of these taxa to test for misleading effects on the remaining topology, however, is not appropriate to test whether non-stationarity really is the the causal explanation for a placement incongruent with other analyses. LogDet methods have been applied to compensate for variations of base frequencies (MALLATT & GIRIBET 2006; MALLATT ET AL. 2004; LUAN ET AL. 2005), which leads to some independence of non-stationarity, but among site rate variation (ASRV) cannot be handled efficiently. After detecting compositional base frequency heterogeneity in the data, a non-stationary approach implemented in PHASE-2.0 is chosen. Because no previous study of arthropod phylogeny uses a time-heterogeneous approach including mixed DNA/RNA models, this approach is compared with a "classical" time-homogeneous setup. The results prove that the time-heterogeneous approach produces improved likelihood values with improved branch lengths estimates and more realistic, though not perfect (see taxa discussion), topology estimates. Since modeling of general time-heterogeneous processes is in its infancy and since its behavioural effect on real data is relatively unknown (BLANQUART & LARTILLOT 2006; GOWRI-SHANKAR & RATTRAY 2007), a set up is favored accounting for the three different "submodels" corresponding to three base frequency categories in the dataset. The application of the three submodels to individual branches in a tree by the MCMC

process is not further constrained. This scheme allows for a maximum of flexibility without losing the proper mix of parameters.

4.1.3 Analyses [C]: Do phylogenomic data enlight crustacean phylogeny within arthropods - or do old problems stick to the analysis of this new large-scale data?

Comparing the unreduced with the condensed EST dataset one can see that the trees do not differ too widely. An essential difference is the improvement of most node support values in the trees of the reduced dataset, in general for all included taxa. Comparing the two trees of the condensed data subset against the "unreduced tree" for crustacean taxa this becomes obvious. In the PHYLOBAYES tree (fig. 3.11) only one posterior probability is 0.81 all other node values are supported maximally or with 0.99 posterior probability. The likelihood values of the RAXML topology (fig. 3.12) are similarly high. This result confirms the utility of matrix reduction heuristics. Discussing the effects of the heuristic on the topologies becomes a bit harder. The resulting topologies of the data subset are not resolved for the Chelicerata and Myriapoda compared to the unreduced dataset. Obviously the choice of genes with high relative information content resulted in a more reliable topology regarding the node values. The clade "Myriochelata" (excluding *Julida*, see figure 3.10) that is reconstructed in the unreduced dataset with a bootstrap value of 89 can be discussed as an artifact by non-informative signal in the data. The tree of the unreduced dataset suggests that a clade "Myriochelata" exists in the data constituted by a rather well supported node. Using the condensed dataset the Chelicerata, Myriapoda and Pycnogonida each are monophyletic clades, but form together a polytomy. This can be interpreted as the result of conflicting or missing information after the exclusion of noisy signal by the reduction heuristics. The tree is unresolved in this point but more reliable because it relies only on genes with potential relative information content.

Importance of gene choice – the supermatrix approach: The advantage of the supermatrix approach is that the dataset can be analyzed within a single tree search. The different substitution patterns among the partitioned genes (matrix partitions) can be estimated in an analysis applying mixed/partitioned substitution models. The reliability for the trees of the total dataset can be addressed with standard procedures like bootstrapping or posterior probabilities. One main disadvantage of supertree methods is for example that single partitions might not contain sufficient signal to resolve the relationships between the taxa (BININDA-EMONDS 2004; DE QUEIROZ & GATESY 2006).

All recent phylogenomic studies choosing the supermatrix approach (ROKAS ET AL. 2005, PHILIPPE ET AL. 2005, BAURAIN ET AL. 2007) apply rather subjective parameters for the gene choice to reconstruct the final data matrix. The first study that established an

applied matrix reduction method recently published by DUNN ET AL. (2008). These authors compare the reduced with their original data matrix by including genes that follow a predefined threshold. This threshold is the percentage of genes available for all taxa and was set to 50 percent. However, an objective parameter choice like relative information content of each gene is not applied, the data matrix is additionally reduced by excluding taxa based on the leaf stability index from the software PHYUTILITY.

Missing data is often observed for phylogenomic and respectively EST data. Its importance or influence on a resulting topology is not yet to estimate. Some studies demonstrate that missing data influences an unstable placement of taxa with incomplete or missing protein sequences (WIENS 1998; HARTMANN & VISION 2008). Contrary, several studies show that taxa with missing data have a minimal effect if the total number of positions is large (WIENS 2003; WIENS 2005; PHILIPPE ET AL. 2004; WIENS & MOEN 2008) or can even improve the results by breaking long branches (WIENS 2006). Thus, at the moment the total number of existent positions seems to be important and absent or missing data has less impact on tree reconstructions. This is of course strongly dependent on the dataset and the included genes. The present study deals with this issue in the way that by relying on the relative information content for each gene the quality of the implemented genes is increased.

Orthology prediction: The task to implement exclusively orthologous genes is a challenging computationally intensive task for phylogenomic data and several software programs and databases exist to identify and administrate clusters of true orthologs avoiding the trap to include closely related but non-orthologous proteins (see e.g. O'BRIEN ET AL. 2005: INPARANOID database; KIM ET AL. 2008: COG database in NCBI; SCHREIBER ET AL. 2009: Orthoselect software; EBERSBERGER ET AL. 2009: HaMStR orthology prediction). Most software tools rely on the results of reciprocal BLAST hits among genomes to reconstruct orthologous protein cluster (KIM ET AL. 2008, EBERSBERGER ET AL. 2009). Different approaches exist to solve the difficulty to find an optimal solution in the BLAST procedure since proteins and genomes of phylogenetically distant species show normally low similarities in their protein sequences. The approaches to predict gene orthology differ in general between the different phylogenomic studies. For the phylogenomic analysis in this thesis the HaMStR approach by EBERSBERGER ET AL. (2009) is used (see chapter 2 for details on its basic protocol). To discuss this part properly a comparison of the dataset with different strategies of orthology predictions should have been performed, but this was impossible due to time limitations. Starting the analysis the HaMStR approach was the most sophisticated approach that existed at the time the analyses were started. Including the Hidden Markov Model trained search for inparalogous with a reciprocal BLAST this procedure was ahead of other simple reciprocal BLAST procedures. Anyhow, the comparison with recently (while finishing this thesis) presented software packages for orthology prediction (e.g. SCHREIBER ET AL. 2009) could be interesting.

Gene overlap with other studies: The present gene selection overlaps with other phylogenomic studies focused on metazoans (BAURAIN ET AL. 2007, DUNN ET AL. 2008, PHILIPPE ET AL. 2009) or chordates (DELSUC ET AL. 2008). The most extensive overlap exists with

PHILIPPE ET AL. (2009) (complete dataset: 51 genes, optimal data subset: 46 genes), BAURAIN ET AL. (2007) (complete dataset: 50 genes, optimal dataset: 45 genes) and DELSUC ET AL. (2008) (complete dataset: 53 genes, optimal dataset: 44 genes). There is little difference in the overlapping gene numbers with latter studies comparing the unreduced and the condensed subset of present data. Less genes are shared with DUNN ET AL. (2008) (complete dataset: 37 genes, optimal dataset: 19 genes) with a remarkable difference between the selected subset and the original data (supplementary table S15). This implicates that DUNN ET AL. (2008) included many genes in their study that were identified in the present approach as less informative.

The role of ribosomal genes: The reduced data subset of the present study includes only 32 ribosomal genes (supplementary table S15), which is a rather low number (for included ribosomal genes) compared to other phylogenomic studies. For example BAURAIN ET AL. (2007) use 68 ribosomal genes (from 133 genes in total), ROKAS ET AL. (2005) 11 ribosomal genes (49 genes in total). DUNN ET AL. (2008) include 40 ribosomal genes in their reduced data matrix (150 genes in total) but state for the first time, that ribosomal genes might lead to systematic errors if included. DUNN ET AL. exclude 30 ribosomal genes of the previous studies (ROKAS ET AL. 2005, PHILIPPE ET AL. 2005) because they were problematic in the orthology prediction.

Still to survey is also the impact of the completely sequenced genome species. Because most of the genome or proteome species are rather derived (e.g. *Daphnia, Artemia, Drosophila*) it might be possible that they influence phylogenetic reconstructions.

Computational limitations were in general a problem with the present dataset. The RAXML analysis of the unreduced dataset crashed several times. To encompass that and to examine possible biases ten single ML tree searches were conducted and separate bootstrapping searches (100 replicates) are applied for the original concatenated supermatrix (350,356 aa positions). Due to the restricted computational power, calculation of branch lengths was not possible. The ML tree with the best likelihood value was chosen to plot bootstrap values.

The software PHYLOBAYES is suitable for large-scale datasets but while performing the analysis the author of the present study became a bit suspicious about the implemented algorithms in PHYLOBAYES. The software relies similar to MrBayes on MCMC sampling but works differently saving parameter values in a dynamic process. The authors of this software define the number of generations (PHYLOBAYES 2.3 manual) „as the number of elementary topological updates tried during a cycle. This number is not constant across cycles, because PHYLOBAYES implements recursive ‚waves' of topological updates along the tree [...]". What that exactly means was not to clarify.

4.2 Discussion of the phylogeny of crustaceans and related arthropods

Non monophyletic Crustacea are reconstructed in all analyses included in this thesis. Strong evidence for paraphyletic crustaceans regarding Hexapoda is found in all trees supporting the Pancrustacea concept. To the knowledge of the author no molecular study exists that reveals a monophyletic clade Crustacea. A better understanding of the phylogeny of Crustacea is hindered by a problematic morphological classification of crustaceans. Many characters of the taxon Crustacea are more diagnostic characters e.g. two pairs of antennae (Diantennata) or biramous limbs instead of crustacean synapomorphies. These characters could be interpreted as mandibulate characters (see BUDD 2002; SCHOLTZ & EDGECOMBE 2006; WALOSSEK & MÜLLER 1998A+B*)*.

Only few morphological characters constitute a monophyletic Crustacea: The existence of eyes in the nauplius larvae is one putative synapomorphy of crustaceans (LAUTERBACH 1983, but see also RICHTER 2002). Also nephridia exclusively found in the segment of second antennae and maxillae are given as synapomorphic character (LAUTERBACH 1983; 1986; AX 1999). WALOßEK & MÜLLER (1998A+B) interpret the existence of three pairs of limbs in the first larvae of crustaceans as a synapomorphy for this group while fossilized larvae of representatives from the mandibulate stem lineage show four limbs.

Pancrustacea (Crustacea + Hexapoda), first erected by PAULUS (1979) ae confirmed in several molecular studies relying on nuclear, mitochondrial and combined data (FRIEDRICH & TAUTZ 1995; FRIEDRICH & TAUTZ 2001; GIRIBET ET AL. 2001; HWANG ET AL. 2001; REGIER & SHULTZ 2001; SHULTZ & REGIER 2000; CARAPELLI ET AL. 2007; CARAPELLI ET AL. 2005; NARDI ET AL. 2003). DOHLE (2001) advocates synonymously the Tetraconata concept, which is based on similarities of eye structure (presence of cone cells) and is later supported by further neuroanatomical data (HARZSCH 2006; HARZSCH ET AL. 2005; HARZSCH ET AL. 2006; UNGERER & SCHOLTZ 2008).

The unreduced phylogenomic dataset of analysis [C] recovers this clade with a bootstrap value of only 89. In the reduced dataset both trees support this clade maximally. This is in line with most molecular studies supporting a clade Pancrustacea and corrobates these results with a large scale phylogenomic dataset focused on arthropods. The higher support values of this clade in the trees of the reduced dataset suggest that there is some noise in the unreduced dataset that was encompassed by the reduction heuristics.

The sistergroup of Hexapoda is to date still discussed. Morphological evidence (including neuroanatomical data) place Branchiopoda (SCHRAM & KOENEMANN 2004A), Malacostraca (STRAUSFELD 2009; STRAUSFELD ET AL. 2009:) as SG to Hexapoda or reconstruct a polytomic clade Malacostraca + Remipedia + Hexapoda (FANENBRUCK ET AL. 2004; FANENBRUCK & HARZSCH 2005). Most molecular studies support paraphyly of crustaceans with respect

to hexapods. Molecular studies mostly reveal either Branchiopoda (REGIER & SHULTZ 1997; SHULTZ & REGIER 2000; REGIER ET AL. 2005; BABBIT & PATEL 2005; TIMMERMANS ET AL. 2008) or Copepoda (MALLATT ET AL. 2004; MALLATT & GIRIBET 2006) as SG to Hexapoda. Recently, the clade Cephalocarida + Remipedia is reconstructed as SG to Hexpoda (REGIER ET AL. 2008). The grouping of Cephalocarida + Remipedia is yet critically discussed by SPEARS & ABELE (1998), see also later paragraph Cephalocarida. Analyses using mitochondrial genes add more confusion to the situation by reconstructing non-monophyletic Hexapoda (NARDI ET AL. 2003; CARAPELLI ET AL. 2007), which is in total contrast to morphological evidence constituting a monophyletic Hexapoda.

The results of the present study are ambiguous. Possible SG of Hexapoda are Copepoda (based on the rRNA data, analysis A), Branchiopoda (based on the phylogenomic data, analysis C) or Remipedia (supported by hemocyanin from the EST data).

Copepoda + Hexapoda: Copepoda are placed in several morphological and combined analyses at the basis of Crustacea as SG to Ostracoda (SCHRAM & HOF 1998; WHEELER ET AL. 2004, excluding fossils) or Mystacocarida (AX 1999; BOXSHALL & HUYS 1989; WHEELER ET AL. 2004 including fossils, WILLS 1998). Most authors accept an inclusion of Copepods into the Maxillopoda (BOXSHALL & HUYS 1989; WALOSSEK & MÜLLER 1998A+B; SCHRAM & KOENEMANN 2004; see also MARTIN & DAVIS 2001). In molecular analyses the mostly basal position of copepods in morphological studies is not supported. In these molecular studies Malacostraca + Cirripedia (REGIER ET AL. 2005; REGIER ET AL. 2008) based on nuclear data or Branchiura + Pentastomida (COOK ET AL. 2005; LAVROV ET AL. 2004) relying on mitochondrial data are presented.

The SG relation of Copepoda to the Hexapoda (present analysis [B]) is in line with some molecular studies (MALLATT ET AL. 2004; MALLATT & GIRIBET 2006) but one might doubt this clade respectively considering above cited morphological studies and the discrepancy of the heterogeneous molecular results. MALLATT AND GIRIBET (2006) doubted the SG copepod relationship to Hexapoda. The findings in the present rRNA data show that with the more sophisticated time-heterogeneity model a rather low support value is obtained in contrast to the time homogeneous tree. Thus, the low support value reflects that eventually some conflicting signal is present in this topology. The standard time-homogeneous approach is not able to handle this and a rather good support is obtained.

In the trees reconstructed with the phylogenomic data the Branchiopoda are the SG to hexapods and copepods cluster with the cirripedes.

Branchiopoda + Hexapoda: The Branchiopoda are recovered as monophyletic group in analysis [B], which is in line with recent publications (STENDERUP ET AL. 2006; RICHTER ET AL. 2007; OLESEN 2007). There exists still no consensus on their position within crustaceans. Some incongruent results based on morphological, molecular and combined evidence exist for a SG relation to Malacostraca (BITSCH & BITSCH 2004; HASSANIN 2006; LAVROV ET AL. 2004; GIRIBET ET AL. 2005), Maxillopoda

(WALOSZEK 2003; WHEELER ET AL. 2004) or Hexapoda (SCHRAM & KOENEMANN 2004B; REGIER & SHULTZ 1997; SHULTZ & REGIER 2000; REGIER ET AL. 2005; BABBIT & PATEL 2005; TIMMERMANS ET AL. 2008).

Branchiopoda are recovered with low support as SG to a Copepoda + Hexapoda clade relying on the rRNA data in the present thesis. The analysis of the phylogenomic data supports the results of above cited studies revealing a SG of Branchiopoda + Hexapoda. Interestingly, recent studies using large-scale data (DUNN ET AL. 2008; PHILIPPE ET AL. 2009) find the same topology. It might be seductive to assume a common ancestry of freshwater crustaceans and hexapods as described in GLENNER ET AL. (2006). Morphologically this scenario is under fire by the partly very specific adaptations of branchiopods to their environment, seasonal freshwater ponds. Basically the conclusion of GLENNER ET AL. (2006) is founded on the lack of fossils of branchiopods and insects before the Devonian. This is contradicted by other paleontological data (WALOSZEK 2003) concluding that respectively "Orsten" type fossils show characters of modern eucrustacean groups as Malacostraca and Branchiopoda.

Remipedia + Hexapoda: Remipedia are introduced as a group of special interest to the reader in the material and methods section. Since their rather late description in 1981 (YAGER) no consensus is established about their phylogenetic position within crustaceans. Morphological studies suggest very early plesiomorphic characters defining this group (AX 1999; BRUSCA & BRUSCA 1990, view not changed in BRUSCA & BRUSCA 2003; SCHRAM 1986) and/or place them as SG to all remaining crustaceans (AX 1999; SCHRAM 1986; EMERSON & SCHRAM 1991; SCHRAM & HOF 1998; WHEELER ET AL. 2004, combined evidence). These results are not supported by SCHRAM & KOENEMANN (2004A) who place the Remipedes near to the Malacostraca. This clade is also supported by KOENEMANN ET AL. (2009) based on first descriptions of remipede larvae. A rather derived position of Remipedia is also supported by recent neuroanatomical studies (FANENBRUCK ET AL. 2004; FAHNENBRUCK & HARZSCH 2005) that reveal a polytomic clade (Malacostraca + Remipedia + Hexapoda) and indicate a possible but not clearly favored clade Malacostraca + Remipedia.

The molecular studies based on mitochondrial markers are quite controversial (even contradicting within the same studies). Published topologies show a SG relation to the Cirripedia (CARAPELLI ET AL. 2007; HASSANIN 2006; LAVROV ET AL. 2004; LIM & HWANG 2006), Ostracoda (COOK ET AL. 2005), Collembola (COOK ET AL. 2005; HASSANIN 2006) or Diplura (CARAPELLI ET AL. 2007). Studies relying on nuclear markers (SPEARS & ABELE 1998; REGIER ET AL. 2005; REGIER ET AL. 2008; SHULTZ & REGIER 2000) reveal mostly a clade Cephalocarida + Remipedia (see later paragraph, "Cephalocarida") if both taxa are included. Relying on combined evidence, GIRIBET ET AL. (2001) find the same result or a clade Remipedia + Cephalocarida+ Ostracoda + Cirripedia.

The suspicious long branch grouping of Remipedia + Cephalocarida is also surveyed in analysis [A] of this study. Analysis [B] relying on the 18S and 28S rRNA genes recovers the

Remipedia as SG to a Cirripedia + Ostracoda clade in both approaches. Interestingly, this result is confirmed in many studies using mt-marker genes (CARAPELLI ET AL. 2007; HASSANIN 2006; LAVROV ET AL. 2004; LIM & HWANG 2006). To the knowledge of the author no morphological study exists supporting a clade composed of Cirripedia + Remipedia. This result has to be doubted which is underpinned by the differences between the time-homo vs. time-heterogeneous run. The pP value for the time-homogeneous approach is 0.73 while the time-heterogeneous tree shows only 0.65. Comparing these values it can be discussed that the heterogeneous approach is probably calculating a more realistic, lower value of this unlikely clade. The cause for this clustering is unclear and reasons can be only speculative. Especially the recent neuroanatomical studies (FANENBRUCK ET AL. 2004; FANENBRUCK & HARZSCH 2005) and the analysis of KOENEMANN ET AL. (2009) relying on characters of the remipede larvae are serious hints for the derived position of the Remipedia in a still unclear scenario of a polytomic clade Remipedia + Malacostraca + Tracheata. The inclusion of Remipedia into the phylogenomic analysis unfortunately failed, but one finding using the EST data is that the hemocyanin of the Remipedia is close related to hexapod hemocyanin (ERTAS ET AL. 2009). This result is supporting a derived position of Remipedia and is in line with an eventual scenario that the Remipedia might have shared ancestors with the Hexapoda (FANENBRUCK ET AL. 2004; FANENBRUCK & HARZSCH 2005) or Tracheata (FANENBRUCK 2003). Thus the **Archilabiata hypothesis** (FANENBRUCK 2003; BÄCKER ET AL. 2008) is indirectly supported. Anyhow this scenario needs to be enlightened by more molecular data of Remipedia. Hopefully soon, a phylogenomic study with Remipedia and additionally Ostracoda and Leptostraca can be started using the tissue collected within this study. Due to technical problems with the 454 sequencing technology an analysis including samples of these species was not yet possible.

Additionally further morphological studies would be important to test the molecular studies. Based on collected material for this study, work started already in this direction. Some individuals of *Speleonectes tulumensis* were analyzed by the present author and colleagues via X-ray tomography at the German Synchroton (DESY) in Hamburg to contribute to this challenge with collaborating partners (BLANKE; HARZSCH; KOENEMANN).

The Entomostraca concept is not supported in the present study. The term Entomostraca was used the first time by LATREILLE (1806) and describes a rather typological than phylogenetic assemblage of non-malacostracan crustaceans. The term Entomostraca today describes a concept namely used and defined by WALOSSEK & MÜLLER (1998A+B WALOSSEK 1999) based on morphological characters. The synapomorphies of this group (Cephalocarida (Maxillopoda + Branchiopoda)) are e.g. an abdomen with at least four limbless somites and the shape of the second maxilla, the maxillula (= fourth pair of head limbs). The maxillulae of Entomostraca develope a stem that is subdivided into four endites and the proximal one is called the proximal endite (e.g. WALOSSEK & MÜLLER 1998A+B). However, several studies criticize this concept and its constituting synapomorphies (FAHNENBRUCK 2003; SCHRAM & KOENEMANN 2001).

None of the present analyses reveal the grouping of Malacostraca + Entomostraca, see figure 1.1. The Malacostraca never form a separated lineage in the present study within crustaceans constituting a SG to Entomostraca. The Entomostraca are not supported by other molecular or combined evidence studies but in some morphological analyses that partly include fossils (WALOSSEK 1999; WALOSZEK 2003; WHEELER ET AL. 2004). With the completely sequenced 18S and 28S rRNA data the Cephalocarida group as SG to the Branchiopoda a finding that is actually in line with the Entomostraca concept and morphological data (WHEELER ET AL. 2004; WALOSZEK 2003; SCHRAM & HOF 1998) but a monophyletic clade Maxillopoda is not revealed at all. Another argument against the Entomostraca is that maxillopodan taxa (copepods=analysis [B]; branchiopods= analysis [C]) constitute a paraphyletic Crustacea in respect to Hexapoda, which will be discussed later in more detail.

The Thoracopoda concept based on the synapomorphy of an epipodite ((Cephalocarida + Branchiopoda) + Malacostraca) is incompatible with the Entomostraca grouping. The presence of epipodites is discussed controversial, see for an overview RICHTER (2002). In short, homology of the epipodites for these taxa is doubtful (e.g. WÄGELE 1993), and WALOßEK (1993) mentions different insertions of the coxopodite of Cephalocarida and the protopodite of Branchiopoda and Malacostraca concluding a convergent development of epipodites. ZHANG ET AL. (2007) discuss this character to be plesiomorphic describing its presence in Cambrian eucrustacean fossils (e.g. in *Yicaris dianensis*). EDGECOMBE ET AL. (2000) confirms the Thoracopoda with a cladistic analysis of arthropods including in total 211 morphological characters for the dataset, but combining this with molecular data (H3 and U2 snRNA) the Thoracopoda clade is no longer reconstructed and instead Cephalocarida cluster together with Branchiopoda or as SG to all crustaceans.

The present thesis does not support the Thoracopoda concept either with molecular data. As stated above the clade Cephalocarida + Branchiopoda (inferred in the time-heterogeneous tree relying on the 18S and 28S rRNA data) is partly also in line with the Thoracopoda concept and the result of EDGECOMBE ET AL. (2000). Anyhow, Thoracopoda are contradicted by the position of the Malacostraca in the present analyses that were neither positioned as SG to the Cephalocarida + Branchiopoda clade nor as the most derived crustacean group.

The Maxillopoda concept is highly debated and several morphological (EDGECOMBE ET AL. 2000; BITSCH & BITSCH 2004; SCHRAM & KOENEMANN 2004A; FANENBRUCK 2003; GIRIBET ET AL. 2005) and molecular analyses based on nuclear and mitochondrial data (SPEARS & ABELE 1998; COOK ET AL. 2005; HASSANIN 2005; 2006; REGIER ET AL. 2005; 2008; MALLATT & GIRIBET 2006) contradict this clade. The present study supports this contradiction. Especially in analysis [B] the widespread taxon sampling includes at least one specimen as representative for each maxillopodan clade (following MARTIN & DAVIS 2001), but a Maxillopoda clade is not recovered despite the sophisticated setting.

The phylogenomic data in this study lacks unfortunately most important taxa like Pentastomida, Mystacocarida and Ostracoda, only Cirripedia and copepods are included as maxillopodan taxa. A well-founded discussion is hardly to achieve concerning these data.

Malacostraca are generally accepted to be monophyletic (RICHTER & SCHOLTZ 2001; JENNER ET AL. 2009) in morphological and combined analysis including molecular data (WHEELER ET AL. 2004; GIRIBET ET AL. 2001; GIRIBET ET AL. 2005) although some early morphological studies (SCHRAM & HOF 1998; WILLS 1998) suppose a non monophyly of this group. Yet SPEARS & ABELE (1998) reconstruct a monophyletic clade Malacostraca in one of the first studies based exclusively on molecular markers.

The position of the Malacostraca within crustaceans is still unclear, and nearly all, possible SG relationships to all remaining crustacean groups have been reconstructed in published literature. Malacostraca are suggested by some authors as the SG to Entomostraca based on morphological and fossil data (WALLOSSEK 1999; WALOSZEK 2003; WHEELER ET AL. 2004). Other studies support a SG relationship to Maxillopoda (WILLS 1998), Anostraca (BITSCH & BITSCH 2004), Phyllopoda (Ax 1999) or a clade of Copepoda + Ostracoda + Thecostraca (SCHRAM & KOENEMANN 2004A). Neuroanatomical data supports a rather terminal position of Malacostraca as SG to the Archilabiata (Tracheata + Remipedia, FANENBRUCK 2003) or in a polyphylum to Remipedia + Malacostraca + Hexapoda (FANENBRUCK ET AL. 2004; FANENBRUCK & HARZSCH 2005).

The phylogenetic position of Malacostraca differs among molecular studies. Often, Malacostraca emerge as nested within the remaining crustacean groups (e.g. SHULTZ & REGIER 2000; EDGECOMBE ET AL. 2000). In studies that use complete mitochondrial genomes for phylogenetic reconstruction Malacostraca are placed close to insects (LIM & HWANG 2006; WILSON ET AL. 2000). Studies based on rRNA or other nuclear genes (MALLATT & GIRIBET 2006; MALLATT ET AL. 2004; GLENNER ET AL. 2006) find Malacostraca positioned more basally within crustaceans and often as SG to Cirripedes (REGIER ET AL. 2005; REGIER ET AL. 2008; MALLATT & GIRIBET 2006). The grouping with Cirripedia lacks any morphological support. In line with the Thoracopoda hypothesis are studies using mitochondrial and combined evidence reconstruction a SG to Branchiopoda (LAVROV ET AL. 2004; HASSANIN 2006; GIRIBET ET AL. 2005). This is additionally confirmed by PAPST & SCHOLTZ (2009) who argue that foliaceaus limbs of Branchiopoda and Leptostraca are probably homologous.

This study is more focused on the position of malacostracans among other crustaceans and aims not to solve internal relationships that are still partly unclear (JENNER ET AL. 2009). In analysis [A] malacostracans were reconstructed for the dataset aligned by hand (figure 3.5), but in the automatically processed data some problematic taxa were placed outside the paraphyletic malacostracans. A good example for problematic malacostracan taxa is *Dyastilis.* This taxon is very likely (as the published mystacocarid sequence) a contamination, misleading tree reconstruction. Some taxa are extremely biasing analysis due to their long branches.

Since in the time-homogeneous tree (analysis [B]) monophyletic Malacostraca branch off at a more basal split within crustaceans (WALOSSEK 1998; ZHANG ET AL. 2007), forming a sister group relationship to Ostracoda while they are branching off at a more terminal split in the time-heterogeneous tree. A final conclusion cannot be drawn about the placement of Malacostraca in analysis [B]. The reconstructed position for both methodologically approaches makes one very suspicious (see figures 3.9 and 3.10) because it is not supported by morphological data and different to previous molecular studies. The question which effect underlies this result is not to be answered here. Probably the phylogenetic signal in the rRNA data is not strong or clear enough to single out the correct placement of the Malacostraca. This issue grows even more complex if the phylogenomic data (analysis [C]) is included. In that analysis the Malacostraca are again placed rather basally as SG to Cirripedia + Copepoda. A final conclusion of the malacostracan position cannot be made here, either. Too many major crustacean groups are not represented yet in phylogenomic data (supplementary table S14). It might be speculated that the inclusion of Leptostraca and additional crustacean species to phylogenomic data will result in a more robust analysis, which is in process yet.

Cephalocarida represent one of the crustacean groups discovered relatively late (SANDERS 1955). Early morphological studies (LAUTERBACH 1979, 1980; MOURA & CRISTOFFERSEN 1996) discuss the Cephalocarida as "primitive" crustaceans because they possess a ventral food channel, primarily missing carapax and the similarity of the second maxilla with the thoracic limbs. These characters all-together suggest a rather basal position as also stated in SCHRAM & HOF (1998). Based on cephalo-skeleton structure (FAHNENBRUCK 2003) the Cephalocarida are recognized as more derived crustaceans and placed as SG to the Branchiopoda. This is congruent with fossil evidence (WALOßEK 1993). In a combined evidence study of WHEELER ET AL. (2004), Cephalocarida are positioned as SG to a Branchiopoda + Maxillopoda clade.

In recent molecular studies the Cephalocarida constitute a problematic group. It has to be mentioned that in all those studies only *Hutchinsoniella macracantha* is sequenced and analyzed. Unfortunately this taxon shows rather long branches in all analyses. For example, relying on nuclear data (REGIER ET AL. 2005; REGIER ET AL. 2009; SPEARS & ABELE 1998) the clustering of the two long branch taxa Remipedia + Cephalocarida is most commonly observed which is also the case in the present study (for analysis [A]). Yet, already SPEARS & ABELE (1998) suggest that this result is created by a long branch artifact. Despite the inclusion of *Lightiella incisa,* a second cephalocarid in analysis [A], this long branch is not to brake down. Studies relying on mitochondrial markers (CARAPELLI ET AL. 2007; COOK ET AL. 2005; HASSANIN 2006; LAVROV ET AL. 2004; PODSIADLOWSKI & BARTHOLOMAEUS 2006; 2007) are also obviously affected by systematic errors like reverse strand bias, meaning more T and G content relative to A and C (HASSANIN 2006) and place the Cephalocarida to nearly all crustacean groups (Cirripedia, Copepoda, Malacostraca) and even Hexapoda. The gene order information also finds a rather ambiguous result (see LAVROV ET AL. 2004): namely SG to

Hexapoda based on tRNA translocation data. PODSIADLOWSKI & BARTHOLOMAEUS (2006) suggest a SG relation to Cirripedia based on gene rearrangements. These results based on mitochondrial data suggest that the recently analyzed mt-data is not suitable to assess at least the phylogenetic position of the Cephalocarida.

In the present study, the crustacean *Hutchinsoniella* (Cephalocarida) clusters in the time-homogeneous approach (analysis [B]) with *Lepisma* (Zygentoma, Hexapoda) within Enthognatha as sister group to Nonoculata (Protura+Diplura, figure 3.10). This led to the polyphyly or paraphyly of several major groups (e.g. Hexapoda, Entognatha, Ectognatha). In the time-heterogeneous analysis, Cephalocarida clustered as sister group to Branchiopoda. This result, although marginally supported, is congruent, at least, with several studies relying on morphological data (SCHRAM & HOF 1998; WHEELER ET AL. 2004), fossil data (WHEELER ET AL. 2004; WALOSSEK & MÜLLER 1998A+B; WALOßEK 1993) and neuroanatomical data (FANENBRUCK 2003). This result also agrees with HESSLER, who states in an essay (1992) about the phylogenetic position of cephalocarids: "[…] If cephalocarids have any special affinity with other living groups, it is with the branchiopods and malacostracans". Most recent molecular studies have not included Cephalocarida (e.g. MALLATT & GIRIBET 2006; MALLATT ET AL. 2004). REGIER ET AL. (2005) reconstruct a sister group relationship of Remipedia and Cephalocarida (likewise represented by *Hutchinsoniella macracantha*), but his result also receives only moderate bootstrap support. The same clade is presented in GIRIBET ET AL. (2001) based on morphological and molecular data. This clade characterized by long branches of involved taxa is also found in analysis [A]. Taking a heterogeneous base frequency into account it was possible for the first time to recover the rather plausible position of the Cephalocarida as SG to Branchiopoda, but the support value of 0.59 pP is rather low. The reason for this might be conflicting signal or long branch effects but to investigate this in detail better tools are needed.

The correction of the misplacement of *Hutchinsoniella*, by allowing for non-stationary processes in the models for sequence evolution, has a major effect on the heuristic value of the analyses. The inference of crustacean phylogeny, however was in general not improved, especially many nodes within Crustacea are rather low supported. Anyhow, not only the monophyletic status of Hexapoda, Entognatha and Ectognatha is supported after the correction, but likewise a causal explanation is given for the misplacement in the time-homogeneous approach, which cannot be accomplished by alternatively excluding a taxon. The time-heterogeneous analyses results in a sister group relationship of Diplura and Protura, which lends support to a monophyletic Nonoculata within a monophyletic Entognatha. This result is congruent with trees published by KJER (2004), LUAN ET AL. (2005), MALLATT AND GIRIBET (2006), and DELL'AMPIO ET AL. (2009).

Cirripedia constitute together with the Facetotecta and Ascothoracica a monophyletic **Thecostraca** (GRYGIER 1987). Synapomorphies are five pairs of lattice organs and prehensile antennulae of the cypris larvae (GLENNER & HEBSGARD 2006; HØEG & KOLBASOV 2002; HØEG ET AL. 2004; HØEG ET AL. 2009; PEREZ-LOSADA ET AL. 2008). Cirripedes as monophlylum are recovered by several morphological and molecular studies as monophyletic group (HØEG & KOLBASOV 2002; HØEG ET AL. 2009; PEREZ-LOSADA ET AL. 2008) while the internal relationships of thecostracan groups are partly still debated. The present study includs Cirripedia (e.g. *Pollicipes pollicipes, Semibalanus balanoides*) as representatives of Thecostraca, tissue from members of Ascothoracica or Facetotecta were not to obtain as material.

The Thecostraca are reconstructed in most morphological studies within a clade Maxillopoda (WALOSSEK & MÜLLER 1998B: possible grouping with Branchiura, BOXSHALL & HUYS 1989; SCHRAM & HOF 1998) or as SG to a Branchiura + Pentastomida clade (GRYGIER 1983; 1987). The grouping with branchiurans has recently also been inferred in molecular studies (MØLLER ET AL. 2008) and is supported by both approaches in the present analysis [B] with maximal support. Though it should be kept in mind that the clade Maxillopoda is not confirmed in the present study. Other molecular studies relying on different markers (e.g. MALLATT & GIRIBET 2006; REGIER ET AL. 2005) reveal a clade Cirripedia + Malacostraca, with cirripedes as thecostracan representatives. Interestingly, the study of MALLATT & GIRIBET (2006) uses also rRNA genes like the present study and their different result is also highly supported in the Bayesian reconstruction.

A close relationship of cirripedes and remipedes is often observed when both groups are included into analyses (CARAPELLI ET AL. 2007; HASSANIN 2006; LAVROV ET AL. 2004; LIM & HWANG 2006). This is also the case for the present single gene analyses, see analyses [A] and [B].

In contrast, cirripedes cluster in the phylogenomic dataset with low support either with copepods (ML) or malacostracans (Bayesian). Looking more closely into each PHYLOBAYES chain one can see that the cirripede *Pollicipes pollicipes* is ambiguously positioned as sistergroup to Malacostraca and Copepoda. REGIER ET AL. (2005; 2008) and MALLATT & GIRIBET (2006) recover a clade Cirripedia + Malacostraca either, but from a morphological point of view this finding is not supported. It can be speculated if the relative information content for this taxon was biased by contradicting signal or specific peculiarity of cirripede sequences.

Ostracoda represent a species rich crustacean group that has eminent stratigraphic relevance. Morphologically is contradicting evidence supporting monophyly (COHEN ET AL. 1998) or paraphyly (VANNIER & ABE 1995) of the two main ostracod lineages, the Myodocopa and Podocopa. Molecular studies reveal a non monophyly of the Ostracoda (SPEARS & ABELE 1998). In many morphological studies the Ostracoda are either the SG to Thecostraca or to a clade Ascothoracica + Cirripedia or positioned near cirripedian taxa (AX 1999; SCHRAM & KOENEMANN 2004; BOXSHALL & HUYS 1989; SCHRAM & HOF 1998). Molecular studies reveal a SG relation to all remaining Pancrustacea (REGIER ET AL. 2005;

2008), to Branchiura + Pentastomida (MØLLER ET AL 2008; MALLATT & GIRIBET 2006) or to Cirripedia (GIRIBET ET AL. 2005, combined evidence). Ambiguous results are found in WHEELER ET AL. (2004) using combined evidence. The result is the clade (Ostracoda + Copepoda) or (Ostracoda + (Mystacocarida + Copepoda)).

In the present study only the single gene analyses included Ostracoda of which only analysis [A] included both lineages. Most trees in this analysis suggest a close relationship to Cephalocarida + Remipedia for the podocopan ostracods while the myodocopa fall out of the analysis mostly grouping not even together.

In analysis [B] only podocopan species are included (*Heterocypris incongruens* and *Pontocypris mytiloides*) resulting in an ambiguous position. The time-homogeneous tree shows an Ostracoda + Malacostraca clade (pP 0.99) while the time-heterogeneous reveals a Mystacocarida + Ostracoda clade. This result would confirm WHEELER ET AL. (2004) and SCHRAM (1996) but the support value for the latter clade is rather low (0.62). Anyhow an Ostracoda + Malacostraca clade that is found for the stationary approach nearly maximally supported is rather unlikely and not to defend by morphological data. Thus this finding is discussed as an artifact of the stationary approach. The question, if one can rely on the reconstructed clade Ostracoda + Mystacocarida in the non-stationary approach is complex to address. On the one hand this is the first time that Mystacocarida, Cephalocarida, Remipedia and Ostracoda are included in one complex molecular analysis. On the other hand a low support value of 0.62 is not really giving a trustworthy topological solution. To solve the position of the Ostracoda more species should be included and more, different markers should be tested. For further phylogenomic analyses a study is planned including data of a pyrosequencing project (Roche, 454 Titanium) of at least one podocopan ostracod. The sequencing is in progress yet.

Mystacocarida belong to the smallest crustaceans inhabiting the interstitial of marine sandy beaches and were first described by PENNAK & ZINN (1943). Since then hitherto most morphological studies conclude a SG relationship to the Copepoda (BOXSHALL & HUYS 1989; AX 1999; WILLS 1998) or Copepoda + Ostracoda (SCHRAM & HOF 1998).

Only one published molecular study includes Mystacocarida, that of SPEARS & ABELE (1998). One has to note that the published 18S rRNA sequence (L81937) in this study is probably a contamination as demonstrated and discussed in analysis [A]. Due to their small body size the amplification of the Mystacocarida genes is difficult because one has to work extremely precisely in the laboratory to avoid contaminations. Probably this is one reason why not many sequences for this group exist.

The present study is the first including this taxon in a broader crustacean sampling for molecular analyses since SPEARS & ABELE (1998). In analyses [A] (figure 3.5), the Mystacocarida are placed in a polytomic clade composed of ((Branchiura + Pentastomida) + Ostracoda) Ostracoda, Mystacocarida. A similar topology but with even worse resolution is reconstructed for the processed MAFFT-based dataset, see figure 3.6. Interestingly, analysis

[B] reconstructs in the time-homogenous tree a clade of Mystacocarida + Pentastomida with a pP of 0.73. In the time-heterogeneous tree a clade of Mystacocarida + Ostracoda is revealed. The support value of 0.62 is rather low, but this finding is in line with findings of WHEELER ET AL. (2004) relying on combined evidence, and SCHRAM & HOF (1998) using morphological characters. The position of the Mystacocarida in this thesis is hard to discuss based on the rather low support value of the time-heterogeneous tree and the inadequate markers in analysis [A]. But it can be argued that the finding in the time-heterogeneous tree is more reliable because it is in line with the morphological data despite the low support value. Anyhow a SG relationship to Copepoda as dominating in morphological studies is not reconstructed in both analyses [A] + [B]. Unfortunately, analysis [B] including a sophisticated modeling and broad taxon sampling is also not able to answer the position of the Mystacocarida within crustaceans unambiguously. Especially the low support value could be a hint for conflicting signal. For further studies in the framework of this thesis the collection of more than 1500 Mystacocarida was conducted to start a 454 project of this group and to elucidate their position within crustaceans in the light of phylogenomic data.

Pentastomida are the parasitic crustacean group described more detailed in the material and method section. This group is morphologically hard to comprise due to their parasitic life style as described in the previous section (see also figure 2.2). Their morphological characters are partly difficult to interpret in a context of a possible affiliation of Pentastomida to the clade Crustacea. The first study indicating that pentastomids are crustaceans was from WINGSTRAND (1972) who compares the development and structure of pentastomids and branchiurans and finds a close similarity of the sperm structure for these two taxa. In contrast some morphologists exclude the Pentastomida from the crustaceans or even crown group arthropods and assign pentastomids to the stem arthropod lineage together with Cambrian fossils, which are suggested to represent members of the extant Pentastomida (MAAS & WALOSZEK 2001; WALOSZEK ET AL. 2006; DE OLIVIERA ALMEIDA ET AL. 2008).

The first molecular study of Pentastomida by ABELE ET AL. (1989) reconstructs a SG relationship between Branchiura and Pentastomida therefore supporting WINGSTRAND (1972). This result is also recovered by several later studies using mitochondrial, nuclear or combined nuclear and mt-data (COOK ET AL. 2005; KILPERT & PODSIADLOWSKI 2006; LAVROV ET AL. 2004; LIM & HWANG 2006; MALLATT & GIRIBET 2006; MØLLER ET AL. 2008). A combined evidence study using morphological and molecular data by GIRIBET ET AL. (2005) supports the previous studies.

The present study reveals the Pentastomida within the Pancrustacea and does not support an exclusion of this group from euarthropods. However, the position within the several present analyses is contradicting. In both trees of analysis [A] the clade Branchiura + Pentastomida is supported. In the time-homogeneous tree pentastomids group with the mystacocarids (pP 0.73), while in the time-heterogeneous tree the clade Pentastomida + Cephalocarida+ Branchiopoda is reconstructed, also low supported (pP 0.60). Again, an

interpretation and discussion is complicated. On the one hand the result (Banchiura + Pentastomida) is in line with previous findings of above cited molecular studies relying on different markers but most of these studies did not implement time-heterogeneity and sophisticated modeling. On the other hand in analysis [B] sophisticated models are used, but support values are rather low. It has to be noted that the 28S sequence of *Raillietiella* sp. could not be sequenced completely in this thesis, thus necessary information (positions) was eventually lacking for a correct positioning of the Pentastomida in analysis [B]. This would also explain the low support value. Anyhow, for a future phylogenomic study a pyrosequencing project (Roche, 454 Titanium) of this taxon is planned to include Pentastomida in a phylogenomic analysis.

Hexapoda are not the topic of this thesis but have to be discussed briefly as they appear as a pancrustacean in-group. The monophyly of Hexapoda is supported by most morphological studies and is in general accepted by most scientists (e.g. BOUDREAUX 1979; HENNIG 1969; KRISTENSEN 1998; KUKALOVÁ-PECK 1998). The generally accepted synapomorphy of Hexapoda is the tagmosis of the body with a thorax constituted of three limb bearing segments and an abdomen constituted of originally 11 segments and a telson.

A rather contradicting, paraphyletic scenario of hexapod origin is created by recent molecular studies that mostly rely on mitochondrial data. In most cases taxa of the Entognatha are reconstructed within Crustacea or other arthropod groups, e.g. Collembola: (COOK ET AL. 2005; HASSANIN 2006; LAVROV ET AL. 2004; NARDI ET AL. 2003) or Diplura: (CARAPELLI ET AL. 2007). However recent studies demonstrate that some of these findings are based on insufficient taxon sampling and that mitochondrial data alone is problematic to solve internal arthropod relationships (CAMERON ET AL. 2004; DELSUC ET AL. 2003; HASSANIN 2006). A good example is the reply of DELSUC ET AL. (2003) to NARDI ET AL. (2003). NARDI ET AL. (2003) state that hexapods are paraphyletic, revealing Collembola (Entognatha) as crustacean in-group. DELSUC ET AL. (2003) could nicely demonstrate that applied taxon sampling and methods were problematic and they reveal at least a monophyletic Hexapoda in their reanalyzes of the data including more taxa. Confirming this result, Collembola are placed within a monophyletic clade Hexapoda by TIMMERMAN ET AL. (2008) relying on nuclear ribosomal proteins revealing the discrepancy between mitochondrial and nuclear data once more.

The present study supports Pancrustacea with paraphyletic Crustacea regarding to Hexapoda. In line with TIMMERMAN ET AL. (2008) and DELSUC ET AL. (2003) are monophyletic Hexapoda and furthermore monophyletic Entognatha and Ectognatha revealed in analyses [B] + [C].

The rRNA data in analysis [B] shows clearly how strong biasing effects of mismodeling can be even with a broad taxon sampling. Ignoring time-heterogeneity affects dramatically the reconstruction for crustacean and hexapod taxa. The time-homogeneous tree demonstrates this by the grouping of *Hutchinsoniella* and *Lepisma* (Zygentoma). This unlikely clustering is

an artifact that creates bias in several clades in the time-homogeneous tree and has a major impact on the reconstructed phylogeny. In the time-homogenous approach non-monophyletic Hexapoda and Entognatha are recovered. However, comparing the results for Hexapoda and Crustacea it can be shown that most major hexapod clades are robustly revealed in the non-stationary approach contrary to most crustacean clades, despite this time-heterogeneous approach. That eventually indicates the limitation of rRNA data for crustacean phylogeny while hexapods are in comparison reconstructed with relatively robust support values.

In ML and Bayesian approaches of the phylogenomic data (analysis [C]), Entognatha (Protura, Diplura and Collembola) are recovered albeit weakly supported. Since the phylogenomic analysis [C] included all critical members of primarily wingless hexapods and showed that hexapods are monophyletic, the present study supports that hexapods are most likely monophyletic and not paraphyletic in respect to crustaceans as revealed in above cited studies. Relationships among hexapods and pterygote insects are still disputed but not topic of this thesis.

4.3 General arthropod phylogeny discussion

The Tracheata are commonly presented and discussed as monophylum based on morphological data in many text books (AX 1999; BOUDREAUX 1979; DUNGER, in GRUNER 1993; PAULUS, in WESTHEIDE & RIEGER 1996). Interestingly the new edition of the "RIEDER & WESTHEIDE" presents the competing hypotheses Tracheata and Pancrustacea discussing both rather equally as potential evolutionary scenario (PAULUS 2007). In addition to recent molecular studies (FRIEDRICH & TAUTZ 1995; FRIEDRICH & TAUTZ 2001; GIRIBET, EDGECOMBE & WHEELER 2001; HWANG ET AL. 2001; REGIER & SHULTZ 2001; SHULTZ & REGIER 2000) some morphological data (PAULUS 1979; DOHLE 2001; RICHTER 2002) contradict the Tracheata combining the Crustacea and Hexapoda to the Pancrustacea (ZRZAVÝ & ŠTYS 1997; ZRZAVÝ & ŠTYS 1998A) or synonymously Tetraconata (DOHLE 2001).

The results of this thesis contradict in all analyses for rRNA and phylogenomic data the Tracheata hypothesis supporting instead a clade Pancrustacea (as discussed previously).

Mandibulata (versus Myriochelata): Instead of the Mandibulata, recent molecular studies relying on mitochondrial, nuclear and phylogenomic data suppose a clade of Chelicerata + Myriapoda (HASSANIN 2006; HASSANIN ET AL. 2005; MALLATT ET AL. 2004; ROTA-STABELLI & TELFORD 2008; PISANI 2004; DUNN ET AL. 2008; ROEDING ET AL. 2007; ROTA-STABELLI AND TELFORD 2008). Analyses of rRNA sequences up till now are held to favor Myriochelata over Mandibulata (FRIEDRICH & TAUTZ 1995; MALLATT & GIRIBET 2006; MALLATT ET AL. 2004). This clade Myriochelata (PISANI 2004) or Parodoxopoda (MALLATT ET

AL. 2004) is generally not supported by morphological data (see e.g. BÄCKER ET AL. 2008). STOLLEWERK & CHIPMAN (2006) find a support from nervous system pattern but this is conflicting with extensive anatomical data that affirm Mandibulata (HARZSCH ET AL. 2005; BÄCKER ET AL. 2008).

The results of this thesis support a "modified" clade Mandibulata with a grouping of Pancrustacea + Myriapoda (see figure 1.4, B).

Myriapoda problematics: The present analysis [B] provides no final conclusion with respect to the conflict Mandibulata versus Myriochelata. The position of Pauropoda and Symphyla causes Myriapoda beeing polyphyletic. To evaluate the impact on the topology of the very likely incorrect positions of Symphyla and Pauropoda, the time-heterogeneous analysis was repeated using a reduced dataset excluding these taxa. The analysis was limited to ten chains with 7,000,000 generations each (2, 000, 000 burn-in). Differences occurring in the inferred consensus topology (not shown) of the final three chains (15, 000, 000 generations) show that some nodes are still sensitive to taxon sampling, since e.g. Pycnogonida cluster with (Chilopoda + Diplopoda) after exclusion of pauropod and symphylan sequences. Also the crustacean topology changes. The remaining long branch taxa *Hutchinsoniella* and *Speleonectes* cluster together in the reduced dataset, forming a clade with (Branchiura + Cirripedia). Despite the endeavor to break down long branches by a dense taxon sampling, some long-branch problems persist. The reason cannot be clearly addressed but, due to the symptoms, it can be assumed that saturation by multiple substitutions caused signal erosion in the rRNA data (class II effect, WÄGELE & MAYER 2007). The exact reconstruction of the position of Myriapoda within the Euarthropoda and a final conclusion supporting Myriochelata or Mandibulata is not possible. But it seems that a clade Myriochelata might be based on a systematic error in molecular analyses. Recent studies demonstrated for example a high sensitivity with respect to gene choice, taxon sampling and out-group choice (BOURLAT ET AL. 2008; PHILIPPE ET AL 2009).

Myriapoda emerged with weak support in the Bayesian trees in the phylogenomic data (analysis [C]) as sister group to chelicerates. In the ML tree reconstructions the relationships of myriapods, sea spiders, euchelicerates and pancrustaceans is not resolved. Applying the matrix reduction heuristics the previous robust support value obtained for this clade in the unreduced dataset vanishes and confirms above cited morphological studies. Thus the application of new markers and suitable phylogenetic strategies like those applied in the phylogenomic analysis (C) have to be applied and developed further.

Chelicerata are in general accepted as a SG clade to Myriapoda + Crustacea + Hexapoda based on morphological, developmental and paleontological data (HARZSCH 2004; HARZSCH ET AL. 2005; RICHTER 2001; SCHOLTZ & EDGECOMBE 2006; WEYGOLD 1998). The Schizoramia concept (TCC) is to date obsolete and was mainly based on fossil data (CISNE 1974; see e.g. SCHOLTZ & EDGCOMBE 2005). This TCC hypothesis is rejected

in all trees of the present thesis. Pycnogonida are placed by morphological and recent neuroanatomical studies mostly in two competing positions (see DUNLOP 2005; DUNLOP & ARANGO 2005), either as SG to Euchelicerata (BRENNEIS ET AL. 2008) or as basal euarthropods (MAXMEN ET AL. 2005, but see contrary SCHOLTZ & EDGECOMBE 2006).

In molecular studies the Chelicerata are more ambiguous and mostly revealed in a Myriochelata clade (see previous section). The Pycnogonida (=Pantopoda) are in most studies the SG to Euarthropoda (GIRIBET ET AL. 2001; REGIER & SHULTZ 2001; SHULTZ & REGIER 2001), thus supporting the "Cormogonida" hypothesis (ZRZAVÝ ET AL. 1998A). Yet, a frequent finding is also a SG relation to Euchelicerata (MALLATT ET AL. 2004; DUNN ET AL. 2008; WHEELER & HAYASHI 1998; REGIER ET AL. 2005) which is supported by developmental data (BRENNEIS ET AL. 2008). Only TELFORD ET AL. (2008) AND BOURLAT ET AL. (2008) published until today a Euarthropoda clade sensu SNODGRASS (1935) with the Pycnogonida as SG to Euchelicerata, relying on molecular data (see figure 1.4, B).

The rRNA data of analysis [B] supports in both approaches a monophyletic Chelicerata with Pycnogonida as SG to Euchelicerata, rejecting the Cormogonida hypothesis. Both approaches reveal a clade "Myriapoda partim", the unusual position of symphylans and pauropods creates a polyphyletic Myriapoda which is obviously an artifact.

While the position of sea spiders is not resolved in the ML tree (fig. 3.13) with the phylogenomic data (analysis [C]), the Bayesian tree (fig. 3.12) shows monophyletic chelicerates with high support (posterior probability, pP 0.99), including sea spiders. This result corroborates recent molecular analyses (DUNN ET AL. 2008) and neuroanatomical studies, which demonstrate the homology of deuterocerebral appendages of Pycnogonida and Euchelicerata (BRENNEIS ET AL. 2008). It further implies that chelicerae and pedipalpi as head appendages evolved only once and are a diagnostic character of chelicerates. The strongly supported clades Ixodida (ticks) and Astigmata (mites) belonging to Acari, Chelicerata, Decapoda, Copepoda (crustaceans) corroborate results of studies based on single nuclear genes (KJER 2004; LUAN 2005; MISOF ET AL. 2007). This is contrary to studies based on (as previously discussed) problematic mitochondrial protein coding genes (CARAPELLI ET AL. 2007; NARDI ET AL. 2003).

Euarthropoda: Most authors confirm on morphological (BUDD 2001; HUGHES ET AL. 2008; WALOSZEK ET AL. 2007), molecular (TELFORD 2005; ROEDING ET AL. 2007; BLEIDORN ET AL. 2009) and developmental data (HARZSCH ET AL. 2005; HARZSCH 2006; LOESEL 2005; SCHOLTZ & EDGECOMBE 2005) the commonly accepted Euarthropoda including Chelicerata, Crustacea, Hexapoda and Myriapoda. For internal relationships see previous sections. The Euarthropoda are a group of the widely accepted Ecdysozoa (TELFORD 2005; TELFORD 2006; TELFORD ET AL. 2008; BOURLAT ET AL. 2008; PAPS ET AL. 2009; GIRIBET 2008) including Euarthropoda and the Cycloneuralia (Priapulida, Kinorhyncha, Loricifera, Nematoda and Nematomorpha). To reveal ecdysozoan relationships was not the focus of this thesis.

The origin of the arthropod bauplan, concerning for example the evolution of segmentation, appendages and the central nervous system, can only be understood if the

phylogenetic positions of tardigrades (water bears) and onychophorans to the Euarthropoda can be resolved.

Onychophora strongly resemble arthropod-like animals. They are characterized by lobopods with pads and claws, one set of antennae, reminiscents of primary segmentation and a ladder-like central nervous system. An extensive fossil record exists of Onychophora-like taxa, the Lobopodia. Morphologically the classification of these taxa within the Arthropoda is still discussed, but it seems that evidence for a close relation to other Lobopodia and Euarthropoda outbalances other interpretations (RAMSKÖLD & JUNYUAN 1998). But some authors conclude an ambiguous position of the Onychophora constituting polyphyletic, unresolved relationships of the taxa Onychophora, Lobopodia and Tardigrada as SG to Euarthropods (WALOSZEK ET AL. 2005; BUDD 2009; BUDD 2001).

Molecular analyses place Onychophora either as sister group to Tardigrada + Euarthropoda (REGIER ET AL. 2005; SHULTZ & REGIER 2000) or sistergroup to Euarthropoda (BRUSCA & BRUSCA 2003; DUNN ET AL. 2008; GIRIBET 2008; MALLATT ET AL 2006; TELFORD ET AL. 2008) which is the most common finding.

The rRNA data based analyses ([A] + [B]) cannot enlight the position of Onychophora within arthropods. The Onychophora drop into the euarthropods within these analyses, except the tree based on the hand-optimized dataset (figure 3.5).

Congruent to cited molecular and phylogenomic studies a strong support for a clade Onychophora + Euarthropoda, excluding tardigrades is received with the phylogenomic data (analysis [C]). It appears that onychophorans are the sistergroup of euarthropods. The extensive fossil record of onychophorans (EDGECOMBE 2009) can thus profitably be used to improve our time scale of arthropod evolution.

Tardigrada are tiny animals of which no consensus exists at the moment regarding their phylogenetic position (RAMSKÖLD & JUNYUAN 1998; EDGECOMBE 2009). BRUSCA & BRUSCA (1990) place Tardigrada in the lineage leading to Euarthropoda above the Onychophora. Anyhow, some morphological characters of Tardigrada are reminiscent of both Arthropoda and Cycloneuralia (BRUSCA & BRUSCA 2003, GIRIBET 2003). Arthropod-like characters of tardigrades include the segmented body, limbs and ladder-like central nervous system (BRUSCA & BRUSCA 2003, GIRIBET 2003). On the other hand, structures of mouth, pharynx, cuticle and sensory organs of tardigrades resemble those of some Cycloneuralia (GIRIBET 2003). Traditionally, tardigrades are allied with arthropods (BRUSCA & BRSUCA 1990; BRUSCA & BRUSCA 2003), an arrangement that is recovered also by molecular phylogenetic studies based on ribosomal RNA (MALLATTT ET AL. 2004). Unfortunately some molecular studies are not able to resolve the position of tardigrades revealing a polytomic clade Tardigrada + Onychophora + Euarthropoda (REGIER ET AL. 2005; REGIER ET AL. 2008). A clade of Tardigrada + (Onychophora + Euarthropoda) (DUNN ET AL. 2008; ROEDING ET AL. 2007) would support the evolution of segmentation, segmented appendages and a ladder-

like central nervous system within this group. A sistergroup relationship of tardigrades with Cycloneuralia (LARTILLOT ET AL. 2008, but without Onychophora!; BLEIDORN ET AL. 2009) would imply a loss of these characteristics within Cycloneuralia and a very ancient evolution of a segmented body plan as the most parsimonious explanation.

The not fully understood position of Onychophora and Tardigrada and the possible grouping of Tardigrada as SG to euarthropods or "Tactopoda" (BUDD 2001) makes an out-group choice of Tardigrada at least questionable to infer arthropod phylogeny. However, the rRNA based analyses [A] + [B] are basically focused on the internal relationships of the Euarthropoda. Thus the selection of tardigrades can be justified as chosen out-group. For the phylogenomic analysis the out-group is changed to Mollusca and a massively broadened taxon sampling is applied.

In the phylogenomic analysis tardigrades (*Hypsibius* and *Richtersius*) are recovered as sister group of nematodes (BS 100%, pP 1.0), corroborating results of recent phylogenomic studies (ROEDING ET AL. 2007; BLEIDORN ET AL. 2009). In contrast, DUNN ET AL. (2008) found tardigrades as SG of arthropods (including onychophorans) with the CAT model of amino acid evolution (LARTILLOT & PHILIPPE 2004), whereas applying the WAG model the result also suggests an association of tardigrades and nematodes. It might therefore be speculated whether some of the arthropod characteristics are actually plesiomorphic for arthropods and are shared character states of a much more inclusive group. However, this interpretation is still very preliminary since data of several important representatives of ecdysozoan taxa are missing.

4.4 General methodological discussion

Based on the previous specific methodological and phylogenetic discussions some methodological issues and problems are summarized and discussed here.

Taxon choice

In recent publications taxon sampling and the choice of taxa, which are included into analysis are discussed as influential parameters. One reason is that some species have shown long branches in molecular analyses that bias the tree reconstruction (AGUINALDO ET AL. 1997; PHILIPPE ET AL. 2005B; BRINKMANN ET AL. 2005, BRINKMANN & PHILIPPE 2008; HEATH ET AL. 2008). The underlying effect of long branch attraction (FELSENSTEIN 1978) is already mentioned in previous paragraphs. This effect is more complex as commonly thought and different types of long branch effects are described in WÄGELE & MAYER (2007). The often-used term and subsequently the practice of "…breaking long branches by adding more taxa…" (BERGSTEN 2005; BRINKMANN & PHILIPPE 2008; Zwickl & Hillis 2002; Hendy & Penny 1989) is a method to minimize long branch artifacts that are caused by multiple substitutions. This effect is named a class III artifact in WÄGELE & MAYER (2007). Addition of taxa to shorten internal long branches of a topology circumvents a second case of LBA,

namely the class I LBA (WÄGELE & MAYER 2007). In this case symplesiomorphic positions reconstruct paraphyletic groups. This effect causes a false interpretation of the (in fact) plesiomorphic positions as synapomorphies. Adding more taxa that are positioned in between the internal and terminal taxon or the out-group and the internal taxa (or both) can minimize this effect.

In this study it was tried to sample at least two species for each crustacean group that are not too derived from the ground pattern of the group. The aim is to minimize class III effects by collecting taxa that show no long branches. Two taxa per group are collected to avoid or minimize class I effects. However, as seen in figures 3.8 and 3.9 especially the two species *Speleonectes tulumensis* (Remipedia) and *Hutchinsoniella macracantha* (Cephalocarida) show extremely long branches. It was not possible for those two groups to include further species in the analysis. It could be achieved unfortunately not until the end of the thesis to collect two further species for both groups. Despite extensive laboratory effort it was not possible to sequence these in time.

In analysis [B] it is clearly demonstrated that for *Hutchinsoniella* obviously the compositional heterogeneity plays a more important role than long branch effects. The position of *Hutchinsoniella* in the time-heterogeneous tree is morphologically rather plausible. So it seems that this is one example for a long branch taxon that is not affected by long branch artifacts but instead extremely biased by other problematic effects, namely the compositional heterogeneity of base frequencies (see later paragraph).

However, analysis [B] also reveales that rRNA data is problematic for crustaceans regarding signal erosion in the sequences (see later paragraph). Some stem lineages show extremely short branches, e.g. for the Malacostraca. It is possible that in these cases class II LBA effects occur, but this is not to handle by taxon choice.

Composite or hybrid taxa

A commonly used approach to compile larger datasets (SHOSHANI & MCKENNA 1998; BAURAIN ET AL. 2006; PHILIPPE ET AL. 2006; DELSUC ET AL. 2008) is the creation of "composite" or "chimerical" taxa (SPRINGER ET AL. 2004; MALIA ET AL. 2003) to minimize missing data or taxa.

For the single gene analyses [A] and [B] chimerical taxa are also constructed. MALIA ET AL. (2003) object that the construction of chimerical taxa is only acceptable if the combined taxa belong to a monophyletic group. This objection is expanded in this thesis and if possible only sequences of species of the same genus are combined. Of course the best solution would be to include only sequences from single species, but in some cases this is not to achieve and the use of chimerical taxa is to be favored over missing data. This is especially the case for smaller datasets comprising only few genes, as it is the case for analyses [A] and [B] of this study. For analysis [C] a different approach is chosen applying the software MARE. The choice and inclusion of genes and taxa depends here on the quality of the relative information content of each gene per taxon.

Quality of the raw data (raw sequences)

Contamination is always an omnipresent problem in molecular data. Even working very clean and focused in the molecular laboratory contaminations can easily occur. Thus, the first step should always be a BLAST analysis of the own sequences after receiving the sequenced nucleotide fragment to check for possible contamination. Especially using published sequences a BLAST analysis should ensure that the sequences are not contaminated. Though the problem of contamination is well known, only few studies (COLGAN ET AL. 2008; WÄGELE ET AL. 2009) reveal contamination of published sequences.

A good example of a single gene phylogeny reconstruction biased by contamination is analysis [A] in which the sequences of *Derocheilocaris* (mystacocarid sequence no. 2) and *Dyastilis* are revealed as very likely contaminated. This example is described in chapter 3.1.4.

For the EST data the contamination problem is transferred to another dimension. The own EST projects included in the phylogenomic analysis pass an especially designed pipeline to detect vector sequences, poly-A tails and bacterial contamination, using a BLAST analysis (see figure 2.9). Contamination by other species is checked using the TGICL package. A check by eye or Blast2Blast procedures is not possible for EST libraries because of the large number of sequences.

A problem remains using total RNA for EST data. If species of very small body size are used it is not possible to dissect tissue samples. This procedure would ensure that only tissue is used that is not contaminated by DNA/RNA of other organisms. The risk is high in the case of the digestive tract. For the samples of *Polllicipes pollicipes* and *Triops cancriformis* it was possible to prepare "clean" tissue. However for small specimens like the copepods whole specimens had to be fixed. Of the copepod *Tigriopus californicus* (average body size of 0.7 mm) about 7500 specimens were pooled. For that reason a special strain of cultured specimes was ordered to ensure that only one species is used for the cDNA library reconstruction.

Quality of the alignment (aligned sequences)

Sequence alignment: The importance of sequence alignment and of a careful choice of alignment programs is clearly demonstrated with analysis [A]. The choice of the best, appropriate alignment software from a collection of over 50 MSA programs (NOTREDAME 2002; WALLACE ET AL. 2006) depends of course on the chosen molecular markers. It is generally believed that in most phylogenetic single gene studies a careful choice of genes makes the alignment problem less difficult (Wong et al. 2008), compared to genomic studies. In analysis [A] it is revealed that a difference in topology is the result using comparatively MAFFT or MUSCLE as alignment programs. But as demonstrated in the network reconstructions (prior to the phylogenetic reconstruction) for that analysis (figures 3.1-3.2), the signal for the chosen (but commonly often used) 16S, 18S and COI markers is strongly eroded for deeper nodes. It can be reasoned here that if the resulting topologies of a dataset vary significantly using different alignment software, an accurate inspection of the signal within the dataset is demanded. If the signal is eroded, too many ambiguous trees can

be inferred despite the fact that they are extremely distant from the "more realistic" evolutionary tree. The author suggests to accomplish this procedure always as a standard routine by applying network reconstructions independently from the resulting topologies.

Secondary structure guided sequence alignment: Most studies relying on rRNA data use only fragments of the included rRNA genes (DELL AMPIO ET AL. 2009; KJER 2004; MISOF ET AL. 2007; D'HEASE 2002; GIRIBET ET AL. 2004; LUAN ET AL. 2005; EDGECOMB & GIRIBET 2002; KJER ET AL. 2006; YAMAGUCHI & ENDO 2003), while only few studies use nearly complete 18S and 28S rRNA sequences (MALLATT & GIRIBET 2006; MALLATT ET AL. 2004; GAI ET AL. 2006). The results of analyses [A] reveal that much efforts need to be done for an optimal secondary structure constraint choice. Obviously this seems to be one crucial point for rRNA studies that implement a secondary structure based alignment procedure relying on rRNA fragments. If the constraint is not especially adapted it might be that too many positions have to be excluded while the aligning process is performed. Furthermore, for later phylogenetic analysis all stem positions have to be transformed in "single, not paired" positions if the corresponding, paired stem position is not present in the sequence fragment. That procedure is reducing the number of stem positions and might end in a chaotic condition of the mixed models, by non-convergence of many of the mixed model parameter values. In analysis [B] only completely sequenced 18S and 28S rRNA genes are implemented to prevent biasing effects introduced by incomplete sequence fragments.

Alignment processing and evaluation: Alignment errors that bias tree reconstructions attract attention in recent studies (DRESS ET AL. 2008; LÖYTYNOJA & GOLDMAN 2005; OGDEN & ROSENBERG 2006; WÄGELE & MAYER 2007; DRESS ET AL. 2008; WONG ET AL. 2008; MISOF & MISOF 2009; HARTMANN & VISION 2008). To avoid noise caused by ambiguous alignment positions an automated alignment evaluation is conducted with the software ALISCORE. The automated alignment processing and evaluation (see figure 2.6: single genes and figure 2.10: phylogenomic data) improves phylogenetic analyses by founding the phylogenetic reconstructions on data with a higher probability of positional homology. However, automated processes should always be used and handled with a suspicious mind. The software ALISCORE (MISOF & MISOF 2009) was revealed to improve the reliability of topologies by excluding ambiguous aligned positions in the alignments but it can be discussed why the resolution of the trees might decrease. This was especially the case for analysis [A]. For the author reliability of the resulting topologies based on identification of ambiguous alignment positions is always to prefer over an eventually better, but suspicious resolution. Manual alignment procedures (as described in 2.4.3), even following a constraint (like a secondary structure constraint) cannot guarantee an objective alignment based on positional homology. Indication for misalignment in the manually aligned dataset is the clustering of the mystacocarids. Anyhow, for analysis [A] the main problem was signal erosion of the chosen markers and not misalignment, which is demonstrated and discussed in previous sections.

Quality of the signal in the aligned sequences – networks versus trees: Several studies advise and promote the use of phylogenetic networks or split decomposition to infer alignment and data quality (BANDELT & DRESS 1992; HUSON & BRYANT 2006; WÄGELE & MAYER 2007). Also the estimation and visualization of conflicts in the alignment or between different genes (HUSON 1998; HOLLAND & MOULTON 2003; HOLLAND ET AL. 2004; WHITFIELD & KJER 2008; WÄGELE & MAYER 2007) is possible relying on these methods. These existing methods to estimate data or alignment quality are still insufficiently used in publications (WÄGELE ET AL. 2009).

The extensive use and discussion of network reconstruction in the present analyses should have demonstrated that believing in bifurcation is a dangerous trap that can be avoided by using more objective software. Resolved trees are good, but reliable ones are better. To date network reconstruction is the only tool to make conflicts or eroded signal in the data visible. The software SAMS presented in WÄGELE & MAYER (2007) relying on split decomposition needs still some improvement. At the moment a graphical user interface is not available for this software, which makes the use very time consuming and not feasible for the included analyses. The task to build a graphical user interface is in progress (personal communication).

Phylogenetic reconstruction – aspects of modeling

General aspects of complex modeling: All models and methods that are used in molecular phylogenetic analysis are approximations or simplifications of the processes of molecular evolution. Each phylogenetic estimation requires assumptions, which are made about the process that finally results in the observed dataset (KELCHNER & THOMAS 2007). These assumptions are the fundament for the applied model. The early substitution models used in phylogenetic analyses assume that the evolutionary process along lineages followes a constant rate and pictures a homogenous process (see review of: LIO & GOLDMAN 1998, but also: WHELAN ET AL. 2001B; FOSTER 2004; COX ET AL. 2008). It was soon discovered that this assumption is violating the real evolution of genes, which is very likely a heterogeneous process (LIO & GOLDMAN 1998; TARRIO ET AL. 2001; FOSTER 2004; GOWRI-SHANKAR & RATTRAY 2007; COX ET AL. 2008). In other words, the underlying assumption of most phylogenetic models that nucleotide or amino acid frequencies do not change over time along lineages and that rates are constant is not correct. There exists a compositional heterogeneity along the tree (FOSTER 2004; COX ET AL. 2008).

A common opinion is that models only need to be good approximations of the reality (KELCHNER & THOMAS 2007) and that too complex models can also mislead the phylogenetic reconstruction. In some datasets compositional heterogeneity exists among lineages but the phylogenetic signal is strong and not in interference with compositional heterogeneity effects (FOSTER 2004). In these cases a simple model assuming compositional homogeneity can reconstruct a topology close to the "true" evolutionary tree and complex modeling is not necessary.

However, SULLIVAN & SWOFFORD (2001) discuss that the impact of parameters, which are violated in the model is important in subsequent tree reconstructions. A badly fitting model can sometimes more efficiently infer the correct tree. That is possible if the bias introduced by the model violations supports the reconstruction of the "true" tree rather than an incorrect one. The crux of the matter is that we normally do not know for empirical data if bias in the data increases or decreases the accuracy of the reconstruction, which is dependent on the model assumptions. For several published datasets misleading effects and biases evoked by compositional base heterogeneity are reported (TARRIO ET AL. 2000; TARRIO ET AL. 2001; Foster 2004; PHILIPPE ET AL. 2005B; GOWRI-SHANKAR & RATTRAY 2007; COX ET AL. 2008; DAVALOS & PERKINS 2008; FOSTER ET AL. 2009). Referring to those, an observed compositional base heterogeneity can compromise a phylogenetic reconstruction that relies only on simple, standard models. The best choice in that case is to implement time-heterogeneity in the model assumptions or to compare both approaches using a complex and a standard model, similar to analysis [B] in this thesis.

Modeling for the single gene data: From the phylogenetic point of view the extremely sophisticated analysis [B] ends in a phylogeny below the author's expectations regarding a better resolution of crustacean phylogeny. However, as discussed and demonstrated, this is an important test case to identify the improvements of complex modeling considering compositional heterogeneity along the tree compared to standard procedures (REUMONT ET AL. 2009). The improvements and obvious artifacts evoked in standard methods relying on compositional homogeneity along the tree are underlined. Complex modeling is important. In this light, some often cited specific results of previous studies relying on rRNA data (e.g. SPEARS & ABELE 1998; GIRIBET ET AL. 1996; GIRIBET & RIBERA 2000) obtain a different flavor for the interpretation of crustacean phylogeny. However, many processes of molecular sequence evolution are still not fully understood, like LBA (WÄGELE & MAYER 2007) and need further, more investigation. Especially for groups like Remipedia, Pentastomida and Myriapoda, it remains unclear which effects mislead the reconstruction in analysis [B].

Modeling and gene-choice for the phylogenomic data: The complex setting for analysis [B] is a test case for the influence of inhomogeneous base compositions. Regarding the phylogenomic data it is assumed that compositional heterogeneity along the tree influences phylogenetic reconstructions to a greater extent than previously assumed (JEFFROY ET AL. 2006). Implementing and developing heterogeneous modeling to protein data might be another point of eminent importance to avoid misleading reconstructions as demonstrated exemplarily for the rRNA data. This is demonstrated by SEO & KISHINO (2008) for synonymous substitutions processes, see also WHELAN (2008). FOSTER ET AL. (2009) recently apply a new approach that employs "composition heterogeneous-methods" to consider for phylogenomic data compositional heterogeneity along the tree (they used two datasets, protein and rRNA data).

For the phylogenomic data an even more complicated situation exists regarding the choice of suitable genes. There is a consensus that at the moment the supermatrix approach is to be preferred over the supertree method (DE QUEIROZ 2006; SANDERSON & DRISKELL 2003,

MISOF ET AL. submitted). However, so far the impact of chosen genes is not really understood. An enormous amount of different methods used to identify orthologous genes (e.g. ZHOU & LANDWEBER 2007; EBERSBERGER ET AL. 2009; SCHREIBER ET AL. 2009) is borne in the meantime, but the impact of single genes on the reconstructed topology remains in the darkness. This is unfortunately also the case for the present phylogenomic study. Handling phylogenomic data to date is like using a black box producing the final topology. With the presented MARE-approach (MISOF ET AL., in prep) a great step is made towards a more specific selection of genes, based on potential relative information content. Still demanded are tools that identify conflicts in the data and respectively between single genes. Identifying conflicting splits within the data respectively within the alignment will be an essential task. Tools like SAMS (WÄGELE & MAYER 2007) that can accomplish this need further development and adaptation to protein data.

Proceeding the analyses for this thesis it was unsatisfying that working with such sophisticated settings and extensively enlarged datasets the computational power and the existing software is limited. Many aspects and interesting further questions of each analysis could not be addressed in detail for this reason and due to the time limitation.

Crustacean phylogeny today not to recover with single gene data? Existing studies using single genes or rRNA data show ambiguous and partly unresolved results for crustacean phylogeny as previously shown and discussed. It is one hope borne from the experience of the highly sophisticated rRNA analysis presented in this thesis and REUMONT ET AL. (2009) to transfer this approach of "complex modeling" to phylogenomic data, as attempted in first studies in this field (FOSTER ET AL .2009).

However, with new tools developed to identify contradicting signal it might turn out that in the evolutionary process of crustaceans for some groups the signal in sequence data is eroded. The best and most complex and sophisticated analysis is useless in that case. It might be speculated that this scenario will not change applying genomic approaches based on the sequence data. A totally new approach would be the search for "patterns" or "phrases" within the sequences, which would be used as single characters with higher complexity. The first studies of these "word-orientated" alignments (BEIKO ET AL. 2005; DIDIER ET AL. 2007) are promising. Eventually, this approach is one solution to recover the phylogeny of such old groups like crustaceans in which the signal might be eroded at the level of nucleotides.

4.5 Conclusions and further aspects

Synergistic collaboration for the future (taxa & genes)

The author of the present thesis cannot avoid one sentence very common in most molecular studies, namely that more and further taxa should be included in future analyses. Unfortunately, at least for crustaceans and probably myriapods this is absolutely essential and demanded.

Regarding the rRNA study it would be interesting to extend taxon sampling to more species per group. Unfortunately the problematic taxa (Pentastomida, Cephalocarida, Remipedia) are very hard to sequence. Based on the revealed conflict in the data this might be using a cannon to kill a mosquito. Anyhow, a future collaboration with the MALLATT-laboratory (Washington State University, Washington) could be established and plans for an extensive further rRNA analyses are sketched.

As mentioned earlier, the existing phylogenomic data on crustaceans is recently summarized by STILLMAN ET AL. (2008) (see supplementary table S14). A clearly unbalanced batch of malacostracan EST sequences overbalances underrepresented non-malacostracan EST projects. Future plans are to collect Cephalocarida and Pentastomida to conduct pyrosequencing projects (Roche, 454 Titanium). The tissue of newly collected Mystacocarida is in preparation. 454 runs of Remipedia, Ostracoda and Leptostraca are in progress and were started while finishing the thesis. To include a broader and more even malacostracan sampling a cooperation with the JENNER–group (Natural History Museum, London) is settled.

Further development of phylogenomic data analysis and data quality assessment tools

Just adding more and more taxa or genes is obviously not the way to improve future molecular studies (PHILIPPE ET AL. 2005), except to correct the unbalanced taxon sample as stated before). It seems more important to investigate the impact of different methods and strategies to analyze molecular data. The identification of conflicts and contradicting signal is one of the major topics in future analyses.

A further aim is the identification of suitable marker genes using phylogenomic data. Eventually phylogenomic data brings one back to the point of single to multi gene analyses relying only on a few genes. Suitable genes can be identified with tools like MARE. This would be a strategy to identify promising molecular markers with less conflict revealing a clear phylogenetic signal. A step toward this direction is made within this thesis. In the so called "primer toolbox" specific DNA primers are constructed from EST data (single gene alignments) to amplify interesting, new marker genes. This enables one to sequence promising genes for rare species of which only few tissue exists without the need of an EST or pyrosequencing project. A similar approach is in parallel developed in the lab of CUNNINGHAM and coworkers (http://www.biology.duke.edu/cunningham/DeepArthropod.html) but they amplify the genes with degenerated primers via mRNA isolation and cDNA

generation (see REGIER ET AL. 2005; REGIER ET AL. 2008) instead directly from DNA. The approach in this thesis works for the crustaceans with one gene (rpl11b, a nuclear ribosomal protein) but is very time consuming and not successful for all tested crustacean taxa. Designing primers for such a heterogeneous group like the Crustacea demands an intensive work in the laboratory, for many major crustacean groups different specific primers have to bee designed. Also the settings for PCR reactions have to be changed between some crustaceans groups. The genes were chosen by hand, main criteria are the overlap with most arthropod and crustacean taxa and the existence of suitable conservative regions for the primer design. With a software like MARE the selection can be automated and based on potential information content of the chosen genes, which is much more promising and effective. The "primer toolbox" is tested and extended with several working groups, the author created a common platform to enable exchange of laboratory protocols and procedures (see figure 4.1).

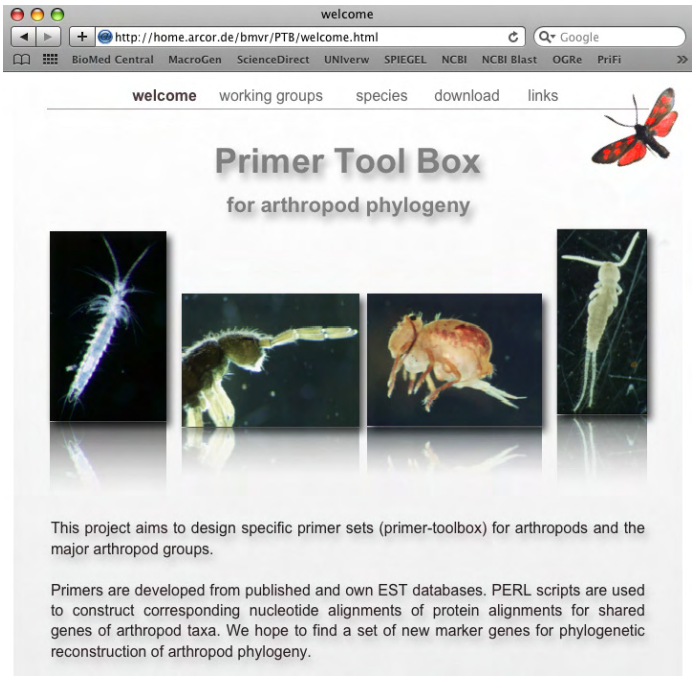


Figure 4.1: Primer toolbox platform. The open sourced internet homepage with diverse information of the, cooperating international working groups, included arthropod species and the laboratory protocols and procedures. Alignments of EST based genes are also to download.

http://home.arcor.de/bmvr/PTB/

Also, a comparison of nucleotide vs. protein matrices could be promising. At the moment protein models are less understood. Site variability (LE ET AL. 2008; QUANG ET AL. 2008) and also time-heterogeneity (FOSTER ET AL. 2009) should be implemented. A comparison of both, nucleotide and protein sequence levels and the further development of protein models in combination with the software MARE (extended to these issue) would be a preferred strategy of the author. Working and collecting experience at the nucleotide level of phylogenomic data would additionally result in one advantage: It provides the basis for a different phylogenetic approach like the use of word orientated alignment procedures as previously mentioned (see BEIKO ET AL. 2005; DIDIER ET AL. 2007).

Towards a sophisticated total evidence analysis

If the molecular analyses are grown in the direction that the best possible modeling is to apply – a goal that is not yet achieved - combined total evidence analysis including morphological and fossil data would be a further final goal. But until this can be reached a lot needs to be done.

5. REFERENCES

ABASCAL, F. (2005). ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21,** 2104-2105.

ABELE, L. G., KIM, W. & FELGENHAUER, B. E. (1989). Molecular evidence for inclusion of the phylum Pentastomida in the Crustacea. *Molecular Biology and Evolution* **6,** 685-691.

ABELE, L. G., SPEARS, T., KIM, W. & APPLEGATE, M. (1992). Phylogeny of selected maxillopodan and other crustacean taxa based on 18S ribosomal nucleotide sequences: a preliminary analysis. *Acta Zoologica (Stockholm)* **73,** 373-382.

ADACHI, J. & HASEGAWA, M. (1996). Model of amino acid substitution in proteins encoded by mitochondrial DNA. *Journal of Molecular Evolution* **42,** 459-68.

ADAMS, M. D., KELLEY, J. M., GOCAYNE, J. D., DUBNICK, M., POLYMEROPOULOS, M. H., XIAO, H., MERRIL, C. R., WU, A., OLDE, B. & MORENO, R. F. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252,** 1651-6.

AGUINALDO, A. M., TURBEVILLE, J. M., LINFORD, L. S., RIVERA, M. C., GAREY, J. R., RAFF, R. A. & LAKE, J. A. (1997). Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* **387,** 489-93.

AX, P. (1999). *Das System der Metazoa II. Ein Lehrbuch der phylogenetischen Systematik*.

BABBIT, C. C. & PATEL, N. H. (2005). Relationships within the Pancrustacea: Examining the influence of additional malacostracan 18S and 28S rDNA. In *Crustacea and arthropod relationships* (ed. S. Koenemann and R. A. Jenner). Taylor and Francis, London.

BÄCKER, H., FANENBRUCK, M. & WÄGELE, J. W. (2008). A forgotten homology supporting the monophyly of Tracheata: The subcoxa of insects and myriapods re-visited. *Zoologischer Anzeiger-A Journal of Comparative Zoology* **247,** 185-207.

BALLARD, J. W., OLSEN, G. J., FAITH, D. P., ODGERS, W. A., ROWELL, D. M. & ATKINSON, P. W. (1992). Evidence from 12S ribosomal RNA sequences that onychophorans are modified arthropods. *Science* **258,** 1345-8.

BAN, N., NISSEN, P., HANSEN, J., MOORE, P. B. & STEITZ, T. A. (2000). The complete atomic structure of the large ribosomal subunit at 2.4 Å Resolution. *Science* **289,** 905-920.

BANDELT, H. J. & DRESS, A. W. (1992). Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Molecular Phylogenetics and Evolution* **1,** 242-52.

BAURAIN, D., BRINKMANN, H. & PHILIPPE, H. (2007). Lack of Resolution in the Animal Phylogeny: Closely Spaced Cladogeneses or Undetected Systematic Errors? *Molecular Biology and Evolution* **24,** 6-9.

BEIKO, R., KEITH, J., HARLOW, T. & RAGAN, M. (2006). Searching for Convergence in Phylogenetic Markov Chain Monte Carlo. *Systematic Biology* **55,** 553-565.

BEIKO, R. G., CHAN, C. X. & RAGAN, M. A. (2005). A word-oriented approach to alignment validation. *Bioinformatics* **21,** 2230-2239.

BEKLEMISHEV, W. N. (1952). *Principles of Comparative Anatomy of Invertebrates.*, 2nd edition Moscow: Nauka.

BERGSTEN, J. (2005). A review of long-branch attraction. *Cladistics* **21,** 163-193.

BININDA-EMONDS, O. (2004). The evolution of supertrees. *Trends in Ecology & Evolution* **19,** 315-322.

BININDA-EMONDS, O. R. (2004). The evolution of supertrees. *Trends in Ecology & Evolution (Amst)* **19,** 315-22.

BIRNEY, E., CLAMP, M. & DURBIN, R. (2004). GeneWise and Genomewise. *Genome Research* **14,** 988-95.

BITSCH, C. & BITSCH, J. (2004). Phylogenetic relationships of basal hexapods among the mandibulate arthropods: a cladistic analysis based on comparative morphological characterscomparative morphological characters. *Zoologica Scripta* **33,** 511-50.

BLANQUART, S. & LARTILLOT, N. (2006). A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. *Molecular Biology and Evolution* **23,** 2058-2071.

BLEIDORN, C., PODSIADLOWSKI, L., ZHONG, M., EECKHAUT, I., HARTMANN, S., HALANYCH, K. M. & TIEDEMANN, R. (2009). On the phylogenetic position of Myzostomida: Can 77 genes get it wrong? *BMC Evolutionary Biology* **9,** 150.

BOGUSKI, M. S. (1995). The turning point in genome research. *Trends in Biochemical Sciences* **20,** 295-6.

BOORE, J. L., COLLINS, T. M., STANTON, D., DAEHLER, L. L. & BROWN, W. M. (1995). Deducing the pattern of arthropod phylogeny from mitochondrial DNA rearrangements. *Nature* **376,** 163-5.

BOSCH, H. (1986). Experimental life-cycle studies of Raillietiella Sambon, 1910 (Pentastomida: Cephalobaenida): the fourth-stage larvae is infective for the definitive host. *Parasitology Research* **72,** 673-680.

BOSCH, H. (1987). Vergleichende experimentelle Untersuchungen zur Biologie der Pentastomidengattungen Raillietiella (Cephalobaenida) und Elenia (Porocephalida) unter BerÜcksichtigung der Erforschungsgeschichte der Pentastomiden. *Dissertation, Abteilung fÜr Parasitologie des Instituts fÜr Zoologie der Universität Hohenheim*

BOUCK, A. & VISION, T. (2007). The molecular ecologist's guide to expressed sequence tags. *Molecular Ecology* **16,** 907-924.

BOUDREAUX, H. B. (1979). *Arthropod phylogeny, with special reference to insects*. Wiley, New York.

BOURLAT, S. J., NIELSEN, C., ECONOMOU, A. & TELFORD, M. (2008). Testing the new animal phylogeny: A phylum level molecular analysis of the animal kingdom. *Molecular Phylogenetics and Evolution* **49,** 23.31.

BOXSHALL, G. A. (1983). A comparative functionaly analysis of the major maxillopodan groups. *Crustacean Issues* **1,** 121-143.

BOXSHALL, G. A. & HUYS, R. (1989). New tantulocarid, Stygotantulus stocki, parasitic on harpacticoid copepods, with an analysis of the phylogenetic relationships within Maxillopoda. *Journal of Crustacean Biology* **9,** 126-140.

BRABAND, A., RICHTER, S., HIESEL, R. & SCHOLTZ, G. (2002). Phylogenetic relationships within the Phyllopoda (Crustacea, Branchiopoda) based on mitochondrial and nuclear markers. *Molecular Phylogenetics and Evolution* **25,** 229-44.

BRENNEIS, G., UNGERER, P. & SCHOLTZ, G. (2008). The chelifores of sea spiders (Arthropoda, Pycnogonida) are the appendages of the deutocerebral segment. *Evolution & Development* **10,** 717-724.

BRENNER, S. (1990). The human genome: the nature of the enterprise. *Ciba Found Symp* **149,** 6-12; discussion 12-7.

BRINKMANN, H. & PHILIPPE, H. (2008). Animal phylogeny and large-scale sequencing: progress and pitfalls. *Journal of Systematics and Evolution* **46,** 274-286.

BRINKMANN, H., VAN DER GIEZEN, M., ZHOU, Y., DE RAUCOURT, G. & PHILIPPE, H.

(2005). An Empirical Assessment of Long-Branch Attraction Artefacts in Deep Eukaryotic Phylogenomics. *Systematic Biology* **54,** 743-757.

BROWN, J. & LEMMON, A. (2007). The Importance of Data Partitioning and the Utility of Bayes Factors in Bayesian Phylogenetics. *Systematic Biology* **56,** 643-655.

BRUSCA, R. C. & BRUSCA, G. J. (1990). *Invertebrates*, first edition. Sinauer Associates.

BRUSCA, R. C. & BRUSCA, G. J. (2003). *Invertabrates*, second edition. Sinauer Associates.

BUCKLEY, T. R., SIMON, C., FLOOK, P. K. & MISOF, B. (2000). Secondary structure and conserved motifs of the frequently sequenced domains IV and V of the insect mitochondrial large subunit rRNA gene. *Insect Molecular Biology***9,** 565-80.

BUDD, G. E. (1996). Progress and problems in arthropod phylogeny. *Trends in Ecology & Evolution* **11,** 356-358.

BUDD, G. E. (2001). Tardigrades as 'stem-group arthropods': the evidence from the Cambrian fauna. *Zoologischer Anzeiger* **240,** 260-279.

BUDD, G. E. (2002). A palaeontological solution to the arthropod head problem. *Nature* **417,** 271-5.

BUDD, G. E. & TELFORD, M. (2009). The origin and evolution of arthropods. *Nature* **457,** 812-7.

BUHAY, J. E. (2009). "COI-like" Sequences are Becoming Problematic in Molecular Systematic and DNA Barcoding Studies. *Journal of Crustacean Biology* **29,** 96-110.

BURMESTER, T. (2001). Molecular evolution of the arthropod hemocyanin superfamily. *Molecular Biology and Evolution* **18,** 184-195.

BURMESTER, T. (2002). Origin and evolution of arthropod hemocyanins and related proteins. *Journal of Comparative Physiology B, Biochemical, Systemic and Environmental Physiology* **172,** 95-107.

CAMERON, S. L., MILLER, K. B., D'HAESE, C., WHITING, M. F. & BARKER, S. C. (2004). Mitochondrial genome data alone are not enough to unambiguously resolve the relationships of Entognatha, Insecta and Crustacea sensu lato (Arthropoda). *Cladistics* **20,** 534-557.

CANNONE, J. J., SUBRAMANIAN, S., SCHNARE, M. N., COLLETT, J. R., D'SOUZA, L. M., DU, Y., FENG, B., LIN, N., MADABUSI, L. V., MÜLLER, K. M., PANDE, N., SHANG, Z., YU, N. & GUTELL, R. R. (2002). The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics* **3,** 2.

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 Evolutionary Biology* **7,** 1-13.

CARAPELLI, A., NARDI, F., DALLAI, R., BOORE, J., LìO, P. & FRATI, F. (2005). Relationships between hexapods and crustaceans based on four mitochondrial genes. In *Crustacea and arthropod relationships* (ed. S. Koenemann and R. A. Jenner), pp. 295-306. Taylor and Francis, CRC press, Boca Raton.

CASTRESANA, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution* **17,** 540-52.

CHOU, H. & HOLMES, M. H. (2001). DNA sequence quality trimming and vector removal. *Bioinformatics* **17,** 1093.

CISNE, J. L. (1974). Evolution of the world fauna of aquatic free-living arhtropods. *Evolution* **28,** 337-366.

COBBETT, A., WILKINSON, M. & WILLS, M. A. (2007). Fossils impact as hard as living taxa in parsimony analyses of morphology. *Systematic Biology* **56,** 753-66.

COHEN, A. C., MARTIN, J. W. & KORNICKER, L. S. (1998). Homology of holocene ostracode biramous appendages wiith those of other crustaceans: the protopod, epipod, exopod and endopos. *Lethaia* **31,** 251-265.

COLGAN, D., HUTCHINGS, P. & BEACHAM, E. (2008). Multi-Gene Analyses of the Phylogenetic Relationships among the Mollusca, Annelida, and Arthropoda. *Zoological Studies* **47,** 338-351.

COOK, C., YUE, Q. & AKAM, M. (2005). Mitochondrial genomes suggest that hexapods and crustaceans are mutually paraphyletic. *Proceedings of the Royal Society B: Biological Sciences* **272,** 1295-1304.

COX, C., FOSTER, P., HIRT, R., HARRIS, S. R. & EMBLEY, T. M. (2008). The archaebacterial origin of eukaryotes. *Proceedings of the national Academy of Sciences, USA***105,** 20356-61.

D'HAESE, C. (2002). Were the first springtails semi-aquatic? A phylogenetic approach by means of 28S rDNA and optimization alignment. *Proceedings of the Royal Society B: Biological Sciences* **269,** 1143-1151.

DAHL, E. (1956). *Some crustacean relationships*. Zoological Institute, Lund.

DAHL, E. (1963). Main evolutionary lines among recent Crustacea. In *Phylogeny and Evolution of Crustacea* (ed. H. B. Whittington and W. D. I. Rolfe), pp. 1-15. Museum of Comparative Zoology, Cambridge.

DÁVALOS, L. M. & PERKINS, S. L. (2008). Saturation and base composition bias explain phylogenomic conflict in Plasmodium. *Genomics* **91,** 433-442.

DAYHOFF, M. O., SCHWARTZ, R. M. & ORCUTT, B. C. (1978). A model for evolutionary change in proteins. In *Collection*, pp. 345-352.

DE OLIVEIRA ALMEIDA, W., CHRISTOFFERSEN, M. L., SOUSA AMORIN, D. & COSTA ELOY, E. C. (2008). Morphological support for the phylogenetic positioning of Pentastomida and related fossils. *Revista Biotemas* **3,** 81-90.

DE QUEIROZ, A. & GATESY, J. (2006). The supermatrix approach to systematics. *Trends in Ecology & Evolution (Amst)* **22,** 34-41.

DE RIJK, P., WUYTS, J., VAN DE PEER, Y. & WINKELMANS, T. (2000). The European large subunit ribosomal RNA database. *Nucleic Acids Research* **28,** 177-178.

DELL'AMPIO, E., SZUCSICH, N., 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. *Zoologica Scripta* **38,** 155-170.

DELSUC, F., BRINKMANN, H. & PHILIPPE, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet* **6,** 361-375.

DELSUC, F., PHILLIPS, M. J. & PENNY, D. (2003). Comment on "Hexapod origins: monophyletic or paraphyletic?". *Science* **301,** 1481-1482.

DELSUC, F., TSAGKOGEORGA, G., LARTILLOT, N. & PHILIPPE, H. (2008). Additional molecular support for the new chordate phylogeny. *Genesis* **46,** 592-604.

DEWAARD, J., SACHEROVA, V., CRISTESCU, M., REMIGIO, E., CREASE, T. & HEBERT, P. (2006). Probing the relationships of the branchiopod crustaceans. *Molecular Phylogenetics and Evolution* **39,** 491-502.

DIDIER, G., DEBOMY, L., PUPIN, M., ZHANG, M., GROSSMANN, A., DEVAUCHELLE, C. & LAPREVOTTE, I. (2007). Comparing sequences without using alignments: application to HIV/SIV subtyping. *BMC Bioinformatics* **8,** 1.

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. *Annales de la Societe Entomologique de France (New Series)* **37,** 85-103.

DRESS, A. W., FLAMM, C., FRITZSCH, G., GRÜNEWALD, S., KRUSPE, M., PROHASKA, S. J. & STADLER, P. F. (2008). Noisy: identification of problematic columns in multiple sequence alignments. *Algorithms for Molecular Biology:* **3,** 7.

DRUMMOND, A. J. & RAMBAUT, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* **7,** 214.

DUNGER, W. (1993). Überklasse Antennata. In *Lehrbuch der speziellen Zoologie*, vol. I Wirbellose Tiere (ed. H.-E. Gruner, M. Moritz and W. Dunger), pp. 1-1279. Gustav Fischer Verlag, Jena.

DUNLOP, J. A. (2005). New ideas about the euchelicerate stem-lineage. *European Arachnology***,** 9-23.

DUNLOP, J. A. & ARANGO, C. P. (2005). Pycnogonid affinities: a review. *Journal of Zoological Systematics and Evolutionary Research* **43,** 8-21.

DUNN, C., HEJNOL, A., MATUS, D., PANG, K., BROWNE, W., SMITH, S., SEAVER, E., ROUSE, G., OBST, M., EDGECOMBE, G., SØRENSEN, M., HADDOCK, S. H., SCHMIDT-RHAESA, A., OKUSU, A., KRISTENSEN, R., WHEELER, W., MARTINDALE, M. & GIRIBET, G. (2008). Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* **452,** 745-9.

EBERSBERGER, I., STRAUSS, S. & VON HAESELER, A. (2009). HamsTr: Profile Hidden markov Model Based Search for Orthologs in ESTs. *BMC Evolutionary Biology* **9,** 1-9.

EDGAR, R. C. (2004a). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5,** 1-19.

EDGAR, R. C. (2004b). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32,** 1792-7.

EDGECOMBE, G. (2000). Arthropod Cladistics: Combined Analysis of Histone H3 and U2 snRNA Sequences and Morphology. *Cladistics* **16,** 155-203.

EDGECOMBE, G. (2009). Palaeontological and Molecular Evidence Linking Arthropods, Onychophorans, and other Ecdysozoa. *Evolution: Education and Outreach* **2,** 178-190.

EDGECOMBE, G. & GIRIBET, G. (2002). Myriapod phylogeny and the relationships of Chilopoda. *Collection*, vol. III, pp. 143-168.

EDGECOMBE, G. D. (2004). Morphological data, extant Myriapoda, and the myriapod stem-group. *Contibutions to Zoology* **73,** 207-252.

EISEN, J. A. (1998). Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Research* **8,** 163-7.

EMERSON, M. J. & SCHRAM, F. (1991). Remipedia; part II, Palaeontology. *Proceedings of the San Diego Soiety of Natural History* **7,** 1-52.

ERTAS, B., V REUMONT, B. M., WÄGELE, J. W., MISOF, B., BURMESTER, T. (2009). Hemocyanin suggests a close relationship of Remipedia and Hexapoda. *Molecular Biology and Evolution*, doi:10.1093/molbev/msp186 (advanced access).

FANENBRUCK, M. (2003). Die Anatomie des Kopfes und des cephalen Skelett-Muskelsystems der Crustacea, Myriapoda und Hexapoda: Ein Beitrag zum phylogenetischen System der Mandibulata und zur Kenntnis der Herkunft der Remipedia und Tracheata, Ruhr Universität Bochum.

FANENBRUCK, M. & HARZSCH, S. (2005). A brain atlas of Godzillognomus frondosus Yager, 1989 (Remipedia, Godzilliidae) and comparison with the brain of Yager, 1987 (Remipedia, Speleonectidae): implications for arthropod relationships. *Arthropod Structure & Development* **34,** 343-378.

FANENBRUCK, M., HARZSCH, S. & WÄGELE, J. W. (2004). The brain of the Remipedia (Crustacea) and an alternative hypothesis on their phylogenetic relationships. *Proceedings of the national Academy of Sciences, USA* **101,** 3868-3873.

FELSENSTEIN, J. (1978). Cases in which Parsimony or Compatibility Methods Will be Positively Misleading. *Systematic Zoology* **27,** 401-410.

FELSENSTEIN, J. (1988). Phylogenies from molecular sequences: inference and reliability. *Annual Review of Genetics* **22,** 521-65.

FITCH, W. M. (1970). Further improvements in the method of testing for evolutionary homology among proteins. *Journal of Molecular Biology* **49,** 1-14.

FITCH, W. M. (2000). Homology a personal view on some of the problems. *Trends in Genetics* **16,** 227-31.

FORTEY, R. A., THOMAS, R. H. & (EDS). (1998). *Arthropod relationships*. Chapman and Hall, London.

FOSTER, P. (2004). Modeling Compositional Heterogeneity. *Systematic Biology* **53,** 485-495.

FOSTER, P., COX, C. & EMBLEY, T. (2009). The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. *Philosophical Transactions of the Royal Society B: Biological Sciences* **364,** 2197-2207.

FOX, G. E. & WOESE, C. R. (1975). The architecture of 5S rRNA and its relation to function. *Journal of Molecular Evolution* **6,** 61-76.

FRIEDRICH, M. & TAUTZ, D. (1995). Ribosomal DNA phylogeny of the major extant arthropod classes and the evolution of myriapods. *Nature* **376,** 165-7.

FRIEDRICH, M. & TAUTZ, D. (1997). An episodic change of rDNA nucleotide substitution rate has occurred during the emergence of the insect order Diptera. *Molecular Biology and Evolution* **14,** 644-53.

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.

GAI, Y., SONG, D., SUN, H. & ZHOU, K. (2006). Myriapod Monophyly and Relationships Among Myriapod Classes Based on Nearly Complete 28S and 18S rDNA Sequences. *Zoo Sci* **23,** 1101.

GALTIER, N. (2004). Sampling Properties of the Bootstrap Support in Molecular Phylogeny: Influence of Nonindependence Among Sites. *Systematic Biology* **53,** 38-46.

GALTIER, N. & GOUY, M. (1995). Inferring phylogenies from DNA sequences of unequal base compositions. *Proceedings of the national Academy of Sciences, USA* **92,** 11317-11321.

GARCíA-MACHADO, E., PEMPERA, M., DENNEBOUY, N., OLIVA-SUAREZ, M., MOUNOLOU, J. C. & MONNEROT, M. (1999). Mitochondrial genes collectively suggest the paraphyly of Crustacea with respect to Insecta. *Journal of Molecular Evolution* **49,** 142-9.

GERHOLD, D. & CASKEY, C. T. (1996). It's the genes! EST access to human genome content. *Bioessays* **18,** 937-981.

GILLESPIE, J. (2005). A Secondary Structural Model of the 28S rRNA Expansion Segments D2 and D3 for Chalcidoid Wasps (Hymenoptera: Chalcidoidea). *Molecular Biology and Evolution* **22,** 1593-1608.

GILLESPIE, J., JOHNSTON, J. S., CANNONE, J. J. & GUTELL, R. R. (2006). Characteristics of the nuclear (18S, 5.8S, 28S and 5S) and mitochondrial (12S and 16S) rRNA genes of Apis mellifera (Insecta : Hymenoptera): structure, organization, and retrotransposable elements. *Insect Molecular Biology* **15,** 657-686.

GIRIBET, G. (2008). Assembling the lophotrochozoan (=spiralian) tree of life. *Philosophical Transactions of the Royal Society B, Biological Science* **363,** 1513-22.

GIRIBET, G., CARRANZA, S., BAGUNA, J., RIUTORT, M. & RIBERA, C. (1996). First molecular evidence for the existence of a Tardigrada plus arthropoda clade. *Molecular Biology and Evolution* **13,** 76-84.

GIRIBET, G., EDGECOMB, G. D., CARPENTER, J. M., D'HAESE, C. & WHEELER, W. C. (2004). Is Ellipura monophyletic? A combined analysis of basal hexapod relationships with emphasis on the origin of insects. *Organisms Diversity & Evolution* **4,** 319-340.

GIRIBET, G., EDGECOMBE, G. D. & WHEELER, W. C. (2001). Arthropod phylogeny based on eight molecular loci and morphology. *Nature* **413,** 157-61.

GIRIBET, G. & RIBERA, C. (2000). A review of arthropod phylogeny: New data based on ribosomal DNA sequences and direct character optimization. *Cladistics* **16,** 204-231.

GIRIBET, G., RICHTER, S., EDGECOMBE, G. D. & WHEELER, W. C. (2005). The position of crustaceans within the Arthropoda - evidence from nine molecular loci and morphology. In *Crustacea and arthropod relationships* (ed. S. Koenemann and R. A. Jenner). Taylor and Francis, CRC press, Boca Raton.

GLENNER, H. & HEBSGAARD, M. B. (2006). Phylogeny and evolution of life history strategies of the parasitic barnacles (Crustacea, Cirripedia, Rhizocephala). *Molecular Phylogenetics and Evolution* **41,** 528-38.

GLENNER, H., THOMSEN, P., HEBSGAARD, M., SORENSEN, M. & WILLERSLEV, E. (2006). EVOLUTION: The Origin of Insects. *Science* **314,** 1883-1884.

GOWRI-SHANKAR, V. & JOW, H. (2006). PHASE: a software package for Phylogenetics And Sequence Evolution. In *manual*.

GOWRI-SHANKAR, V. & RATTRAY, M. (2006). On the correlation between composition and site-specific evolutionary rate: implications for phylogenetic inference. *Molecular Biology and Evolution* **23,** 352-64.

GOWRI-SHANKAR, V. & RATTRAY, M. (2007). A Reversible Jump Method for Bayesian Phylogenetic Inference with a Nonhomogeneous Substitution Model. *Molecular Biology and Evolution* **24,** 1286-1299.

GRYGIER, M. (1983). Ascothoracida and the unity of Maxillopoda. In *Crustacean Phylogeny*, vol. 1 Crustacean Issues (ed. F. R. Schram). A.A. Balkema, Rotterdam.

GRYGIER, M. (1987). New records, external and internal anatomy, and systematic position of Hansen's Y-larvae(Crustacea, Maxillopoda, Facetotecta). *Sarsia* **72,** 261-278.

GUTELL, R. R., LEE, J. C. & CANNONE, J. J. (2002). The accuracy of ribosomal RNA comparative structure models. *Current Opinion in Structural Biology* **12,** 301-10.

HAECKEL, E. H. P. A. (1866). *Generelle Morphologie der Organismen: allgemeine GrundzÜge derr organischen Formen-Wissenschaft, mechanisch begrÜndet durch die von Charles Darwin reformierte Descendenz-Theorie*. Georg Reimer, Berlin.

HAGNER-HOLLER, S., SCHOEN, A., ERKER, W., MARDEN, J., RUPPRECHT, R., DECKER, H. & BURMESTER, T. (2004). A respiratory hemocyanin from an insect. *Proceedings of the national Academy of Sciences, USA* **101,** 871-4.

HALL, T. A. (1999). Bioedit: a user-friendly biological sequence alignment editor and anaylses program for Windos 95/98/NT. *Nucleic Acids Symposium Series* **41,** 95-98.

HANCOCK, J. M., TAUTZ, D. & DOVER, G. A. (1988). Evolution of the secondary structures and compensatory mutations of the ribosomal RNAs of Drosophila melanogaster. *Molecular Biology and Evolution* **5,** 393-414.

HARTMANN, S. & VISION, T. J. (2008). Using ESTs for phylogenomics: Can one accurately infer a phylogenetic tree from a gappy alignment? In *BMC Evolutionary Biology*, vol. 8, pp. 95.

HARZSCH, S. (2004). Phylogenetic comparison of serotonin-immunoreactive neurons in

representatives of the Chilopoda, Diplopoda, and Chelicerata: Implications for arthropod relationships. *Journal of Morphology* **259,** 198-213.

HARZSCH, S. (2006). Neurophylogeny: Architecture of the nervous system and a fresh view on arthropod phyologeny. *Integrative and Comparative Biology* **46,** 162-194.

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. *Development, Genes and Evolution* **215,** 53-68.

HARZSCH, S., VILPOUX, K., BLACKBURN, D., PLATCHETZKI, D., BROWN, N., MELZER, R., KEMPLER, K. & BATTELLE, B. (2006). Evolution of arthropod visual systems: Development of the eyes and central visual pathways in the horseshoe crabLimulus polyphemus Linnaeus, 1758 (Chelicerata, Xiphosura). *Developmental Dynamics* **235,** 2641-2655.

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. *Molecular Phylogenetics and Evolution* **38,** 100-16.

HASSANIN, A., LEGER, N. & DEUTSCH, J. (2005). Evidence for Multiple Reversals of Asymmetric Mutational Constraints during the Evolution of the Mitochondrial Genome of Metazoa, and Consequences for Phylogenetic Inferences. *Systematic Biology* **54,** 277-298.

HAUG, J. T., (2009). Arthropod ontogeny – fossil record of development and 4-dimensional data (morphologtical changes during ontogeny) in phylogenetic systematics. Dissertation, Biosystematische Dokumentation, Universität Ulm.

HEATH, T. A., HEDTKE, S. M. & HILLIS, D. M. (2008). Taxon sampling and the accuracy of phylogenetic analyses. *Journal of Systematics and Evolution* **46,** 239-257.

HENDY, M. & PENNY, D. (1989). A framework for the quantitative study of evolutionary trees. *Systematic Zoology* **38,** 297-309.

HENNIG, W. (1969). *Die Stammesgeschichte der Insekten*, Frankfurt a. M.

HESSLER, R. (1982). The structural morphology of walking mechanisms in eumalacostracan crustaceans. *Philosophical Transactions of the Royal Society B* **296,** 245-298.

HESSLER, R. (1992). Reflections on the phylogenetic position of the Cephalocarida. *Acta Zoologica (Stockholm)*.

HESSLER, R. & NEWMAN, W. (1975). A trilobitomorph origin for the Crustacea. *Fossils and Strata* **4**.

HEYMONDS, R. (1901). Die Entwicklungsgeschichte derr Scolopender. *Zoological Studies* **33,** 1-244.

HICKSON, R. E., SIMON, C. & PERREY, S. W. (2000). The performance of several multiple-sequence alignment programs in relation to secondary-structure features for an rRNA sequence. *Molecular Biology and Evolution* **17,** 530-9.

HIGGINS, R. P. & THIEL, H. (1988). *Introducton to the study of Meiofauna*. Smithonian Institution Press, Washington, D.C.

HOLLAND, B. & MOULTON, V. (2003). *Consensus networks: A Method for Visualising Incompatibilities in Collections of Trees*. Springer Berlin / Heidelberg.

HOLLAND, B. R., HUBER, K. T., MOULTON, V. & LOCKHART, P. J. (2004). Using Consensus Networks to Visualize Contradictory Evidence for Species Phylogeny. *Molecular Biology and Evolution* **21,** 1459-1461.

HOLTHUIS, L. B. (1973). *Caridean Shrimps Found in Land-locked Saltwater Pools at Four Indo-West Pacific localities (Sinai Peninsula, Funafuti Atoll, Maui and Hawaii Islands),*

with the description of one new genus and four new species. Brill.

HOMBERG, U. (2008). Evolution of the central complex in the arthropod brain with respect to the visual system. *Arthropod Structure & Development* **37,** 347-62.

HORNE, D. J., SCHÖN, I., SMITH, R. J. & MARTENS, C. (2005). What are Ostracoda? A cladistic analysis of the extant superfamilies of the subclasses Myodocopa and Podocopa (Crustacea: Ostracoda). In *Crustacea and Arthropod Relationships* (ed. S. Koenemann and R. A. Jenner), pp. 249-273. Taylor and Francis CRC PRess, Boca Raton.

HØEG, J. T., ACHITUV, Y., CHAN, B. K., CHAN, K., JENSEN, P. G. & PÉREZ-LOSADA, M. (2009). Cypris morphology in the barnacles Ibla and Paralepas (Crustacea: Cirripedia Thoracica) implications for cirripede evolution. *Journal of Morphology* **270,** 241-55.

HØEG, J. T. & KOLBASOV, G. A. (2002). Lattice organs in y-cyprids of the Facetotecta and their significance in the phylogeny of the Crustacea Thecostraca. *Acta Zoologica* **83,** 67-79.

HØEG, J. T., LARSON, N. C. & GLENNER, H. (2004). The complete cypris larva and its significance in thecostracan phylogeny. In *Evolutionary Developmental Biology of Crustacea*, vol. 15 Crustacean Issues (ed. G. Scholtz), pp. 197-215.

HUDSON, M. E. (2008). Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources* **8,** 3-17.

HUELSENBECK, J. P. & RONQUIST, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17,** 754-5.

HUGHES, J., LONGHORN, S. J., PAPADOPOULOU, A., THEODORIDES, K., DE RIVA, A., MEJIA-CHANG, M., FOSTER, P. & VOGLER, A. (2006). Dense taxonomic EST sampling and its applications for molecular systematics of the Coleoptera (beetles). *Molecular Biology and Evolution* **23,** 268-78.

HUGHES, N., HAUG, J. & WALOSZEK, D. (2008). Basal euarthropod development: a fossil-based perspective. In *Evolutionary Pathways: Key Themes in Evolutionary Developmental Biology* (ed. A. Minelli and G. Fusco). Cambridge University Press.

HUSON, D. H. (1998). SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* **14,** 68-73.

HUSON, D. H. & BRYANT, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* **23,** 254-267.

HUSON, D. H., KLOEPPER, T., LOCKHART, P. J. & STEEL, M. (2005). Reconstruction of reticulate networks from gene trees. In *Ninth International Conference on Research in Computational Molecular Biology*. *Proceedings of the Ninth International Conference on Research in Computational Molecular Biology* (ed. S. Miyano).

HWANG, U. W., FRIEDRICH, M., TAUTZ, D., PARK, C. J. & KIM, W. (2001). Mitochondrial protein phylogeny joins myriapods with chelicerates. *Nature* **413,** 154-7.

ILIFFE, T. M. (1992a). Anchialine Cave Biology. *The Natural History of Biospeleology***,** 24.

ILIFFE, T. M. (1992b). An annotated list of the troglobitic anchialine and freshwater fauna of Quintana Roo. *Diversidad Biologica en la Reserva de la Biosfera de Sian ka#an Quintana Roo, Mexico* **II,** 20.

ILIFFE, T. M. & SARBU, S. (1990). Anchialine Caves and Cave fauna of the South Pacific. *NSS***,** 88-96.

JEFFROY, O., BRINKMANN, H., DELSUC, F. & PHILIPPE, H. (2006). Phylogenomics: the beginning of incongruence? *Trends in Genetics* **22,** 225-231.

JENNER, R. A. & LITTLEWOOD, D. T. J. (2008). Problematica old and new. *Philosophical Transactions of the Royal Society B* **363,** 1503-1512.

JENNER, R. A., Ní DHUBHGHAILL, C., FERLA, M. P. & WILLS, M. A. (2009). Eumalacostracan phylogeny and total evidence: limitations of the usual suspects. *BMC Evolutionary Biology* **9,** 21.

JENSEN, P. G., MOYSE, J., HØEG, J. & AL-YAHYA, H. (1994). Comparative SEM studies of lattice organs: Putative sensory structures on the carapace of larvae from Ascothoracida and Cirripedia (Crustacea Maxillopoda Thecostraca). *Acta Zoologica* **75,** 125-142.

JERMIIN, L. S., HO, S. Y., ABABNEH, F., ROBINSON, J. & LARKUM, A. W. D. (2004). The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. *Systematic Biology* **53,** 638-643.

JONES, D. T., TAYLOR, W. R. & THORNTON, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosci*ences **8,** 275-82.

JONGENEEL, C. V. (2000). Searching the expressed sequence tag (EST) databases: panning for genes. *Briefings in Bioinformatics* **1,** 76-92.

JORDAL, B., GILLESPIE, J. J. & COGNATO, A. I. (2008). Secondary structure alignment and direct optimization of 28S rDNA sequences provide limited phylogenetic resolution in bark and ambrosia beetles (Curculionidae: Scolytinae). *Zoologica Scripta* **37,** 43-56.

JOW, H., HUDELOT, C., RATTRAY, M. & HIGGS, P. G. (2002). Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution. *Molecular Biology and Evolution* **19,** 1591-601.

JURKA, J., KAPITONOV, V. V., PAVLICEK, A., KLONOWSKI, P., KOHANY, O. & WALICHIEWICZ, J. (2005). Repbase update, a database of eukaryotic repetitive elements. *Cytogenet Genome Research* **110**, 462-467.

KASS, R. E. & RAFTERY, A. E. (1993). Bayes Factors and model unceratinty. *Technical Report* **254,** 73.

KASS, R. E. & RAFTERY, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association* **90,** 773-795.

KATOH, K., MISAWA, K., KUMA, K. & MIYATA, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30,** 3059-66.

KATOH, K. & TOH, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics* **9,** 286-298.

KELCHNER, S. & THOMAS, M. (2007). Model use in phylogenetics: nine key questions. *Trends in Ecology & Evolution* **22,** 87-94.

KILPERT, F. & PODSIADLOWSKI, L. (2006). The complete mitochondrial genome of the common sea slater, Ligia oceanica (Crustacea, Isopoda) bears a novel gene order and unusual control region features. *BMC Genomics* **7,** 241.

KIM, S., KANG, J., CHUNG, Y. J., LI, J. & RYU, K. H. (2008). Clustering orthologous proteins across phylogenetically distant species. *Proteins* **71,** 1113-22.

KJER, K. (2004). Aligned 18S and Insect Phylogeny. *Systematic Biology* **53,** 506-514.

KJER, K., GILLESPIE, J. J. & OBER, K. A. (2007). Opinions on Multiple Sequence Alignment, and an Empirical Comparison of Repeatability and Accuracy between POY and structural alignment. *Systematic Biology* **56,** 133-146.

KJER, K. M. (1995). Use of rRNA secondary structure in phylogenetic studies to identify homologous positions: an example of alignment and data presentation from the frogs. *Molecular Phylogenetics and Evolution* **4,** 314-30.

KJER, K. M., CARLE, F. L., LITMAN, J. & WARE, J. (2006). A Molecular Phylogeny of

Hexapoda. *Arthropod Systematics & Phylogeny* **64,** 35-44.

KLUGE, A. G. (1989). A concern for evidence and a phylogenetic hypotheis of relationships among epicrates (Boidae, serpentes). *Systematic Zoology* **38,** 7-25.

KOENEMANN, S., ILIFFE, T. M. & VAN DER HAM, J. (2003). Three new sympatric species of Remipedia (Crustacea) from Great Exuma Island, Bahamas Islands. *Contributions to Zoology* **72,** 227-252.

KOENEMANN, S., SCHRAM, F. R., ILIFFE, T. & HINDERSTEIN, L. M. (2007). Behavior of Remipedia in the laboratory, with suppoting field observations. *Journal of Crustacean Biology* **27,** 534-542.

KOONIN, E. (2005). Orthologs, paralogs and evolutionary genomics. *Annual Review of Genetics* **39,** 309-338.

KRISTENSEN, N. P. (1998). The groundplan and basal diversification of the hexapods. In *Arthropod Relationships* (ed. R. A. Fortey and R. H. Thomas), pp. 281-295. Chapman and Hall, London.

KUKALOVÁ-PECK, J. (1998). Arthropod phylogeny and 'basal' morphological structures. In *Arthropod Relationships* (ed. R. A. Fortey and R. H. Thomas), pp. 249-269. Chapman and Hall, London.

KUMAR, S. & FILIPSKI, A. (2007). Multiple sequence alignment: In pursuit of homologous DNA positions. *Genome Research* **17,** 127-135.

KUSCHE, K. & BURMESTER, T. (2001). Diplopod hemocyanin sequence and the phylogenetic position of the Myriapoda. *Molecular Biology and Evolution* **18**, 1566-1573.

KUSCHE, K., RUHBERG, H. & BURMESTER, T. (2002). A hemocyanin from the Onychophora and the emergence of respiratory proteins. *Proceedings of the national Academy of Sciences, USA* **99,** 10545-8.

LANKESTER, E. R. (1904). the structure and classification of the Arthropoda. *Microscopical Society (London), Quaterly Journal* **47,** 523-582.

LARKIN, M. A., BLACKSHIELDS, G., BROWN, N. P., CHENNA, R., MCGETTIGAN, P. A., MCWILLIAM, H., VALENTIN, F., WALLACE, I. M., WILM, A., LOPEZ, R., THOMPSON, J. D., GIBSON, T. J. & HIGGINS, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* **23,** 2947-8.

LARTILLOT, N., BRINKMANN, H. & LEPAGE, T. (2007b). Phylobyes 2.3.c manual.

LARTILLOT, N., BRINKMANN, H. & PHILIPPE, H. (2007a). Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evolutionary Biology* **7,** 1-14.

LARTILLOT, N. & PHILIPPE, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution* **21,** 1095-1109.

LARTILLOT, N. & PHILIPPE, H. (2008). Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. *Philosophical Transactions of the Royal Society B, Biological Science* **363,** 1463-1472.

LATREILLE, P. A. (1817). *Les crustaces, les arachnides et les insectes.*, Paris.

LAUTERBACH, K. E. (1983). Zum Problem derr Monophylie der Crustacea. *Verh. naturwiss. Ver. Hamburg* **26,** 293-320.

LAUTERBACH, K. E. (1986). Zum Grundplan der Crustacea. *Verh. naturwiss. Ver., Hamburg (N.F.)* **28,** 27-63.

LAVROV, D., BROWN, W. & BOORE, J. (2004). Phylogenetic position of the Pentastomida and (pan)crustacean relationships. *Proceedings of the Royal Society B: Biological Sciences* **271,** 537-544.

LE, S. Q., LARTILLOT, N. & GASCUEL, O. (2008). Phylogenetic mixture models for proteins. *Philosophical Transactions of the Royal Society B, Biological Science* **363,** 3965-76.

LIM, J. T. & HWANG, U. W. (2006). The complete mitochondrial genome of Pollicipes mitella (Crustacea, Maxillopoda, Cirripedia): non-monophylies of maxillopoda and crustacea. *Molecules and Cells* **22,** 314-22.

LIÒ, P. & GOLDMAN, N. (1998). Models of molecular evolution and phylogeny. *Genome Research* **8,** 1233-44.

LOESEL, R. (2005). The arthropod brain: retracing six hundred million years of evolution. *Arthropod Structure & Development* **34,** 207-209.

LÖYTYNOJA, A. & GOLDMAN, N. (2005). An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the national Academy of Sciences, USA* **102,** 10557-62.

LUAN, Y., MALLATT, J., XIE, R., YANG, J. M. & YIN, W. (2005). The Phylogenetic Positions of Three Basal-Hexapod Groups (Protura, Diplura, and Collembola) Based on Ribosomal RNA Gene Sequences. *Molecular Biology and Evolution* **22,** 1579-1592.

MAAS, A. & WALOSZEK, D. (2001). Cambrian Derivatives of the Early Arthropod Stem Lineage, Pentastomids, Tardigrades and Lobopodians An 'Orsten'Perspective. *Zoologischer Anzeiger-A* **240,** 451-459.

MAAS, A. & WALOSZEK, D. (2005). Phosphatocopina – ostracode-like sister group of Eucrustacea. *Hydrobiologia* **538,** 139-152.

MADSEN, O., SCALLY, M., DOUADY, C. J., KAO, D. J., DEBRY, R. W., ADKINS, R., AMRINE, H. M., STANHOPE, M. J., DE JONG, W. W. & SPRINGER, M. S. (2001). Parallel adaptive radiations in two major clades of placental mammals. *Nature* **409,** 610-4.

MALIA, M. J., LIPSCOMB, D. L. & ALLARD, M. W. (2003). The misleading effects of composite taxa in supermatrices. *Molecular Phylogenetics and Evolution* **27**, 522-527.

MALLATT, J. & GIRIBET, G. (2006). Further use of nearly complete 28S and 18S rRNA genes to classify Ecdysozoa: 37 more arthropods and a kinorhynch. *Molecular Phylogenetics and Evolution* **40,** 772-794.

MALLATT, J. M., GAREY, J. R. & SHULTZ, J. (2004). Ecdysozoan phylogeny and Bayesian inference: first use of nearly complete 28S and 18S rRNA gene sequences to classify the arthropods and their kin. *Molecular Phylogenetics and Evolution* **31,** 178-91.

MANGUM, C. P. (1985). Oxygen transport in invertebrates. *American Journal of Physiology* **248**.

MANTON, S. M. (1973). Arthropod phylogeny - a modern synthesis. *Journal of the Zoological Society of London* **171,** 111-130.

MARKL, J. & DECKER, H. (1992). Molecular structure of the arthropod hemocyanins. *Advances in Comparative Environmental Physiology* **13,** 325-376.

MARLÉTAZ, F., GILLES, A., CAUBIT, X., PEREZ, Y., DOSSAT, C., SAMAIN, S., GYAPAY, G., WINCKER, P. & LE PARCO, Y. (2008). Chætognath transcriptome reveals ancestral and unique features among bilaterians. *Genome Biolology* **9,** R94.

MARTIN, J. & DAVIS, G. E. (2001). An update classification of the recent Crustacea. *Natural History Museum of Los Angeles County Science Series* **39,** 1-124.

MAXMEN, A., BROWNE, W., MARTINDALE, M. & GIRIBET, G. (2005). Neuroanatomy of sea spiders implies an appendicular origin of the protocerebral segment. *Nature* **437,** 1144-1148.

MELAND, K. & WILLASSEN, E. (2007). The disunity of "Mysidacea" (Crustacea). *Molecular Phylogenetics and Evolution* **44,** 1083-104.

MELDRUM, D. (2000). Automation for genomics, part two: sequencers, microarrays, and

future trends. *Genome Research* **10,** 1288-303.

MICHOT, B., BACHELLERIE, J. P. & RAYNAL, F. (1983). Structure of mouse rRNA precursors. Complete sequence and potential folding of the spacer regions between 18S and 28S rRNA. *Nucleic Acids Research* **11,** 3375-91.

MISOF, B. & MISOF, K. (2009). A Monte Carlo Approach Successfully Identifies Randomness in Multiple Sequence Alignments: A More Objective Means of Data Exclusion *Systematic Biology* **58,** 1: 21-34.

MISOF, B., NIEHUIS, O., BISCHOFF, I., RICKERT, A., ERPENBECK, D. & STANICZEK, A. (2006). A Hexapod nuclear SSU rRNA secondary-structure model and catalog of taxon-specific structural variation. *Journal of Experimental Zoology* **306B,** 70-88.

MISOF, B., NIEHUIS, O., BISCHOFF, I., RICKERT, A., ERPENBECK, D. & STANICZEK, A. (2007). Towards an 18S phylogeny of hexapods: Accounting for group-specific character covariance in optimized mixed nucleotide/doublet models. *Zoology* **110,** 409-429.

MISOF, B., MEYER, B., V. REUMONT, B. M., KÜCK, P., MEUSEMANN, K. Selecting informative subsets of sparse super matrices increases the chance to find correct trees. *In preparation.*

MORRISON, D. (2006). Multiple sequence alignments for the phylogenetic purposes. *Aust Systematic Botany* **19,** 479-539.

MOURA, G., CHRISTOFFERSEN, M., L. (1996). The system of the mandibulate arthropods: Tracheata and Remipedia as sister groups, "Crustacea" non-monoiphyletic. Journal of Comparative Biology **I** 3,4.

MØLLER, O. S., OLESEN, J., AVENANT-OLDEWAGE, A., THOMSEN, P. F. & GLENNER, H. (2008). First maxillae suction discs in Branchiura (Crustacea): development and evolution in light of the first molecular phylogeny of Branchiura, Pentastomida, and other "Maxillopoda". *Arthropod Structure & Development* **37,** 333-46.

MÜLLER, C. (2007). *Vergleichend-ultrastrukturelle Untersuchungen an Augen ausgewÄhlter HundertfÜßer (Mandibulata: Chilopoda) und zur Bedeutung von Augenmerkmalen fur die phylogentische Rekonstruktion der Euarthropoda.* Inaugural-Dissertation, Cuvillier Verlag GÖttingen.

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. *Molecular Biology and Evolution* **18,** 1204-19.

NISHIHARA, H., OKADA, N. & HASEGAWA, M. (2007). Rooting the eutherian tree: the power and pitfalls of phylogenomics. *Genome Biology* **8,** R199.

NOLLER, H. (2005). RNA Structure: Reading the Ribosome. *Science* **309,** 1508-1514.

NOTREDAME, C. (2002). Recent progress in multiple sequence alignment: a survey. *Pharmacogenomics* **3,** 131-144.

NOTREDAME, C., HIGGINS, D. G. & HERINGA, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology* **302,** 205-17.

NUIN, P. A., WANG, Z. & TILLIER, E. R. (2006). The accuracy of several multiple sequence alignment programs for proteins. *BMC Bioinformatics* **7,** 471.

NYLANDER, J. A. A., RONQUIST, F., HUELSENBECK, J. P. & NIEVES-ALDREY, J. L. (2004). Bayesian phylogenetic analysis of combined data. *Systematic Biology* **53,** 47-67.

O'BRIEN, K. P., REMM, M. & SONNHAMMER, E. L. (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Research* **33,** D476-80.

O'BRIEN, S. J. & STANYON, R. (1999). Phylogenomics. Ancestral primate viewed. *Nature*

402, 365-6.

OGDEN, T. & ROSENBERG, M. (2006). Multiple sequence alignment accuracy and phylogenetic inference. *Systematic Biology* **55,** 314-328.

OGDEN, T. H., WHITING, M. F. & WHEELER, W. C. (2005). Poor taxon sampling, poor character sampling, and non-repeatable analyses of a contrived dataset do not provide a more credible estimate of insect phylogeny: a reply to Kjer. *Cladistics* **21,** 295-302.

OLESEN, J. (2007). Monophyly and phylogeny of Branchiopoda, with focus on morphology and homologies of branchiopod phyllopodus limbs. *Journal of Crustacean Biology* **27,** 165-183.

OTT, M., ZOLA, J., STAMATAKIS, A. & ALURU, S. (2007). Large-scale maximum likelihood-based phylogenetic analysis on the IBM BlueGene/L. In *Proceedings of the 2007 ACM/IEEE conference on Supercomputing*. ACM, Reno, Nevada.

PABST, T. & SCHOLTZ, G. (2009). The development of phyllopodous limbs in Leptostraca and Branchiopoda. *Journal of Crustacean Biology* ***29***: 1, 1-12.

PAPS, J., BAGUñà, J. & RIUTORT, M. (2009). Lophotrochozoa internal phylogeny: new insights from an up-to-date analysis of nuclear ribosomal genes. *Proc Biol Sci* **276,** 1245-54.

PAULUS, H. (2007). Euarthropoda. In *Spezielle Zoologie*, vol. I (ed. W. Westheide and R. Rieger). Elsevier, Spektrum Akademischer Verlag, MÜnchen.

PAULUS, H. F. (1979). Eye structure and the monophyly of the Arthropoda. In *Arthropod Phylogeny* (ed. A. P. Gupta), pp. 299-383. Van Nostrand Rheinold Co., New York.

PENNY, D., LOCKHART, P. J., STEEL, M. A. & HENDY, M. D. (1994). The role of models in recontructing evolutionary trees. In *Models in phylogeny estimation*. *The Systematics Association Special Volume Series* (ed. S. R.W., D. D.J. and W. D.M.), pp. 211-230. Oxford University Press.

PÉREZ-LOSADA, M., HARP, M., HØEG, J. T., ACHITUV, Y., JONES, D., WATANABE, H. & CRANDALL, K. A. (2008). The tempo and mode of barnacle evolution. *Molecular Phylogenetics and Evolution* **46,** 328-46.

PÉREZ-LOSADA, M., HØEG, J. T. & CRANDALL, K. (2004). Unraveling the evolutionary radiation of the thoracican barnacles using molecular and morphological evidence: a comparison of several divergence time estimation approaches. *Systematic Biology* **53,** 244-64.

PERTEA, G., HUANG, X., LIANG, F., ANTONESCU, V., SULTANA, R., KARAMYCHEVA, S., LEE, Y., WHITE, J., CHEUNG, F., PARVIZI, B., TSAI, J. & QUACKENBUSH, J. (2003). TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* **19,** 651-2.

PHILIPPE, H., DELSUC, F., BRINKMANN, H. & LARTILLOT, N. (2005a). Phylogenomics. *Annual Review of Ecology, Evolution and Systematics* **36,** 541-562.

PHILIPPE, H., DERELLE, R., LOPEZ, P., PICK, K., BORCHIELLINI, C., BOURY-ESNAULT, N., VACELET, J., RENARD, E., HOULISTON, E., QUÉINNEC, E., DA SILVA, C., WINCKER, P., LE GUYADER, H., LEYS, S., JACKSON, D. J., SCHREIBER, F., ERPENBECK, D., MORGENSTERN, B., WÖRHEIDE, G. & MANUEL, M. (2009). Phylogenomics revives traditional views on deep animal relationships. *Current Biology* **19,** 706-12.

PHILIPPE, H., LOPEZ, P., BRINKMANN, H., BUDIN, K., GERMOT, A., LAURENT, J., MOREIRA, D., MULLER, M. & LE GUYADER, H. (2000). Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions. *Proceedings of the Royal Society B: Biological Sciences* **267**, 1213-1221.

PHILIPPE, H., SNELL, E. A., BAPTESTE, E., LOPEZ, P., HOLLAND, P. W. & CASANE, D.

(2004). Phylogenomics of eukaryotes: impact of missing data on large alignments. *Molecular Biology and Evolution* **21,** 1740-52.

PHILIPPE, H. & TELFORD, M. (2006). Large-scale sequencing and the new animal phylogeny. *Trends in Ecology & Evolution (Amst)* **21,** 614-20.

PHILIPPE, H., ZHOU, Y., BRINKMANN, H., RODRIGUE, N. & DELSUC, F. (2005b). Heterotachy and long-branch attraction in phylogenetics. *BMC Evolutionary Biology* **5,** 50.

PHILLIPS, M. J. & PENNY, D. (2003). The root of the mammalian tree inferred from whole mitochondrial genomes. *Molecular Phylogenetics and Evolution* **28,** 171-185.

PICK, C., HAGNER-HOLLER, S. & BURMESTER, T. (2008). Molecular characterization of hemocyanin and hexamerin from the firebrat Thermobia domestica (Zygentoma). *Insect Biochemistry and Molecular Biology* **38,** 977-83.

PICK, C., SCHNEUER, M. & BURMESTER, T. (2009). The occurrence of hemocyanin in Hexapoda. *FEBS Journal* **276,** 1930-41.

PISANI, D. (2004). Identifying and removing fast-evolving sites using compatibility analysis: An example from the arthropoda. *Systematic Biology* **53**, 978-989.

POCOCK, R. I. (1893). On the classification of the tracheate Arthropoda. *Zoologischer Anzeiger* **16,** 271-275.

POCOCK, R. I. (1893). On the classification of the tracheate Arthropoda- A correction. *Nature* **49,** 124.

PODSIADLOWSKI, L. & BARTOLOMAEUS, T. (2005). Organization of the Mitochondrial Genome of Mantis Shrimp Pseudosquilla ciliata (Crustacea: Stomatopoda). *Marine Biotechnology* **7,** 618-624.

PODSIADLOWSKI, L. & BARTOLOMAEUS, T. (2006). Major rearrangements characterize the mitochondrial genome of the isopod Idotea baltica (Crustacea: Peracarida). *Molecular Phylogenetics and Evolution* **40,** 893-9.

PODSIADLOWSKI, L., BRABAND, A. & MAYER, G. (2008). The complete mitochondrial genome of the onychophoran Epiperipatus biolleyi reveals a unique transfer RNA set and provides further support for the ecdysozoa hypothesis. *Molecular Biology and Evolution* **25,** 42-51.

PODSIADLOWSKI, L., KOHLHAGEN, H. & KOCH, M. (2007). The complete mitochondrial genome of Scutigerella causeyae (Myriapoda: Symphyla) and the phylogenetic position of Symphyla. *Molecular Phylogenetics and Evolution* **45,** 251-260.

PUTNEY, S. D., HERLIHY, W. C. & SCHIMMEL, P. (1983). A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. *Nature* **302,** 718-21.

QUANG, L. S., GASCUEL, O. & LARTILLOT, N. (2008). Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* **24,** 2317-2323.

RAMSKÖLD, L. & JUNYUAN, C. (1998). Cambrian Lobopodians: Morphology and Phylogeny. In *Arthropod fossils and phylogeny* (ed. G. D. Edgecomb), pp. 107-150. Columbia University Press, New York.

RANNALA, B. & YANG, Z. (2008). Phylogenetic Inference Using Whole Genomes. *Annual Review of Genomics and Human Genetics* **9,** 217-231.

REGIER, J., SHULTZ, J. & KAMBIC, R. (2005). Pancrustacean phylogeny: hexapods are terrestrial crustaceans and maxillopods are not monophyletic. *Proceedings of the Royal Society B: Biological Sciences* **272,** 395-401.

REGIER, J. C. & SHULTZ, J. W. (1997). Molecular phylogeny of the major arthropod groups indicates polyphyly of Crustaceans and a new hypothesis for the origin of hexapods. *Molecular Biology and Evolution* **14,** 902-913.

REGIER, J. C. & SHULTZ, J. W. (2001). Elongation factor-2: a useful gene for arthropod phylogenetics. *Molecular Phylogenetics and Evolution* **20,** 136-48.

REGIER, J. C., SHULTZ, J. W., GANLEY, A. R., HUSSEY, A., SHI, D., BALL, B., ZWICK, A., STAJICH, J. E., CUMMINGS, M. P., MARTIN, J. W. & CUNNINGHAM, C. W. (2008). Resolving arthropod phylogeny: exploring phylogenetic signal within 41 kb of protein-coding nuclear gene sequence. *Systematic Biology* **57,** 920-38.

RICHTER, S. (2002). The Tetraconata concept: hexapod-crustacean relationships and the phylogeny of Crustacea. *Organisms Diversity & Evolution* **2,** 217-237.

RICHTER, S., OLESEN, J. & WHEELER, W. C. (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). *Journal of Zooogical Systematic and Evolutionary Research* **39,** 113-116.

RICHTER, S. & WIRKNER, C. S. (2004). Kontroversen in der phylogenetische Systematik der Euarthropoda. In: Kontroversen der phylogenetischen Systematik der Metazoa, vol. 43 (ed. S. Richter and W. Sudhaus), pp. 73-102. Sitzungsberichte der Gesellschaft Naturforschender Freunde zu Berlin.

RODRíGUEZ-EZPELETA, N., BRINKMANN, H., ROURE, B., LARTILLOT, N., LANG, B. F. & PHILIPPE, H. (2007). Detecting and overcoming systematic errors in genome-scale phylogenies. *Systematic Biology* **56,** 389-399.

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. *Molecular Phylogenetics and Evolution* **45,** 942-51.

ROKAS, A., KRÜGER, D. & CARROLL, S. B. (2005). Animal evolution and the molecular signature of radiations compressed in time. *Science* **310,** 1933-8.

ROKAS, A., WILLIAMS, B. L., KING, N. & CARROLL, S. B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425,** 798-804.

RONAGHI, M. (2001). Pyrosequencing Sheds Light on DNA Sequencing. *Genome Research* **11,** 3-11.

RONAGHI, M., KARAMOHAMED, S., PETTERSSON, B., UHLÉN, M. & NYRÉN, P. (1996). Real-time DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry* **242,** 84-9.

RONAGHI, M., UHLÉN, M. & NYRÉN, P. (1998). A sequencing method based on real-time pyrophosphate. *Science* **281,** 363-365.

RONQUIST, F. & HUELSENBECK, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19,** 1572-1574.

ROTA-STABELLI, O. & TELFORD, M. (2008). A multi criterion approach for the selection of optimal outgroups in phylogeny: recovering some support for Mandibulata over Myriochelata using mitogenomics. *Molecular Phylogenetics and Evolution* **48,** 103-11.

RUDD, S. (2003). Expressed sequence tags: alternative or complement to whole genome sequences? *Trends in Plant Science* **8,** 321-9.

SANDERS, H. L. (1955). The Cephalocarida, a new subclass of Crustacea from Long Island sound. *Proceedings of the national Academy of Sciences, USA***41,** 61-6.

SANDERSON, M. J. (1998). Phylogenetic supertrees: assembling the trees of life. *Trends in Ecology & Evolution (Amst)* **13,** 105-109.

SANDERSON, M. J. & DRISKELL, A. C. (2003). The challenge of constructing large phylogenetic trees. *Trends in Plant Science* **8,** 374-9.

SANDERSON, M. J. & MCMAHON, M. M. (2007). Inferring angiosperm phylogeny from EST data with widespread gene duplication. *BMC Evolutionary Biology* **7,** 1-14.

SANGER, F., NICKLEN, S. & COULSON, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the national Academy of Sciences, USA***74,** 5463-7.

SCHMIDT, H. A., STRIMMER, K., VINGRON, M. & VON HAESELER, A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18,** 502-4.

SCHOLTZ, G. & EDGECOMBE, G. (2006). The evolution of arthropod heads: reconciling morphological, developmental and palaeontological evidence. *Development, Genes and Evolution* **216,** 395-415.

SCHOLTZ, G., MITTMANN, B. & GERBERDING, M. (1998). The pattern of Distal-less expression in the mouthparts of crustaceans, myriapods and insects: new evidence for a gnathobasic mandible and the common origin of Mandibulata. *International Journal of Developmental Biology* **42,** 801-10.

SCHÖNINGER, M. & VON HAESELER, A. (1994). A stochastic model for the evolution of autocorrelated DNA sequences. *Molecular Phylogenetics and Evolution* **3,** 240-247.

SCHRAM, F. R. (1983). Remipedia and crustacean phylogeny. *Crustacean Issues* **1,** 23-28.

SCHRAM, F. R. (1986). *Crustacea*. Oxford Univeristy Press, Oxford.

SCHRAM, F. R. & HOF, C. H. J. (1998). Fossils and interrelationships of major crustacean groups. In *Arthropod Fossils and Phylogeny* (ed. G. D. Edgecomb), pp. 233-302. Columbia University Press, New York.

SCHRAM, F. R. & KOENEMANN, S. (2001). Developmental genetics and arthropod evolution: part I, on legs. *Evolution & Development* **3,** 343-354.

SCHRAM, F. R. & KOENEMANN, S. (2004). Are the crustaceans monophyletic? In *Assembling the tree of life* (ed. J. Cracraft and M. J. Donoghue), pp. 319-329. Oxford University Press, New York.

SCHRAM, F. R. & KOENEMANN, S. (2004). Developmental genetics and arthropod evolution: on body regions of Crustacea. In *Evolutionary Developmental Biology of Crustacea, Crustacean Issues*, vol. 15 (ed. G. Scholotz), pp. 75-92.

SCHREIBER, F., WORHEIDE, G. & MORGENSTERN, B. (2009). OrthoSelect: a web server for selecting orthologous gene alignments from EST sequences. *Nucleic Acids Research* **434,** 1-4.

SCHUSTER, S. C. (2008). Next-generation sequencing transforms today's biology. *Nature Methods* **5,** 16-18.

SEO, T. & KISHINO, H. (2008). Synonymous substitutions substantially improve evolutionary inference from highly diverged proteins. *Systematic Biology* **57,** 367-377.

SHENDURE, J., MITRA, R. D., VARMA, C. & CHURCH, G. M. (2004). Advanced sequencing technologies: methods and goals. *Nature Reviews Genetics* **5,** 335-344.

SHOSHANI, J. & MCKENNA, M. C. (1998). Higher taxonomic relationships among extant mammals based on morphology, with selected comparisons of results from molecular data. *Molecular Phylogenetics and Evolution* **9,** 572-84.

SHULTZ, J. W. & REGIER, J. C. (2000). Phylogenetic analysis of arthropods using two nuclear protein-encoding genes supports a crustacean + hexapod clade. *Proceedings of the Royal Society B: Biological Sciences* **267,** 1011-9.

SIMON, C., BUCKLEY, T. R., FRATI, F., STEWART, J. B. & BECKENBACH, A. T. (2006). Incorporating Molecular Evolution into Phylogenetic Analysis, and a New Compilation of Conserved Polymerase Chain Reaction Primers for Animal Mitochondrial DNA. *Annual Review of Ecology, Evolution, and Systematics* **37,** 545-579.

SIVETER, D. J., WILLIAMS, M. & WALOSZEK, D. (2001). A phosphatocopid crustacean with appendages from the Lower Cambrian. *Science* **293,** 479-81.

SMIT, A. F. A. (1996). The origin of interspersed repeats in the human genome. *Current Opinion in Genetics and Development* **6,** 743-748.

SMITH, S. & DUNN, C. (2008). Phyutility: a phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics* **24,** 715-6.

SNODGRASS, R. E. (1935). *The principles of insect morphology*. McGraw-Hill, New York.

SNODGRASS, R. E. (1938). Evolution of the Annelida, Onychophora, and Arthropoda. *Smithsonian Miscellaneous Collection* **97,** 1-159.

SPEARS, T. & ABELE, L. G. (1998). Crustacean phylogeny inferred from 18S rDNA. In *Arthropod Relationships* (ed. R. A. Fortey and R. H. Thomas), pp. 169-187. Chapman and Hall, London.

SPEARS, T., ABELE, L. G. & KIM, W. (1992). The monophyly of brachyuran crabs: a phylogenetic study based on 18S rRNA. *Systematic Biology*.

SPEARS, T., DEBRY, R. W., ABELE, L. G. & CHODYLA, K. (2005). Peracarid monophyly and interordinal phylogeny inferred from nuclear small-subunit …. *Proceedings of the Biological Society of Washington*.

SPRINGER, M. S., SCALLY, M., MADSEN, O., DE JONG, W. W., DOUADY, C. J. & STANHOPE, M. J. (2004). The use of composite taxa in supermatrices. *Molecular Phylogenetics and Evolution* **30,** 883-884.

STAMATAKIS, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22,** 2688-90.

STEEL, M., HUSON, D. H. & LOCKHART, P. J. (2000). Invariable sites models and their use in phylogeny reconstruction. *Systematic Biology* **49,** 225-32.

STEEL, M. & PENNY, D. (2000). Parsimony, likelihood, and the role of models in molecular phylogenetics. *Molecular Biology and Evolution* **17,** 839-50.

STENDERUP, J., OLESEN, J. & GLENNER, H. (2006). Molecular phylogeny of the Branchiopoda (Crustacea)–Multiple approaches suggest a 'diplostracan' ancestry of the Notostraca. *Molecular Phylogenetics and Evolution* **41,** 182-194.

STEPHAN, W. (1996). The Rate of Compensatory Evolution. *Genetics* **144,** 419-426.

STILLMAN, J. H., COLBOURNE, J. K., LEE, C. E. & PATEL, N. H. (2008). Recent advances in crustacean genomics. In *Annual meeting of the Society for Integrative and Comparative Biology*, pp. 1-17, San Antonio, Texas.

STOCSITS, R., LETSCH, H., HERTEL, J., MISOF, B. & STADLER, P. (2009). Accurate and efficient reconstruction of deep phylogenies from structured RNAs. *Nucleic Acids Research***,** 1-10.

STOLLEWERK, A. & CHIPMAN, A. D. (2006). Neurogenesis in myriapods and chelicerates and its importance for understanding arthropod relationships. *Integrative Computational Biology* **46,** 195-206.

STRAUSFELD, N., SINAKEVITCH, I., BROWN, S. M. & FARRIS, S. M. (2009). Ground plan of the insect mushroom body: functional and evolutionary implications. *Journal of Comparative Neurology.* **513,** 265-91.

STRAUSFELD, N., STRAUSFELD, C. M., LOESEL, R., ROWELL, D. & STOWE, S. (2006a). Arthropod phylogeny: onychophoran brain organization suggests an archaic relationship with a chelicerate stem lineage. *Proceedings of the Royal Society B: Biological Sciences* **273,** 1857-66.

STRAUSFELD, N., STRAUSFELD, C. M., STOWE, S., ROWELL, D. & LOESEL, R. (2006b). The organization and evolutionary implications of neuropils and their neurons in the brain

of the onychophoran Euperipatoides rowelli. *Arthropod Structure & Development* **35,** 169-96.

STRAUSFELD, N. J. (2009). Brain organization and the origin of insects: an assessment. *Proceedings of the Royal Society B: Biological Sciences* **276,** 1929-37.

SULLIVAN, J. & SWOFFORD, D. L. (2001). Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? *Systematic Biology* **50,** 723-9.

SUSKO, E., SPENCER, M. & ROGER, A. (2005). Biases in phylogenetic estimation can be caused by random sequence segments. *Journal of Molecular Evolution* **61,** 351-359.

SWOFFORD, D. L. (2002). *PAUP*4.0b10. Phylogenetic analysis using parsimony (*and other methods).* Sinauer Associates, Sunderland, Massachusetts.

TARRIO, R., RODRIGUEZ-TRELLES, F. & AYALA, F. J. (2000). Tree rooting with outgroups when they differ in their nucleotide composition from the ingroup: the Drosophila saltans and willistoni groups, a case study. *Molecular Phylogenetics and Evolution* **16,** 344-9.

TARRIO, R., RODRIGUEZ-TRELLES, F. & AYALA, F. J. (2001). Shared nucleotide composition biases among species and their impact on phylogenetic reconstructions of the drosophilidae. *Molecular Biology and Evolution* **18,** 1464-1473.

TATENO, Y., NEI, M. & TAJIMA, F. (1982). Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. *Journal of Molecular Evolution* **18,** 387-404.

TELFORD, M. (2005). Consideration of RNA secondary structure significantly improves likelihood-based estimates of phylogeny: examples from the Bilateria. *Molecular Biology and Evolution* **22,** 1129-1136.

TELFORD, M. (2006). Animal phylogeny. *Current Biology* **16,** R981-R985.

TELFORD, M., BOURLAT, S., ECONOMOU, A., PAPILLON, D. & ROTA-STABELLI, O. (2008). The evolution of the Ecdysozoa. *Philosophical Transactions of the Royal Society B: Biological Sciences* **363,** 1529-1537.

THOMPSON, J. D., GIBSON, T. J., PLEWNIAK, F., JEANMOUGIN, F. & HIGGINS, D. G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research* **25,** 4876-82.

THOMPSON, J. D., HIGGINS, D. G. & GIBSON, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22,** 4673-80.

THORLEY, J. L. & WILKINSON, M. (1999). Testing the Phylogenetic Stability of Early Tetrapods. *Journal of Theoretical Biology* **200,** 343-344.

THORNE, J. (2000). Models of protein sequence evolution and their applications. *Current Opinion in Genetics and Development* **10,** 602-605.

TIMMERMANS, M., ROELOFS, D., MARIEN, J. & VAN STRAALEN, N. M. (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 Evolutionary Biology* **8,** 1-10.

UNGERER, P. & SCHOLTZ, G. (2008). Filling the gap between identified neuroblasts and neurons in crustaceans adds new support for Tetraconata. *Proceedings of the Royal Society B: Biological Sciences* **275,** 369-376.

VACCARI, N. E., EDGECOMBE, G. D. & ESCUDERO, C. (2004). Cambrian origins and affinities

of an enigmatic fossil group of arthropods. *Nature* **430,** 554-7.

VAN DE PEER, Y., DE RIJK, P., WUYTS, J., WINKELMANS, T. & DE WACHTER, R. (2000). The European small subunit ribosomal RNA database. *Nucleic Acids Research* **28,** 175-176.

VAN HOLDE, K. E. & MILLER, K. I. (1995). Hemocyanins. *Advances in Protein Chemistry* **47,** 1-81.

VANNIER, J. & ABE, K. (1995). Size, body plan and respiration in the Ostracoda. *Palaeontology* **38,** 843-873.

V. REUMONT, B.M., MEUSEMANN K., SZUCSICH, N. U., DELL'AMPIO, E., GOWRI_SHANKAR, V., BARTEL, D., SIMON, S., LETSCH, H. O., STOCSITS, R., LUAN, Y. X., WÄGELE J. W., PASS G., HADRYS, H., MISOF, B. (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.*

WADDELL, P. J., PENNY, D. & MOORE, T. (1997). Hadamard Conjugations and Modeling Sequence Evolution with Unequal Rates across Sites. *Molecular Phylogenetics and Evolution* **8,** 33-50.

WAEGELE, J. W., LETSCH, H., KLUSSMANN-KOLB, A., MAYER, C., MISOF, B. & WAGELE, H. (2009). Phylogenetic support values are not necessarily informative: the case of the Serialia hypothesis (a mollusk phylogeny). *Front Zool* **6,** 12.

WÄGELE, J., HOLLAND, B., DREYER, H. & HACKETHAL, B. (2003). Searching factors causing implausible non-monophyly: ssu rDNA phylogeny of Isopoda Asellota (Crustacea: Peracarida) and faster evolution in marine than in freshwater habitats. *Molecular Phylogenetics and Evolution* **28,** 536-51.

WÄGELE, J. W. (1993). Rejection of the 'Uniramia' hypothesis and implications of the Mandibulata concept. *Zoologische JahrbÜcher. Abteilung fÜr Systematik, Ökologie und Geographie der Tiere* **120,** 253-288.

WÄGELE, J. W. & MAYER, C. (2007). Visualizing differences in phylogenetic information content of alignments and distinction of three classes of long-branch effects. *BMC Evolutionary Biology* **7,** 1-24.

WALLACE, I. M., O'SULLIVAN, O. & HIGGINS, D. G. (2005). Evaluation of iterative alignment algorithms for multiple alignment. *Bioinformatics* **21,** 1408-14.

WALLACE, I. M., O'SULLIVAN, O., HIGGINS, D. G. & NOTREDAME, C. (2006). M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Research* **34,** 1692-9.

WALOSSEK, D. (1993). The upper Cambrian Rehbachiella and the phylogeny of Branchiopoda and Crustacea. *Fossils and Strata* **32,** 1-202.

WALOSSEK, D. (1999). On the Cambrian diversity of Crustacea. In *Crustaceans and the biodiversity crisis* (ed. F. Schram and J. Vaupel Klein). Brill Acadamic Publishers, Leiden, Amsterdam.

WALOSSEK, D. & MÜLLER, K. (1998a). Cambrian'Orsten'-type arthropods and the phylogeny of Crustacea. In *Arthropod relationships* (ed. R. A. Fortey and R. H. Thomas), pp. 139-153. Chapman and Hall, London.

WALOSSEK, D. & MÜLLER, K. J. (1998b). Early arthropod phylogeny in light of the Cambrian "Orsten" fossils. In *Arthropod fossils and phylogeny* (ed. G. D. Edgecomb), pp. 185-231. Columbia University Press, New York.

WALOSZEK , D. (2003). Cambrian 'Orsten'-type preserved arthropods and the phylogeny of Crustacea. In *The new panorama of animal evolution., Proc. 18th Congr. Zoology* (ed. A. Legakis, R. Sfenthourakis, R. Polymeni and M. Thessaqlou-Legaki), pp. 69-87. Pensoft Publishers.

WALOSZEK, D., CHEN, J., MAAS, A. & WANG, X. (2005). Early Cambrian arthropods—new insights into arthropod head and structural evolution. *Arthropod Structure & Development* **34,** 189-205.

WALOSZEK, D., MAAS, A., CHEN, J. & STEIN, M. (2007). Evolution of cephalic feeding structures and the phylogeny of Arthropoda. *Palaeogeography, Paleoclimatology, Paleoecology* **254,** 273-287.

WALOSZEK, D., REPETSKI, J. E. & MAAS, A. (2006). A new Late Cambrian pentastomid and a review of the relationships of this parasitic group. *Transactions of the Royal Society Edingurgh: Earth Science* **96,** 163-176.

WEYGOLDT, P. (1998). REVIEW Evolution and systematics of the Chelicerata. *Experimental and Applied Acarology* **22,** 63-79.

WHEELER, W. & HAYASHI, C. (1998). The phylogeny of the extant chelicerate orders. *Cladistics* **14,** 173-192.

WHEELER, W. C. (1998). Sampling, groundplans, total evidence and the systematics of arthropods. In *Arthropod relationships* (ed. R. A. Fortey and R. H. Thomas), pp. 87-97. Chapman and Hall, London.

WHEELER, W. C., GIRIBET, G. & EDGECOMBE, G. D. (2004). Arthropod systematics. The comparative study of genomic, anatomical, and paleontological information. In *Assembling the tree of life* (ed. J. Cracraft and M. J. Donoghue), pp. 281-295. Oxford University Press, New York.

WHELAN, S. (2008). The genetic code can cause systematic bias in simple phylogenetic models. *Philosophical Transactions of the Royal Society B: Biological Sciences* **363,** 4003-4011.

WHELAN, S. & GOLDMAN, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution* **18,** 691-699.

WHELAN, S., LIÒ, P. & GOLDMAN, N. (2001b). Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends in Genetics* **17,** 262-72.

WHITFIELD, J. B. & KJER, K. M. (2008). Ancient rapid radiations of insects: Challenges for phylogenetic analysis. *Annual Review of Entomology* **53,** 449-472.

WIENS, J. J. (1998). Does adding characters with missing data increase or decrease phylogenetic accuracy? *Systematic Biology* **47,** 625-40.

WIENS, J. J. (2003). Missing Data, Incomplete Taxa, and Phylogenetic Accuracy. *Systematic Biology* **52,** 528-538.

WIENS, J. J. (2005). Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? *Systematic Biology* **54,** 731-42.

WIENS, J. J. (2006). Missing data and the design of phylogenetic analyses. *Journal of Biomedical Informatics* **39,** 34-42.

WIENS, J. J., KUCZYNSKI, C. A., SMITH, S. A., MULCAHY, D. G., SITES, J. W., TOWNSEND, T. M. & REEDER, T. W. (2008). Branch lengths, support, and congruence: testing the phylogenomic approach with 20 nuclear loci in snakes. *Systematic Biology* **57,** 420-31.

WIENS, J. J. & MOEN, D. S. (2008). Missing data and the accuracy of Bayesian phylogenetics. *Journal of Systematics and Evolution* **46,** 307-314.

WILLS, M. A. (1998). A phylogeny of recent and fossil crustacea derived from morphological characters. In *Arthropod Relationships*, vol. The systematics Association Special Volume Series (ed. R. A. Fortey and R. H. Thomas). Chapman and Hall, London.

WILSON, G. D. F. (1992). Computerized analysis of crustacean relationships. *Acta Zoologica* **72,** 383-389.

WILSON, K., CAHILL, V., BALLMENT, E. & BENZIE, J. (2000). The complete sequence of the mitochondrial genome of the crustacean Penaeus monodon: Are malacostracan crustaceans more closely related to insects than to branchiopods? *Molecular Biology and Evolution* **17,** 863-874.

WINGSTRAND, K. G. (1972). Comparative spermatology of a pentastomid Raillietiella hemidactyli and a branchiuran crustacean Argulus foliacaetus with a discussion of pentstomid relationships. *Biologiske Skrifter* **19,** 1-72.

WONG, K. M., SUCHARD, M. A. & HUELSENBECK, J. P. (2008). Alignment uncertainty and genomic analysis. *Science* **319**, 473-476.

WUYTS, J., DE RIJK, P., VAN DE PEER, Y. & PISON, G. (2000). Comparative analysis of more than 3000 sequences reveals the existence of two pseudoknots in area V4 of eukaryotic small subunit ribosomal RNA. *Nucleic Acids Research* **28,** 4698-4708.

WUYTS, J., PERRIèRE, G. & VAN DE PEER, Y. (2004). The European ribosomal RNA database. *Nucleic Acids Research* **32,** D101-3.

WUYTS, J., VAN DE PEER, Y., WINKELMANS, T. & DE WACHTER, R. (2002). The European database on small subunit ribosomal RNA. *Nucleic Acids Research* **30,** 183-5.

YAGER, J. (1981). Remipedia, a new class of Crustacea from a marine cave in the Bahamas. *Journal of Crustacean Biology* **1,** 328-333.

YAMAGUCHI, S. & ENDO, K. (2003). Molecular phylogeny of Ostracoda (Crustacea) inferred from 18S ribosomal DNA sequences: implication for its origin and diversification. *Marine Biology* **143,** 23-38.

YANG, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution* **39,** 306-14.

YANG, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology & Evolution* **11,** 367-372.

ZANTKE, J., WOLFF, C. & SCHOLTZ, G. (2008). Three-dimensional reconstruction of the central nervous system of Macrobiotus hufelandi (Eutardigrada, Parachela): implications for the phylogenetic position of Tardigrada. *Zoomorphology* **127,** 21-36.

ZHANG, X., SIVETER, D., WALOSZEK, D. & MAAS, A. (2007). An epipodite-bearing crown-group crustacean from the Lower Cambrian. *Nature* **449,** 595-598.

ZHOU, Y. & LANDWEBER, L. (2007). BLASTO: a tool for searching orthologous groups. *Nucleic Acids Research* **35,** W678-W682.

ZRZAVÝ, J., HYPSA, V. & VLÁSKOVÁ. (1998a). Arthropod phylogeny: taxonomic congruence, total evidence and conditional combination approaches to morphological and molecular data sets. In *Arthropod relationships* (ed. R. A. Fortey and R. H. Thomas), pp. 97-107. Columbia University Press, New York.

ZRZAVÝ, J., MIHULKA, S., KEPKA, P. & TIETZ, D. (1998b). Phylogeny of the Metazoa Based on Morphological and 18S Ribosomal DNA Evidence. *Cladistics* **14,** 249-285.

ZRZAVÝ, J. & STYS, P. (1997). The basic body plan of arthropods: insights from evolutionary morphology and developmental biology. *Journal of Evolutionary Biology* **10,** 353-367.

ZWICKL, D. J. & HILLIS, D. M. (2002). Increased taxon sampling greatly reduces phylogenetic error. *Systematic Biology* **51,** 588-98.

ZWICKL, D. J. & HOLDER, M. T. (2004). Model parameterization, prior distributions, and the general time-reversible model in bayesian phylogenetics. *Systematic Biology* **53,** 877-888.

ELECTRONICAL REFERENCES (LINKS)

SOFTWARE - PROGRAMS

ALICUT	http://utilities.zfmk.de
ALISCORE	http://aliscore.zfmk.de
CROSSMATCH	http://www.incogen.com/public_documents/vibe/details/crossmatch.html
INPARANOID	http://inparanoid6.sbc.su.se
RNASALSA	http://rnasalsa.zfmk.de
SEQCLEAN	http://www.tigr.org/tdb/tgi/software/
TRACER	http://tree.bio.ed.ac.uk/software/tracer/
UNIVEC	http://www.ncbi.nlm.nih.gov/VecScreen/ UniVec.html

HOMEPAGES

Arthropod "primer toolbox"	http://home.arcor.de/bmvr/PTB/
Deep Arthropod Phylogeny Project (CUNNINGHAM lab)	http://arthropods.nhm.org/
NCBI dbEST database	http://www.ncbi.nlm.nih.gov/dbEST/
NCBI nucleotide BLAST	http://blast.ncbi.nlm.nih.gov/Blast.cgi
SPP: "Deep Metazoan Phylogeny"	http://www.deep-phylogeny.org/

ABBREVIATIONS

AIC	Akaike Information Criterion
AIC	Akaike information criterion
ASRV	Among site rate variation
BFT	Bayes Factor Test
cDNA	Copy desoxyribonucleine acid
COI	Cytochrome oxidase subunit I
DNA	Desoxyribonucleine acid
DNA-POL I	DNA Polymerase I
e.g.	For example
EST	Expressed sequence tag
GHz	Gigahertz
GB	Gigabyte
HDD	Hard disc drive
HGP	Human Genome Project
HP	Hewlet Packard (corporation)
IANTD	International Association for Nitrox and Trimix Diving
kDa	Kilo Dalton
LBA	Long Branch Artifact
LSU rRNA	Large subunit rRNA
MB	Megabyte
Mb	Megabase
MCMC	Markov Chains Monte Carlo
ML	Maximum Likelihood
mRNA	Messenger ribonucleine acid
MSA	Multiple sequence alignment
NACD	National Association for Cave Diving
NCBI	National Center for Biotechnology Information
NSS	National Speleological Society
PCR	Polymerase chain reaction
pP	Posterior probability
RAM	Random Access Memory
rRNA	Ribosomal ribonuceline acid
rRNA	Ribosomal ribonucleine acid
RT-PCR	Reverse transcriptase PCR
SG	Sistergroup
SSU rRNA	Small subunit rRNA
TCC	Trilobita-Chelicerata-Crustacea concept
TiHo	Medizinische Tier-Hochschule Hannover
ZFMK	Zoologisches Forschungsmuseum Alexander Koenig

INDEX OF FIGURES AND TABLES

FIGURES

TABLES

ACKNOWLEDGEMENT

If my doctor told me I had only six minutes to live, I wouldn't brood. I'd type a little faster (ISAAC ASIMOV).

I have the need to thank so many persons that I hope nobody is forgotten in this section. I would like to thank first all those persons who provided and determined specimens in the framework of this thesis: E. EDER, A. BRABAND, D. WALOßEK, G. STRIESO, H. BOSCH, P. FRENZEL, M. RAUPACH, J. OTT, F. KRAPP. The present taxon sample was only possible with their help. Of course many people supported me in the field to collect successfully and with fun (which is at least equally important) specimens, T. ILIFFE, J. MOREIRA-MORALEZ, K. MEUSEMANN, K. HÄNDELER are to mention here. You will meet the Karen later again. Very special appearance while respectively cave diving, featured B. GONZALES, J. PAKES and B. MALONEY. So much fun and serious diving combined. Puh –what a task. Howdy! A word is to be said regarding my cave diving instructors, T. WÄLDE & P. WIDGET and my buddy S. LIER. Yes, I know, I am not a full cave diver yet. We will see us at latest in summer 2010 for that issue – hope so. You got the bottles I got the brain – lets start nitrogen flooding the vein. Thank you so much for drilling me without that I would not have managed to see the remipedes in their habitat. Well, at least not in the combination with coming back alive and telling about it. Unforgettable! The impressions of the Mexican underworld were in general - unforgettable^2.

T. ILLIFE is really a great buddy and indeed a special, luminary diver to meet. Thank you so much for everything, including the addiction to cave diving! I will never forget the time at your house and especially your hospitality in the "emergency" flat after the destruction of your house by the hurricane IKE in 2008. In the context of the Remipedes I have to thank so much S. KOENEMANN, he is really great and open minded for cooperation that I am really happy about this cooperation and I hope that it proceeds for a long time. And I really hope that I will meet more scientists like him in future ☺.

The time, discussions and talks within the meetings and framework of the "Deep Metazoan Phylogeny" program were really great, hard work and stress but also damn pushing and bringing awesome experiences in diverse issues. Especially the contact and network with Co-PhD students, postdocs (A. BRABAND, M. HELMKAMPF, A. MEYER, M. PERSEKE, K. PICK, P. UNGERER, C. MÜLLER, L. PODSIADLOWSKI) and cooperating partners (T. BURMESTER, M. WIENS, S. HARZSCH) was very fruitful. Anke, thank you so much for the comments to this thesis ☺ !

A great influence had also the discussions with D. WALOSZEK, A. MAAS AND J. HAUG. The importance of morphology and respectively the fossil record was since that always in my mind and hopefully our molecular results are soon good enough to start "the" total evidence paper(s) bringing it all together.

I will remember the PhD time as a time of traveling physically and mentally. A very influencing and pushing experience was the visit to K. KJER in New Jersey and the visit of the "Evolution 2008" meeting in Minneapolis together. Thank you so much Karl, also for the great barbecues and the first "own" baseball gaming. The evenings on the bridge over the

Mississippi at the UoM joining the poster sessions while drinking a cold beer watching in parallel the skyline of Minneapolis. Awsome! Best poster session I ever participated.

Also the "Vienna gang" was really great and always a challenge for me being per definiton not a hexapod expert. But who knows. Lets see… Ingroup strangers feature: G. PASS, N. SUCSICZ, E. DELL AMPIO and D. BARTEL. Emiliano, thank you so much for your primer setting help! It worked fdg.

Regarding the molecular work I have to thank C. ETZBAUER our famous "lab chief in charge". Or "head of anything". Every lab should have a Claudia. Perhaps soon available as a non smoking version. ☺ Jokes apart. THANK YOU. Of course also to the rest of the Lab crew (B. ULLRICH, O. SCHULTZ, J. DAMBACH, P. KÜCK, H. LETSCH, K. LANGEN, J. SCHWARZER, C. GREVE), it was always fun to work with you guys. My two scientific assistents ☺ S. BECKER and B. ULLRICH really were a great help. And yes, I was an official member of the famous lab inauguration party finally ending outside the roof. And here she comes. KAREN "TINY TANTE" MEUSEMANN. The colleague who never sleeps. Especially if looking for typos. The Duracell Bunny. Our discussions were really great. Sometimes hard - but always very professional ☺ It was great joining all the meetings together and keeping talks after each other. The Big B. and tiny t. … Karen, my officially apologies at this place for hiding your cigarettes at the ceiling (it did not work anyway…).

At the end I would like to thank the two captains and chiefs of the boat: PROF. WÄGELE for the great opportunity and the confidence in me to conduct this thesis. He was always open for my plans and suggestions and also pushed me to "feel" and collect the specimens. He supported all the fieldtrips and meetings I joined. By this I got really fascinated in the crustaceans and as by product also to diving. Thank you so much! Together with PROF. MISOF he opened and sharpened the mind for analyses, interpretation and argumentation of data. A suspicious mind is extended. And I am really grateful for the critical remarks and suggestions of both while practicing talks and discussions. I know from so many Co-students that this is not the "standard" procedure.

I am also very grateful to PROF. RUST and PROF. HOCH from the University Bonn who agreed to co-correct this thesis.

The Alexander Koenig Stiftung and the Raiffeisenbank Rhein Sieg I have to thank for financial support for fieldtrips. The travel grant from the Raiffeisenbank was mostly arranged by U. SCHÄKEL and R. MÜLLER, I am very grateful for their trust in me to succeed in collecting the remipedes in Mexican caves systems.

Last but not least I thank my family. My parents, my sister and Phillip supported me in so many ways that I really had a guideline in some times of trouble. If there were some really troubles, I do not really feel so. Stressy times, okay… But all the awesome meals and energy sessions at Rheidt and Rheinbreitbach were unforgettable and they always helped a lot to come down. A last word is directed to all of my good old friends, I know sometimes I am hardly available, sorry for that and thank you so much for your respect!

9. SUPPLEMENT

2. MATERIAL AND METHODS

Table S1 | Collection plan of crustacean and pycnogonid outgroup species. Species of groups that were aimed but impossible to collect are marked by xxx.

	Major group	Infraorder	Family	Species
Thecostraca	**Mystacocarida**		Derocheilocaridae	*Derocheilocaris typicus* (Pennak & Zinn 1943)
	Copepoda	Gymnoplea	Platicopidae	***xxx***
			Calanoida	*Tigriopus fulvus* (Fischer 1860)
		Podoplea	Canuellidae	*Canuella perplexa* (Scott & Scott 1893)
			Cyclopidae	*Cyclops* (Muller 1785) sp.
	Ostracoda	Paleocopida	Punciidae	***xxx***
		Myodocopida		***xxx***
		Podocopida	Cyprididae	*Heterocypris incongruens* (Ramdohr 1808)
				Pontocypris mytiloides (Norman 1862)
	Cirripedia	Thoracica	Lepadomorpha	*Pollicipes pollicipes* (Gmelin (1789)
			Balanomorpha	*Semibalanus balanoides* (Linneaus 1758)
		Acrothoracica		***xxx***
		Rhizocephala		*too derived*
	Facetotecta			*too derived*
	Ascothoracica			*not available*
Branchiopoda	**Cephalocarida**		Lightiella	*Lightiella incisa* (Gooding 1963)
	Anostraca		Artemiidae	*Artemia* (Leach 1819) sp.
			Branchiopodidae	*Branchipus schaefferi* (Fischer 1834)
	Notostraca		Triopsidae	*Triops cancriformis* (Bosc 1801)
	Cladocera	Haplopoda	Leptodoridae	*Leptodora kindtii* (Focke 1844)
		Eucladocera	Daphniidae	*Daphnia magna* (Straus 1820)
				Polyphemus pediculus (Linneaus 1761)
			Bosmiidae	*Bosminia* (Baird, 1845) sp.
	Spinicaudata			*Caenestheria* (Daday 1914) sp.
	Laevicaudata			*Lynceus brachyurus* (Muller 1776)
	Remipedia		Speleonectidae	*Speleonectes tulumensis* (Yager 1987)
	Branchiura			*Argulus foliacaeus* (Linnaeus 1758)
	Pentastomida			*Raillitiella* (Sambon 1910) sp.
	Leptostraca		Nebaliidae	*Sarsinebalia urgorrii* (Moreira, Gestoso & Troncoso 2003)
	Stomatopoda			*Squilla mantis* (Linneaus 1758)
	Pantopoda			*Nymphon stroemii* (Kroyer 1844)
				Colossendeis (Jarzinsky 1870) sp.

Table S2 | Localities and fixation methods for collected specimens. Collectors of specimes and collection date are additionally given.

Major group	Taxon	Collection Locality	date	Collector	Fixation
Mystacocarida	*Derocheilocaris typicus*	Playa dos ninos, Ferrol, Galicia, Spain	2006	v. Reumont	98% Ethanol
Copepoda	*Tigriopus fulvus*	Vigo, Galicia, Spain	2006	v. Reumont	98% Ethanol
	Canuella perplexa	Hooksiel, Niedersachsen, Germany	2005	v. Reumont	98% Ethanol
	Cyclops sp.	Wahner Heide, Nord-Rhein-Westfalia, Germany	2005	v. Reumont	98% Ethanol
Ostracoda	*Heterocypris incongruens*	Hirschweiher, Röttgen, Nord-Rhein-Westfalia, Germany	2005	v. Reumont	98% Ethanol
	Pontocypris mytiloides	Wilhelmshaven, Niedersachsen, Germany	2007	v. Reumont	98% Ethanol
	Podocopida	Campese Bay, Isla di Giglioi, Italy (bait trap, 15m depth)	2008	v. Reumont	98% Ethanol RNAlater
Cirripedia	*Pollicipes pollicipes*	Ferrol supermercado, Galicia, Spain	2006	v. Reumont	98% Ethanol RNAlater
	Semibalanus balanoides	Hooksiel, Niedersachsen, Germany	2006	v. Reumont	98% Ethanol
Cephalocarida	*Lightiella incisa*	Carry Bow Cay, Barrier Reef, Belize	2008	Ott	98% Ethanol
Anostraca	*Artemia sp.*	Tegler See, Berlin, Germany	2005	Braband	DNA-sample
	Branchipus schaefferi	Marchauen, Austria	2006	Eder	98% Ethanol
Notostraca	*Triops cancriformis*	Marchauen, Austria	2005	Eder	98% Ethanol RNAlater
Cladocera	*Leptodoria*	Tegler See, Berlin, Germany	2005	Braband	DNA-sample
	Daphnia sp.	Bonn, Nord-Rhein-Westfalia, Germany	2005	v. Reumont	98% Ethanol
	Polyphemus pediculus	Tegler See, Berlin, Germany	2005	Braband	DNA-sample
	Bosminia	Tegler See, Berlin, Germany	2005	Braband	DNA-sample
Spinicaudata	*Caenestheria berneyi*	Tegler See, Berlin, Germany	2005	Braband	DNA-sample
Laevicaudata	*Lynceus brachyurus*	Tegler See, Berlin, Germany	2005	Braband	DNA-sample
Remipedia	*Speleonectes tulumensis*	Cenote Eden, Puerto Aventuras, Quintana Roo, Mexico	2006	Koenemann	98% Ethanol
	Speleonectes tulumensis	Cenote Crustacea, Akumal, Quintana Roo, Mexico	2007 2008	v. Reumont	98% Ethanol RNAlater
Branchiura	*Argulus foliacaetus*	Sweden	2007	Waloßek	98% Ethanol
	Argulus foliacaetus	Bochum, fishpond, Nord-Rhein-Westfalia, Germany	2006	Strieso	98% Ethanol
Pentastomida	*Raillitiella sp.*	Asia, host: *Hemidactylus* cf. *frenatus*	2007	v. Reumont	98% Ethanol
Leptostraca	*Sarsinebalia urgorrii*	Ria Ferrol, Ferrol, Galicia, Spain	2006 2009	v. Reumont	98% Ethanol RNAlater
Stomatopoda	*Squilla mantis*	Porto San Stefano, Italy	2007	v. Reumont	98% Ethanol
Pantopoda	*Nymphon stroemii*	Hinlopen Svalbard, Arctica	2003	Krapp	98% Ethanol
	Colosseides spec.	ANDEEP I Expedition, Ant XIX-3, Antarctica	2002	Raupach	94% Ethanol

Table S3 | Primer list for all amplified and sequenced fragments. Primer sequences are given in 5' to 3' prime end direction. References and modifications are given for each primer. "PCR" indicates the use of primers in the PCR reaction, "CS" indicates that primer were applied for cycle sequencing. For combination of used primers see Figure S1.

Marker	Primer name	Reaction	Sequence (in 5' - 3' direction)	Direction	Reference
16S rRNA	16Sa	PCR & CS	CGCCTGTTTATCAAAAACAT	forward	Palumbi et al. (1991)
16S rRNA	16Sb	PCR & CS	CCGGTCTGAACTCAGATCACG	reverse	Palumbi et al. (1991)
16S rRNA	LRJ12887	PCR & CS	CCGGTCTGAACTCAGATCACGT	forward	Simon et al. (1994)
16S rRNA	LRN13398	PCR & CS	CGCCTGTTTAACAAAAACAT	reverse	Simon et al. (1994)
18S rRNA	18A1	PCR & CS	CTGGTTGATCCTGCCAGTCATATGC	forward	Dreyer & Wägele (2001)
18S rRNA	1800	PCR & CS	GATCCTTCCGCAGGTTTCACCTACG	reverse	Dreyer & Wägele (2001)
18S rRNA	700 F-MR	CS	GCCGCGGTAATTCCAGC	forward	Raupach unpubl.
18S rRNA	700R	CS	CGCGGCTGCTGGCACCAGAC	reverse	Dreyer & Wägele (2001)
18S rRNA	1000F	CS	CGATCAGATACCGCCCTAGTTC	forward	Dreyer & Wägele (2001)
18S rRNA	1155R	CS	CCGTCAATTCCTTTAAGTTTCAG	reverse	Dreyer & Wägele (2001)
18S rRNA	1250 FN-MR	CS	GGCCGTTCTTAGTTGGTGGAG	forward	Raupach unpubl.
18S rRNA	1500R	CS	CATCTAGGGCATCACAGACC	reverse	Wollscheid et al. unpubl.
28S rRNA	CS632	PCR & CS	CGATGAAGAACGCAGC	forward	Schlötterer et al. (1994)
28S rRNA	427 or D1a	PCR & CS	CCC(C/G)CGTAA(T/C)TTAAGCATAT	forward	Friedrich & Tautz (1997)
28S rRNA	D2a	PCR & CS	GATAGCGAACAAGTACC	forward	Dell'Ampio et al. (2009)
28S rRNA	D3a	PCR & CS	GACCCGTCTTGAAACACGGA	forward	Nunn et al. (1996)
28S rRNA	D3b.rev.MOD	PCR & CS	TAGTAGCTGGTTCCTTCCG	forward	Nunn et al. (1996), modif. reverse D3b, Dell'Ampio et al. (2009)
28S rRNA	742 or D5a	PCR & CS	CTCAAACTTTAAATGG	forward	Friedrich & Tautz (1997)
28S rRNA	28ee.mod	PCR & CS	CCGCTAAGGAGTGTGTAAC	forward	Hillis & Dixon (1991), modif. Dell'Ampio, unpubl. (PHDthesis)
28S rRNA	476 or D7a1	PCR & CS	CTGAAGTGGAGAAGGGT	forward	Friedrich & Tautz (1997)
28S rRNA	D7aN	PCR & CS	AGAACCTGGTGACGGAAC	forward	Dell'Ampio, unpubl. (PHDthesis)
28S rRNA	D7b.rev	PCR & CS	ATGTAGGTAAGGGAAGTC	forward	Friedrich & Tautz (1997), reverse D7b, Dell'Ampio et al. (2009)
28S rRNA	D7b.rev.MOD	PCR & CS	GATCCGTAACTTCG	forward	Friedrich & Tautz (1997), reverse D7b modif., Dell'Ampio et al. (2009)
28S rRNA	D8aN	PCR & CS	TCAGAACTGGCACGGACCGG	forward	Dell'Ampio, unpubl. (PHDthesis)
28S rRNA	28v	PCR & CS	AAGGTAGCCAAATGCCTCATC	forward	Hillis & Dixon (1991)
28S rRNA	28w	PCR & CS	CCT(G/T)TTGAGCTTGACTCTAATCTG	forward	Hillis & Dixon (1991)
28S rRNA	D10aPC	PCR & CS	GGGGAGTTTGACTGGGGCGG	forward	Dell'Ampio et al. (2009)
28S rRNA	D12aN	PCR & CS	GAGCAAGAGGTGTCAGAAAAGTTAC	forward	Dell'Ampio, unpubl. PHDthesis
28S rRNA	D1a.rev	PCR & CS	ATATGCTTAAATTAAGCGGG	reverse	Friedrich & Tautz (1997), reverse D1a, Dell'Ampio
28S rRNA	D1b2	PCR & CS	CGTACTATTGAACTCTCTCTT	reverse	Dell'Ampio et al. (2002)
28S rRNA	D3a.rev	PCR & CS	TCCGTGTTTCAAGACGGGAC	reverse	Nunn at al. (1996), reverse D3a, Dell'Ampio, unpubl. (PHDthesis)
28S rRNA	D3b	PCR & CS	TCCGGAAGGAACCAGCTACTA	reverse	Nunn et al. (1996)

28S rRNA	706 or D5b2	PCR & CS	CGCCAGTTCTGCTTACC	reverse	Friedrich &Tautz (1997)
28S rRNA	689 or D5b1	PCR & CS	ACACACTCCTTAGCGGA	reverse	Friedrich & Tautz (1997
28S rRNA	D7a1.rev	PCR & CS	AAACCCTTCTCCACATCGG	reverse	Friedrich & Tautz (1997), reverse D7a1.rev, Dell'Ampio et al. (2009)
28S rRNA	477 or D7b	PCR & CS	GACTTCCCTTACCTACAT	reverse	Friedrich & Tautz (1997)
28S rRNA	D7bNLe	PCR & CS	GGACCCGACGGATTCTC	reverse	Dell'Ampio
28S rRNA	23 or 28f	PCR & CS	CAGAGCACTGGGCAGAAATCAC	reverse	Dell'Ampio, unpubl. (PHDthesis)
28S rRNA	28w.rev	PCR & CS	CAGATTAGAGTCAAGCTCAACAGG	reverse	Hillis & Dixon (1991), reverse 28w, Dell'Ampio et al. subm
28S rRNA	28jj	PCR & CS	AGTAGGGTAAAACTAACCT	reverse	Hillis & Dixon (1991)
28S rRNA	D10bN	PCR & CS	TTTGACAGATGTACCCCCCC	reverse	Dell'Ampio, unpubl. (PHDthesis)
28S rRNA	D12b.PLANB	PCR & CS	GAGTACGACACCCC	reverse	Dell'Ampio et al. (2009)
28S rRNA	D12bN	PCR & CS	TATGGCAGCTGCTCTACC	reverse	Dell'Ampio, unpubl. (PHDthesis)
28S rRNA	Mallat.Rv1	PCR & CS	ACTTTCAATAGATCGCAG	reverse	Mallat & Sullivan (1998)
COI	HCO	PCR & CS	TAAACTTCAGGGTGACCAAAAAATCA	forward	Folmer et al. (1994)
COI	LCO	PCR & CS	GGTCAACAAATCATAAAGATATTGG	reverse	Folmer et al. (1994)

PRIMER REFERENCES

DELL`AMPIO, E., CARAPELLI, A. & FRATI, F. (2002). Secondary structure and sequence variation of the 28S rRNA gene in the Neanuridae, and its utility as a phylogenetic marker: Proceedings of the Xth international Colloquium on Apterygota, Ceske Budejovice 2000: Apterygota at the Beginning of the Third Millennium. *Pedobiologia (Jena)* **46,** 274-283.

DELL'AMPIO, E., SZUCSICH, N., 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. *Zoologica Scripta*, **38**, 155-170.

DREYER, H. & WÄGELE, J. W. (2001). Parasites of crustaceans (Isopoda: Bopyridae) evolved from fish parasites: molecular and morphological evidence. *Zoology (Jena)* **103,** 157-178.

FOLMER, O., BLACK, M.; HOEH, W., LUTZ, R., VRIJENHOEK, R. (1994). DANN primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Molecular Marine Biology Biotech **3: 294-299.**

FRIEDRICH, M. & TAUTZ, D. (1997). An episodic change of rDNA nucleotide substitution rate has occurred during the emergence of the insect order Diptera. *Molecular Biology and Evolution* **14,** 644-53.

HILLIS, D. M. & DIXON, M. T. (1991). Ribosomal DNA: molecular evolution and phylogenetic inference. *The Quarterly review of biology* **66,** 411-53.

MALLATT, J. & SULLIVAN, J. (1998). 28S and 18S rDNA sequences support the monophyly of lampreys and hagfishes. *Molecular Biology and Evolution* **15,** 1706-1718.

NUNN, G. B., THEISSEN, B. F., CHRISTENSEN, B. & ARCTANDER, P. (1996). Simplicity-correlated size growth of the nuclear 28S ribosomal RNA D3 expansion segment in the crustacean order Isopoda. *Journal of Molecular Evolution* **42,** 211-223.

PALUMBI, S. R., MARTIN, A., ROMANO, S., MCMILLAN, W. O., STICE, L., GRABOWSKI, G. (1991). The simple fools guide to PCR. *A collection of PCR protocols, version 2. Honolulu*, University of Hawai.

SCHLÖTTERER, C., HAUSER, M. T., VON HAESELER, A. & TAUTZ, D. (1994). Comparative evolutionary analysis of rDNA ITS regions in Drosophila. *Molecular Biology and Evolution* **11,** 513-22.

SIMON, C., FRATI, F., BECKENBACH, A., CRESPI, B., LIU, H., FLOOK, P. (1994). Evolution, weighting and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. *Annals of the Emtomological Society of America* **87**, 651-701.

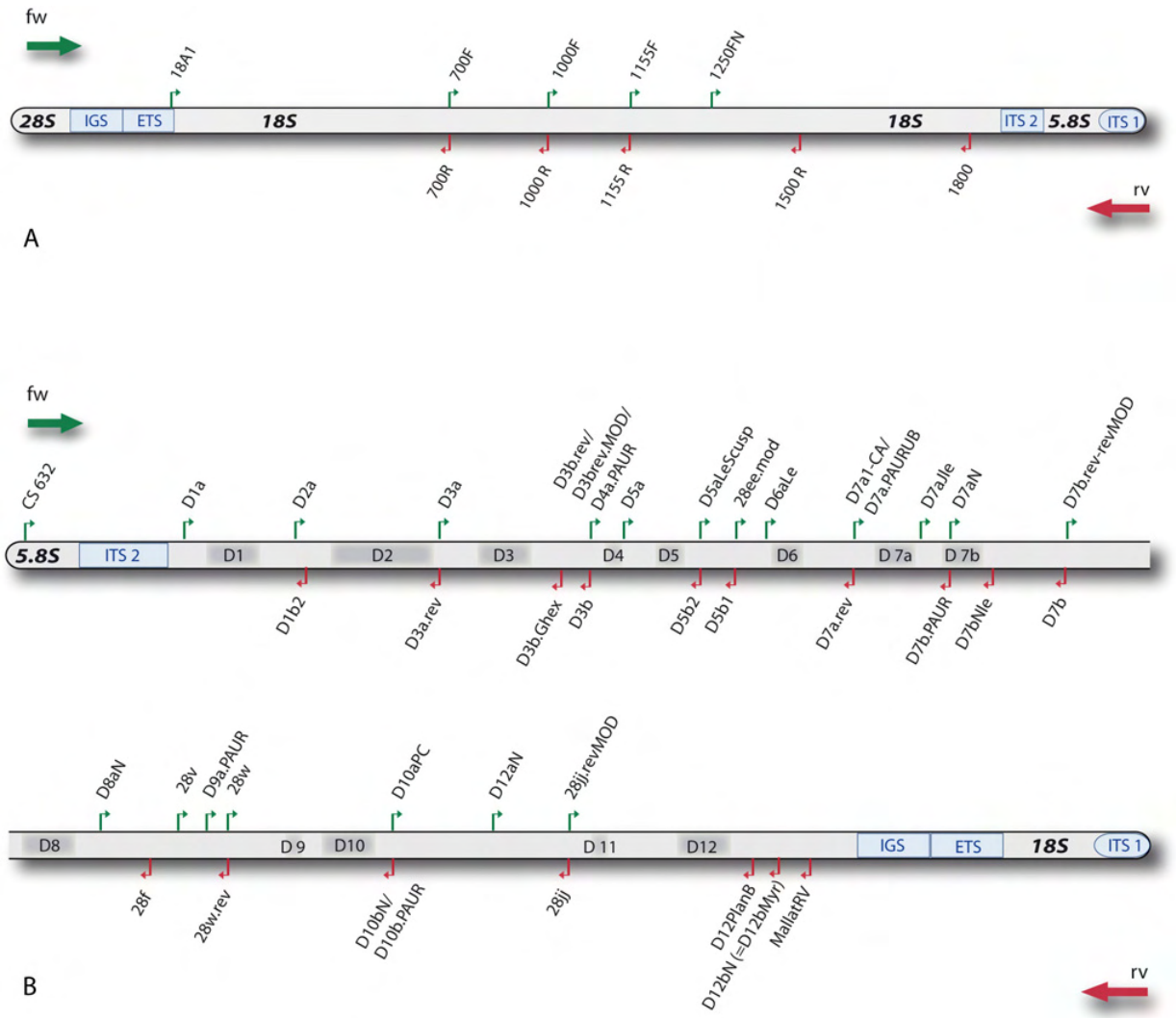


Figure S1 | Primer card. Used primers and their positions for the 18S (A) and the 28S rRNA genes are shown. Green arrows mark the position of forward primers, red arrows the position of reverse primers. In case that different primers were used for the same position combined primers are given on the specific arrow.

The 18S in crustaceans was amplified in one PCR product (18A1/1800) and sequenced with eight primers (700F, 1000F, 1155F, 1250FN, 700R, 1000R, 1155R and 1500R), see table S3. Unless otherwise noted all applied protocols refer to manufacturers advices. The PCR-Multiplex-Kit (Qiagen) was used to prevent pooling of weak PCR products, in case this failed, weak PCR products were pooled for purificaton. PCR Products were purified with the NucleoSpin Extract II (Macherey-Nagel). Purified products were checked on agarose gel. To estimate the DNA concentration a mass marker (BioRad) and Nanodrop Spectrophotometer ND-1000 (peqLab) was used.

The nuclear 28S rRNA gene was amplified in nine overlapping fragments using following primer combinations: CS632/D1b2, D1a/D3b, D2a/D3a.rev, D3a/D5b1, D5a/D7b, D7a1/28f, D7b.rev/28w.rev, 28v–28jj and D12aN/D12bN or alternatively D12aN/D12bPLANB, D12aN/D12bMYR or D12aN/MallatRv1. Different primer combinations were used whenever necessary for specific taxa. Alternative combinations for crustaceans are: D3b.rev/D5b2, D3b.rev/D5b1, D1a/D5b1, D1a/D5b2, D2a/D5b1, D2a/D5b2, D3a/D5b2, D7aN/28f, D7b.rev/28f, D7brev/D10bN, D10aPC/D12bN and D12aN/D12b.PLANB.

Cycle Sequencing reactions were performed using DNA Quick Start Mastermix (Beckman Coulter). CS products were ethanol-precipitated or purified with CleanSeq magnetic bead system (Agencourt) followed by sequencing on Beckman Coulter capillary sequencers CEQTM 8000 and

CEQTM 8800.

Table S4 | Used PCR chemicals. The two different chemical mixes are given. Concentration of primers and the use of DMSO was tested in different gradients and varied in for the reactions. Cycle sequencing was conducted following the Beckmann Coulter protocol for the 8000/8800 capillary sequencers. PCR= Polymerase chain reaction, HPLC= High Performance Liquid Chromatography, dNTPs= di-Nucleotidetriphosphate.

PCR reaction mix	Chemicals	[Concentration]	Volume	Gene	Specifications
(A)	Reagents (SIGMA)			18S, 28S	28S : Different $MgCl_2$-gardients, PCR-profile 1;
	10 x PCR buffer without $MgCl_2$		5.0 µl	12S, 16S,	18S : DMSO replaced
	$MgCl_2$	[25 mM]	5.0 µl	H3, COI,	by sterile water
	dNTPs	[2 mM]	4.0 µl		
	DMSO		2.5 µl		generally test gradients for each gene fragment were performed to settle the optimal PCR reaction set. DMSO and MgCl2 concentrations varied
	Primer forward	[10 pmol/µl]	0.8 µl		
	Primer reverse	[10 pmol/µl]	0.8 µl		
	Taq-Polymerase	[5 u/µl]	0.15 µl		
	HPLC-H_2O		30.75 µl		
	DNA template		1.0 – 2.0 µl		
	total volume		50 µl		
(B)	Reagents (Qiagen)			18S, 28S	generally test gradients for each gene fragment were performed to settle the optimal PCR reaction set. DMSO and MgCl2 concentrations varied
	Multiplex Mastermix (incl. mixture of taq, dNTPs, $MgCl_2$, reaction buffer		10.0 µl	12S, 16S,	
	2 µl Q-solution		2.0 µl	H3, CO I	
	1.6 µl Primer forward	[10 pmol/µl]	1.6 µl		
	1.6 µl Primer reverse	[10 pmol/µl]	1.6 µl		
	HPLC-H_2O		4.3 µl		
	DNA template		0.5 – 1.0 µl		
	total volume		20 µl		

Table S5 | PCR temperature profiles. Temperature is given in Celsius °C, the runtime in minutes. TD= touchdown.

Profile	Temperature profile	Cycles	Gene	Thermocycler	Remarks / Primer specification
1	94°C 3:00 min		16S, COI	GeneAmp PCR System 2720, GeneAmp PCR System 2700, (Applied Biosystems) T3000 Thermocycler	Depending on fragments and taxa the 1st annealing temperature varied from 60°C-45°C or 55°C-40°C or 50°C-35°C. In each cycle the temperature was decreased by 1°C.
	94°C 0:35 min 60°C 0:30 min, TD -1°C to 45°C 72°C 1:30 min	15 cycles	18S, 28S		
	94°C 0:35 min 50°C 0:30 min	25 cycles			

	72°C 1:30 min			(Biometra)	
	72°C 10:00 min				
	4°C				
2	94°C 3:00 min		18S, 28S	GeneAmp PCR System 2720, GeneAmp PCR System 2700, (Applied Biosystems) T3000 Thermocycler (Biometra)	Depending on fragments and taxa the 1st annealing temperature varied from 60°C-45°C or 55°C-40°C or 50°C-35°C. In each cycle the temperature was decreased by 1°C.
	94°C 1:00 min	15 cycles			
	60°C 0:30 min, TD -1°C to 45°C				
	72°C 1:30 min				
	94°C 0:35 min	25 cycles			
	50°C 0:30 min				
	72°C 1:30 min				
	72°C 12:00 min				
	4°C				
3	94°C 5:00 min		18S, 28S	GeneAmp PCR System 2720, GeneAmp PCR System 2700, (Applied Biosystems) T3000 Thermocycler (Biometra)	Depending on fragments and taxa the 1st annealing temperature varied from 60°C-45°C or 55°C-40°C or 50°C-35°C. In each cycle the temperature was decreased by 1°C.
	94°C 1:00 min	15 cycles			
	60°C 0:35 min, TD -1°C to 45°C				
	72°C 1:30 min				
	94°C 0:35 min	25 cycles			
	50°C 0:30 min				
	72°C 1:30 min				
	72°C 10:00 min				
	4°C				
4	94°C 4:00 min		18S, 28S	GeneAmp PCR System 2720, GeneAmp PCR System 2700, (Applied Biosystems) T3000 Thermocycler (Biometra)	Depending on fragments and taxa the 1st annealing temperature varied from 60°C-45°C or 55°C-40°C or 50°C-35°C. In each cycle the temperature was decreased by 1°C.
	94°C 1:00 min	15 cycles			
	60°C 1:00 min, TD -1°C to 45°C				
	72°C 1:30 min				
	94°C 1:00 min	25 cycles			
	50°C 0:35 min				
	72°C 1:30 min				
	72°C 12:00 min				
	4°C				
5	94°C 3:00 min		16S, COI 18S, 28S	Applied Biosystems 9600	Depending on fragments and taxa the 1st annealing temperature varied from 60°C-45°C or 55°C-40°C or 50°C-35°C. In each cycle the temperature was decreased by 1°C.
	94°C 1:00 min	15 cycles			
	60°C 0:30 min, TD -1°C to 45°C				
	72°C 1:30 min				
	94°C 0:35 min	25 cycles			
	50°C 0:30 min				
	72°C 1:30 min				
	72°C 10:00 min				
	4°C				

Table S6 | Overview of sequenced genes for the collected specimens. Green plus indicates the sequencing and implementation into analyses of the sequence. A circle represents sequences that were not successfully sequenced. A minus incdicates that the amplification was not successful at all. For the NCBI accession numbers see the taxon lists of each analysis.

taxon		group	species	sequenced genes			
Standard Fragments and additional Genes				**16S rRNA**	**18S rRNA**	**28S rRNA**	**CO I**
	C02	Ostracoda	*Heterocypris incongruens*	+	+	+	+
	C01	Notostraca	*Triops cancriformis*	-	+	+	+
Crustacea	C10	Mystacocarida	*Derocheilocaris typica*	+	+	+	+
	C16	Copepoda	*Canuella perplexa*	+	+	+	+
	C21		*Tigriopus fulvus*	+	+	+	-
	C18		*Pontocypris mytiloides*	+	+	+	-
	C17	Cirripedia	*Pollicipes pollicipes*	+	+	+	+
	C37		*Semibalanus balanoides*	+	+	+	+
	C38	Branchiura	*Argulus foliaceus*	+	+	+	+
	C44	Anostraca	*Branchipus schaefferii*	+	+	o	+
	C32	Cladocera	*Daphnia sp.*	+	+	+	-
	C34		*Bosmina sp.*	+	+	+	+
	C56		*Polyphemus pediculus*	+	-	+	+
	C53	Spinicaudata	*Caenestheria berneyi*	+	+	+	-
	C51	Laevicaudata	*Lynceus brachyurus*	+	+	+	-
	C42	Remipedia	*Speleonectes tulumensis*	+	+	o	+
	C14	Leptostraca	*Nebalia spec.*	-	+	+	+
	C63	Stomatopoda	*Squilla mantis*	+	+	+	+
	C68	Cephalocarida	*Lightiella incisa*	+	o	o	+
	C65	Pentastomida	*Raillietiella sp.*	+	o	o	+
out group	C28	Pantopoda	*Nymphon stroemii*	+	+	+	+
	C40		*Colosseides spec.*	+	+	+	+

Table S7 | Taxa list for analysis A. List of genetic markers, specimens and taxa used for the phylogenetic analysis. Sequences obtained from GenBank are shown by accession numbers (Acc. no.); new sequences from this thesis are colored in bold green. Gene sequences that we were unable to acquire are marked as "n/a". Classification of Crustacea is according to Martin and Davis (2001). Sequences marked with [1] were sequenced by the group Koenemann in Hannover. * 18S sequences for *Labidura riparia* consisted of three, non-contiguous fragments.

CRUSTACEA		**Acc. no. COI**	**Acc. no. 16S**	**Acc. no. 18S**
Remipedia				
Speleonectidae	*Speleonectes tulumensis*	NC_005938	NC_005938	L81936
Godzilliidae	*Pleomotra apletocheles*	**xxx[1]**	**xxx[1]**	**xxx[1]**
Branchiopoda				
Anostraca	*Artemia franciscana*	NC_001620	NC_001620	AJ238061
Notostraca	*Triops cancriformis*	**GQ328960**	**GQ328946**	**EU370422**
Diplostraca				
Laevicaudata	*Lynceus brachyurus* +*L. macleyanus*	DQ467706	**GQ328954**	**GQ328957**
Spinicaudata	*Eulimnadia braueriana*	EF189667	EF189604	EF189621
Cyclestherida	*Cyclestheria hislopi*	DQ889093	EF189603	AF144209
Cladocera				
Anomopoda	*Daphnia magna* + *D.* cf. *magna*	AY803061	**GQ328951**	**EU370423**
Ctenopoda	*Sida crystallina*	AF277889	DQ470594	AM490294
Onychopoda	*Polyphemus pediculus*	**GQ328966**	**GQ328955**	EF189633
Haplopoda	*Leptodora kindtii*	DQ310659	**GQ328950**	AF144214
Malacostraca				
Stomatopoda	*Squilla mantis*	**GQ328967**	**GQ328956**	**GQ328957**
Leptostraca	*Paranebalia longipes*	n/a	AY744909	EF189630
Syncarida				
Anaspidacea	*Anaspides tasmaniae*	DQ889076	AF133685	L81948
Eucarida				
Euphausiacea	*Euphausia pacifica*	AF177184	AF177176	AY141010
	Meganyctiphanes norvegica	AY601091	AY744910	DQ900731
Decapoda				
Dendrobranchiata	*Penaeus monodon* + *P. semisulcatus*	NC_002184	NC_002184	DQ079766
Pleocyemata				
Stenopodidea	*Stenopus hispidus*	AF125441	AY583884	AY743957
Anomura	*Eumunida sternomaculata*	EU243561	AY351260	AF436011
Palinura	*Jasus verreauxi*	AF192883	AF192874	AF498665
Astacidea	*Parastacus pugnax*	EF599157	AF175239	AF235969
Brachyura	*Carcinus maenas*	FJ159028	AJ130811	AY583974
Caridea	*Atyoida bisulcata*	n/a	EF489995	DQ079738
Peracarida				
Mysida	*Mysis oculata*	EF609269	DQ189194	AM422510
Lophogastrida	*Neognathophausia ingens*	DQ889115	n/a	AM422475
Mictacea	*Thetispelecaris remex*	n/a	n/a	AY781416
Amphipoda	*Orchestia cavimana*	EF989708	AY744911	AY826953
	Gammarus pulex	EF570334	AJ269626	EF582923
Isopoda	*Ligia oceanica*	NC_008412	NC_008412	AF255698
	Colubotelson thomsoni	AF255775	AF259531	AF255703
Cumacea	*Diastylis sculpta* + *D.* sp.	AF137510	U81512	Z22519
Tanaidacea	Tanaidacea sp.	AF520452	n/a	AY743939
Spelaeogriphacea	*Spelaeogriphus lepidops*	n/a	n/a	AY781414
Thermosbaenacea	*Tethysbaena argentarii*	n/a	DQ470612	AY781415

Maxillopoda				
Mystacocarida	*Derocheilocaris typicus* 1	**GQ328961**	n/a	**EU370429**
	Derocheilocaris typicus 2	**GQ328961**	n/a	L81937
Copepoda				
Calanoida	*Calanus pacificus*	AF315013	AF315006	L81939
Harpacticoida	*Cletocamptus deitersi*	AF315010	AF315003	n/a
	Tigriopus cf. *fulvus*	n/a	n/a	**EU370430**
Cyclopoida	*Stellicola* sp.	DQ889130	n/a	AY627004
Cyclopoida	*Thermocyclops inversus + T.* sp.	EU770558	n/a	DQ107580
Siphonostomatoida	*Caligus elongatus*	EF452647	AY660020	AY627020
Ostracoda				
Myodocopa				
Myodocopida	*Parasterope gamurru*	n/a	EU587255	EU591819
Halocyprida	*Polycope japonica*	n/a	n/a	AB076657
Podocopa				
Platycopida	*Cytherella leizhouensis*	n/a	n/a	AB076611
Podocopida				
Cypridocopina	*Heterocypris incongruens*	n/a	**GQ328947**	**EU370424**
Branchiura				
Arguloida	*Argulus nobilis*	n/a	n/a	M27187
Arguloida	*Dolops ranarum + D.* sp.	DQ889096	n/a	DQ813453
Pentastomida				
Cephalobaenida	*Raillietiella* sp.	n/a	n/a	**EU370434**
Thecostraca				
Facetotecta	*Hansenocaris itoi*	n/a	n/a	AF439393
Ascothoracida	*Dendrogaster asterinae*	n/a	n/a	AF057560
	Ulophysema oeresundense	n/a	n/a	L26521
Cirripedia				
Sessilia	*Semibalanus balanoides*	**GQ328964**	**GQ328952**	**EU370426**
Pedunculata	*Pollicipes pollicipes*	**GQ328962**	**GQ328948**	**EU370427**
Kentrogonida	*Heterosaccus californicus*	n/a	AY520756	AY265359
Cephalocarida				
	Hutchinsoniella macracantha	AY456189	AY456189	L81935
	Lightiella incisa	**GQ328968**	n/a	**GQ328959**
HEXAPODA				
Protura				
Sinentomata	*Fujientomon dicestum*	n/a	n/a	AY596359
Acerentomata	*Neocondeellum dolichotarsum*	n/a	n/a	AY037170
Diplura	*Campodea fragilis + C. tillyardi*	DQ529236	NC_008233	AF173234
Collembola	*Sminthurinus bimaculatus*	AY555545	AY555555	AY555522
Insecta				
Archaeognatha	*Trigoniophthalmus alternatus*	NC_010532	NC_010532	U65106
Zygentoma	*Tricholepidion gertschi*	AY191994	AY191994	AF370789
Pterygota	*Callibaetis ferrugineus*	AY326804	AF370873	AF370791
	Lestes rectangularis	n/a	EF044271	FJ010011
Neoptera	*Echinosoma yorkense*	n/a	AY144636	AY144626
	*Labidura riparia**	AB435163	AY144640	U65114, AY707333, AY707356
	Ceuthophilus gracilipes + C. uthaensis	AY793593	AY793561	AY521870
	Tipula sp.	AY165639	EU005437	X89496
	Anopheles gambiae + A. albimanus	DQ465336	L20934	L78065
MYRIAPODA				
Chilopoda	*Thereuopoda clunifera*	AY288739	AY288716	AF119088

Symphyla	*Scutigerella causeyae* 1	DQ666065	DQ666065	AY336742
	Scutigerella causeyae 2	DQ666065	DQ666065	AF007106
Pauropoda	*Allopauropus* sp.	n/a	n/a	DQ399857
Diplopoda	*Polyxenus lagurus*	AF370840	n/a	X90667
CHELICERATA				
Pycnogonida	*Austrodecus glaciale*	DQ390048	DQ389994	DQ389890
	Nymphon sp.	**GQ328963**	**GQ328949**	EU420136
	Colossendeis sp.	**GQ328965**	**GQ328953**	EU420135
Xiphosura	*Limulus polyphemus*	AF216203	AF373606	L81949
Arachnida				
Araneae	*Atrax* sp.	n/a	AF370857	AF370784
	Hypochilus thorelli + *H. pococki*	NC_010777	NC_010777	AF062951
Scorpiones	*Pandinus imperator*	AY156582	AY156567	AY210831
	Androctonus australis	AJ506919	AJ506868	X77908
Acari	*Amblyomma triguttatum*	AB113317	AB113317	AF018641
Opiliones	*Siro valleorum*	AY639580	AY639552	AY639492
Uropygi	*Mastigoproctus giganteus*	NC_010430	NC_010430	AF005446
ONYCHOPHORA				
Peripatidae	*Epiperipatus biolleyi* + *Euperipatoides leuckarti*	NC_009082	NC_009082	U49910
TARDIGRADA				
Heterotardigrada	*Echiniscus testudo*	EF620375	n/a	DQ839607

Table S8 | Pretests for the Bayes Factor Tests (BFT). Best harmonic means of the model likelihood was reached in runs D1 and F1. D1 was consequently the 0-hypothesis in the BFT (A) and F1 the 0-hypothesis in BFT (B). All other hypotheses (ln likelihoods of remaining runs) were tested against these.

Testruns	**run A1**	**run A2**	**run B1**	**run B2**	**run C1**	**run C2**	**run C3**	**run D1**	**run D2**	**run F1**	**run F3**
Align. prog.	muscle	muscle	muscle	muscle	mafft	mafft	mafft	mafft	mafft	mafft	mafft
Optimization			RNAsalsa Aliscore (-r) Alicut	RNAsalsa Aliscore (-r) Alicut	RNAsalsa Aliscore (-r) Alicut	RNAsalsa Aliscore (-r) Alicut	RNAsalsa Aliscore (-r) Alicut	RNAsalsa Aliscore (-r) Alicut	RNAsalsa Aliscore (-r) Alicut	RNAsalsa Aliscore (-r) Alicut RY-coded	RNAsalsa Aliscore (-r) Alicut RY-coded
Model	nst=6	nst=6, r-diff	nst=6 gamma	nst=6, r-diff gamma	nst=6 gamma	nst=6, r-diff gamma	nst=2, r-diff gamma	nst=6 (all) gamma 5 partitions, unlinked	nst=2 (all) gamma 5 partitions, unlinked	nst=2 + 1 gamma 5 partitions, unlinked	nst=2 + 1 gamma 5 partitions
Iterations	20 mio.	20 mio.	20 mio.	20 mio.	20 mio.	20 mio.	30 mio	20 mio	20 mio	30 mio	30 mio
Excluded burnin	7 mio.	7 mio.	7 mio.	7 mio.	7 mio.	7 mio.	7 mio	7 mio	7 mio	7 mio	7 mio
n-chains	2 x 4	2 x 4	2 x 4	2 x 4	2 x 4	2 x 4	2 x 4	2 x 4	2 x 4	2 x 4	2 x 4
|HarMean|	120431.044	118111.4569	55826.7935	54538.5661	55541.4968	54273.211	54569.5579	54054.4693	54158.5021	40863.4973	41160.4058
BFT	**not included**	**not included**	**A**	**A**	**A**	**A**	**A**	**A**	**A**	**B**	**B**
Hypothesis 1			55826.7935	54538.5661	55541.4968	54273.211	54569.5579	54054.4693	54158.5021	40863.4973	41160.4058
Hypothesis 0			54054.4693	54054.4693	54054.4693	54054.4693	54054.4693		54054.4693		40863.4973
ln (B10)			1772.3242	484.0968	1487.0275	218.7417	515.0886		104.0328		296.9085
2 ln (B10)			3544.6484	968.1936	2974.055	437.4834	1030.1772	**H_0**	**208.0656**	**H_0**	**296.9085**

Table S9 | Finalruns for the final Bayes Factor Tests (BFT). Best harmonic means of the model likelihood was reached in finalruns 4 and 5. But the topologies of these runs showed no resolution, obviously affected by the MUCLE alignment reconstruction. Thus finalrun 1 and finalrun 3 were chosen as best models. Finalrun 1 was consequently the 0-hypothesis in the BFT (A) and finalrun 3 the 0-hypothesis in BFT (B). For resulting topologies of all runs see figures S2-S11.

Testruns	**finalrun 1**	**finalrun 2**	**finalrun 3**	**finalrun 4**	**finalrun 5**	**finalrun 6**	**finalrun 7**	**finalrun 8**	**finalrun 9**	**finalrun 10**
Align. prog.	mafft	mafft	mafft	muscle	muscle	by hand	by hand	by hand	by hand	mafft
Optimization	RNAsalsa	RNAsalsa	RNAsalsa	RNAsalsa	RNAsalsa		alignment masked by hand	RNAsalsa	RNAsalsa	RNAsalsa
	Aliscore (r)	Aliscore (r)	Aliscore (r)	Aliscore (r)	Aliscore (r)			Aliscore (r)	Aliscore (r)	Aliscore (r)
	Alicut	Alicut	Alicut	Alicut	Alicut			Alicut	Alicut	Alicut
Model	RYcoded	RYcoded	RYcoded	RYcoded	RYcoded				RYcoded	RYcoded
	nst=2 + 1	nst=2 + 1	nst=2 + 1	nst=2 + 1	nst=2 + 1	nst=2 + 1	nst=2 + 1	nst=2 + 1	nst=2 + 1	
	gamma	gamma	invgamma	gamma	gamma					
	5 partitions	5 partitions	5 partitions	5 partitions	5 partitions	3 partitions	3 partitions	3 partitions	3 partitions	RAXML
iIerations	40 mio	40 mio	40 mio	40 mio	40 mio	40 mio	40 mio	40 mio	40 mio	
Excluded burnin	unlinked				unlinked			unlinked	unlinked	unpartitioned
Alignment positions	2528	2528	2528	2547	2547	4902	3288	2449	2449	
|HarMean| BFT	**41813.19444**	42127.45431	**42076.51351**	*41502.80506*	*41196.4578*	109819.3987	97364.26521	54675.27796	44070.40893	
Hypothesis 1		42127.45431	42076.51351	Not included	Not included					
Hypothesis 0	**H_0**	41813.19444	41813.19444							
ln (B10)		314.25987	263.31907							
2 ln (B10)		628.51974	**526.63814**							

Table S10 | Taxa list of sampled sequences for analysis B. An aterisk (*) at species names indicates concatenated 18S and 28S rRNA sequences form different species. For combinations of genes to reconstruct concatenated sequences of chimeran taxa see table S11. Two aterisks (**) indicate contributed sequences in the present study from colleagues (author names of sequences are given). Own sequences are marked in bold green.

Order	Taxon	Accession number 28S rRNA	Length (bp) 28S rRNA	Accession number 18S rRNA	Length (bp) 18S rRNA
Arachnida	*Amblyomma americanum*	AF291874	4005	AF291874	1815
	Dermacentor sp. *	AY859582	3920	L76340	1784
	Chalocheiridius cf. *termitophilus*	AY859558	3394	AY859559	1773
	Pandinus imperator	AY210830	3777	AY210831	1762
	Siro rubens	AY859602	3762	U36998	1809
	Eremobates sp.	AY859572	3833	AY859573	1767
	Aphonopelma hentzi *	AY210803	3819	DQ639776	1750
	Misumenops asperatus	AY210461	3467	AY210445	1786
	Mastigoproctus giganteus	AY859587	3796	AF005446	1790
	Paraphrynus sp.	AY859594	3785	AF005445	1777
Xiphosura	*Limulus polyphemus*	AF212167	3772	L81949	1807
Pycnogonida	*Callipallene* sp.	AY210807	3900	AF005439	1817
	Colossendeis sp.	**EU420133** ** (v. Reumont)	3864	**EU420135** ** (v. Reumont)	1798
	Anoplodactylus portus	AY859550	3893	AY859551	1809
	Nymphon stroemii	**EU420134** ** (v. Reumont)	3818	**EU420136** ** (v. Reumont)	1825
Anostraca	*Artemia* sp. *	AY210805	3628	AJ238061	1809
Notostraca	*Triops cancriformis*	**EU370435** ** (v. Reumont)	3420	**EU370422** ** (v. Reumont)	1784
	Triops longicaudatus	AY157606	3458	AF144219	1809
Diplostraca	*Daphnia* cf. *magna*	**EU370436** ** (v. Reumont)	3823	**EU370423** ** (v. Reumont)	2291
	Bosmina sp. *	**EU370437** ** (v. Reumont)	3332	Z22731	1875
	Eulimnadia texana	AY859574	3665	AF144211	1813
Ostracoda	*Heterocypris incongruens*	**EU370438** ** (v. Reumont)	3279	**EU370424** ** (v. Reumont)	1786
	Pontocypris mytiloides	**EU370439** ** (v. Reumont)	3672	**EU370425** ** (v. Reumont)	1897
Cirripedia	*Semibalanus balanoides*	**EU370440** ** (v. Reumont)	3274	**EU370426** ** (v. Reumont)	1847
	Megabalanus californicus	AY859588	3720	AY520632	1812
	Pollicipes pollicipes	**EU370441** ** (v. Reumont)	3549	**EU370427** ** (v. Reumont)	1852
Branchiura	*Argulus foliaceus*	**EU370442** ** (v. Reumont)	3512	**EU370428** ** (v. Reumont)	1851
Mystacocarida	*Derocheilocaris typicus*	**EU370443** ** (v. Reumont)	3663	**EU370429** ** (v. Reumont)	2171
Copepoda	Cyclopidae sp. *	AY210813	3536	AJ746334	1808
	Chondracanthus lophii	DQ180341	3465	L34046	1810
	Tigriopus cf. *fulvus*	**EU370444** ** (v. Reumont)	3532	EU370430 ** (v. Reumont)	1792
	Canuella perplexa	**EU370445** ** (v. Reumont)	3462	**EU370432** ** (v. Reumont)	1573
	Lepeophtheirus salmonis	DQ180342	3692	AF208263	1799
Remipedia	*Speleonectes tulumensis*	**EU370446** ** (v. Reumont)	3797	**EU370431** ** (v. Reumont) / L81936	1302 / 1965
Cephalocarida	*Hutchinsoniella macracantha*	EF189645	2480	L81935	2018
Leptostraca	*Nebalia* sp.	**EU370447** **	3519	**EU370433** **	1789

		(v. Reumont)		(v. Reumont)	
Anaspidacea	*Anaspides tasmaniae*	AY859549	3997	L81948	1827
Mysidacea	*Heteromysis* sp.	AY859578	3400	AY743946	1724
Decapoda	*Homarus americanus*	AY859581	4351	AY743945	1758
	Penaeus vannamei *	AF124597	5820	DQ079766	1781
Stomatopoda	*Squilla empusa*	AY210842	3913	L81946	1817
Pentastomida	*Raillietiella* sp. *	**EU370448** ** (v. Reumont) AY744894	1286 / 1983	**EU370434** ** (v. Reumont)	1814
Chilopoda	*Craterostigmus tasmanianus*	EU376009 ** (Bartel)	4024	EU368617 ** (Meusemann)	1786
	Otostigmus politus	DQ666180	4170	DQ666177	1868
	Scolopendra mutilans	DQ666181	4174	DQ666178	1848
	Scutigera coleoptrata	AY859601	4024	AF000772	1865
	Lithobius forficatus	EF199984	3913	EU368618 ** (Meusemann)	1752
Diplopoda	*Polyxenus lagurus*	EU376011 ** (Bartel)	3967	EU368619 ** (Meusemann)	1733
	Monographis sp.	EF192437 ** (Bartel / Luan)	3866	AY596371	1744
	Paradoxosomatidae sp.	DQ666182	4288	DQ666179	1797
	Polydesmus complanatus	EU376010 ** (Bartel)	4271	EU368620 ** (Meusemann)	1689
	Cherokia georgiana	AY859562	4225	AY859563	1781
	Orthoporus sp.	AY210828	4124	AY210829	1791
	Cylindroiulus caeruleocinctus	EF199985	4084	EU368621 ** (Meusemann)	1753
Pauropoda	*Allopauropus* sp.	DQ666185	4406	DQ399857	2227
	Pauropodidae sp.	EU376012 ** (Bartel)	4238	EU368622 ** (Meusemann)	2250
Symphyla	*Scutigerella* sp.	DQ666184	4471	DQ399856	1902
	Hanseniella sp.	AY210821-22	4539	AY210823	1925
	Symphylella sp.	DQ666183	4558	DQ399855	2057
Protura	*Acerentomon franzi*	EF199976	4099	EU368597 ** (Meusemann)	1790
	Baculentulus densus *	EU376049	4100	AY037169	1984
	Eosentomon sp.	EU376047 ** (Dell'Ampio)	3654	EU368598 ** (Meusemann)	1860
	Eosentomon sakura	EF192434 ** (Dell'Ampio / Luan)	3789	AY596355	1948
	Sinentomon erythranum	EF192442 ** (Dell'Ampio / Luan)	4043	AY596358	1934
Diplura	Campodeidae sp.	AY859560	3718	AY859561	1866
	Campodea augens	EF199977	4010	EU368599 ** (Meusemann)	1788
	Lepidocampa weberi	EU376050	4061	AY037167	1878
	Catajapyx aquilonaris	EF199978	5016	EU368600 ** (Meusemann)	2154
	Parajapyx emeryanus	EF192440 ** (Dell'Ampio / Luan)	4143	AY037168	2120
	Octostigma sinensis	EF192439 ** (Dell'Ampio / Luan)	4001	AY145134	2138
Collembola	*Tetrodontophora bielanensis*	EU376051	3868	AY555519	1760
	Gomphiocephalus hodgsoni	EF199969	3893	EU368601 ** (Meusemann)	1746
	Triacanthella sp.	AY859609	3823	AY859610	1758
	Bilobella aurantiaca	AJ251729	3934	EU368602 **	1759

				(Meusemann)	
	Anurida maritima	AJ251738	3965	EU368603 ** (Meusemann)	1680
	Podura aquatica	EF199970	3899	EU368604 ** (Meusemann)	1696
	Cryptopygus antarcticus	EF199971	3862	EU368605 ** (Meusemann)	1724
	Isotoma viridis	EU376052	3866	AY596361	1748
	Orchesella villosa	EF199972	3867	EU368606 ** (Meusemann)	1739
	Pogonognathellus flavescens	EU376053	3874	EU368607 ** (Meusemann)	1688
	Megalothorax minimus	EF199975	3868	EU368608 ** (Meusemann)	1703
	Sminthurus viridis	EF199973	3912	EU368609 ** (Meusemann)	1695
	Allacma fusca	EU376054	3877	EU368610 ** (Meusemann)	1759
	Dicyrtomina saundersi	EF199974	3871	EU368611 ** (Meusemann)	1739
Archaeognatha	*Machilis hrabei*	EF199981	3750	EU368612 ** (Meusemann)	1703
	Lepismachilis y-signata	EF199980	3826	EU368613 ** (Meusemann)	1679
	Pedetontus okajimae	EU376055	3800	EU368614 ** (Meusemann)	1742
Zygentoma	*Lepisma saccharina*	EU376048 ** (Dell'Ampio)	3506	EU368615 ** (Meusemann)	1703
	Ctenolepisma longicaudata	AY210810	3907	EU368616 ** (Meusemann)	1744
Odonata	*Brachytron pratense*	EU424323 ** (Letsch)	3738	AF461232	1737
	Aeshna juncea	EU424324 ** (Letsch)	3736	AF461231	1767
	Oxygastra curtisi	EU424325 ** (Letsch)	3736	DQ008194	1787
	Cordulia aenea	EU424326 ** (Letsch)	3795	AF461236	1768
	Somatochlora flavomaculata	EU424327 ** (Letsch)	3795	AF461242	1757
	Epiophlebia superstes	EU424328 ** (Letsch)	3736	AF461247	1835
	Progomphus obscurus	EU424329 ** (Letsch)	3756	AY749909	1843
	Sympetrum danae	EU424330 ** (Letsch)	3756	AF461243	1754
	Leucorrhinia sp.	AY859583	4114	AY859584	1815
	Lestes viridis	EU424331 ** (Letsch)	3747	AJ421949	1867
Ephemeroptera	*Callibaetis ferrugineus*	AY859557	3887	AF370791	1812
	Epeorus sylvicola *	EU414715 ** (Simon)	3680	AY749837	1808
	Siphlonura aestivalis *	EU414716 ** (Simon)	4151	DQ008181	1784
Phasmatodea	*Carausius morosus*	EU426878 ** (Simon)	3737	X89488	1899
	Bacillus rossius	EU426879 ** (Simon)	3889	AY121180	1891
Mantophasmatodea	*Mantophasma zephyra* *	EU414719 ** (Simon)	3383	DQ874153	2018
	Tyrannophasma gladiator	EU426875 ** (Simon)	3878	AY521863	2074
Mantodea	*Mantis religiosa*	AY859585	3990	AY491153	1734

	Hierodula membranacea *	EU414720 ** (Simon)	3603	AY491194	1734
Blattaria	*Gromphadorhina laevigata*	AY210819	4015	AY210820	1877
	Ectobius lapponicus	EU426877 ** (Simon)	4006	DQ874125	1808
	Blattella germanica	AF005243	3931	AF005243	1964
Isoptera	*Zootermopsis angusticollis*	AY859614	4183	AY859615	1873
Dermaptera	*Forficula auricularia*	EU426876 ** (Simon)	4016	Z97594	1873
Plecoptera	*Isoperla* sp. *	EU414717 ** (Simon)	4299	AF461256	2054
	Nemoura flexuosa *	EU414718 ** (Simon)	3256	AF461257	1763
Hemiptera	*Pyrrhocoris apterus* *	EU414725 ** (Simon)	3389	AY627318	1829
	Rhaphigaster nebulosa	EU426880 ** (Simon)	3983	X89495	1924
	Harpocera thoracica *	EU414726 ** (Simon)	3405	AY252388	1895
	Cercopis vulnerata *	EU414724 ** (Simon)	3615	AY744798	1856
	Clastoptera obtusa	AF304569	3201	AY744784	1859
	Pectinariophyes reticulata	AF304570	3259	AY744778	1848
Orthoptera	Gomphocerinae sp.	AY859546	4187	AY859547	1864
	Anacridium aegypticum *	EU414723 ** (Simon)	3819	AY379759	1833
	Acheta domesticus	AY859544	4092	X95741	1802
	Leptophyes punctatissima *	EU414721 ** (Simon)	3918	AY521867	1897
	Pholidoptera griseoaptera *	EU414722 ** (Simon)	3950	Z97587	1884
Hymenoptera	*Myrmecia croslandi*	AB052895	3460	AB121786	1766
	Vespula pensylvanica	AY859612	3912	AY859613	1871
	Nomada sp. *	EU414727 ** (Simon)	3386	AY703484	1854
	Scolia sp. *	EU414728 ** (Simon)	3405	EF012932	1851
	Tenthredinidae sp. *	EU414729 ** (Simon)	3472	AF423781	1836
Coleoptera	*Tenebrio sp.* *	AY210843	4459	X07801	2083
	Silpha obscura	EU426881 ** (Simon)	2783	AJ810737	1930
Siphonaptera	*Ctenocephalides felis* *	EU414732 ** (Simon)	3333	AF423914	1878
Mecoptera	*Merope tuber*	DQ202351	3736	AF286287	1886
	Boreus hyemalis	EU426882 ** (Simon)	3534	AF423882	1881
Lepidoptera	*Pieris napi* *	EU414731 ** (Simon)	3743	AF423785	1856
Trichoptera	*Oxyethira rossi* *	DQ202352	3869	AF423801	1848
	Triaenodes sp. *	EU414730 ** (Simon)	3095	AF286300	1897
Diptera	*Acricotopus lucens*	AJ586562	3910	AJ586561	1939
	Chironomus tentans	X99212	3973	X99212	1528
	Anopheles albimanus	L78065	4022	L78065	1977
	Aedes albopictus	L22060	4102	X57172	1950
	Drosophila melanogaster	M21017	3900	M21017	1995
	Simulium sanctipauli	AF403805	3733	AF403800	1912
Onyphora	*Peripatus* sp.	AY210836	3297	AY210837	2476
	Peripatoides novaezealandiae	AF342793	4570	AF342794	2064
Tardigrada	*Milnesium* sp. *	AY210826	3579	U49909	1844

Table S11 | List of chimeran species reconstructed for concatenated 18S and 28S rRNA sequences. Crustaceans are marked in bold green. Given subgroups have not necessarily the same hierarchical level.

in Table S 10 listed as:	Species 28S rRNA	Species 18S rRNA	subgroup*
Dermacentor sp. *	*Dermacentor* sp.	*Dermacentor andersoni*	Ixodoidea
Aphonopelma hentzi *	*Aphonopelma hentzi*	*Aphonopelma reversum*	Mygalomorphae
***Artemia* sp. ***	***Artemia* sp.**	***Artemia franciscana***	**Anostraca**
***Bosmina* sp. ***	***Bosmina* sp.**	***Bosmina longirostris***	**Cladocera**
Cyclopidae sp. *	**Cyclopidae sp.**	***Macrocyclops albidus***	**Cyclopoida**
***Penaeus vannamei* ***	***Penaeus vannamei***	***Penaeus semisulsatus***	**Dendrobranchiata**
***Raillietiella* sp. ***	***Raillietiella* sp.**	***Raillietiella* sp.**	**Pentastomida**
Baculentulus densus *	*Baculentulus densus*	*Baculentulus tianmushanensis*	Acerentomata
Epeorus sylvicola *	*Epeorus sylvicola*	*Epeorus longimanus*	Setisura
Siphlonura aestivalis *	*Siphlonura aestivalis*	*Siphlonura croaticus*	Pisciforma
Mantophasma zephyra *	*Mantophasma zephyra*	*Mantophasma* cf. *zephyra*	Mantophasmatodea
Hierodula membranacea *	*Hierodula membranacea*	*Hierodula schultzei*	Mantodea
Isoperla sp. *	*Isoperla sp.*	*Isoperla obscura*	Perloidea
Nemoura flexuosa *	*Nemoura flexuosa*	*Nemoura cinerea*	Nemouroidea
Pyrrhocoris apterus *	*Pyrrhocoris apterus*	*Dysdercus poecilus*	Heteroptera
Harpocera thoracica *	*Harpocera thoracica*	*Polymerus castilleja*	Heteroptera
Cercopis vulnerata *	*Cercopis vulnerata*	*Mahanarva costaricensis*	Cercopoidea
Anacridium aegypticum *	*Anacridium aegypticum*	*Acrida cinerea*	Caelifera
Leptophyes punctatissima *	*Leptophyes punctatissima*	*Microcentrum rhombifolium*	Ensifera
Pholidoptera griseoaptera *	*Pholidoptera griseoaptera*	*Tettigonia viridissima*	Ensifera
Nomada sp. *	*Nomada* sp.	*Apis mellifera*	Aculeata
Scolia sp. *	*Scolia* sp.	*Scolia verticalis*	Aculeata
Tenthredinidae sp. *	Tenthredinidae sp.	*Dolerus* sp.	Tenthredinoidea
Tenebrio sp. *	*Tenebrio* sp.	*Tenebrio molitor*	Polyphaga
Ctenocephalides felis *	*Ctenocephalides felis*	*Ctenocephalides canis*	Pulicomorpha
Pieris napi *	*Pieris napi*	*Anthocharis sara*	Glossata
Oxyethira rossi *	*Oxyethira rossi*	*Oxyethira dualis*	Spicipalpia
Triaenodes sp. *	*Triaenodes* sp.	*Oecetis avara*	Integripalpia
Milnesium sp. *	*Milnesium* sp.	*Milnesium tardigradum*	Apochela

Table S12 | Setting of exchangeability parameters used for the pre-runs. See GOWRI-SHANKAR & RATTRAY (2007).

Parameters	time-homogeneous preruns (500,000 and 3,000,000 generations)
Model	MIXED
Tree, proposal priority	1
Model, proposal priority	5
Topology changes, proposal priority	10
Branch lengths, proposal priority	40

Model 1, 3, proposal priority	7
Model 2, 4 proposal priority	8
Average rates, proposal priority	1
Frequencies, proposal priority	2
Rate ratios, proposal priority	1
Gamma parameter, proposal priority	1
random seed	new seed set for each run

Table S 13 | Taxa included in analyses. Species are written in capitals represent used proteome data. [1] Taxa included in the optimal data subset were selected by reduction heuristics. [2] represents taxa used to train Hidden Markov Models (HMMs) to predict putative orthologous gene loci. Sources were: dbEST = http://www.ncbi.nlm.nih.gov/dbEST, Gene Index Project (gendix) = http://compbi.dfci.harvard.edu/tgi/cgi-bin/magic/r1.pl, NCBI Trace Archive =ftp://ftp.ncbi.nih.gov/pub/TraceDB, JGI = http://www.jgi.doe.gov, InParanoidx v6.1 = http://inparanoid6.sbc.su.se, VectorBase = http://www.vectorbase.org, BeetleBase = http://beetlebase.org/, SilkDB = http://silkworm.genomics.org.cn/, UniProt (integr8) = http://www.ebi.ac.uk/integr8/, UCSC = http://hgdownload.cse.ucsc.edu; Data of *C. pipiens quinquefasciatus* was kindly provided by the Broad Institute of MIT and Harvard (USA).
No. of EST contigs = number of assembled EST contigs. No. of genes orig. data set = number of orthologous genes per taxon in the original data set. No. of genes data subset = number of orthologous genes per taxon in the optimal data subset after performing reduction heuristics. Crustacean taxa are colored in green, onw EST species in bold green.

species	**group**	**source**	**no. of EST contigs**	**no. of genes unreduced dataset**	**no. of genes data subset**
Hypsibius dujardini[1]	Tardigrada	dbEST	2386	140	81
Richtersius coronifer[1]	Tardigrada	NCBI Trace Archive	1537	99	52
Epiperipatus sp. TB-2001	Onychophora	dbEST	825	49	
Peripatopsis sedgwicki[1]	Onychophora	Burmester	3452	142	72
Euperipatoides kanangrensis[1]	Onychophora	NCBI Trace Archive	1449	110	53
Julida sp. APV-2005	Myriapoda	dbEST	231	13	
Archispirostreptus gigas[1]	Myriapoda	Burmester	2299	117	58
Scutigera coleoptrata[1]	Myriapoda	NCBI Trace Archive	807	54	35
Anoplodactylus eroticus[1]	Chelicerata	NCBI Trace Archive	1281	91	55
Endeis spinosa[1]	Chelicerata	Burmester	2672	174	69
Limulus polyphemus[1]	Chelicerata	Burmester	4050	210	89
Carcinoscorpius rotundicauda	Chelicerata	dbEST	512	21	
Mesobuthus gibbosus	Chelicerata	dbEST	587	38	
Loxosceles laeta	Chelicerata	dbEST	1209	66	
Dysdera erythrina	Chelicerata	dbEST	279	22	
Cupiennius salei	Chelicerata	dbEST	208	30	
Araneus ventricosus	Chelicerata	dbEST	204	11	
Acanthoscurria gomesiana[1]	Chelicerata	dbEST	3713	234	90
Chilobrachys jingzhao	Chelicerata	dbEST	230	22	

Ixodes scapularis[1]	Chelicerata	genidx	38275	578	128
Ixodes ricinus	Chelicerata	dbEST	1300	53	
Amblyomma variegatum[1]	Chelicerata	genidx	2109	162	62
Amblyomma americanum[1]	Chelicerata	dbEST	2798	88	44
Amblyomma cajennense	Chelicerata	dbEST	1165	71	
Dermacentor andersoni[1]	Chelicerata	dbEST	752	63	38
Dermacentor variabilis	Chelicerata	dbEST	1075	49	
Boophilus microplus[1]	Chelicerata	dbEST	14507	425	112
Rhipicephalus appendiculatus[1]	Chelicerata	genidx	7359	321	92
Argas monolakensis	Chelicerata	dbEST	1620	51	
Ornithodoros porcinus porcinus	Chelicerata	dbEST	771	29	
Ornithodoros parkeri	Chelicerata	dbEST	689	37	
Ornithodoros coriaceus	Chelicerata	dbEST	702	19	
Glycyphagus domesticus[1]	Chelicerata	dbEST	2511	97	56
Blomia tropicalis[1]	Chelicerata	dbEST	1331	80	37
Psoroptes ovis	Chelicerata	dbEST	281	18	
Sarcoptes scabiei	Chelicerata	dbEST	817	38	
Dermatophagoides pteronyssinus	Chelicerata	dbEST	1258	67	
Dermatophagoides farinae	Chelicerata	dbEST	1046	59	
Suidasia medanensis[1]	Chelicerata	dbEST	2083	139	73
Tyrophagus putrescentiae	Chelicerata	dbEST	881	46	
Acarus siro	Chelicerata	dbEST	652	57	
Aleuroglyphus ovatus	Chelicerata	dbEST	1440	58	
Gammarus pulex[1]	Crustacea	dbEST	4241	102	63
Eurydice pulchra	Crustacea	dbEST	562	26	
Euphausia superba	Crustacea	dbEST	1101	43	
Homarus americanus[1]	Crustacea	dbEST	14147	383	111
Pacifastacus leniusculus	Crustacea	dbEST	175	14	
Petrolisthes cinctipes[1]	Crustacea	dbEST	27086	416	119
Callinectes sapidus[1]	Crustacea	dbEST	2239	114	56
Carcinus maenas[1]	Crustacea	dbEST	4567	233	76
Cancer magister	Crustacea	dbEST	445	14	
Celuca pugilator	Crustacea	dbEST	1482	64	
Gecarcoidea natalis	Crustacea	dbEST	656	23	
Ilyoplax pusilla	Crustacea	dbEST	251	2	
Eriocheir sinensis	Crustacea	dbEST	1136	58	
Marsupenaeus japonicus[1]	Crustacea	dbEST	1944	61	46
Fenneropenaeus chinensis[1]	Crustacea	dbEST	3458	114	74
Penaeus monodon[1]	Crustacea	dbEST	4097	129	81
Litopenaeus vannamei[1]	Crustacea	dbEST	3774	126	75
Litopenaeus stylirostris	Crustacea	dbEST	314	12	
Litopenaeus setiferus	Crustacea	dbEST	642	50	
***Tigriopus californicus*[1]**	**Crustacea**	**own EST**	2598	65	39
Calanus finmarchicus[1]	Crustacea	dbEST	4906	189	49
Lepeophtheirus salmonis[1]	Crustacea	dbEST	5102	339	98
***Pollicipes pollicipes*[1]**	**Crustacea**	**own EST**	1721	107	59
Artemia franciscana[1]	Crustacea	dbEST	10330	323	116
***Triops cancriformis*[1]**	**Crustacea**	**own EST**	2542	115	54

Daphnia magna[1]	Crustacea	dbEST	5307	207	85
DAPHNIA PULEX[1, 2]	Crustacea	JGI	30939	775	129
Acerentomon franzi[1]	Hexapoda	Meusemann	1995	99	52
Campodea cf. *fragilis*[1]	Hexapoda	Meusemann	6407	150	68
Folsomia candida[1]	Hexapoda	dbEST	5955	143	41
Anurida maritima[1]	Hexapoda	Meusemann	3504	131	53
Onychiurus arcticus[1]	Hexapoda	dbEST	9981	309	89
Lepismachilis y-signata[1]	Hexapoda	Meusemann	2288	123	66
Tricholepisma aurea	Hexapoda	dbEST	344	34	
Ischnura elegans[1]	Hexapoda	Simon	3194	177	66
Baetis sp.[1]	Hexapoda	Simon	3035	144	49
Locusta migratoria[1]	Hexapoda	dbEST	12255	303	107
Allonemobius fasciatus	Hexapoda	dbEST	116	10	
Laupala kohalensis[1]	Hexapoda	dbEST	8371	292	90
Gryllus bimaculatus[1]	Hexapoda	dbEST	3945	238	93
Gryllus pennsylvanicus	Hexapoda	dbEST	338	30	
Gryllus firmus	Hexapoda	dbEST	271	14	
Periplaneta americana[1]	Hexapoda	dbEST	1577	84	58
Blattella germanica[1]	Hexapoda	dbEST	1546	75	38
Diploptera punctata	Hexapoda	dbEST	666	20	
Hodotermopsis sjoestedti[1]	Hexapoda	dbEST	1471	73	46
Reticulitermes flavipes	Hexapoda	dbEST	113	1	
Sphodromantis centralis	Hexapoda	dbEST	120	4	
PEDICULUS HUMANUS[1]	Hexapoda	VectorBase	11198	636	122
Pediculus humanus corporis	Hexapoda	dbEST	472	55	
Pediculus humanus capitis	Hexapoda	dbEST	2868	147	
Homalodisca coagulata[1]	Hexapoda	dbEST	5661	237	96
Graphocephala atropunctata[1]	Hexapoda	dbEST	1827	97	63
Oncometopia nigricans[1]	Hexapoda	dbEST	1772	114	63
Lygus lineolaris	Hexapoda	dbEST	371	21	
Oncopeltus fasciatus	Hexapoda	dbEST	448	11	
Rhodnius prolixus	Hexapoda	dbEST	735	48	
Triatoma infestans	Hexapoda	dbEST	908	39	
Triatoma brasiliensis	Hexapoda	dbEST	1897	33	
Bemisia tabaci[1]	Hexapoda	dbEST	4548	61	40
Aleurothrixus sp. APV-2005	Hexapoda	dbEST	288	18	
Pachypsylla venusta[1]	Hexapoda	dbEST	4631	118	56
Diaphorina citri	Hexapoda	dbEST	2257	66	
Aphis gossypii[1]	Hexapoda	dbEST	3716	210	88
Myzus persicae[1]	Hexapoda	dbEST	9946	447	107
Acyrthosiphon pisum[1]	Hexapoda	dbEST	18253	413	110
Rhopalosiphum padi	Hexapoda	dbEST	335	34	
Toxoptera citricida[1]	Hexapoda	dbEST	2196	143	74
Sogatella furcifera	Hexapoda	dbEST	122	9	
Nilaparvata lugens	Hexapoda	dbEST	167	7	
Maconellicoccus hirsutus[1]	Hexapoda	dbEST	3929	217	85
Nasonia giraulti[1]	Hexapoda	dbEST	6764	277	101
Nasonia vitripennis[1]	Hexapoda	dbEST	2999	160	86

Copidosoma floridanum	Hexapoda	dbEST	216	9	
Lysiphlebus testaceipes[1]	Hexapoda	dbEST	3881	210	84
Microctonus hyperodae	Hexapoda	dbEST	545	22	
Vespula squamosa[1]	Hexapoda	dbEST	1227	70	50
Solenopsis invicta[1]	Hexapoda	dbEST	12252	297	95
Camponotus festinatus	Hexapoda	dbEST	149	8	
Lasius niger	Hexapoda	dbEST	347	3	
Bombus ignitus	Hexapoda	dbEST	213	22	
APIS MELLIFERA[1, 2]	Hexapoda	Inparanoid	13448	775	129
Melipona quadrifasciata	Hexapoda	dbEST	321	2	
Eoxenos laboulbenei	Hexapoda	dbEST	345	32	
Mengenilla chobauti	Hexapoda	dbEST	297	27	
Micromalthus debilis	Hexapoda	dbEST	157	13	
Carabus granulatus	Hexapoda	dbEST	177	16	
Melademа coriacea	Hexapoda	dbEST	328	23	
Cicindela litorea	Hexapoda	dbEST	232	5	
Cicindela campestris	Hexapoda	dbEST	340	24	
Cicindela littoralis	Hexapoda	dbEST	236	12	
Sphaerius sp. APV-2005	Hexapoda	dbEST	396	29	
Eucinetus sp. APV-2005	Hexapoda	dbEST	344	27	
Dascillus cervinus	Hexapoda	dbEST	354	28	
Georissus sp. APV-2005	Hexapoda	dbEST	408	33	
Trox sp. JH-2005	Hexapoda	dbEST	223	9	
Scarabaeus laticollis	Hexapoda	dbEST	328	30	
Julodis onopordi	Hexapoda	dbEST	337	24	
Hister sp. APV-2005	Hexapoda	dbEST	358	35	
Agriotes lineatus	Hexapoda	dbEST	452	22	
Tenebrio molitor	Hexapoda	dbEST	100	3	
TRIBOLIUM CASTANEUM[1, 2]	Hexapoda	BeetleBase	16421	775	129
Mycetophagus quadripustulatus	Hexapoda	dbEST	419	28	
Biphyllus lunatus	Hexapoda	dbEST	260	28	
Hypothenemus hampei	Hexapoda	dbEST	844	64	
Diaprepes abbreviatus[1]	Hexapoda	dbEST	1921	65	42
Curculio glandium	Hexapoda	dbEST	241	25	
Sitophilus zeamais	Hexapoda	dbEST	82	8	
Ips pini	Hexapoda	dbEST	565	58	
Platystomus albinus	Hexapoda	dbEST	145	5	
Diabrotica virgifera virgifera[1]	Hexapoda	dbEST	7871	336	114
Timarcha balearica	Hexapoda	dbEST	272	21	
Leptinotarsa decemlineata[1]	Hexapoda	dbEST	2668	122	56
Callosobruchus maculatus	Hexapoda	dbEST	561	58	
Anoplophora glabripennis	Hexapoda	dbEST	386	31	
Limnephilus flavicornis	Hexapoda	dbEST	117	2	
Hydropsyche sp. T20	Hexapoda	dbEST	203	23	
Plutella xylostella[1]	Hexapoda	dbEST	1048	72	55
Tineola bisselliella	Hexapoda	dbEST	188	7	
Danaus plexippus[1]	Hexapoda	dbEST	9930	470	114
Bicyclus anynana[1]	Hexapoda	dbEST	5575	165	68

Heliconius erato[1]	Hexapoda	dbEST	3327	219	93
Heliconius melpomene[1]	Hexapoda	dbEST	1820	104	64
Papilio dardanus	Hexapoda	dbEST	310	52	
Plodia interpunctella[1]	Hexapoda	dbEST	3808	175	81
Ostrinia nubilalis	Hexapoda	dbEST	489	25	
Epiphyas postvittana[1]	Hexapoda	dbEST	2895	154	88
Choristoneura fumiferana	Hexapoda	dbEST	589	17	
Trichoplusia ni	Hexapoda	dbEST	417	42	
Agrotis segetum	Hexapoda	dbEST	812	58	
Spodoptera litura	Hexapoda	dbEST	61	3	
Spodoptera frugiperda[1]	Hexapoda	dbEST	8362	309	123
Heliothis virescens[1]	Hexapoda	dbEST	1723	167	73
Helicoverpa armigera[1]	Hexapoda	dbEST	692	70	54
Euclidia glyphica	Hexapoda	dbEST	187	16	
Bombyx mandarina	Hexapoda	dbEST	207	12	
BOMBYX MORI[1, 2]	Hexapoda	SilkDB	16329	775	129
Manduca sexta[1]	Hexapoda	dbEST	2197	120	68
Lonomia obliqua	Hexapoda	dbEST	610	58	
Samia cynthia ricini[1]	Hexapoda	dbEST	5721	254	105
Antheraea yamamai	Hexapoda	dbEST	421	27	
Antheraea assama[1]	Hexapoda	dbEST	8927	292	108
Antheraea mylitta[1]	Hexapoda	dbEST	1478	93	58
Panorpa cf. *vulgaris* APV-2005	Hexapoda	dbEST	322	21	
Ctenocephalides felis	Hexapoda	dbEST	1775	82	
Xenopsylla cheopis	Hexapoda	dbEST	283	26	
Culicoides sonorensis[1]	Hexapoda	dbEST	1405	90	62
Chironomus tentans[1]	Hexapoda	dbEST	3445	216	97
ANOPHELES GAMBIAE[1]	Hexapoda	Uniprot (integr8)	12463	726	126
Anopheles aquasalis	Hexapoda	dbEST	121	4	
Anopheles darlingi	Hexapoda	dbEST	461	24	
Anopheles albimanus[1]	Hexapoda	dbEST	3096	94	53
Anopheles anthropophagus	Hexapoda	dbEST	141	5	
Anopheles funestus	Hexapoda	dbEST	1224	59	
AEDES AEGYPTI[1, 2]	Hexapoda	Inparanoid	15419	654	112
Armigeres subalbatus[1]	Hexapoda	NCBI Trace Archive	7770	329	97
CULEX PIPIENS QUINQUEFASCIATUS[1]	Hexapoda	Broad Institute	20306	721	128
Culex pipiens pallens	Hexapoda	dbEST	76	3	
Toxorhynchites amboinensis	Hexapoda	dbEST	199	7	
Lutzomyia longipalpis[1]	Hexapoda	dbEST	19739	478	126
Phlebotomus papatasi[1]	Hexapoda	dbEST	10797	422	125
Rhynchosciara americana[1]	Hexapoda	dbEST	3449	112	66
Mayetiola destructor[1]	Hexapoda	dbEST	1482	81	48
Sitodiplosis mosellana	Hexapoda	dbEST	1100	64	
Orseolia oryzae	Hexapoda	dbEST	976	29	
Glossina morsitans morsitans[1]	Hexapoda	dbEST	12444	512	124
Musca domestica	Hexapoda	dbEST	296	14	

Stomoxys calcitrans	Hexapoda	dbEST	296	31	
Haematobia irritans	Hexapoda	dbEST	196	13	
Haematobia irritans irritans	Hexapoda	dbEST	189	13	
Ceratitis capitata[1]	Hexapoda	dbEST	11132	475	123
Rhagoletis suavis	Hexapoda	dbEST	370	27	
Rhagoletis pomonella	Hexapoda	dbEST	160	7	
Drosophila arizonae[1]	Hexapoda	dbEST	770	88	55
DROSOPHILA ANANASSAE[1]	Hexapoda	USCS	29704	673	113
DROSOPHILA ERECTA[1]	Hexapoda	USCS	17531	673	117
DROSOPHILA MELANOGASTER[1, 2]	Hexapoda	Inparanoid	13854	752	129
Meloidogyne hapla[1]	Nematoda	dbEST	7802	252	92
CAENORHABDITIS ELEGANS[1, 2]	Nematoda	Inparanoid	20084	749	127
CAENORHABDITIS REMANEI[1]	Nematoda	Inparanoid	25595	719	126
CAENORHABDITIS BRIGGSAE[1, 2]	Nematoda	Inparanoid	19334	711	126
Haemonchus contortus[1]	Nematoda	dbEST	5842	262	98
Ascaris suum[1]	Nematoda	dbEST	9165	197	84
Xiphinema index[1]	Nematoda	dbEST	4824	228	89
Trichinella spiralis[1]	Nematoda	dbEST	8843	373	111
CAPITELLA CAPITATA[1, 2]	Annelida	JGI	32415	724	122
HELOBDELLA ROBUSTA[1]	Annelida	JGI	23432	730	126
Lumbricus rubellus[1]	Annelida	dbEST	10386	196	94
LOTTIA GIGANTEA[1, 2]	Mollusca	JGI	23851	672	120
Crassostrea gigas[1]	Mollusca	dbEST	14857	339	102
Argopecten irradians[1]	Mollusca	dbEST	3610	95	59

Table S14 | Current Genome and EST projects of Crustacea. At the moment existing genome-scale and EST sequencing projects of crustaceans and own projects are given. In the EST analysis C integrated new crustacean species are colored in dark green. Ongoing own projects are colored in lighter green.

Nr	Species	major group	group	Scource	ESTs Raw	EST processed	EST contigs
1	*Litopeneaus stylirostris*	Decapoda	Penaeoidea	NCBI	416	416	314
2	*Pacifastacus leniusculus*	Decapoda	Astacidea	NCBI	392	390	175
3	*Homarus americanus*	Decapoda	Astacidea	NCBI	29,558	29,680	28,280
4	*Litopenaeus vannamei*	Decapoda	Penaeoidea	NCBI	155,411	155,411	11,280
5	*Penaeus monodon*	Decapoda	Penaeoidea	NCBI	8,073	7,800	7,928
6	*Marsupenaeus japonicus*	Decapoda	Penaeoidea	NCBI	3156	3,152	1,944
7	*Fenneropaeneus chinensis*	Decapoda	Penaeoidea	NCBI	10446	10,446	3,458
8	*Litopeneaus setiferus*	Decapoda	Penaeoidea	NCBI	1042	1,042	642
9	*Carcinus maenas*	Decapoda	Brachyura	NCBI	15,558	15,558	4,567
10	*Celuca pugilator*	Decapoda	Brachyura	NCBI	3,646	3,656	2,345
11	*Callinectes sapidus*	Decapoda	Brachyura	NCBI	10563	10,563	2,239
12	*Eriocheir sinensis*	Decapoda	Brachyura	NCBI	3,153	3,152	1,136
13	*Daphnia pulex*	Branchiopoda	Cladocera	NCBI	152,659	1,548	855

14	*Daphnia magna*	Branchiopoda	Cladocera	NCBI	13,183	13,134	10,078
15	*Artemia franciscana*	Branchiopoda	Anostraca	NCBI	37,607	37,579	10,649
16	*Lepeophtheirus salmonis*	Copepoda	Siphonostomatoida	NCBI	15,018	15,018	5,102
17	*Calanus finmarchicus*	Copepoda	Calanoida	NCBI	10,049	6,822	11,93
18	*Eurydice pulchra*	Perarcarida	Isopoda	NCBI	1026	1026	562
19	*Gammarus pulex*	Perarcarida	Amphipoda	NCBI	12345	12,645	4,241
20	*Cancer magister*	Decapoda	Brachyura	NCBI	1,137	1,137	445
21	*Euphausa superba*	Euphausiacea	Euphausia	NCBI	1,770	1,770	1,101
22	*Gecarcoidea*	Decapoda	Brachyura	NCBI	2,118	2,118	656
23	*Ilyoplax pusilla*	Decapoda	Brachyura	NCBI	438	438	251
24	*Petrolisthes cinctipes*	Decapoda	Anomura	NCBI	97,806	97,806	27,086
25	***Triops cancriformis***	Branchiopoda	Notostraca	own ESTs		3,981	2,542
26	***Pollicipes pollicipes***	Cirripedia	Thecostraca	own ESTs		4,224	1,721
27	***Tigriopus californicus***	Copepoda	Harpacticoida	own ESTs		5,024	2,816
28	*Speleonectes tulumensis*	Remipedia	Speleonectidae	own ESTs	1,384	1,282	in progress
28	*Speleonectes tulumensis*	Remipedia	Speleonectidae	own "454"	400,000		in progress
29	*Sarsinebalia urgorrii*	Hoplocarida	Leptostraca	own "454"	400,000		in progress
30	*Ostracoda sp.*	Ostracoda	to determine	own "454"	400,000		in progress

Table S15 | Genes selected by HaMStR and used in phylogenetic analyses. Gene ID = numerical internal identifier that corresponds to the partition number (gene number) of the data matrix. Protein ID = FlyBase-ID from Ensembl Archive February 2007 (Ensembl Arch. 02/07) for *Drosophila melanogaster*, http://feb2007.archive. ensembl.org/ respectively AEE-ID from Inparanoidxx v6.1 for *Aedes aegypti*, http://inparanoid6. sbc.su.se. Gene / Description = Description of genes as determined from the Ensemble Archive / Flybase for *D. melanogaster* (Dmel), from InParanoid v6.1 for *A. aegypti* (Aaeg) or from *HomoloGene* for *Homo sapiens (Hsap),* http://www.ncbi.nlm.nih.gov/homologene. Other studies = genes shared with other studies: ph= Philippe *et al.*x; de= Delsuc *et al.*x; du= Dunn *et al.*x; ba= Baurain *et al.*x; name is assigned to the genes in previous studies are given in squared brackets. Rib. Protein = gene is characterised as ribosomal protein (x). Pot. rel. info. content = potential relative information content calculated by new reduction heuristics (MARE). No. of taxa in data set = amount of taxa in the original data set. present in data subset = indicates the presence of that gene in the optimal data set safter performing matrix reduction. No. of taxa in data subset = amount of taxa in the optimal data subset.

Gene ID	**Protein ID**	**Gene / Description** Ensemble Arch. 02/07 / InParanoid v6.1	**Gene / Description** InParanoid v6.1 (Aaeg) / *HomoloGene (Hsap)*	**Other studies [abbrivated]**	**Rib. Protein**	**Pot. rel. info. content**	**No. of taxa in data set**	**present in data subset**	**No. of taxa in data subset**
12061	FBpp0078222	ADP-ribosylation factor 1				0.92	107	x	88
11924	FBpp0081153	Tubulin alpha-1 chain				0.92	149	x	102
11899	FBpp0078664	26S proteasome non-ATPase regulatory subunit 14				0.91	70	x	61
11735	FBpp0076890	26S protease regulatory subunit 8		ph, de, ba [nsf1-G]		0.90	81	x	76
11806	FBpp0081524	Beta-2 tubulin				0.90	127	x	96
11491	FBpp0083502	AP-2	clathrin coat assembly protein ap17			0.90	51	x	46

11394	FBpp0073292	Rpt3	26S protease regulatory subunit 6b*Proteasome 26S subunit ATPase 4*	de, ba [nsf1-L]		0.89	70	x	66
12024	FBpp0083645	AP-50, isoform A	clathrin coat associated protein ap-50			0.89	56	x	54
11846	FBpp0082140	Vacuolar ATP synthase subunit B		ph, de, ba [vatb]		0.89	74	x	69
11362	FBpp0084434	Histone H2A				0.88	80	x	70
11958	FBpp0078984	smt3				0.87	74	x	62
11637	FBpp0086701	40S ribosomal protein S23		ph, de, ba [rps23]	x	0.86	142	x	95
11460	FBpp0083906	26S protease regulatory subunit 4		ph, de, ba [nsf1-M]		0.86	74	x	69
11547	FBpp0085265	Elongation factor 2		ph, de, ba [ef2-EF2]		0.86	71	x	65
12071	FBpp0088250	ATP synthase beta chain, mitochondrial precursor				0.86	119	x	95
11624	FBpp0081592	AP-47	clathrin coat assembly protein ap-1			0.85	58	x	55
11511	FBpp0076145	CG6767-PB, isoform B	ribose-phosphate pyrophosphokinase 1			0.85	61	x	59
11609	FBpp0083843	Tat-binding protein-1	26S protease regulatory subunit 6a	ph, de, ba [nsf1-K]		0.85	78	x	71
11484	FBpp0088174	CG1970-PA	NADH-ubiquinone oxidoreductase fe-s protein 2 (ndufs2)			0.84	77	x	67
11902	FBpp0087084	GTP-binding protein 128up				0.84	58	x	55
11442	FBpp0077792	Splicing factor U2af 38 kDa subunit		de [u2snrnp]		0.83	58	x	53
11552	FBpp0071808	60S ribosomal protein L23		ph, du, de, ba [rpl23a]	x	0.83	127	x	91
11760	FBpp0074520	Cdc42 homolog	rac GTPase			0.82	68	x	60
11645	FBpp0073446	Heat shock 70 kDa protein cognate 3 precursor		ph, de, ba [hsp70-E]		0.82	64	x	58
11868	FBpp0079999	Vacuolar ATP synthase catalytic subunit A isoform 2		ph, de, ba [vata]		0.82	51	x	48
11762	FBpp0078847	CG9140-PA	NADH-ubiquinone oxidoreductase flavoprotein 1 (ndufv1)			0.81	71	x	63
11377	FBpp0082724	SF2	arginine/serine-rich splicing factor			0.80	47	x	46
11759	FBpp0081401	CG8351-PA	chaperonin	ph, de, ba [cct-N]		0.78	64	x	60
11635	FBpp0080639	40S ribosomal protein S26		ph, de, ba [rps26]	x	0.78	124	x	93
11983	FBpp0082535	Tropomyosin-2				0.77	129	x	98
11617	FBpp0079992	CG5525-PA	chaperonin *T-complex protein 1 subunit delta*	ph, de, ba [cct-D]		0.77	73	x	67
11639	FBpp0085586	40S ribosomal protein S18		ph, du, de, ba [rps18]	x	0.77	136	x	97
11366	FBpp0077571	Enolase				0.77	94	x	78
11634	FBpp0083684	T-complex protein 1 subunit alpha	chaperonin	ph, de, ba [cct-A]		0.76	64	x	62
11393	FBpp0075700	Eukaryotic translation initiation factor 2 beta subunit		ph, de, ba [if2b]		0.76	70	x	60
11379	FBpp0071226	CG7033-PB, isoform B	chaperonin	ph, de, ba [cct-B]		0.76	66	x	63
11893	FBpp0072197	26S proteasome non-ATPase regulatory subunit 7				0.76	62	x	58
12097	FBpp0085919	Polyadenylate-binding protein				0.76	60	x	54
11514	FBpp0074180	40S ribosomal protein S5a		ph, ba [rps5]	x	0.75	148	x	108
11750	FBpp0073328	GTP-binding nuclear protein Ran				0.75	88	x	77

11681	FBpp0082459	CG3731-PB, isoform B	mitochondrial processing peptidase beta subunit	du [rpl27]		0.75	88	x	79
11874	FBpp0079187	Guanine nucleotide-binding protein beta subunit-like protein				0.75	133	x	104
11962	FBpp0080495	Vacuolar ATP synthase subunit H				0.75	66	x	61
11965	FBpp0077142	60S ribosomal protein L27a		ph, de, ba [rpl27]	x	0.75	136	x	95
11450	FBpp0086603	Proteasome p44.5 subunit, isoform B				0.75	76	x	70
11917	FBpp0082464	VhaPPA1-1	vacuolar ATP synthase proteolipid subunit			0.74	71	x	66
11411	FBpp0082516	Heat shock 70 kDa protein cognate 4				0.74	109	x	92
11660	FBpp0078024	26S proteasome non-ATPase regulatory subunit 4				0.74	62	x	61
11385	FBpp0088565	Eukaryotic initiation factor 3 p66 subunit				0.74	72	x	66
11642	FBpp0077419	Phosphoglycerate kinase				0.74	79	x	72
11695	FBpp0077741	lesswright, isoform A	ubiquitin-conjugating enzyme E2 i			0.74	65	x	60
11587	FBpp0073626	40S ribosomal protein S15Aa		ph, de, ba [rps22a]	x	0.73	119	x	91
11829	FBpp0071794	ATP synthase alpha chain, mitochondrial precursor				0.73	103	x	92
12012	FBpp0074825	Catalase				0.73	63	x	61
12121	FBpp0086269	Ribosomal protein S15, isoform B		ph, du, de, ba [rps15]	x	0.73	133	x	97
11479	FBpp0081234	Probable small nuclear ribonucleoprotein Sm D2		du [small nuclear ribonucleo-protein polypeptide D2]		0.72	57	x	50
11798	FBpp0085483	Vacuolar ATP synthase 16 kDa proteolipid subunit				0.72	98	x	81
11848	FBpp0086468	Vacuolar ATP synthase subunit D 1				0.72	75	x	62
11627	FBpp0080691	Probable 26S proteasome non-ATPase regulatory subunit 3				0.72	65	x	62
11454	FBpp0073847	Adenosylhomocysteinase		ph, de [Sadhchydrolase-E1]		0.72	94	x	81
12054	FBpp0077637	CG5001-PA	DNA-J/hsp40			0.72	65	x	60
11911	FBpp0078134	60S acidic ribosomal protein P0		ph, de, ba [rpp0]	x	0.72	148	x	108
12019	FBpp0076393	Isocitrate dehydrogenase, isoform F				0.71	79	x	66
11375	FBpp0077716	60S acidic ribosomal protein P1		du, de, ba [rla2-B]	x	0.71	134	x	89
11855	FBpp0082571	Surfeit locus protein 4 homolog				0.71	60	x	57
11583	FBpp0082788	T-complex protein 1 subunit gamma		ph, de, ba [cct-G]		0.71	57	x	55
11563	FBpp0081581	Calreticulin precursor				0.70	100	x	87
11429	FBpp0086381	CG8446-PA	*lipoyltransferase 1*			0.69	57	x	55
12033	FBpp0084585	CG5590-PA	short-chain dehydrogenase			0.69	72	x	65
11437	FBpp0078532	CG9769-PA	eukaryotic translation initiation factor 3f eif3f			0.69	73	x	62
11577	FBpp0082062	Proteasome subunit alpha type 2		ph, de, ba [psma-D]		0.69	64	x	54
11534	FBpp0071451	Proteasome subunit alpha type 4				0.69	76	x	64

11591	FBpp0072968	CG32276-PB, isoform B	*stress-associated endoplasmic reticulum protein family member 2*			0.68	104	x	76
11603	FBpp0077740	Signal peptide protease				0.68	58	x	57
11910	FBpp0072801	60S ribosomal protein L8		ph, du, de, ba [rpl2]	x	0.67	134	x	101
11802	Fbpp0071279	Oligosaccharyltransferase 48kD subunit	Dolichyl-diphosphooligosaccharide protein glycosyltransferase			0.67	68	x	64
11555	FBpp0080395	CaBP1	protein disulfide-isomerase A6 precursor			0.67	72	x	64
12120	FBpp0081780	Arginine methyltransferase 1				0.67	67	x	62
11814	Fbpp0076960	CG1532-PA	lactoylglutathione lyase			0.66	61	x	56
11567	FBpp0086066	Proteasome subunit alpha type 5		ph, de, ba [psma-A]		0.66	66	x	61
11451	FBpp0076152	40S ribosomal protein S9		ph, ba [rps9]	x	0.66	118	x	86
11710	FBpp0080724	Ribosomal protein L30, isoform A		ph, du, de, ba [rpl30]	x	0.66	117	x	88
11580	FBpp0110423	ribosomal protein L5		ph, de, ba [rpl5]	x	0.65	133	x	105
11976	FBpp0085489	Succinate dehydrogenase [ubiquinone] iron-sulfur protein, mitochondrial precursor				0.65	73	x	64
12029	FBpp0082985	CG7998-PA	malate dehydrogenase			0.65	86	x	75
11849	FBpp0072312	60S ribosomal protein L19		ph, du, de, ba [rpl19a]	x	0.65	134	x	93
12047	FBpp0084901	CG7834-PB, isoform B	electron transfer flavoprotein beta-subunit			0.65	73	x	68
11350	FBpp0076859	Uev1A, isoform B	ubiquitin-conjugating enzyme			0.64	62	x	58
11386	FBpp0079640	CG5362-PA	malate dehydrogenase			0.64	90	x	78
12112	FBpp0072250	Inorganic pyrophosphatase				0.63	71	x	59
11428	FBpp0073989	Proteasome subunit alpha type 7-1				0.63	72	x	65
11711	FBpp0075766	60S ribosomal protein L10a-2		ph, de, ba [rpl1]	x	0.63	133	x	103
11499	FBpp0070430	CG8636-PA	eukaryotic translation initiation factor *eukaryotic translation initiation factor 3 subunit 4*	du [eukaryotic trans-lation ini-tiation factor 3, subunit 4 delta]		0.63	81	x	71
11652	FBpp0085889	Eip55E	cystathionine beta-lyase			0.63	72	x	63
11378	FBpp0079472	yippee interacting protein 2				0.63	79	x	71
11919	FBpp0083371	40S ribosomal protein S20		ph, du, de, ba [rps20]	x	0.63	135	x	95
11772	FBpp0087186	walrus, isoform B	electron transport oxidoreductase			0.62	68	x	61
11928	FBpp0070047	60S ribosomal protein L10		ph, de, ba [grc5]	x	0.62	155	x	109
11932	FBpp0070871	Lethal (1), isoform A	citrate synthase			0.62	70	x	65
11380	FBpp0084617	60S ribosomal protein L4		ph, de, ba [rpl4B]	x	0.62	134	x	107
11820	FBpp0076804	Thioredoxin-like				0.61	75	x	69
11707	FBpp0088441	40S ribosomal protein S7		ph [rps7]	x	0.61	134	x	103
11663	FBpp0088522	Ubiquitin conjugating enzyme 10				0.61	69	x	62
11912	FBpp0074599	Clathrin light chain				0.61	79	x	68
12046	FBpp0088505	Annexin-B9				0.61	76	x	64
11992	FBpp0087164	Erp60, isoform B	protein disulfide isomerase			0.61	101	x	87
11754	FBpp0071766	40S ribosomal protein S16		ph, du, de, ba [rps16]	x	0.60	138	x	101
11984	FBpp0100039	Voltage-dependent anion-selective channel				0.60	104	x	84

11773	FBpp0075382	Proteasome 2 subunit		ph, de, ba [psmb-K]		0.60	92	x	81
11847	FBpp0088242	40S ribosomal protein S3a		ph, de, ba [rps1]	x	0.59	139	x	103
11649	FBpp0076848	CG4769-PA	cytochrome C1			0.58	93	x	82
12040	FBpp0084306	Ribosomal protein L27			x	0.58	129	x	87
12035	FBpp0073344	Glutamine synthetase 2, cytoplasmic				0.57	90	x	79
12115	FBpp0087972	cathD	cathepsin d			0.57	96	x	81
12093	FBpp0082645	NADH:ubiquinone reductase 23kD subunit precursor				0.57	70	x	66
12081	FBpp0084762	Elongation factor 1-gamma				0.56	137	x	109
11619	FBpp0086103	60S ribosomal protein L18a		ph, du, de, ba [rpl20]	x	0.56	137	x	97
11869	FBpp0077580	Rieske iron-sulfur protein, isoform B	ubiquinol-cytochrome c reductase iron-sulfur subunit	du [Ubiquinol cytochrome c reductase, Rieske iron-sulfur polypeptide I]		0.56	97	x	80
11391	FBpp0075618	40S ribosomal protein S4		ph, ba [rps4]	x	0.55	144	x	105
11584	FBpp0081488	Proteasome subunit beta type 3		ph, du, de, ba [psmb-I]		0.54	81	x	72
11383	FBpp0086973	Nascent polypeptide-associated complex alpha subunit				0.53	94	x	81
11778	FBpp0087608	60S ribosomal protein L31		ph, de, ba [rpl31]	x	0.53	122	x	89
11844	FBpp0099686	40S ribosomal protein S8		ph, du, ba [rps8]	x	0.53	152	x	104
11618	FBpp0085166	Ribosomal protein L6, isoform B		ph, de, ba [rpl6]	x	0.47	145	x	106
11841	FBpp0110173	hydrogen-transporting ATP synthase, G-subunit, putative				0.46	110	x	78
12122	FBpp0078354	60S ribosomal protein L13A			x	0.45	136	x	98
12123	FBpp0083376	Ribosomal protein S30, isoform B		du [Ubiquitin-like FUBI and ribo-ribosomal protein S30 precursor]	x	0.44	136	x	94
12074	FBpp0072084	CG3195-PA, isoform A	60S ribosomal protein L12	ph, du, de, ba [rpl12b]	x	0.44	136	x	100
11793	FBpp0076602	Ribosomal protein L18		ph, du, de, ba [rpl18]	x	0.42	124	x	95
11417	FBpp0087352	Ras-related protein Rab-3				0.88	34		
11816	Fbpp0079447	Pka-C1: cAMP-dependent protein kinase catalytic subunit	cAMP-dependent protein kinase catalytic subunit			0.87	44		
11387	FBpp0075260	diablo				0.86	36		
12009	FBpp0088695	CG2944-PF, isoform F	*splA/ryanodine receptor domain and SOCS box containing 4*			0.84	30		
11405	FBpp0077302	Protein mothers against dpp				0.83	29		
12073	FBpp0074756	reptin				0.83	46		
11755	FBpp0083248	CG10889-PA	*zinc finger CCCH-type containing 12B*			0.82	21		
11783	FBpp0079615	Transcription initiation factor IIB				0.81	46		
11860	FBpp0070208	SNF1A/AMP-activated protein kinase, isoform B				0.81	35		
11803	FBpp0079634	CG5343-PA	orf protein			0.81	47		
12118	FBpp0099616	cAMP-dependent protein kinase type I regulatory subunit				0.80	52		

11398	FBpp0088599	Potassium voltage-gated channel protein Shaker	voltage-gated potassium channel		0.80	19
11954	FBpp0079565	Putative ATP-dependent RNA helicase me31b			0.80	43
11509	FBpp0087094	Small nuclear ribonucleoprotein SM D3			0.79	43
12087	FBpp0082743	COP9 signalosome complex subunit 5			0.79	46
11865	FBpp0070361	Unc-76, isoform B			0.79	31
11680	FBpp0088583	CG11266-PG, isoform G	splicing factor		0.78	49
11989	FBpp0083135	CG5451-PA	WD-repeat protein		0.78	37
11797	AAEL007662-PA	casein kinase			0.78	46
12084	FBpp0079951	Ef1-like factor			0.78	28
11850	FBpp0099884	UGP, isoform A			0.77	49
11496	FBpp0078469	Katanin 60		de [nsf1-N]	0.77	34
11687	FBpp0084528	CG5934-PA			0.76	36
12070	AAEL005833-PA	cytosolic purine 5-nucleotidase			0.76	36
11956	FBpp0086942	Guanine nucleotide-binding protein G(q) subunit alpha			0.76	36
11616	FBpp0086375	Lissencephaly-1 homolog			0.76	46
11673	FBpp0083973	Syntaxin-1A			0.75	38
11991	FBpp0083588	CG6439-PA	isocitrate dehydrogenase		0.75	56
12060	FBpp0074486	6-phosphofructo-2-kinase, isoform I			0.75	41
11384	FBpp0081448	CG11990-PA	cdc73 domain protein		0.75	30
11542	FBpp0080659	Sterol carrier protein X-related thiolase			0.74	54
11763	FBpp0072052	Guanine nucleotide-binding protein G(s), alpha subunit			0.74	27
11700	FBpp0079629	RluA-1, isoform C			0.74	26
11589	FBpp0083573	Probable ATP-dependent RNA helicase pitchoune			0.74	38
11940	FBpp0085430	CG10465-PA	*potassium channel tetramerisation domain containing 10*		0.74	33
12065	FBpp0110163	CAMP-dependent protein kinase catalytic subunit			0.74	50
12007	FBpp0074691	tricornered			0.74	28
11864	FBpp0073387	DNA-directed RNA polymerase II largest subunit		du [DNA directed RNA polymerase II polypeptide C]	0.74	15
11921	FBpp0085737	Succinate dehydrogenase [ubiquinone] flavoprotein subunit, mitochondrial precursor			0.73	35
11406	FBpp0083112	endophilin A, isoform B			0.73	30
11726	FBpp0088499	Protein ariadne-1			0.73	42
11640	FBpp0071553	CG4279-PA	Sm protein G putative		0.73	42
11564	FBpp0085131	CG31005-PA	trans-prenyltransferase		0.73	32
11508	FBpp0080261	Suppressor of hairless protein			0.73	18
11788	FBpp0110435	synaptosomal associated protein			0.73	38
11672	FBpp0071424	Inosine-5'-monophosphate dehydrogenase			0.73	46
11610	FBpp0070859	Spliceosomal protein on the X			0.73	28

11980	FBpp0085902	GTP-binding-protein			0.72	46
11523	FBpp0078624	CG14641-PA	RNA binding motif protein		0.72	37
11990	FBpp0081290	ADP-ribosylation factor-like protein 8			0.72	49
11768	FBpp0074278	CG6842-PA	skd/vacuolar sorting		0.72	43
11934	FBpp0070250	CG32810-PB	*potassium channel tetramerisation domain containing 5*		0.72	34
11578	FBpp0071600	Rae1			0.72	46
11436	FBpp0084036	atlastin, isoform B			0.72	41
11821	FBpp0070883	Serine/threonine-protein phosphatase PP-V		de [stcproptase2a-c]	0.72	47
11490	FBpp0081483	Aryl hydrocarbon receptor nuclear translocator homolog			0.72	20
11796	FBpp0081704	pontin			0.71	34
11549	FBpp0076078	Ard1, isoform A			0.70	51
12038	FBpp0082507	CG4203-PA	*KIAA0892*		0.70	26
12088	FBpp0079676	Stress-activated protein kinase JNK			0.70	26
11718	FBpp0086599	CG32105-PB	*LIM homeobox transcription factor 1, alpha*		0.70	20
11785	FBpp0080801	Tyrosine-protein phosphatase Lar precursor			0.70	17
11572	FBpp0086790	Elongation factor Tu mitochondrial			0.70	61
12068	FBpp0082129	Malic enzyme, isoform A			0.70	47
11402	FBpp0070794	Males-absent on the first protein			0.70	28
11535	FBpp0073872	CG9281-PC, isoform C	ATP-dependent transporter		0.70	38
11536	FBpp0083954	CG31137-PE, isoform E	carbon catabolite repressor protein		0.70	24
11419	FBpp0081437	CG11963-PA	succinyl-coa synthetase beta chain		0.69	57
11859	FBpp0074609	Soluble NSF attachment protein			0.69	52
11641	FBpp0075372	Echinoderm microtubule-associated protein-like CG13466	WD-repeat protein		0.69	16
11815	Fbpp0078880	Cpr: NADPH-cytochrome P450 reductase	NADPH cytochrome P450		0.68	53
11573	FBpp0073010	Succinyl-CoA ligase [GDP-forming] alpha-chain, mitochondrial precursor		ph, de, ba [suca]	0.68	62
12059	FBpp0078764	CG7236-PA	cdkl1/4		0.68	17
11838	FBpp0083687	26S proteasome non-ATPase regulatory subunit 6			0.68	63
12108	FBpp0086605	CG12858-PA	*major facilitator superfamily domain containing 6*		0.68	20
12082	FBpp0073090	Transcription factor IIE			0.68	36
12002	FBpp0081336	steamer duck, isoform C			0.68	39
11733	FBpp0087535	CG1513-PA	oxysterol binding protein 9		0.68	30
12105	FBpp0083549	CG6560-PA	ADP-ribosylation factor arf		0.67	36
11935	FBpp0087870	Protein peanut			0.67	35
11403	FBpp0077414	Congested-like trachea protein			0.67	45
11418	FBpp0080281	Transcription elongation factor S-II	transcription elongation factor s-ii		0.67	51
11480	FBpp0073003	eIF5B			0.67	23

11988	FBpp0077735	Notchless			0.67	38
12010	FBpp0081216	Transcription initiation factor IIF alpha subunit			0.67	43
11823	FBpp0077996	Rab26			0.67	27
11824	FBpp0076647	UDP-glucose 6-dehydrogenase			0.67	32
11482	FBpp0077720	CG4164-PA	DNA-J/hsp40		0.67	48
11787	FBpp0072621	phosphocholine cytidylyltransferase 1, isoform D			0.67	35
11604	FBpp0072122	Calcium-transporting ATPase sarcoplasmic/endoplasmic reticulum type			0.67	25
11392	FBpp0088988	Glutamate dehydrogenase, mitochondrial precursor			0.67	57
11628	FBpp0075684	Probable small nuclear ribonucleoprotein Sm D1			0.67	54
11805	FBpp0088775	CG33096-PB, isoform B	*family with sequence similarity 108, member C1*		0.67	29
11495	FBpp0077171	CG3714-PB, isoform B	nicotinate phosphoribosyltransferase		0.66	28
11854	FBpp0073119	Protein ROP			0.66	36
11993	FBpp0081350	tex			0.66	41
11808	FBpp0082867	Guanine nucleotide-binding protein-like 3 homolog	GTP-binding protein-invertebrate		0.66	56
11545	FBpp0070808	CG32758-PA	sorting nexin		0.66	26
12003	FBpp0074543	Tao-1, isoform E			0.66	30
11857	FBpp0085619	Proliferating cell nuclear antigen		du [Proliferating cell nuclear antigen]	0.66	58
11502	FBpp0073600	CG1640-PA, isoform A	alanine aminotransferase		0.66	63
11757	FBpp0072672	Spectrin alpha chain			0.66	24
11916	FBpp0078400	Splicing factor 3A subunit 3			0.65	48
11507	FBpp0081840	CG17184-PB, isoform B	*ADP-ribosylation factor interacting protein 2*		0.65	27
11588	FBpp0087472	CG12140-PA	electron transfer flavoprotein-ubiquinone oxidoreductase		0.65	34
11728	FBpp0099695	Dystrobrevin-like, isoform A			0.65	19
11716	FBpp0071992	no extended memory, isoform B			0.65	31
12018	FBpp0083611	Pyruvate kinase			0.65	77
11830	FBpp0087865	Rs1			0.65	29
12076	FBpp0078099	CG7145-PD, isoform D	pyrroline-5-carboxylate dehydrogenase		0.65	57
11556	FBpp0079843	CG14939-PA	*cyclin Y*		0.65	31
11553	FBpp0071285	Puff-specific protein Bx42			0.65	43
11955	FBpp0110314	conserved hypothetical protein			0.65	36
11667	FBpp0077214	CG17593-PA	*coiled-coil domain containing 47*		0.65	49
12069	FBpp0071046	Protein bys			0.65	40
11709	FBpp0074022	CG9911-PA, isoform A	endoplasmic reticulum resident protein (ERp44) putative		0.65	44
11389	FBpp0081988	Putative inner dynein arm light chain	axonemal inner arm dynein light chain		0.65	24

12051	FBpp0072144	Probable eukaryotic translation initiation factor 6		ph, de, ba [if6]	0.65	60
11525	FBpp0086098	eIF3-S9, isoform B			0.64	76
11432	FBpp0079642	CG33303-PA	ribophorin		0.64	68
11712	FBpp0070249	CG14782-PA	*pleckstrin homology domain containing, family F (with FYVE domain) member 2*		0.64	31
11905	FBpp0078433	DNA-directed RNA polymerases I, II, and III 14.4 kDa polypeptide			0.64	47
11598	FBpp0075202	CG5284-PA, isoform A	chloride channel protein 3		0.64	26
11853	FBpp0081617	CG8500-PA	MRAS2 putative		0.64	19
11890	FBpp0086340	mrj, isoform D			0.64	49
11801	FBpp0072419	Tudor-SN	ebna2 binding protein P100		0.64	55
11455	FBpp0082728	belphegor			0.64	34
12057	FBpp0076921	lethal (1) G0269			0.64	25
11574	FBpp0077676	Clipper			0.64	30
11971	FBpp0070873	Transmembrane GTPase Marf			0.64	39
11891	FBpp0081958	CG18347-PA	mitochondrial glutamate carrier protein		0.64	32
11786	FBpp0084191	CG11859-PA	serine/threonine-protein kinase rio2 (rio kinase 2)		0.64	35
11629	FBpp0079617	CHIP			0.63	46
11632	FBpp0078997	nop5			0.63	49
11608	FBpp0078606	ATP-dependent RNA helicase abstrakt			0.63	31
11351	FBpp0070651	cap binding protein 80, isoform A			0.63	28
11975	FBpp0080282	crinkled, isoform A			0.63	17
11349	FBpp0083972	4EHP	eukaryotic translation initiation factor 4e type		0.63	56
11529	FBpp0076244	Probable signal recognition particle 68 kDa protein	srp68		0.63	50
11355	FBpp0081374	belle	DEAD box ATP-dependent RNA helicase		0.63	41
12053	FBpp0072788	CG9018-PB, isoform B	*regulation of nuclear pre-mRNA domain containing 1B*		0.63	38
11368	FBpp0075729	RhoGAP68F			0.63	36
11831	FBpp0078191	CG6838-PB, isoform B	*ADP-ribosylation factor GTPase activating protein 2*		0.63	55
11613	FBpp0086590	CG12797-PA	WD-repeat protein		0.63	42
11799	FBpp0078371	MLF1-adaptor molecule			0.63	30
11883	FBpp0089034	Armadillo segment polarity protein			0.63	21
11880	FBpp0071407	Mannosyl-oligosaccharide alpha-1,2-mannosidase isoform 2			0.63	30
11771	FBpp0078161	Tenascin major			0.63	16
11843	FBpp0078887	CG9523-PA	*FIC domain containing*		0.63	28
11531	FBpp0085140	CG31004-PB, isoform B	*sushi domain containing 2*		0.63	23
11571	AAEL012316-PA	arsenical pump-driving ATPase			0.63	31
12031	AAEL014285-PA	growth hormone inducible transmembrane protein		du [growth hormone inducible transmembrane protein]	0.63	59

11501	FBpp0074151	Probable small nuclear ribonucleoprotein G		du [small nuclear ribonucleoprotein polypeptide G]	0.63	61
11520	FBpp0073134	Fumarylacetoacetase			0.62	52
12042	FBpp0082569	CG6194-PA	*ATG4 autophagy related 4 homolog D*		0.62	29
11647	FBpp0078811	Tetraspanin 26A			0.62	34
11964	FBpp0089047	Voltage-dependent calcium channel type D alpha-1 subunit			0.62	13
11739	FBpp0087699	Receptor mediated endocytosis 8			0.62	17
12049	FBpp0100147	conserved membrane protein at 44E, isoform A			0.62	29
11742	FBpp0070418	CG16903-PA	cyclin l		0.62	38
11822	FBpp0078893	CG9547-PA	acyl-CoA dehydrogenase		0.62	51
11424	FBpp0079870	escl, isoform A			0.62	40
11413	FBpp0086223	Flap endonuclease 1			0.62	39
11705	FBpp0075485	Protein frizzled precursor			0.62	25
11737	FBpp0087722	Dynamitin			0.62	53
11701	AAEL010002-PB	5-formyltetrahydrofolate cyclo-ligase			0.62	53
11939	FBpp0076523	Protein henna			0.62	59
11576	FBpp0070104	Beta-amyloid-like protein precursor			0.62	37
11527	FBpp0084894	CG31033-PB, isoform B	*ATG16 autophagy related 16-like 1*		0.62	18
12102	FBpp0081633	CG9461-PA	F-box only protein		0.62	23
12092	FBpp0088153	Eph receptor tyrosine kinase, isoform D			0.62	18
11631	FBpp0070368	6-phosphogluconate dehydrogenase, decarboxylating			0.62	56
12075	FBpp0084499	CG6051-PA	lateral signaling target protein		0.62	29
11903	FBpp0072723	CG1140-PA, isoform A	succinyl-coa: 3-ketoacid-coenzyme a transferase		0.61	43
11896	FBpp0086289	HMG Coenzyme A synthase, isoform A			0.61	42
11922	FBpp0079976	PICK1, isoform B			0.61	30
11996	FBpp0078360	sec23, isoform B			0.61	24
11625	FBpp0088955	Protein tumorous imaginal discs, mitochondrial precursor			0.61	56
12066	FBpp0070793	CG3016-PA	ubiquitin-specific protease		0.61	23
11913	FBpp0083976	Rox8, isoform F			0.61	39
12062	FBpp0088862	Hypothetical protein CG7816			0.61	36
11416	FBpp0088910	CG1732-PB, isoform B	sodium/chloride dependent neurotransmitter transporter		0.61	30
11530	FBpp0070469	Hypothetical protein CG32795 in chromosome 1			0.61	44
11565	FBpp0077998	CG7338-PA	ribosome biogenesis protein tsr1		0.61	50
11881	FBpp0082624	CG4525-PA	*tetratricopeptide repeat domain 26*		0.61	21
12103	FBpp0084774	CG1458-PA	*CDGSH iron sulfur domain 2*		0.61	68
12014	FBpp0075942	CG7628-PA	phosphate transporter		0.61	29

11607	FBpp0071259	CG12135-PA	*CWC15 spliceosome-associated protein homolog*		0.61	51
11930	FBpp0074330	CG6179-PA	*nitric oxide synthase interacting protein*		0.61	44
12016	FBpp0070751	RhoGAP5A, isoform A			0.61	24
11630	FBpp0079643	CG5366-PA	cullin-associated NEDD8-dissociated protein 1		0.60	20
12013	FBpp0087648	RNA-binding protein 8A			0.60	56
11666	FBpp0072660	Hsp90 co-chaperone Cdc37			0.60	56
11483	FBpp0077886	Lipoic acid synthase, isoform B	lipoic acid synthetase		0.60	48
11390	FBpp0075731	Neurexin-4 precursor			0.60	17
11929	FBpp0074822	Aut1			0.60	40
11396	FBpp0070064	Molybdenum cofactor synthesis protein cinnamon	molybdopterin biosynthesis protein		0.60	29
11719	FBpp0081331	CG10153-PA	*trafficking protein particle complex 5*		0.60	35
11978	FBpp0087366	CG11777-PA	cyclophilin-10		0.60	30
11727	FBpp0071303	CG3004-PA	vegetatible incompatibility protein HET-E-1 putative		0.60	39
11600	FBpp0087340	CG7686-PA	*LTV1 homolog*		0.60	53
12107	FBpp0085500	CG3358-PB, isoform B	*TatD DNase domain containing 1*		0.60	33
11372	FBpp0087938	Nup44A, isoform A			0.60	40
11792	FBpp0071262	CG17446-PA	cpg binding protein		0.60	27
11840	FBpp0078381	CG2185-PA	calcineurin b subunit		0.60	61
11926	AAEL002852-PA	conserved hypothetical protein			0.60	16
11357	FBpp0074246	CG8142-PA	replication factor C 37-kDa subunit putative		0.60	42
12011	FBpp0070162	CG11642-PC, isoform C	translocation associated membrane protein		0.60	66
11447	FBpp0086703	CG8394-PA	amino acid transporter		0.60	21
11512	FBpp0074937	NUCB1			0.60	51
11671	FBpp0071392	CG32687-PA	internalin A putative		0.60	46
11606	FBpp0077047	lethal (1) G0196, isoform E			0.60	21
12094	FBpp0079675	CG5676-PA			0.59	51
11518	FBpp0073557	CG4332-PA	*CLPTM1-like*		0.59	35
11925	FBpp0071138	Probable phenylalanyl-tRNA synthetase alpha chain			0.59	37
12079	FBpp0078891	CG9543-PA	*coatomer protein complex, subunit epsilon*		0.59	58
11654	FBpp0077263	Probable tyrosyl-DNA phosphodiesterase			0.59	33
11407	FBpp0076124	Ubiquitin-conjugating enzyme E2-22 kDa	ubiquitin-conjugating enzyme E2-25kDa		0.59	50
11643	FBpp0071688	Protein ariadne-2			0.59	33
11731	FBpp0085222	lethal (3) s1921			0.59	47
12056	FBpp0081800	Sorbitol dehydrogenase-2			0.59	69
11767	FBpp0086875	F-box/SPRY-domain protein 1			0.59	21
11944	FBpp0071269	CG12121-PA	lung seven transmembrane receptor		0.59	26
11561	FBpp0072481	CG13887-PB, isoform B	B-cell receptor-associated protein bap	du [B-cell receptor-associated protein	0.59	70

				31]			
11997	FBpp0079914	Threonyl-tRNA synthetase, isoform C		ba [trs]		0.59	27
11834	FBpp0077129	CG15433-PA	elongator component putative			0.59	35
11538	FBpp0087714	CG8080-PA	*chromosome 5 open reading frame 33*			0.59	32
11478	FBpp0074792	CG6812-PA	sideroflexin 123			0.59	33
12000	FBpp0073354	CG1749-PA	ubiquitin-activating enzyme E1			0.59	46
11694	FBpp0070933	Serine/threonine-protein kinase				0.59	16
11817	FBpp0110272	multiple C2 domain and transmembrane region protein				0.59	20
11570	FBpp0071229	CG7039-PA	ARL3 putative			0.59	42
11692	FBpp0079812	Replication factor C 38kD subunit				0.59	42
11761	FBpp0088881	supercoiling factor, isoform B				0.59	49
11524	FBpp0086373	Cysteinyl-tRNA synthetase				0.59	38
12110	FBpp0099935	CG11919-PA, isoform A	peroxisome assembly factor-2 (peroxisomal-type ATPase 1)			0.59	26
11884	FBpp0071478	CDK5RAP3-like protein				0.59	44
11427	FBpp0075042	rogdi, isoform A				0.59	29
11960	FBpp0087073	CG8841-PC, isoform C	*chromosome 17 open reading frame 28*			0.59	18
11488	FBpp0079946	Probable ribosome production factor 1	U3 small nucleolar ribonucleoprotein protein imp4			0.59	44
12080	FBpp0082525	CG4338-PA	*chromosome 16 open reading frame 42*			0.59	37
11358	FBpp0078721	thickveins, isoform D				0.59	31
11677	FBpp0071189	CG12125-PA	*family with sequence similarity 73, member B*			0.59	24
12109	FBpp0075866	CG11660-PA, isoform A	serine/threonine-protein kinase rio1 (rio kinase 1)			0.58	33
12050	FBpp0081087	CG2656-PA	*GPN-loop GTPase 3*			0.58	41
11356	FBpp0080407	CG5861-PA	*transmembrane protein 147*			0.58	47
11651	FBpp0075139	CG4933-PA	o-sialoglycoprotein endopeptidase			0.58	29
11752	FBpp0088329	Calcium-dependent secretion activator				0.58	13
11878	FBpp0071669	GlcT-1				0.58	28
11657	FBpp0070443	40S ribosomal protein S12, mitochondrial precursor			x	0.58	38
11871	FBpp0074990	UDP-sugar transporter UST74c				0.58	29
11920	FBpp0074662	Rpn1				0.58	30
11953	FBpp0084464	BM-40-SPARC				0.58	68
11863	FBpp0083098	Mekk1, isoform B				0.58	17
11614	FBpp0099673	Tousled-like kinase, isoform D				0.58	24
11438	FBpp0078382	MTA1-like, isoform B				0.58	17
11381	FBpp0081659	lethal (3) IX-14				0.58	20
11686	FBpp0072142	Protein within the bgcn gene intron				0.57	38
11875	FBpp0084118	CG5805-PA	mitochondrial glutamate carrier putative			0.57	17

11360	FBpp0088059	Dpld (Protein dappled)			0.57	17
12090	FBpp0075734	CG6910-PA	myoinositol oxygenase		0.57	53
11734	AAEL010797-PA	RNA polymerase II holoenzyme component			0.57	31
11477	FBpp0078633	CG3756-PA	DNA-directed RNA polymerase		0.57	44
12048	FBpp0086107	Anaphase-promoting complex subunit 10			0.57	36
11987	FBpp0081601	CG9373-PA	myelinprotein expression factor		0.57	43
11882	FBpp0073588	CG1622-PA	*PRP38 pre-mRNA processing factor 38*		0.57	33
11810	FBpp0086757	CG12295-PB: straightjacket	dihydropyridine-sensitive l-type calcium channel		0.57	16
11668	FBpp0070389	mitochondrial ribosomal protein L14		x	0.57	38
11858	FBpp0086063	Ngp			0.56	34
11795	FBpp0084691	CG1646-PC, isoform C	*PRP39 pre-mRNA processing factor 39 homolog*		0.56	43
11544	FBpp0085838	GDI interacting protein 3, isoform C			0.56	33
11400	FBpp0071061	Integrin beta-PS precursor	integrin beta subunit		0.56	45
11550	FBpp0076134	ATP synthase B chain, mitochondrial precursor			0.56	16
11725	FBpp0076486	pebble, isoform D			0.56	18
11804	FBpp0087806	CG8635-PA	*zinc finger CCCH-type containing 15*		0.56	45
11674	FBpp0074121	CG9099-PA	*density-regulated protein*		0.56	54
11839	FBpp0081520	Probable maleylacetoacetate isomerase 2			0.56	43
11559	FBpp0110208	calnexin			0.56	59
11794	FBpp0084559	rapsynoid			0.56	19
11656	FBpp0075168	Tyrosyl-tRNA synthetase			0.56	45
11489	FBpp0071445	CG9236-PA	calcium and integrin-binding protein 1		0.56	22
12111	FBpp0084144	CG11920-PA	U3 small nucleolar ribonucleoprotein protein imp4		0.56	45
11774	FBpp0085422	O-glycosyltransferase, isoform B			0.56	19
11683	FBpp0086129	Fat-spondin, isoform B			0.56	52
11461	FBpp0081481	Protein neuralized			0.55	29
11835	FBpp0079780	CG6724-PA	WD-repeat protein		0.55	45
11974	FBpp0083581	CG6015-PA	pre-mRNA splicing factor prp17		0.55	35
11354	FBpp0081552	CG8286-PA	tetratricopeptide repeat protein putative		0.55	48
12063	FBpp0078685	Probable GDP-mannose 4,6 dehydratase			0.55	38
12078	FBpp0087709	Mystery 45A			0.55	34
11972	FBpp0072382	mrityu, isoform C			0.55	24
11828	FBpp0082895	CG5840-PB, isoform B	pyrroline-5-carboxylate reductase		0.55	51
11866	FBpp0083076	Probable 28 kDa Golgi SNARE protein			0.55	37
11382	FBpp0082888	Sur-8, isoform A			0.55	33
11704	FBpp0081370	CG8036-PD, isoform D	transketolase I		0.55	41
11487	FBpp0083131	Prp18			0.55	38

11691	FBpp0071063	Glutamate--cysteine ligase				0.55	30
11458	FBpp0074562	CG32528-PA	parvin			0.54	48
11861	FBpp0074026	Katanin 80, isoform B				0.54	20
11586	FBpp0074835	CG6841-PA	pre-mRNA splicing factor			0.54	25
11959	FBpp0080906	La protein homolog				0.54	55
11404	FBpp0080622	CG10333-PA	DEAD box ATP-dependent RNA helicase			0.54	20
11467	FBpp0072979	CG11537-PB, isoform B	*hippocampus abundant transcript 1*			0.54	20
11364	FBpp0075238	PDCD-5	*programmed cell death 5* du, de [pace6]			0.54	53
11370	FBpp0081276	pyd3				0.54	49
11722	FBpp0080048	Coatomer subunit beta'				0.54	21
11425	FBpp0076861	Kinesin-like protein at 64D				0.54	27
11675	FBpp0076332	CG7112-PA	rab6 GTPase activating protein gapcena (rabgap1 protein)			0.54	23
11895	FBpp0081475	CG18005-PA	red protein (ik factor) (cytokine ik)			0.54	30
11827	FBpp0081719	Sirt6				0.54	29
12045	FBpp0079832	CG6509-PB, isoform B	discs large protein			0.54	14
11615	FBpp0070367	CG3835-PA, isoform A	D-lactate dehydrognease 2			0.54	31
11937	FBpp0079577	Ubiquitin thioesterase otubain-like protein				0.54	41
11546	FBpp0072404	mitochondrial ribosomal protein L17		ph, du, de, ba [rpl17]	x	0.54	38
12099	FBpp0085155	Coatomer protein, isoform B				0.54	28
11915	FBpp0073739	MRNA-capping-enzyme				0.54	46
11579	FBpp0077551	CG31938-PA	*exosome component 3*			0.54	31
11776	FBpp0072455	Probable UDP-glucose 4-epimerase				0.54	47
11446	FBpp0084351	CG6095-PB, isoform B	exocyst complex-subunit protein 84kDa-subunit putative			0.54	31
11434	FBpp0074582	CG14232-PA	*acyl-Coenzyme A binding domain containing 3*			0.54	29
11809	FBpp0085181	CG1800-PA: partner of drosha	double-stranded binding protein putative			0.53	23
11521	FBpp0071095	CG10932-PA	acetyl-coa acetyltransferase mitochondrial			0.53	65
12106	FBpp0075693	Probable phosphomannomutase				0.53	47
12098	FBpp0086667	CG8531-PA	*DnaJ (Hsp40) homolog, subfamily C, member 11*			0.53	35
11646	FBpp0075111	COP, isoform B				0.53	62
11426	FBpp0083921	CG5991-PC, isoform C				0.53	37
11764	FBpp0073806	CG14407-PA	glutaredoxin			0.53	68
12067	FBpp0088040	CG11107-PA	ATP-dependent RNA helicase			0.53	22
11528	FBpp0080117	CG16865-PA	*chromosome X open reading frame 56*			0.53	34
11769	FBpp0073649	CG11134-PA	*APAF1 interacting protein*			0.53	48
11590	FBpp0085763	Exostosin-3				0.53	24
12083	FBpp0075947	Multidrug-Resistance like Protein 1, isoform B				0.53	26
11894	FBpp0071256	C12.2				0.53	27
11409	FBpp0074844	CG3961-PB, isoform B	long-chain-fatty-acid coa ligase			0.53	35

11441	FBpp0073635	CG11178-PB, isoform B	*AVL9 homolog*			0.53	25
11456	FBpp0072830	misshapen, isoform E				0.53	31
11444	FBpp0071232	AP-1, isoform E				0.53	17
11423	FBpp0078070	CG9391-PA, isoform A	myo inositol monophosphatase			0.53	52
11526	FBpp0074564	CG12703-PA	peroxisomal membrane protein 70 abcd3			0.52	23
11708	FBpp0081576	eclair				0.52	65
12043	FBpp0074736	CG8798-PA, isoform A	ATP-dependent Lon protease putative			0.52	27
11745	FBpp0090943	CG33505-PA	WD-repeat protein			0.52	36
12096	FBpp0087342	CG12343-PA	*SYF2 homolog, RNA splicing factor*			0.52	51
12026	FBpp0084813	CG1907-PA	*solute carrier family 25 (mitochondrial carrier; oxoglutarate carrier), member 11*			0.52	49
12116	FBpp0078694	mitochondrial ribosomal protein L24			x	0.52	50
12114	FBpp0075560	CG10711-PA	conserved hypothetical protein			0.52	46
11596	FBpp0076073	nudE				0.52	38
11698	FBpp0078992	Gas41				0.52	31
11904	AAEL010402-PA	DEAD box ATP-dependent RNA helicase				0.52	20
11970	FBpp0083436	Exocyst complex component 6				0.52	36
11756	FBpp0086641	Lamin-C				0.52	37
11724	FBpp0081283	CG10903-PA	*Williams Beuren syndrome chromosome region 22*			0.52	42
11721	AAEL004763-PA	conserved hypothetical protein				0.52	26
11592	FBpp0070806	Lethal (1), isoform A				0.52	21
11601	FBpp0081834	CG5214-PA	dihydrolipoamide succinyltransferase component of 2-oxoglutarate dehydrogenase			0.52	45
11431	FBpp0074226	CG5703-PA	NADH-ubiquinone oxidoreductase 24 kDa subunit	du [NADH dehydrogenase (ubiquinone) flavoprotein 2, 24kDa]		0.52	71
11811	FBpp0079495	CG5885-PA	translocon-associated protein gamma subunit			0.52	84
11775	FBpp0074517	Glucose-6-phosphate 1-dehydrogenase				0.52	42
11506	FBpp0082642	CG4225-PA	ABC transporter *Mitochondrial ABC transporter 3*			0.51	17
11826	FBpp0110402	eukaryotic translation initiation factor 3, theta subunit				0.51	24
11422	FBpp0085952	Dgp-1, isoform A				0.51	33
11782	FBpp0073828	CG6227-PA	DEAD box ATP-dependent RNA helicase			0.51	17
11685	FBpp0071217	Polycomb protein l(1)G0020				0.51	30
11723	FBpp0081799	CG6465-PA	aminoacylase putative			0.51	52
11376	FBpp0075938	NEDD8-activating enzyme E1 regulatory subunit	app binding protein			0.51	39

11699	FBpp0075034	CG7728-PA	ribosome biogenesis protein		0.51	29
11789	FBpp0084190	CG11858-PA	peptidyl-prolyl cis/trans isomerase putative	du [protein (peptidyl-prolyl cis/trans isomerase) NIMA-interacting 1]	0.51	40
11463	FBpp0087926	drosha			0.51	15
11867	FBpp0074734	CG8793-PA	*KIAA1012*		0.50	17
11697	AAEL013319-PA	conserved hypothetical protein		ph, de [stbproptase2a-b]	0.50	15
11688	FBpp0075069	CG4169-PA	ubiquinol-cytochrome c reductase complex core protein		0.50	82
11952	FBpp0076782	Regulator of chromosome condensation			0.50	30
11740	FBpp0074366	Histidyl-tRNA synthetase, isoform B			0.50	45
11898	FBpp0087506	6-phosphofructokinase			0.50	30
11892	FBpp0083899	Bifunctional aminoacyl-tRNA synthetase			0.50	24
11473	FBpp0084489	DNA polymerase alpha subunit B			0.50	29
12008	FBpp0078184	Secretory Pathway Calcium atpase, isoform C			0.50	17
11513	FBpp0072531	CG9119-PA	*chromosome 11 open reading frame 54*		0.50	44
11669	FBpp0073875	CG9245-PB, isoform B	phosphatidylinositol synthase		0.50	46
11813	Fbpp0089153	smallminded CG8571-PB, isoform B	peroxisome assembly factor-2 (peroxisomal-type ATPase 1)	de [nsf2-B]	0.50	32
12023	FBpp0083842	3-hydroxy-3-methylglutaryl-coenzyme A reductase			0.50	35
11659	FBpp0087353	CG16728-PA	*G protein-coupled receptor kinase interacting ArfGAP 2*		0.50	25
11886	FBpp0080045	Two A-associated protein of 42kDa			0.50	37
11852	FBpp0084032	CG6643-PB, isoform B	synaptotagmin putative		0.50	33
11982	FBpp0072779	CG1317-PB	ssm4 protein		0.50	25
11457	FBpp0082284	falafel, isoform C			0.50	17
11412	FBpp0070643	CG3564-PA	copii-coated vesicle membrane protein P24		0.50	72
11837	FBpp0074510	CG14211-PB	dual-specificity protein phosphatase putative		0.49	30
11449	FBpp0074285	3-hydroxyacyl-CoA dehydrogenase type-2	hydroxyacyl dehydrogenase		0.49	68
11653	FBpp0085609	Mediator complex subunit 8			0.49	38
11941	FBpp0085630	CG11208-PA	2-hydroxyphytanoyl-coa lyase		0.49	42
11914	FBpp0072495	CG13900-PB, isoform B	spliceosomal protein sap		0.49	23
11889	FBpp0099977	CG1410-PA, isoform A	GTP-binding protein lepa		0.49	30
11539	FBpp0074936	CG5589-PA	DEAD box ATP-dependent RNA helicase		0.49	36
11471	FBpp0080509	Aminopeptidase P			0.49	52
11730	FBpp0086954	Chromatin remodelling complex			0.49	20
11459	FBpp0080319	ATPase chain Iswi lethal (2) 35Df			0.49	21

11909	FBpp0078319	CG2051-PC, isoform C	histone acetyltransferase type b catalytic subunit		0.49	38
11825	FBpp0070181	CG3704-PA	xpa-binding protein 1 (mbdin)		0.49	34
12039	FBpp0085873	Late endosomal/lysosomal Mp1-interacting protein homolog			0.49	51
11936	FBpp0086402	CG8386-PA	*ubiquitin-fold modifier conjugating enzyme 1*		0.49	53
12037	FBpp0080062	Ski6		du, de [rrp46-B]	0.49	37
11918	FBpp0087402	Caf1-105			0.49	29
11636	FBpp0084013	Golgin-84			0.49	25
11500	FBpp0088794	CG33298-PB, isoform B	phospholipid-transporting ATPase 1 (aminophospholipid flippase 1)		0.48	20
11566	FBpp0082288	neither inactivation nor afterpotential B			0.48	26
11784	FBpp0072058	Alpha-catenin-related, isoform B			0.48	21
12077	FBpp0077208	Exocyst complex component 2			0.48	33
11363	FBpp0083319	CG5434-PA (Srp72)	*Signal recognition particle 72 kDa protein*		0.48	50
11465	AAEL000324-PA	tyrosine-protein kinase drl			0.48	21
12017	FBpp0087867	Mlh1			0.48	26
11741	FBpp0079735	Vacuolar protein sorting protein 72 homolog			0.48	31
11621	FBpp0072711	CG12091-PA	protein phosphatase 2c		0.48	48
11408	FBpp0082066	Interleukin enhancer-binding factor 2 homolog	interleukin enhancer binding factor		0.48	54
11720	AAEL002870-PA	Dipeptidyl-peptidase 3			0.48	55
12101	FBpp0076771	CG10467-PA	aldose-1-epimerase		0.48	42
11676	FBpp0075209	Signal sequence receptor		du [signal sequence receptor, beta precursor]	0.48	92
11644	FBpp0075535	Ral guanine nucleotide exchange factor 2, isoform A			0.48	21
11352	FBpp0075513	Hsc70Cb, isoform C			0.48	50
11714	FBpp0089006	CG32626-PD, isoform D	AMP deaminase		0.48	22
11833	FBpp0075151	multiprotein bridging factor, isoform B		du [endothelial-differentiation-related factor 1 isoform alpha]	0.48	74
11515	FBpp0083989	Dis3			0.48	32
11998	FBpp0078512	CG1126-PA	*Bardet-Biedl syndrome 5*		0.48	17
11435	FBpp0086042	CG6401-PA	glycosyltransferase		0.48	28
11790	FBpp0072468	CG6905-PA	cell division control protein		0.48	19
11558	FBpp0071277	Zpr1			0.48	43
11862	FBpp0077203	CG31957-PA	translation initiation factor 1A putative		0.48	34
12085	FBpp0086957	CG8632-PB, isoform B	*solute carrier family 30 (zinc transporter), member 9*		0.48	27
11469	FBpp0083440	Uridine 5'-monophosphate synthase	orotidine-5'-phosphate decarboxylase, putative		0.48	46
11474	FBpp0070180	CG3703-PA	*RUN domain containing 1*		0.47	23
11443	FBpp0079162	CG7429-PA	*coiled-coil domain*		0.47	32

			containing 53			
12022	FBpp0084411	CG5484-PC, isoform C	*Yip1 interacting factor homolog B*		0.47	54
11551	FBpp0074964	CG6259-PA	charged multivesicular body protein 5		0.47	55
11887	FBpp0070637	CG6133-PA	*NOL1/NOP2/Sun domain family, member 2*		0.47	39
12058	FBpp0074616	FRG1 protein homolog			0.47	42
11533	FBpp0087458	CG12214-PA, isoform A	tubulin-specific chaperone e		0.47	29
12104	FBpp0085393	CG7791-PA	mitochondrial intermediate peptidase		0.47	34
11938	FBpp0083768	CG13827-PA	*peroxisomal biogenesis factor 11 gamma*		0.47	27
12004	FBpp0071818	Hypothetical UPF0172 protein CG3501			0.47	38
11439	FBpp0085829	CG15087-PA	*chromosome 11 open reading frame2*		0.47	28
11361	FBpp0082332	CG3061-PA	DNA-J, putative		0.47	52
11779	FBpp0070924	COQ7			0.47	46
11421	FBpp0074715	anti-silencing factor 1			0.47	39
11626	FBpp0076459	CG7550-PA	2-aminoethanethiol (cysteamine) dioxygenase		0.47	26
11650	FBpp0080553	Putative conserved oligomeric Golgi complex component 5			0.47	24
11715	FBpp0086992	CG18177-PB, isoform B			0.47	28
11981	FBpp0077333	CG3542-PB, isoform B	U1 small nuclear ribonucleoprotein putative		0.46	37
11961	FBpp0084418	CG6420-PA	WD-repeat protein		0.46	19
11517	FBpp0073082	CG14997-PB, isoform B	sulfide quinone reductase		0.46	52
11957	FBpp0089113	Transcription elongation factor SPT5			0.46	15
11414	FBpp0079258	CG12375-PA	metallo-beta-lactamase putative		0.46	47
11729	FBpp0079469	CG4537-PA	*cysteine-rich PDZ-binding protein*		0.46	29
11807	FBpp0081556	Spermidine Synthase			0.46	47
11464	FBpp0079697	CG6415-PA	aminomethyltransferase		0.46	39
12025	FBpp0082065	Aos1			0.46	41
11684	FBpp0083351	CG4159-PA	pseudouridylate synthase		0.46	43
12036	FBpp0072421	Enhancer of bithorax, isoform C			0.46	14
11743	FBpp0071597	CG9865-PB, isoform B	*phosphatidylinositol glycan anchor biosynthesis, class M (CG9865)*		0.46	32
11452	FBpp0083214	Vacuolar ATP synthase subunit G			0.46	102
11770	FBpp0085121	39S ribosomal protein L32, mitochondrial precursor		x	0.46	42
11856	FBpp0080638	CG12750-PA	cell cycle control protein cwf22		0.46	28
11433	FBpp0079468	FK506-binding protein 59			0.46	57
11623	FBpp0075393	CG6859-PA	perixosomal biogenesis factor		0.46	36
11493	FBpp0085258	CG1416-PC, isoform C	*AHA1, activator of heat shock 90kDa protein ATPase homolog 1*		0.46	50
11522	FBpp0072564	CG9153-PB, isoform B	hect E3 ubiquitin ligase		0.46	36
11415	FBpp0072841	mitochondrial ribosomal protein S35		x	0.46	43

11374	FBpp0085363	SCAP		0.46	17
11365	FBpp0077150	Probable DNA replication complex GINS protein PSF2	*GINS complex subunit 2 (Psf2 homolog)*	0.46	33
11367	FBpp0070302	Myb-interacting protein 130		0.46	20
11494	FBpp0079203	CG8506-PA	*zinc finger, FYVE domain containing 20*	0.46	28
12001	FBpp0099494	C-1-tetrahydrofolate synthase, cytoplasmic		0.45	18
11977	FBpp0075120	CG4098-PA	nudix hydrolase 6	0.45	27
11766	FBpp0100136	1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase		0.45	19
11947	FBpp0072946	CG11526-PB, isoform B	*family with sequence similarity 40, member A*	0.45	20
11877	AAEL006769-PA	tryptophanyl-tRNA synthetase		0.45	25
11670	FBpp0080918	CG2614-PA	*KIAA0859*	0.45	34
11946	FBpp0079699	CG6443-PA	*chromosome 20 open reading frame 43*	0.44	54
11453	FBpp0072961	CG14967-PA	*KIAA0100*	0.44	18
11679	FBpp0077525	tho2		0.44	24
11943	FBpp0078358	CG12170-PA	3-oxoacyl-[acyl-carrier-protein] synthase	0.44	36
11664	FBpp0074808	CG3808-PA	RNA m5u methyltransferase	0.44	40
11900	FBpp0073966	Clathrin heavy chain		0.44	18
11713	FBpp0073663	iodotyrosine dehalogenase	iodotyrosine dehalogenase	0.44	17
11562	FBpp0079897	CG6746-PA	ptpla domain protein	0.44	53
12086	FBpp0071031	Probable mitochondrial import receptor subunit TOM40 homolog		0.44	54
11748	FBpp0087244	CG30022-PA	beta lactamase domain	0.44	55
11655	FBpp0084069	tolkin, isoform B		0.44	16
11732	FBpp0084626	CG4849-PA	116 kDa U5 small nuclear ribonucleoprotein component	0.44	17
11738	FBpp0081355	CG9630-PA	DEAD box ATP-dependent RNA helicase	0.44	29
11505	FBpp0086380	CG8443-PA	eukaryotic translation initiation factor 3 subunit (eif-3)	0.43	16
11678	FBpp0072703	CG13926-PA	*chromosome 11 open reading frame 73*	0.43	37
11466	FBpp0084779	ligatin		0.43	26
11906	FBpp0081810	CG6608-PB, isoform B	mitochondrial carrier protein putative	0.43	32
11605	FBpp0076242	CG5026-PA, isoform A	myotubularin	0.43	28
11445	FBpp0085923	adipose		0.43	25
11948	FBpp0071194	Probable U3 small nucleolar RNA-associated protein 11		0.43	42
11901	FBpp0074131	Integrin alpha-PS2 precursor		0.43	17
11747	FBpp0073491	CG1824-PA	lipid a export ATP-binding/permease protein msba	0.43	26
11942	FBpp0099560	Protein retinal degeneration B		0.43	23
12119	FBpp0077133	CG17840-PA	inositol 5-phosphatase	0.42	26

11818	FBpp0084478	CG5880-PA	*zinc finger, DHHC-type containing 16*		0.42	26
11371	FBpp0077357	okra			0.42	23
11845	FBpp0071543	CG30390-PA	*coiled-coil domain containing 101*		0.42	29
11397	FBpp0083272	Ire-1	Serine threonine-protein kinase *endoplasmic reticulum to nucleus signaling 2*		0.42	19
11979	FBpp0083132	gatA			0.42	29
11510	FBpp0075990	1,2-dihydroxy-3-keto-5-methylthiopentene dioxygenase	acireductone dioxygenase		0.42	53
12064	FBpp0083137	CG14290-PB	*brain protein 44-like*		0.42	34
11472	FBpp0082817	CG16941-PA	spliceosome associated protein		0.42	19
11532	FBpp0070299	CG14805-PA	PAF acetylhydrolase 45 kDa subunit putative		0.42	42
11781	FBpp0073083	pavarotti			0.42	24
12124	FBpp0078448	Probable proteasome subunit beta type 4		ph, du, de, ba [psmb-N]	0.41	92
12100	AAEL010379-PA	ATP-binding cassette transporter			0.41	16
12089	FBpp0072366	3-phosphoinositide-dependent protein kinase 1			0.41	34
12020	FBpp0075609	CG11267-PA	heat shock protein putative	du [Heat shock 10 kDa protein 1 (chaperonin 10)]	0.41	83
11599	FBpp0076280	CG5288-PC, isoform C	galactokinase		0.41	41
11462	FBpp0087323	CG6751-PA	WD-repeat protein		0.41	38
11492	FBpp0110411	conserved hypothetical protein			0.41	18
11949	FBpp0086399	CG8397-PA	actin binding protein putative		0.41	60
11662	FBpp0075280	Homeotic gene regulator			0.41	22
11485	FBpp0083354	Elongin B		du [elongin B isoform a]	0.41	57
11582	AAEL004330-PA	conserved hypothetical protein			0.41	17
11369	FBpp0085431	Transcription-associated protein 1 (Nipped-A)	transformation/transcription domain-associated protein		0.41	15
11986	FBpp0071155	Neuroglian precursor			0.41	15
11969	FBpp0070319	CG4199-PA, isoform A	disulfide oxidoreductase		0.41	34
11486	FBpp0110523	nitrate, fromate, iron dehydrogenase			0.41	34
11765	AAEL011712-PA	diacylglycerol kinase			0.41	12
11973	FBpp0075344	CG7650-PA	viral IAP-associated factor putative		0.41	41
11994	FBpp0072767	CG8993-PA	thioredoxin putative		0.41	61
11758	FBpp0076708	Transportin, isoform A			0.41	19
11682	FBpp0085071	Protein tailless			0.41	16
11638	FBpp0073725	CG1461-PA	tyrosine aminotransferase		0.41	40
11661	FBpp0070304	CG3573-PA	inositol polyphosphate 5-phosphatase		0.40	27
11746	FBpp0088517	CG5009-PA	acyl-CoA oxidase		0.40	28
11933	FBpp0084307	CG4743-PA	mitochondrial carrier protein		0.40	26
11475	FBpp0073983	Actin-like protein 13E			0.40	32

12052	FBpp0080894	cdc23			0.40	27
11689	FBpp0087297	BBS4			0.40	20
11717	FBpp0075685	Protein angel			0.40	24
12030	FBpp0072119	CG3735-PA	*chromosome 1 open reading frame 107*		0.40	29
11999	FBpp0079776	CG6700-PA	leukocyte receptor cluster (lrc) member		0.40	29
11888	FBpp0085589	CG11788-PA	*defective in sister chromatid cohesion 1 homolog*		0.40	34
11780	FBpp0080922	Importin beta subunit			0.40	19
11568	FBpp0074481	CG12203-PA	NADH: ubiquinone dehydrogenase putative	du [NADH dehydrogenase (ubiquinone) Fe-S protein 4]	0.40	72
11897	FBpp0082657	Mitochondrial import inner membrane translocase subunit TIM16			0.40	43
11597	FBpp0087591	Protein preli-like			0.40	55
11540	FBpp0087891	CG8709-PA	lipin		0.40	32
11648	FBpp0080807	Probable phosphomevalonate kinase			0.39	37
11950	FBpp0088810	Protein arginine N-methyltransferase capsuleen			0.39	37
11554	FBpp0071033	Pyruvate dehydrogenase phosphatase (CG12151-PA)			0.39	29
11468	FBpp0079547	Niemann-Pick Type C-1			0.39	24
11611	AAEL006321-PA	1-acylglycerol-3-phosphate acyltransferase			0.39	20
11967	FBpp0078711	CG12512-PA	AMP dependent coa ligase		0.39	43
12072	FBpp0082148	CG5608-PA	*Vac14 homolog*		0.39	33
11541	FBpp0079586	CG31715-PA	*myotrophin*		0.39	37
11968	FBpp0087979	Cytochrome b5			0.38	89
11395	FBpp0076789	Pole2			0.38	27
12117	FBpp0081376	Dihydroorotate dehydrogenase, mitochondrial precursor			0.38	34
11927	FBpp0081157	CG1104-PA, isoform A			0.38	27
11359	FBpp0083514	DNA polymerase alpha catalytic subunit			0.38	17
11966	FBpp0082813	CG3534-PA	xylulose kinase		0.38	30
11876	FBpp0074672	Translocase of outer membrane 20			0.38	67
11908	FBpp0079488	CG13126-PA	*methyltransferase 11 domain containing 1*		0.38	30
11497	FBpp0110309	poly a polymerase			0.38	21
11401	FBpp0083840	CG10365-PA, isoform A	*ChaC, cation transport regulator homolog 1*		0.38	38
11842	FBpp0073355	Probable signal peptidase complex subunit 2		du [signal peptidase complex subunit 2 homolog]	0.38	67
11791	FBpp0077447	CG9867-PA	glycosyltransferase		0.38	29
11872	FBpp0078684	CG8891-PA	inosine triphosphate pyrophosphatase (itpase) (inosine triphosphatase)		0.37	37

11470	FBpp0084728	Protein kinase C			0.37	21
12095	FBpp0070301	mitochondrial ribosomal protein L16		x	0.37	41
12041	FBpp0074227	CG5800-PA	DEAD box ATP-dependent RNA helicase		0.37	34
11923	FBpp0083244	CG4973-PA	zinc finger protein putative		0.37	35
11885	FBpp0084711	CG1951-PA	*SCY1-like 2*		0.37	21
11622	FBpp0083022	CG7146-PA	vacuolar protein sorting 39 homolog		0.37	26
11498	FBpp0081588	CG9399-PA, isoform A	*brain protein 44*		0.37	60
11353	FBpp0074004	CG32579-PA	*XK, Kell blood group complex subunit-related family, member 6*		0.37	21
11870	FBpp0079567	CG31717-PA	*phosphatidic acid phosphatase type 2 domain containing 2*		0.37	39
11951	FBpp0077965	UPF0315 protein			0.37	43
11560	FBpp0078275	jagunal, isoform C			0.37	47
11690	FBpp0077209	Pdsw, isoform B			0.37	71
11777	FBpp0087085	DNA-directed RNA polymerase III 128 kDa polypeptide			0.36	16
12091	FBpp0083854	Probable oligoribonuclease			0.36	36
11593	FBpp0081841	CG17187-PA	*DnaJ (Hsp40) homolog, subfamily C, member 17*		0.36	33
11851	FBpp0071891	Arginine methyltransferase 7			0.36	34
11581	FBpp0073235	Putative 6-phosphogluconolactonase	6-phosphogluconolactonase		0.36	47
11481	FBpp0086877	CG4646-PA	*chromosome 1 open reading frame 123*		0.36	36
12032	FBpp0083853	twister			0.35	17
11569	FBpp0081451	Adenosine deaminase			0.35	29
11703	FBpp0080628	CG15161-PA			0.35	21
11476	FBpp0071426	CG1826-PA	*BTB (POZ) domain containing 9*		0.35	29
11749	FBpp0078844	CG9154-PA	*N-6 adenine-specific DNA methyltransferase 2 (putative)*		0.35	34
11706	FBpp0082314	CG9588-PA	26S proteasome non-ATPase regulatory subunit		0.35	48
11612	FBpp0073995	CG3560-PA	ubiquinol-cytochrome c reductase complex 14 kd protein		0.35	85
12034	FBpp0076111	Laminin gamma-1 chain precursor			0.35	18
11548	FBpp0074104	mitochondrial ribosomal protein L22		x	0.35	44
11620	FBpp0073585	Vesicular-fusion protein Nsf1			0.34	20
11557	FBpp0079620	CG6206-PB, isoform B	lysosomal alpha-mannosidase (mannosidase alpha class 2b member 1)		0.34	42
11543	FBpp0089008	Adenine phosphoribosyltransferase			0.34	49
11602	AAEL007823-	PIWI			0.34	18

	PA						
12015	FBpp0085924	CG10914-PA				0.34	26
11931	FBpp0089163	Cleavage and polyadenylation specificity factor, 160 kDa subunit				0.34	19
11696	FBpp0082172	Xanthine dehydrogenase				0.34	31
11399	FBpp0086640	DNA-directed RNA polymerase I largest subunit	DNA-directed RNA polymerase I largest subunit			0.33	20
12006	FBpp0080203	DNA mismatch repair protein spellchecker 1				0.33	22
11963	FBpp0086591	SMC2				0.33	23
11744	FBpp0073979	Graf, isoform A				0.33	25
11440	FBpp0078583	CG9804-PA	lipoate-protein ligase b			0.33	29
11575	FBpp0084349	Dak1				0.33	56
11594	FBpp0086887	Tripeptidyl-peptidase 2				0.33	19
11832	FBpp0072460	Rhythmically expressed gene 2 protein				0.32	25
12113	FBpp0075106	Probable ATP-dependent RNA helicase Dbp73D				0.32	32
11585	FBpp0071193	Hypothetical protein CG1785				0.32	41
11693	FBpp0083857	Putative succinate dehydrogenase [ubiquinone] cytochrome b small subunit, mitochondrial precursor				0.32	63
11985	FBpp0076589	Signal recognition particle 19 kDa protein	srp19			0.32	50
11751	FBpp0081763	CG4511-PA	viral IAP-associated factor putative			0.32	51
11595	FBpp0075755	lethal (3) neo18		du [NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 5, 16kDa precursor]		0.32	73
11633	FBpp0075399	Probable DNA mismatch repair protein MSH6				0.32	25
12021	FBpp0072426	thoc7, isoform A				0.32	41
11410	FBpp0070403	Probable ATP-dependent RNA helicase	ATP-dependent RNA helicase			0.31	20
12005	FBpp0080475	CG31739-PA	aspartyl-tRNA synthetase			0.31	27
11658	FBpp0081860	mitochondrial ribosomal protein L40			x	0.31	46
11420	FBpp0082522	ATP synthase O subunit, mitochondrial precursor		du [Mitochondrial ATP synthase, O subunit precursor]		0.31	111
11503	FBpp0075148	CG33158-PB	translation elongation factor *longation factor Tu GTP binding domain containing 1 isoform 2*			0.31	28
11819	FBpp0077251	CG33123-PA	leucyl-tRNA synthetase			0.30	18
11504	FBpp0085690	CG11242-PA	tubulin-specific chaperone b (tubulin folding cofactor b)			0.30	48
11373	FBpp0078895	CG9542-PA	*arylformamidase*			0.30	28
11430	FBpp0086226	Superoxide dismutase [Mn], mitochondrial precursor				0.30	81

11879	FBpp0070639	CG6379-PA	*FtsJ methyltransferase domain containing 2*		0.30	26
11702	FBpp0071916	CG11079-PC, isoform C	5-formyltetrahydrofolate cyclo-ligase		0.30	41
11537	FBpp0083226	CG4686-PA			0.30	51
11945	FBpp0073762	Probable mitochondrial 28S ribosomal protein S25		x	0.29	45
12044	FBpp0073196	CG15014-PA	*THUMP domain containing 1*		0.29	42
11800	FBpp0077399	Transportin-Serine/Arginine rich			0.29	23
12028	FBpp0077173	CG31961-PA, isoform A	tubulin folding cofactor c		0.28	36
11873	AAEL011682-PA	nuclear pore complex protein nup93			0.28	17
11907	FBpp0083650	Probable prefoldin subunit 5			0.27	63
11519	FBpp0072615	CG9187-PA	partner of sld5		0.27	32
11388	FBpp0087629	CG1884-PB, isoform B			0.27	16
11736	FBpp0084051	CG13625-PA	*BUD13 homolog*		0.27	34
11812	FBpp0100031	Protein male-less	ATP-dependent RNA helicase		0.26	26
11753	FBpp0079316	CG13397-PA	alpha-n-acetylglucosaminidase		0.26	22
12055	FBpp0072456	Rev1			0.25	20
12027	AAEL011963-PA	conserved hypothetical protein			0.25	18
11836	AAEL009888-PA	WD-repeat protein			0.25	25
11665	AAEL004081-PA	dj-1 protein			0.25	57
11995	FBpp0080305	CG15261-PA	ribonuclease UK114 putative		0.24	70
11516	AAEL005494-PA	conserved hypothetical protein			0.17	9

3. RESULTS

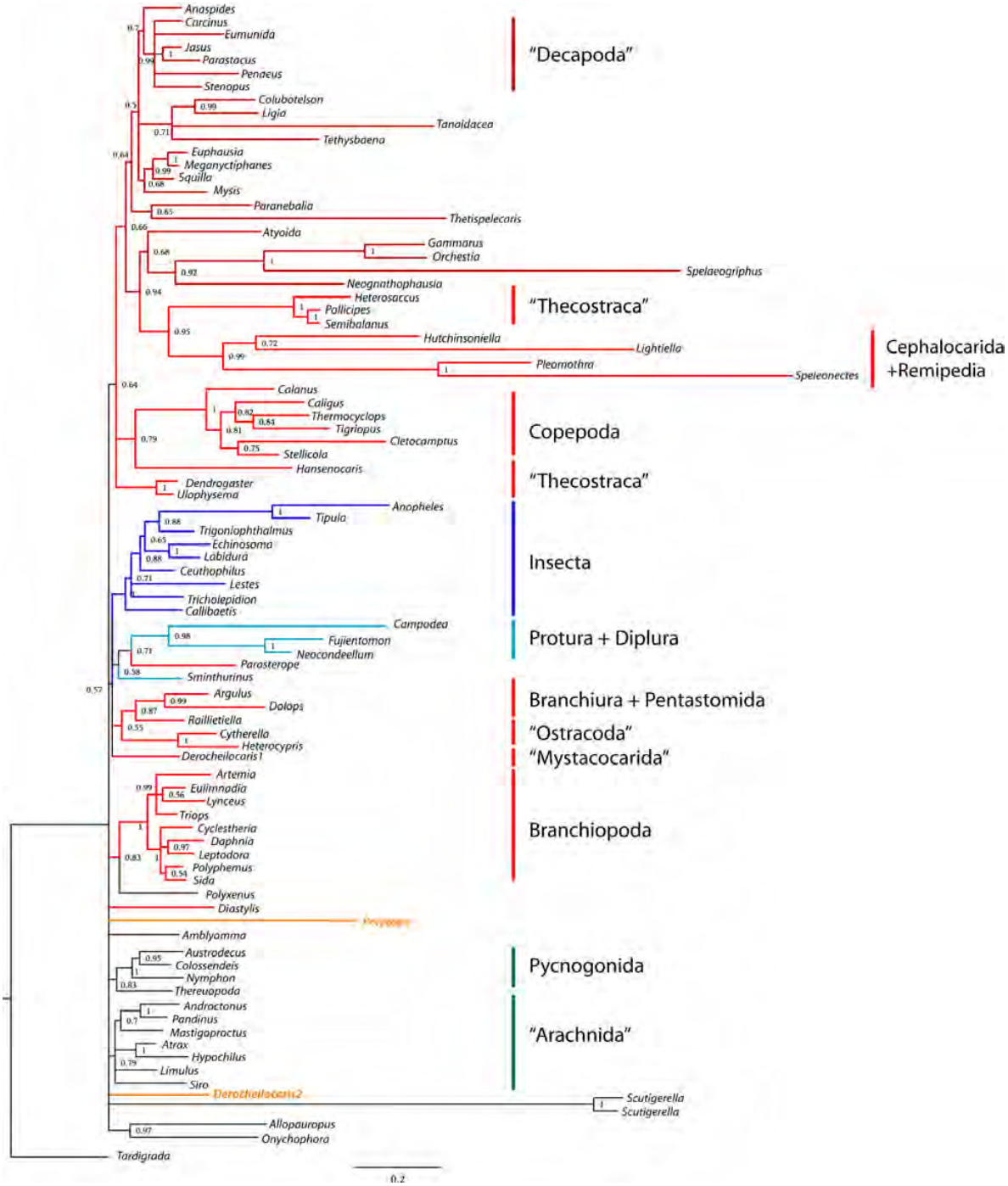


Figure S2 | Resulting topology of final run 1. Pre-alignment conducted with MAFFT, mixed models are applied and the dataset is RY coded. Model settings are (gamma, linked)

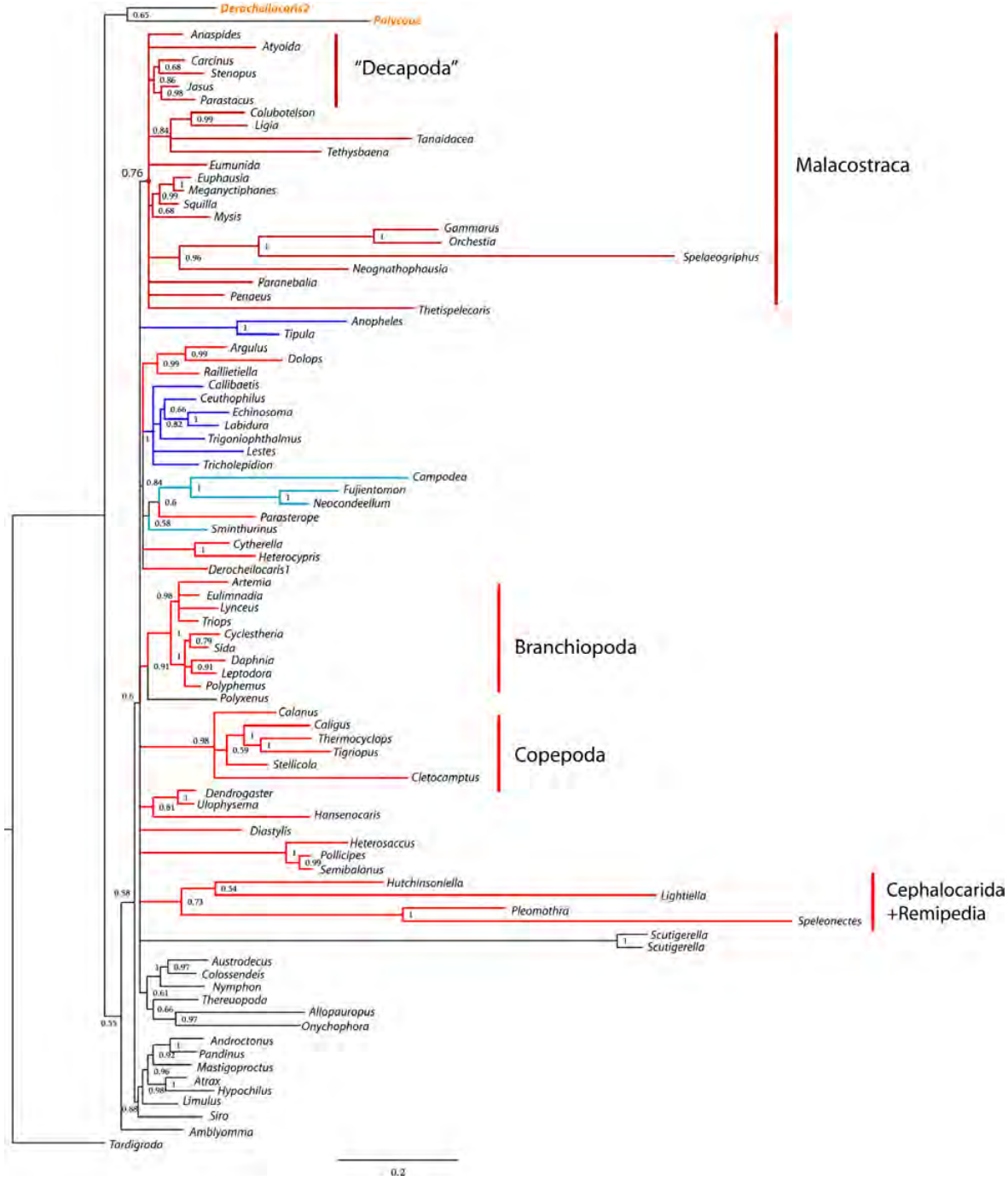


Figure S3 | Resulting topology of final run 3. Pre-alignment conducted with MAFFT, no mixed models are applied and the dataset is RY coded. Model settings are (3 partitions, gamma, unlinked).

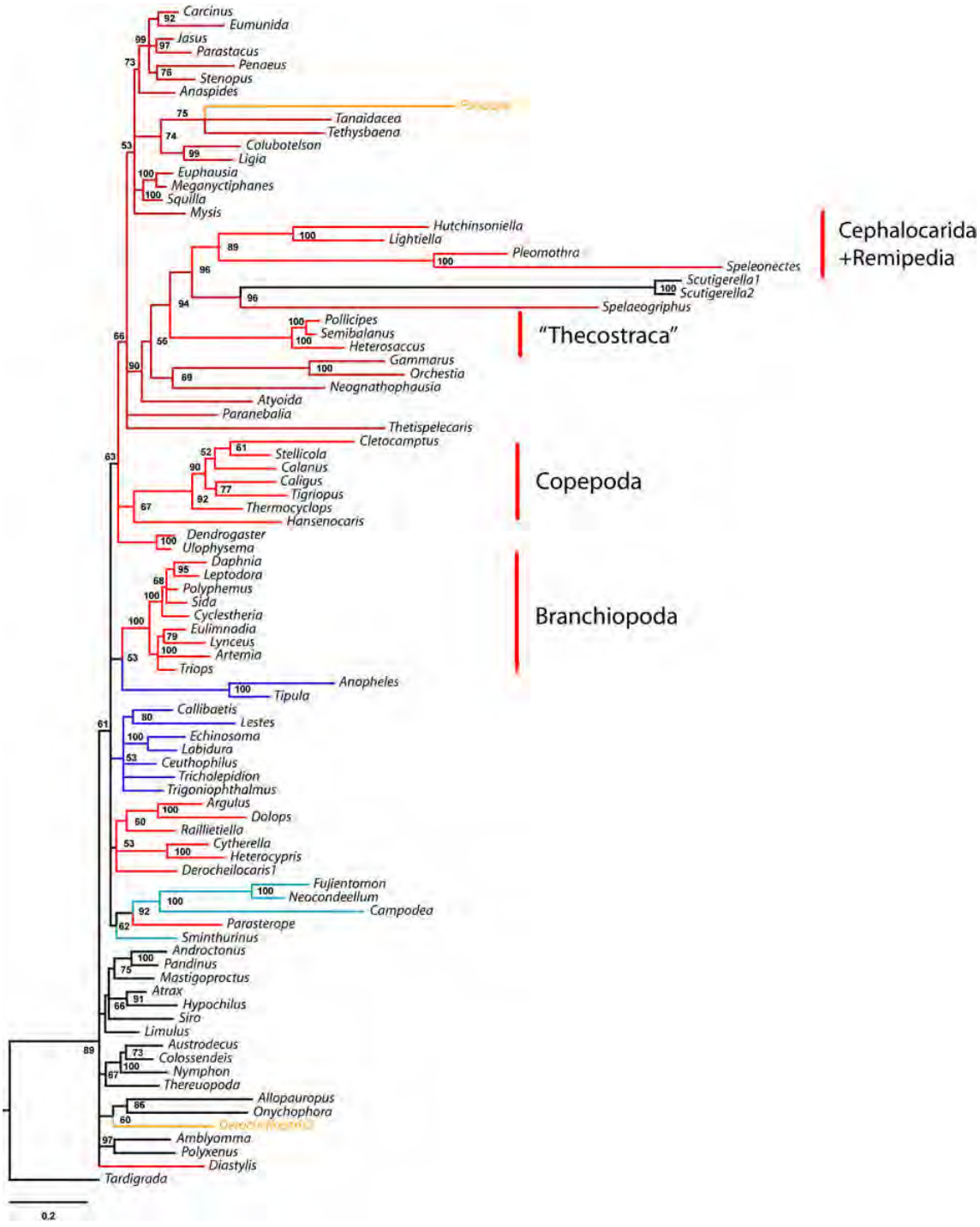


Figure S4 | Resulting topology of final run 4. Pre-alignment conducted with MUSCLE, mixed models are applied and the dataset is RY coded. Model settings are (gamma, linked)

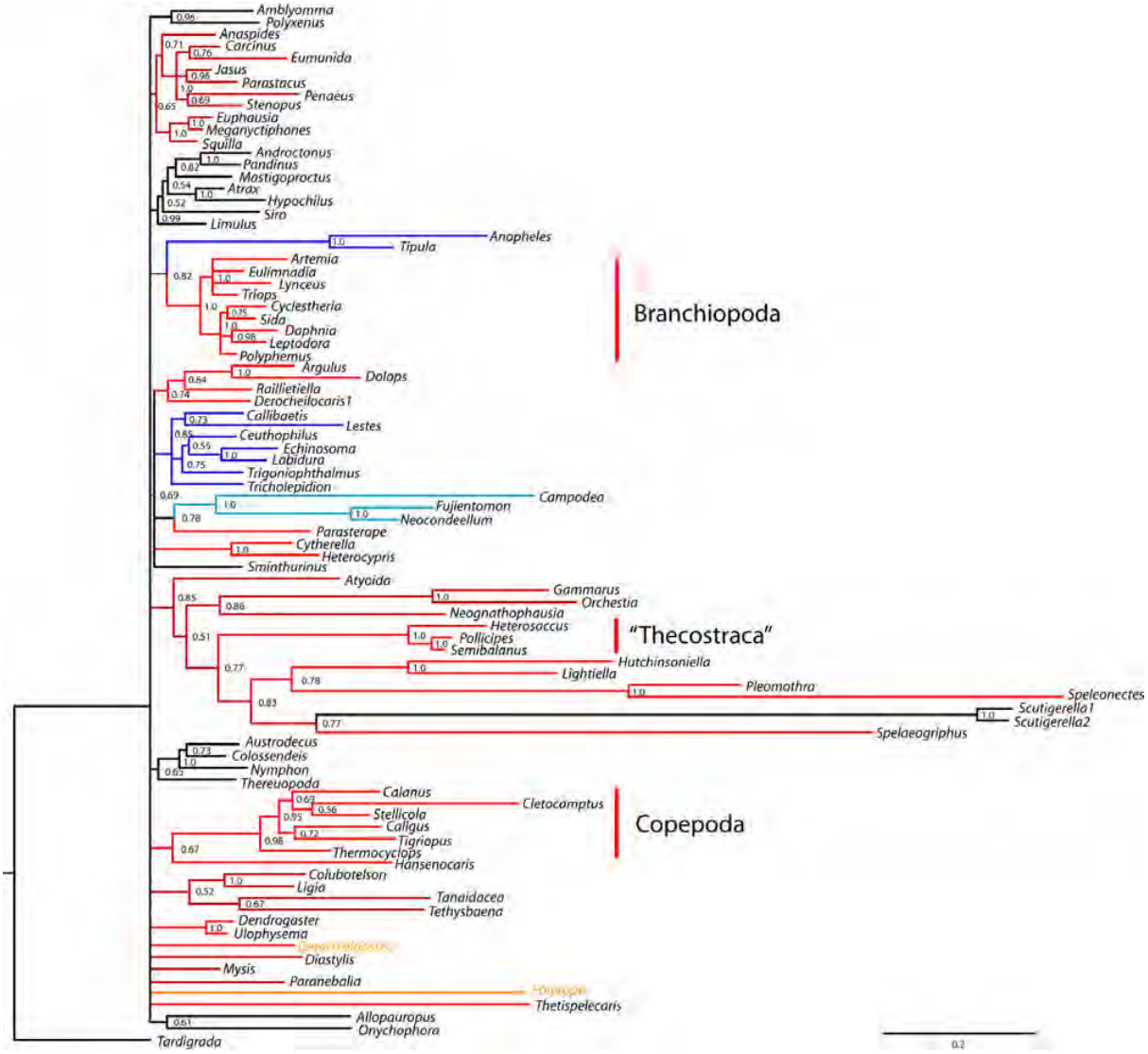


Figure S5 | Resulting topology of final run 5. Prealignment conducted with MUSCLE, mixed models are applied and the dataset is RY coded. Model settings are (gamma, unlinked)

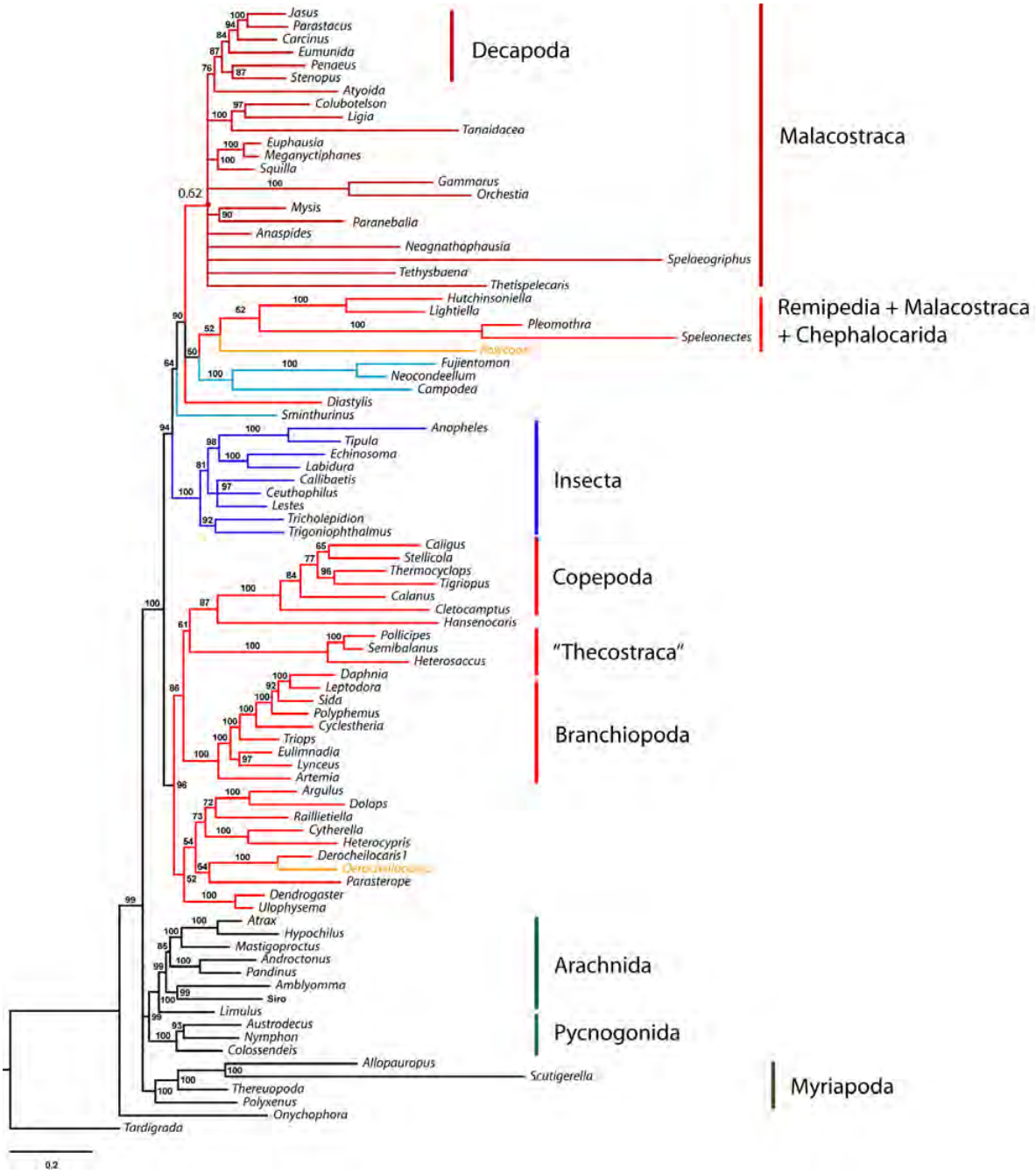


Figure S6 | Resulting topology of final run 6. Manual alignment pre-aligned with MUSCLE, the dataset is parted in three partitions. Model settings are (gamma, linked)

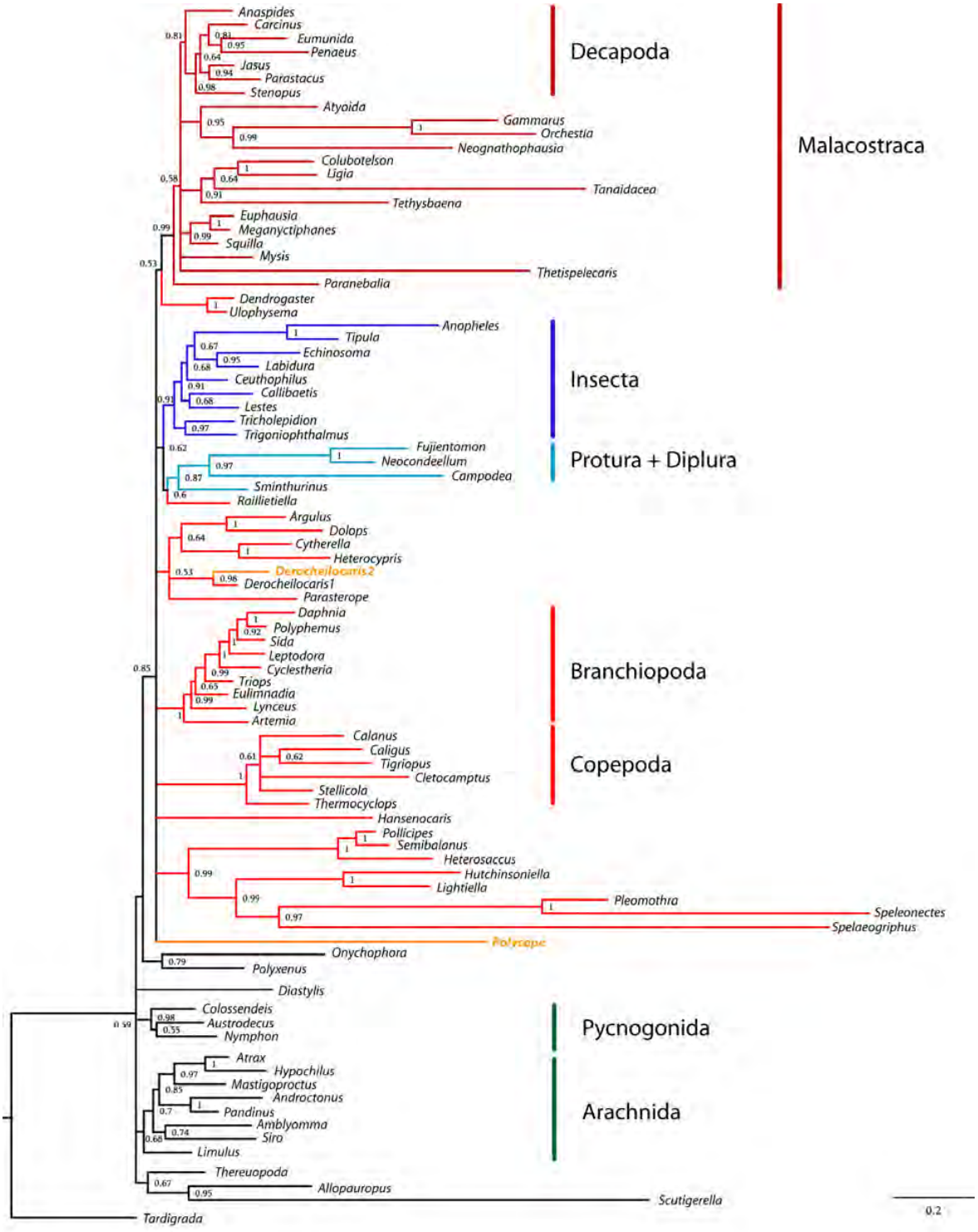


Figure S7 | Resulting topology of final run 8. Manual alignment pre-aligned with MUSCLE, the dataset is parted in three partitions. Model settings are (gamma, linked)

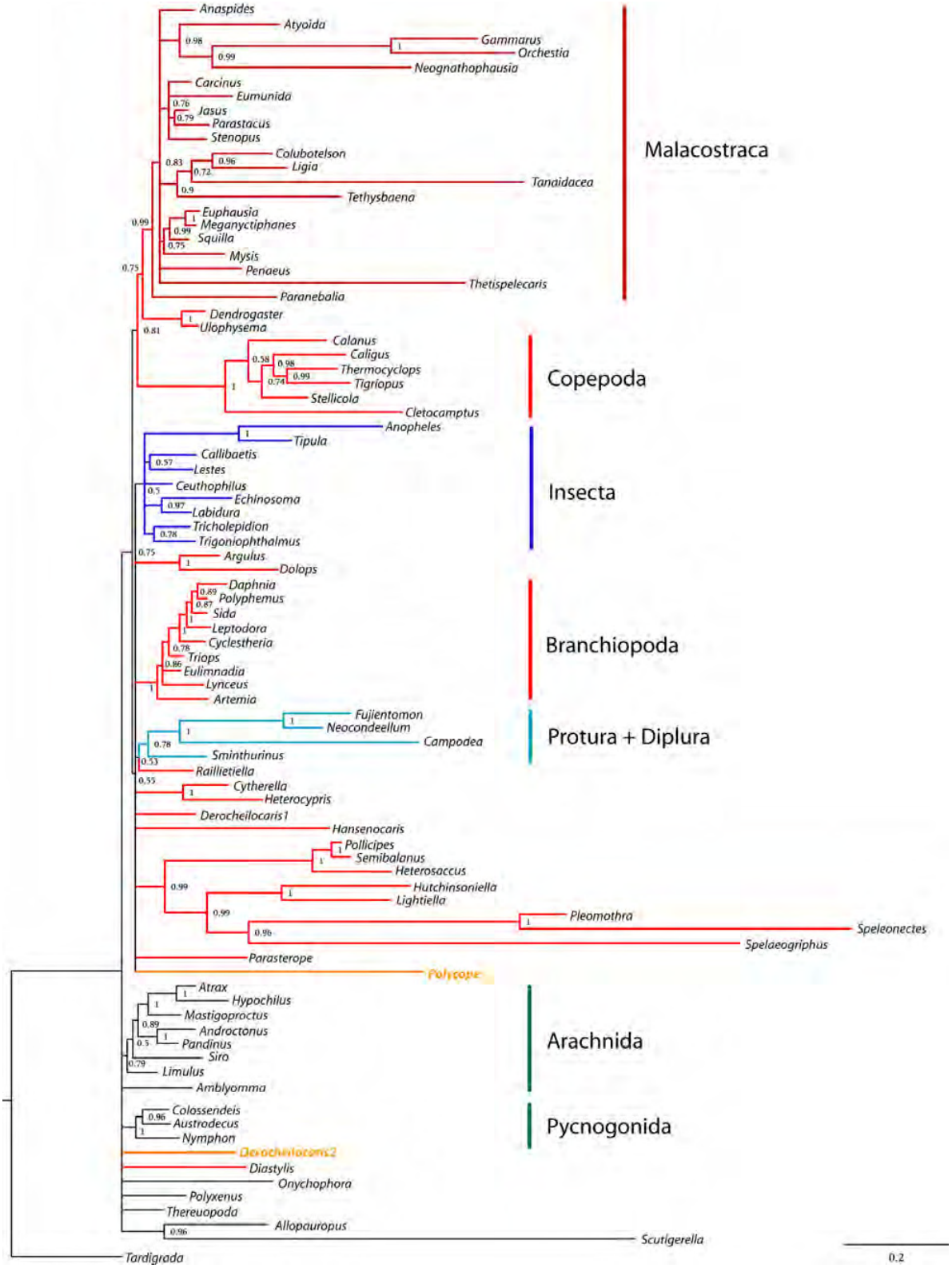


Figure S8 | Resulting topology of final run 9. Manual alignment pre-aligned with MUSCLE, the dataset is parted in three partitions. Model settings are (gamma, linked)

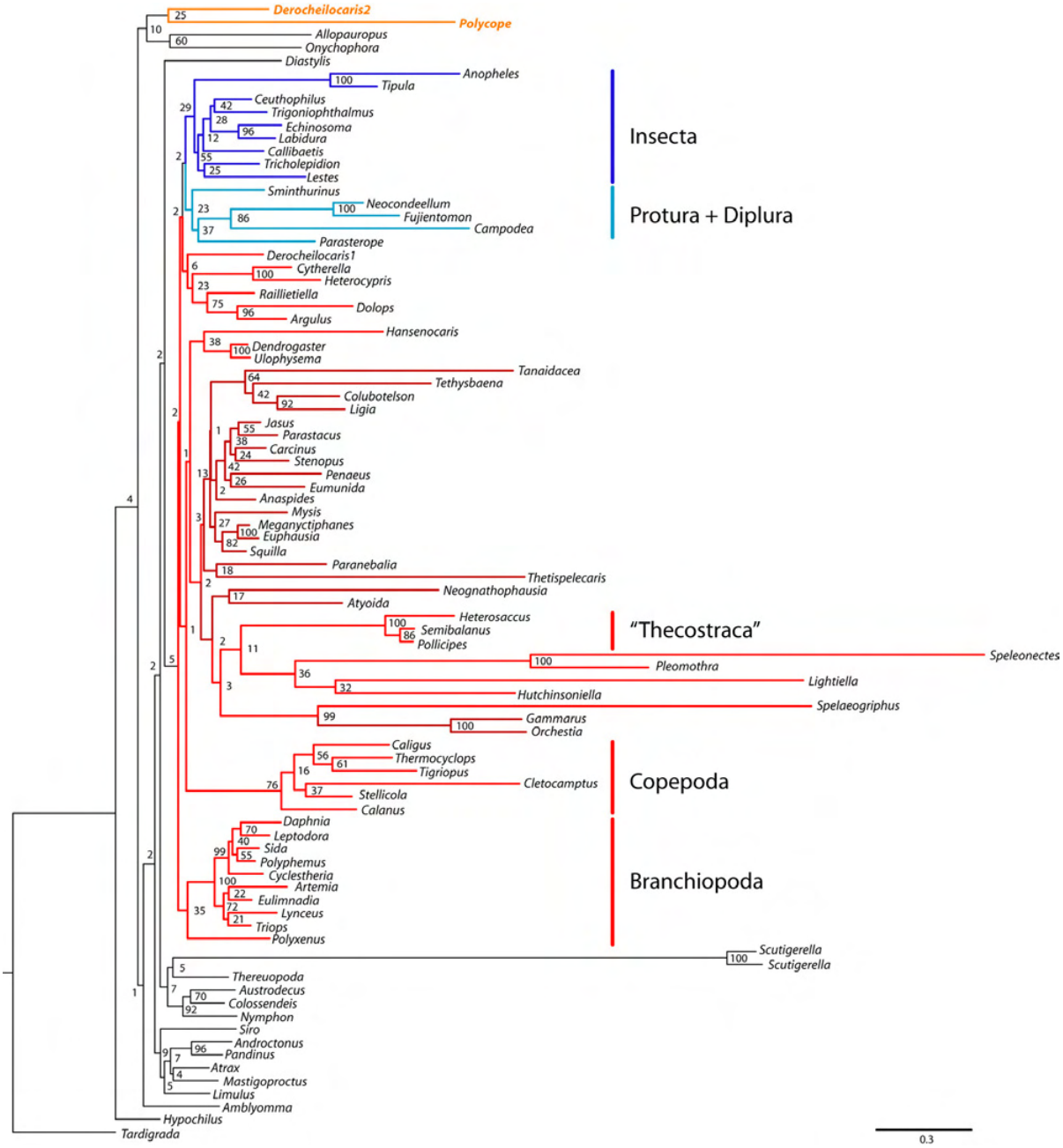
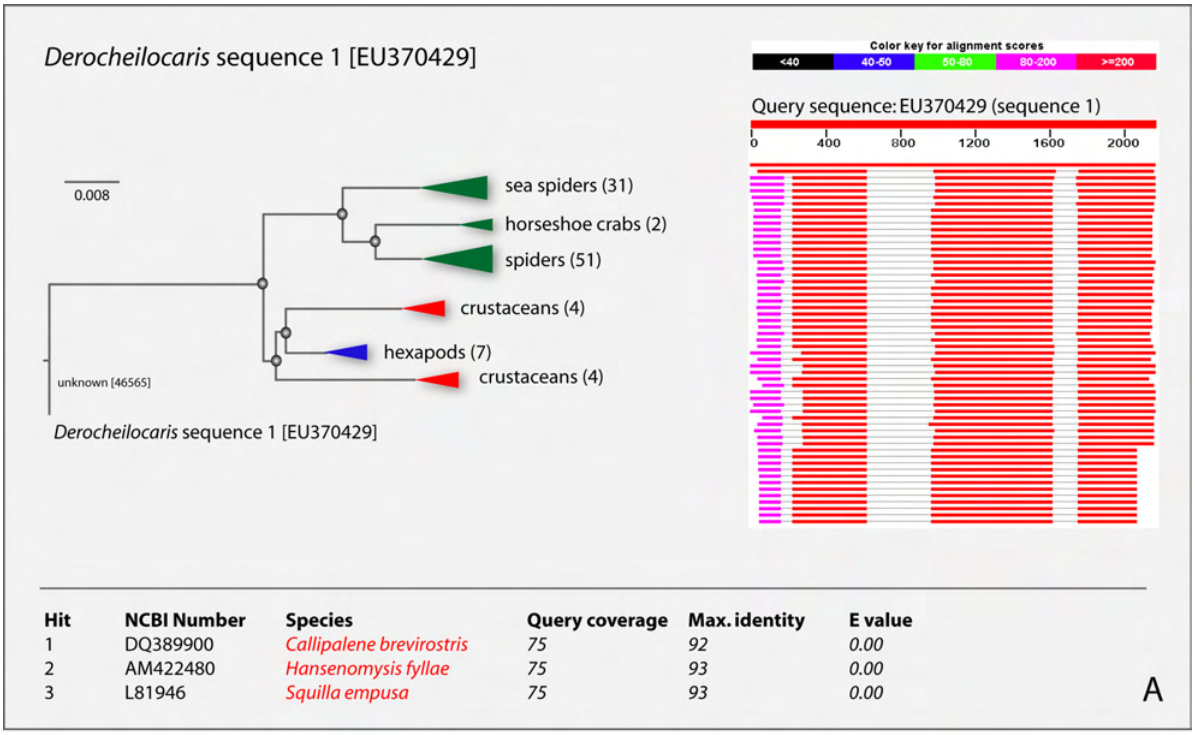
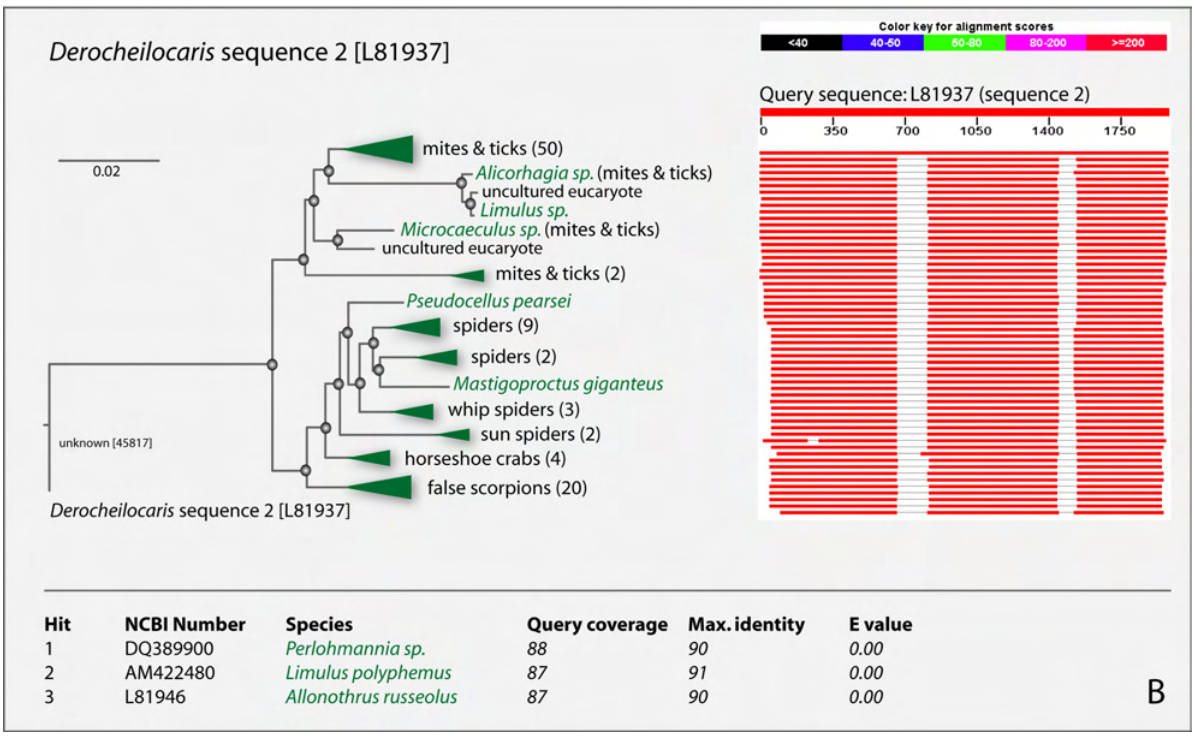


Figure S9 | Resulting topology of final run 10. RAXML analysis (-f, a, GTR+CAT) with 10.000 bootstrap replicates.



Hit	NCBI Number	Species	Query coverage	Max. identity	E value
1	DQ389900	*Callipalene brevirostris*	*75*	*92*	*0.00*
2	AM422480	*Hansenomysis fyllae*	*75*	*93*	*0.00*
3	L81946	*Squilla empusa*	*75*	*93*	*0.00*

A



Hit	NCBI Number	Species	Query coverage	Max. identity	E value
1	DQ389900	*Perlohmannia sp.*	*88*	*90*	*0.00*
2	AM422480	*Limulus polyphemus*	*87*	*91*	*0.00*
3	L81946	*Allonothrus russeolus*	*87*	*90*	*0.00*

B

Figure S10 | MEGABLAST result of the two mystacocarid sequences.

A) shows the MEGABLAST result of the mystacocarid sequence that was obtained in this thesis. B) pictures the result of the published sequence L81937. The species color code is identical to all previous graphics.

The mimimum-evolution trees (left) give a graphical (tree-like) summary of the BLAST results. The hits for highly similar sequences to the query sequence are shown on the right, the matching sequence parts are represented by the red bars. The first three hits are given more detailed in

the tables.

The first hits for A) are crustaceans , for B) chelicerates. As pictured in the tree and in the table the BLAST analysis finds highly similar sequences that are only chelicerates in the case of the published sequence (B).

SUPPLEMENT - MANUSCRIPT I

BMC Evolutionary Biology

Research article

Open Access

Can comprehensive background knowledge be incorporated into substitution models to improve phylogenetic analyses? A case study on major arthropod relationships

Björn M von Reumont*[1], Karen Meusemann[1], Nikolaus U Szucsich[2], Emiliano Dell'Ampio[2], Vivek Gowri-Shankar, Daniela Bartel[2], Sabrina Simon[3], Harald O Letsch[1], Roman R Stocsits[1], Yun-xia Luan[4], Johann Wolfgang Wägele[1], Günther Pass[2], Heike Hadrys[3,5] and Bernhard Misof[6]

Address: [1]Molecular Lab, Zoologisches Forschungsmuseum A. Koenig, Bonn, Germany, [2]Department of Evolutionary Biology, University Vienna, Vienna, Austria, [3]ITZ, Ecology & Evolution, Stiftung Tieraerztliche Hochschule Hannover, Hannover, Germany, [4]Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, PR China, [5]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA and [6]UHH Biozentrum Grindel und Zoologisches Museum, University of Hamburg, Hamburg, Germany

Email: Björn M von Reumont* - bmvr@arcor.de; Karen Meusemann - mail@karen-meusemann.de; Nikolaus U Szucsich - nikola.szucsich@univie.ac.at; Emiliano Dell'Ampio - emiliano.dell.ampio@univie.ac.at; Vivek Gowri-Shankar - gowrishv@cs.man.ac.uk; Daniela Bartel - dani.bartel@chello.at; Sabrina Simon - sabrina.simon@ecolevol.de; Harald O Letsch - hletsch@freenet.de; Roman R Stocsits - roman.stocsitz@gmail.com; Yun-xia Luan - yxluan@sibs.ac.cn; Johann Wolfgang Wägele - w.waegele.zfmk@uni-bonn.de; Günther Pass - guenther.pass@univie.ac.at; Heike Hadrys - heike.hadrys@ecolevol.de; Bernhard Misof - bernhard.misof@uni-hamburg.de

* Corresponding author

Published: 27 May 2009

BMC Evolutionary Biology 2009, **9**:119 doi:10.1186/1471-2148-9-119

Received: 29 September 2008
Accepted: 27 May 2009

This article is available from: http://www.biomedcentral.com/1471-2148/9/119

Abstract

Background: Whenever different data sets arrive at conflicting phylogenetic hypotheses, only testable causal explanations of sources of errors in at least one of the data sets allow us to critically choose among the conflicting hypotheses of relationships. The large (28S) and small (18S) subunit rRNAs are among the most popular markers for studies of deep phylogenies. However, some nodes supported by this data are suspected of being artifacts caused by peculiarities of the evolution of these molecules. Arthropod phylogeny is an especially controversial subject dotted with conflicting hypotheses which are dependent on data set and method of reconstruction. We assume that phylogenetic analyses based on these genes can be improved further i) by enlarging the taxon sample and ii) employing more realistic models of sequence evolution incorporating non-stationary substitution processes and iii) considering covariation and pairing of sites in rRNA-genes.

Results: We analyzed a large set of arthropod sequences, applied new tools for quality control of data prior to tree reconstruction, and increased the biological realism of substitution models. Although the split-decomposition network indicated a high noise content in the data set, our measures were able to both improve the analyses and give causal explanations for some incongruities mentioned from analyses of rRNA sequences. However, misleading effects did not completely disappear.

Conclusion: Analyses of data sets that result in ambiguous phylogenetic hypotheses demand for methods, which do not only filter stochastic noise, but likewise allow to differentiate phylogenetic signal from systematic biases. Such methods can only rely on our findings regarding the evolution of the analyzed data. Analyses on independent data sets then are crucial to test the plausibility of the results. Our approach can easily be extended to genomic data, as well, whereby layers of quality assessment are set up applicable to phylogenetic reconstructions in general.

Background

Most recent studies that focused on the reconstruction of ancient splits in animals, have relied on 18S and/or 28S rRNA sequences, e.g. [1]. These data sets strongly contributed to our knowledge of relationships, however, several nodes remain that are suspected of being artifacts caused by peculiar evolutionary rates which may be lineage specific. Particular unorthodox nodes were discussed as long branch artifacts, others were held to be clusters caused by non-stationary evolutionary processes as indicated by differences in nucleotide composition among the terminals. The reconstruction of ancient splits seems to be especially dependent on taxon sampling and character choice, since in single lineages the signal-to-noise ratio is consistently marginal in allowing a reasonable resolution. Thus, quality assessment of data via e.g. secondary structure guided alignments, discarding of randomly similar aligned positions or heterogeneity of the data set prior to analysis is a crucial step to obtain reliable results. Arthropod phylogeny is especially suitable as a case study, since their ancient and variable phylogenetic history, which may have included intermittent phases of fast radiation, impedes phylogenetic reconstruction.

Major arthropod relationships

While currently there is wide agreement about the monophyly of Arthropoda, relationships among the four major subgroups (Chelicerata, Myriapoda, Crustacea, Hexapoda) remain contested, even the monophyly of each of the subgroups has come under question. The best supported relationship among these subgroups seems to be the clade comprising all crustaceans and hexapods. This clade, named Pancrustacea [2], or Tetraconata [3], is supported by most molecular analyses, e.g. [1,4-14]. Likewise, the clade has increasingly found support from morphological data [3,15-18], especially when malacostracans are directly compared with insects. Most of these studies reveal that crustaceans are paraphyletic with respect to a monophyletic Hexapoda. However, most analyses of mitochondrial genes question hexapod monophyly [19-22]. Additionally, various crustacean subgroups are discussed as potential hexapod sister groups. Fanenbruck et al. [15] favored a derivation of Hexapoda from a common ancestor with Malacostraca + Remipedia based on neuroanatomical data. In recent molecular studies, either Branchiopoda [12] or Copepoda [1,11,23] emerged as the sister group of Hexapoda. The Pancrustacea hypothesis implies that Atelocerata (Myriapoda + Hexapoda) is not monophyletic. In most of the above mentioned molecular studies, the Myriapoda appear at the base of the clade Mandibulata or as the sistergroup to Chelicerata. The combination of Chelicerata + Myriapoda [1,7,13,14,24] was coined Paradoxopoda [11] or Myriochelata [10]. It seems that this grouping can be partly explained by signal erosion [25], and likewise is dependent on outgroup choice [26]. In addition, the most recent morphological data is consistent with the monophyly of Mandibulata [27], but not of Myriochelata. Almost no morphological data corroborate Myriochelata except for a reported correspondence in neurogenesis [28]; this however alternatively may reflect the plesiomorphic state within Arthropoda [29,30]. Within Hexapoda, relationships among insect orders are far from being resolved [31-35]. Open questions concern the earliest splits within Hexapoda, e.g. the monophyly or paraphyly of Entognatha (Protura + Diplura + Collembola) [9,19,22,32,34,36-45].

Goals and methodological background

The aim of the present study is to optimize the phylogenetic signal contained in 18S and 28S rRNA sequences for the reconstruction of relationships among the major arthropod lineages. A total of 148 arthropod taxa representing all major arthropod clades including onychophorans and tardigrades (the latter as outgroup taxa) were sampled to minimize long-branch artifacts [25]. A new alignment procedure that takes secondary structure into account is meant to corroborate the underlying hypotheses of positional homology as accurately as possible. A new tool for quality control optimizes the signal-to-noise ratio for the final analyses. In the final step, we try to improve the analyses by fitting biologically realistic mixed DNA/RNA substitution models to the rRNA data. Time-heterogeneous runs were performed to allow for lineage specific variation of the model of evolution.

The use of secondary structure information both corroborates hypotheses of positional homology in the course of sequence alignment, as well as helps to avoid misleading effects of character dependence due to covariation among sites. It was demonstrated that ignoring correlated variance may mislead tree reconstructions biased by an over-

emphasis of changes in paired sites [34,46,47]. Evolutionary constraints on rRNA molecules are well known, for example constraints resulting from secondary structure interactions. The accuracy of rRNA comparative structure models [48-50] has been confirmed by crystallographic analyses [51,52]. Based on this background knowledge, rRNA sequences are an ideal test case to study the effect of biologically realistic substitution models on tree reconstructions.

Recent studies of genome scale data revealed that a careful choice of biologically realistic substitution models and model fitting are of particular importance in phylogenetic reconstructions [53-55]. The extent, however, to which biological processes can/should be modeled in detail is still unclear. The analyses of rRNA sequences can still deliver new insights in this direction, since the relatively comprehensive background knowledge allows to better separate different aspects of the substitution processes. In order to model covariation in rRNA sequences, we estimated secondary structure interactions by applying a new approach implemented in the software RNAsalsa [56] (download available from http://rnasalsa.zfmk.de/), which helps to accommodate inadequate modeling (e.g. missing covariotide effects) of rRNA substitution processes in deep phylogenetic inference [34,57]. Essentially, this approach combines prior knowledge of conserved site interactions modeled in a canonical eukaryote secondary structure consensus model with the estimation of alternative and/or additional site interactions supported by the specific data. Inferred site covariation patterns were used then to guide the application of mixed substitution models in subsequent phylogenetic analyses.

Finally, we accounted for inhomogeneous base composition across taxa, a frequently observed phenomenon indicating non-stationary substitution processes [58-60]. Non-stationary processes, if present, clearly violate assumptions of stationarity regularly assumed in phylogenetic analyses [60-62]. Thus, we modeled non-stationary processes combined with the application of mixed DNA/RNA substitution models in a Bayesian approach using the *PHASE-2.0* software package [63] to provide a better fit to our data than standard substitution models [60,64]. In *PHASE-2.0* a nonhomogeneous substitution model is implemented [...] "by introducing a reversible jump Markov chain Monte Carlo method for efficient Bayesian inference of the model order along with other phylogenetic parameters of interest" [60].

Application of a new hierarchical prior leads to more reasonable results when only a small number of lineages share a particular substitution process. Additionally *PHASE-2.0* includes specialized substitution models for RNA genes with conserved secondary structure [60].

Results

We contributed 103 new and nearly complete 18S or 28S rRNA sequences and analyzed sequences for 148 taxa (Additional file 1), of which 145 are Arthropoda *sensu stricto*, two onychophorans and *Milnesium* sp. (Tardigrada). The alignment of the 18S rRNA sequences comprised 3503 positions, and the 28S rRNA alignment 8184. The final secondary consensus structures included 794 paired positions in the 18S and 1326 paired positions in the 28S. The consensus structures contained all paired sites that in 60% or more sequences were detected after folding (default *s*3 = 0.6 in RNAsalsa). ALISCORE[65] scored 1873 positions as randomly similar (negative scoring values in the consensus profile) to the 18S and 5712 positions of the 28S alignment (Figure 1).

Alignment filtering and concatenation of data

After the exclusion of randomly similar sections identified by ALISCORE, 1630 (originally 3503) of the 18S rRNA and 2472 (originally 8184) positions of the 28S rRNA remained. Filtered alignments were concatenated and used for analyses in *PHASE-2.0*. The concatenated alignment comprised 4102 positions.

Split supporting patterns

The neighbornet graph, which results from a split decomposition based on uncorrected p-distances (Figure 2) and LogDet correction plus invariant sites model (see Additional file 2) pictured a dense network, which hardly resembles a tree-like topology. This indicates the presence of some problems typical in studies of deep phylogeny: a) Some taxa like Diptera (which do not cluster with ectognathous insects), Diplura, Protura and Collembola each appear in a different part of the network with Diplura and Protura seperated from other hexapods, *Lepisma saccharina* (clearly separated from the second zygentoman *Ctenolepisma* that is nested within Ectognatha), Symphyla, Pauropoda, as well as Remipedia and Cephalocarida have very long branches. Consequently the taxa may be misplaced due to signal erosion or occurrence of homoplasies, and their placement in trees must be discussed critically [25]. The usage of the LogDet distance adjusts the length of some branches but does not decrease the amount of conflicts in deep divergence splits. b) The inner part of the network shows little treeness, which indicates a high degree of conflicting signal.

A remarkable observation seen in both phylogenetic networks is that some taxa have long stem-lineages, which means that the species share distinct nucleotide patterns not present in other taxa. Such well separated groups are Copepoda, Branchiopoda, Cirripedia, Symphyla, Collembola, Diplura, Protura and Diptera, while e.g. Myriapoda partim, Chelicerata and the Ectognatha (bristletails, silver-

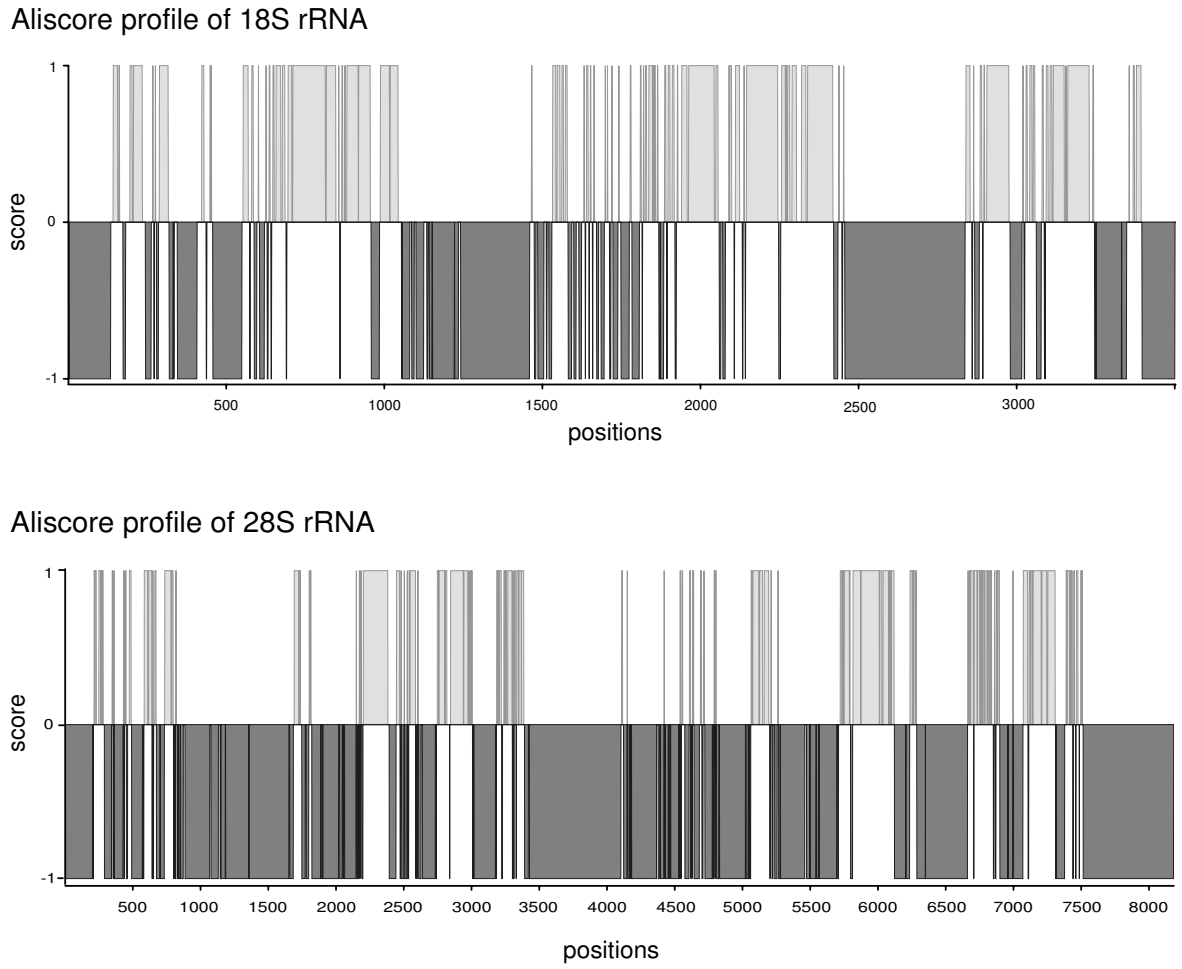


Figure 1
ALISCORE consensus profiles of rRNA alignments. **1A** ALISCORE consensus profile of the 18S rRNA alignment generated from single profiles of aligned positions after applying the sliding window approach based on MC resampling. Randomly similar sections (1873 positions) show negative score values or positive values non-random similarity (y-axis). Sequence length and positions are given on the x-axis. **1B** ALISCORE consensus profile of the 28S rRNA alignment generated from single profiles of aligned positions after applying the sliding window approach based on MC resampling. Randomly similar sections (5712 positions) show negative score values or positive values for non-random similarity (y-axis). Sequence length and positions are given on the x-axis.

fish/firebrats and pterygote insects) excluding Diptera share weaker patterns.

Compositional heterogeneity of base frequency

We excluded in *PAUP* 4.0b10 [66] parsimony uninformative positions explicitely for the base compositional heterogeneity test. Randomly similar alignment blocks identified by ALISCORE were excluded for both, the base compositional heterogeneity test and phylogenetic recontructions. 901 characters of the 18S rRNA and 1152 characters of the 28S rRNA were separately checked for inhomogeneous base frequencies. Results led to a rejection of the null hypothesis (H_0), which assumes homogeneous base composition among taxa (18S: $\chi^2 = 1168.94$, $df = 441$, $P = 0.00$; 28S: $\chi^2 = 1279.98$, $df = 441$, $P = 0.00$). Thus, base frequencies significantly differed across taxa in both 18S and 28S data sets.

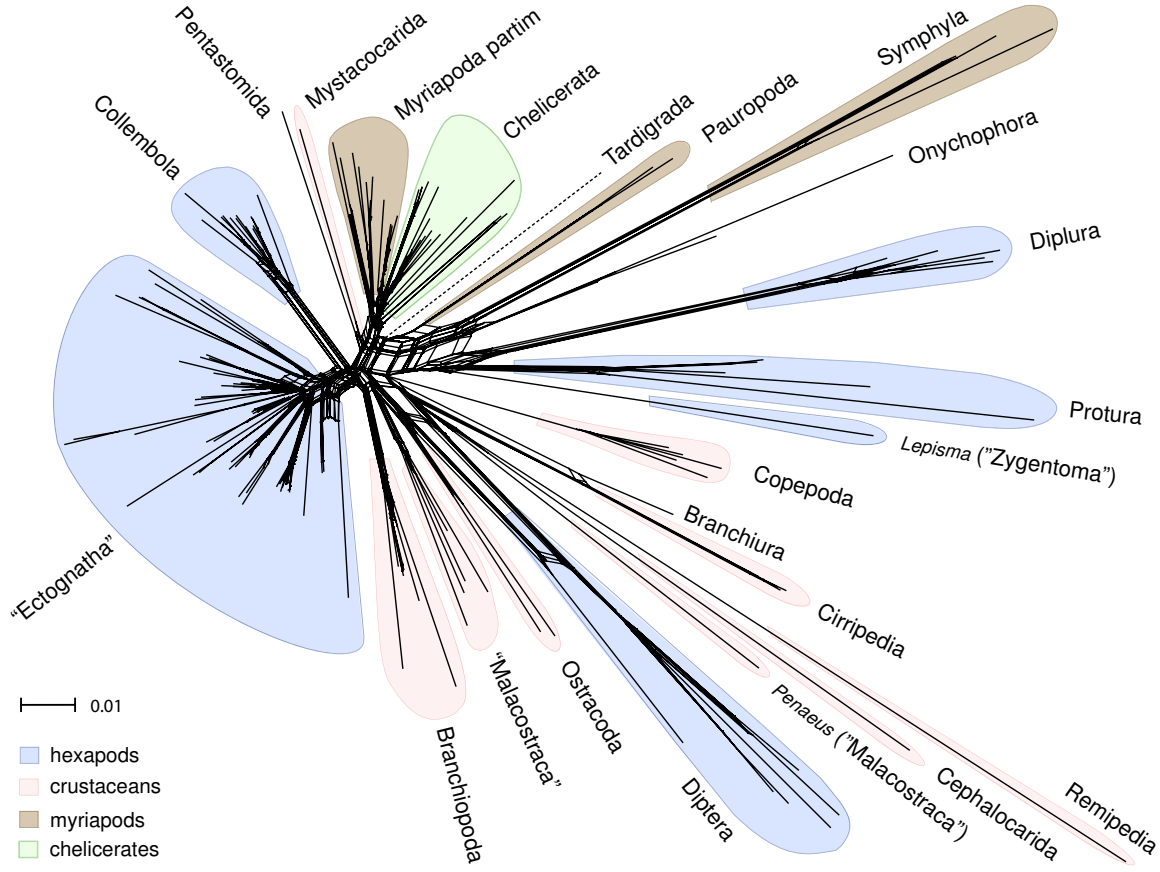


Figure 2
Neighbornet graph of the concatenated 18S and 28S rRNA alignment. Neighbornet graph based on uncorrected p-distances constructed in SplitsTree4 using the concatenated 18S and 28S rRNA alignment after exclusion of randomly similar sections evaluated with ALISCORE. Hexapods are colored blue, crustaceans red, myriapods brown and chelicerates green. Quotation marks indicate that monophyly is not supported in the given neighbornet graph.

A data partition into stems and loops revealed 477 unpaired positions and 424 paired positions in the 18S, and 515 unpaired and 637 paired positions in the 28S. Separate analyses of all four partitions confirmed heterogeneity of base frequencies across taxa in all sets ($P = 0.00$ in all four partitions).

We repeated the homogeneity test for partitions as used in tree reconstruction, if base pairs were disrupted by the identification of the corresponding partner as randomly similar (ALISCORE), remaining formerly paired positions were treated as unpaired. Hence, 1848 characters of the concatenated alignment (18S: 706; 28S: 1142) were treated as paired in all analyses. Again the test revealed heterogeneity in unpaired characters of both the 18S and 28S ($P = 0.00$ for both genes; 18S: 506 characters; 28S: 567 characters). Examination at paired positions also rejected the null hypothesis H_0 (18S, 395 characters included: $P < 0.0003$, 28S, 585 characters included: $P = 0.00$). Since non-stationary processes in all tests were strongly indicated, we chose to apply time-heterogeneous models to account for lineage-specific substitution patterns. To fix the number of "free base frequency sub-models" in time-heterogeneous analyses, we identified the minimal exclusive set of sequence groups. Based on χ^2-tests the dataset could be divided into three groups for both rRNA genes. In both genes Diptera are characterized by a high A/T content and Diplura by a low A/T content. Exclusion of only one of the groups was not sufficient to retain a homogeneous data set (18S: excluding Diptera: $\chi^2 = 972.91$, $df = 423$, $P = 0.00$, excluding Diplura: $\chi^2 = 532.13$, $df = 423$, $P < 0.0003$; 28S: excluding Diptera: $\chi^2 =$

986.72, *df* = 423, *P* = 0.00, excluding Diplura: χ^2 = 813.8, *df* = 423, *P* = 0.00). Simultaneous exclusion of both groups led to acceptance of H_0 for 18S sequences (χ^2 = 342.22, *df* = 405, *P* = 0.99). For the 28S, after exclusion of both groups, H_0 was still rejected (χ^2 = 524.98, *df* = 405, *P* < 0.0001). After sorting taxa according to base frequencies in ascending order, additional exclusion of *Peripatus* sp. and *Sinentomon erythranum* resulted in a homogeneous base composition for the 28S gene (H_0: χ^2 = 434.99, *df* = 399, *P* = 0.1), likewise indicating that three sub-models are suffucient to cover the taxon set. We repeated the homogeneity-test for stem and loop regions of each gene seperately. The exclusion of Diplura was sufficient to obtain homogeneity in the loop regions for both genes (18S: 474 characters, *P* = 0.9757; 28S: 541 characters, *P* = 0.0684). For stem regions in the 18S it likewise was sufficient to exclude either Diptera (378 characters, *P* = 0.6635) or Diplura (385 characters, *P* = 0.99). These partitions would make two sub-models sufficient to cover the data set. However, in the stem regions of the 28S homogeneity was received only after the exclusion of both Diptera and Diplura (547 characters, *P* = 0.99). Since *PHASE-2.0* does not allow to vary the number of chosen sub-models among partitions, we applied and fitted three sub-models to each data partition.

Phylogenetic reconstructions

Three combinations of mixed DNA/RNA models (REV + Γ & RNA16I + Γ, TN93 + Γ & RNA16J + Γ and HKY85 + Γ & RNA16K + Γ) were compared to select the best model set. Overall model *ln* likelihoods converged for all tested mixed models after a burn-in of 250,499 generations in an initial pre-run of 500,000 generations. However, most parameters did not converge for the combined REV + Γ & RNA16I + Γ models, consequently, this set up was excluded from further analyses. For each of the remaining two sets a chain was initiated for 3 million generations, with a burn-in set to 299,999 generations. The applied Bayes Factor Test [[67,68], BFT], favored the TN93 + Γ & RNA16J + Γ model combination ($2lnB_{10}$ = 425.39, harmonic mean lnL_0(TN93 + Γ & RNA16J + Γ) = 79791.08; harmonic mean lnL_1(HKY85 + Γ & RNA16K + Γ) = 80003.78). For each approach (Aditional file 3) all chains which passed a threshold value in a BFT were assembled to a metachain. Each resulting extended majority rule consensus tree was rooted with *Milnesium*. Node support values for clades were deduced from 56,000 sampled trees for the time-heterogeneous set (Figure 3) and from 18,000 sampled trees for the time-homogeneous set (Figure 4), detailed support values are shown in Additional file 3. Harmonic means of the *ln* likelihoods of included time-heterogeneous chains were compared against all *ln* likelihoods of included time-homogeneous chains (burn-in discarded) in a final BFT: the time-heterogeneous model was strongly favored ($2lnB_{10}$ = 1362.13).

Resulting topologies

Representatives of Symphyla and Pauropoda, already identified in the neighbornet graph as taxa with conspicuously long branches (Figure 2), assumed unorthodox positions in both trees which are clearly incongruent with morphological evidence and results obtained from other genes. Symphyla formed the sister group of all remaining arthropod clades, and Pauropoda clustered with Onychophora. Consequently, myriapods always appeared polyphyletic in both analyses. We consider these results as highly unlikely, since they contradict all independent evidence from morphology, development, and partly from other genes. In the following, we focus on major clades and point out differences between time-heterogeneous tree (Figure 3) and time-homogeneous tree (Figure 4) without considering the position of Symphyla and Pauropoda. Possible causes for the misplacement of these groups, however, will be treated in the discussion. Both analyses supported a monophyletic Chelicerata (pP 0.91 in the time-heterogeneous tree and maximal support in the time-homogeneous tree) with Pycnogonida (sea spiders) as sister group to remaining chelicerates. Pycnogonida received maximal support in both analyses. Euchelicerata received highest support in the time-homogeneous approach while this clade in the time-heterogeneous approach received a support of only pP 0.89. *Limulus polyphemus* (horseshoe crab) clustered within arachnids, but some internal relationships within Euchelicerata received only low support. Chilopoda always formed the sister group of a monophyletic Diplopoda in both analyses with high support. Within the latter the most ancient split lied between Penicillata and Helminthomorpha. This myriapod assemblage – Myriapoda partim – formed the sister group of Chelicerata, thus giving support to the Myriochelata hypothesis, respectively Myriochelata partim, when the long-branch clades Symphyla and Pauropoda are disregarded.

Pancrustacea showed always maximal support. The monophyly of Malacostraca and Branchiopoda received highest support in both approaches while their position varied. Branchiopoda was the sister group of the clade consisting of Copepoda + Hexapoda in the homogeneous tree (Figure 4), however the cephalocarid *Hutchinsoniella* nested within hexapods. Among hexapods, monophyly was unambiguously supported for Protura, Diplura, Collembola, Archaeognatha, Odonata, Ephemeroptera, Phasmatodea, Mantophasmatodea, Mantodea, Plecoptera, Hemiptera, Coleoptera, Hymenoptera, Lepidoptera and Diptera. Diplura clustered with Protura, and gave support to a monophyletic Nonoculata. Pterygota occurred in both topologies, well supported in the non-stationary tree (pP 0.97) and with moderate support (pP 0.94) in stationary tree. Within the winged insects, both analyses resolved Odonata as the sister group to a well supported mono-

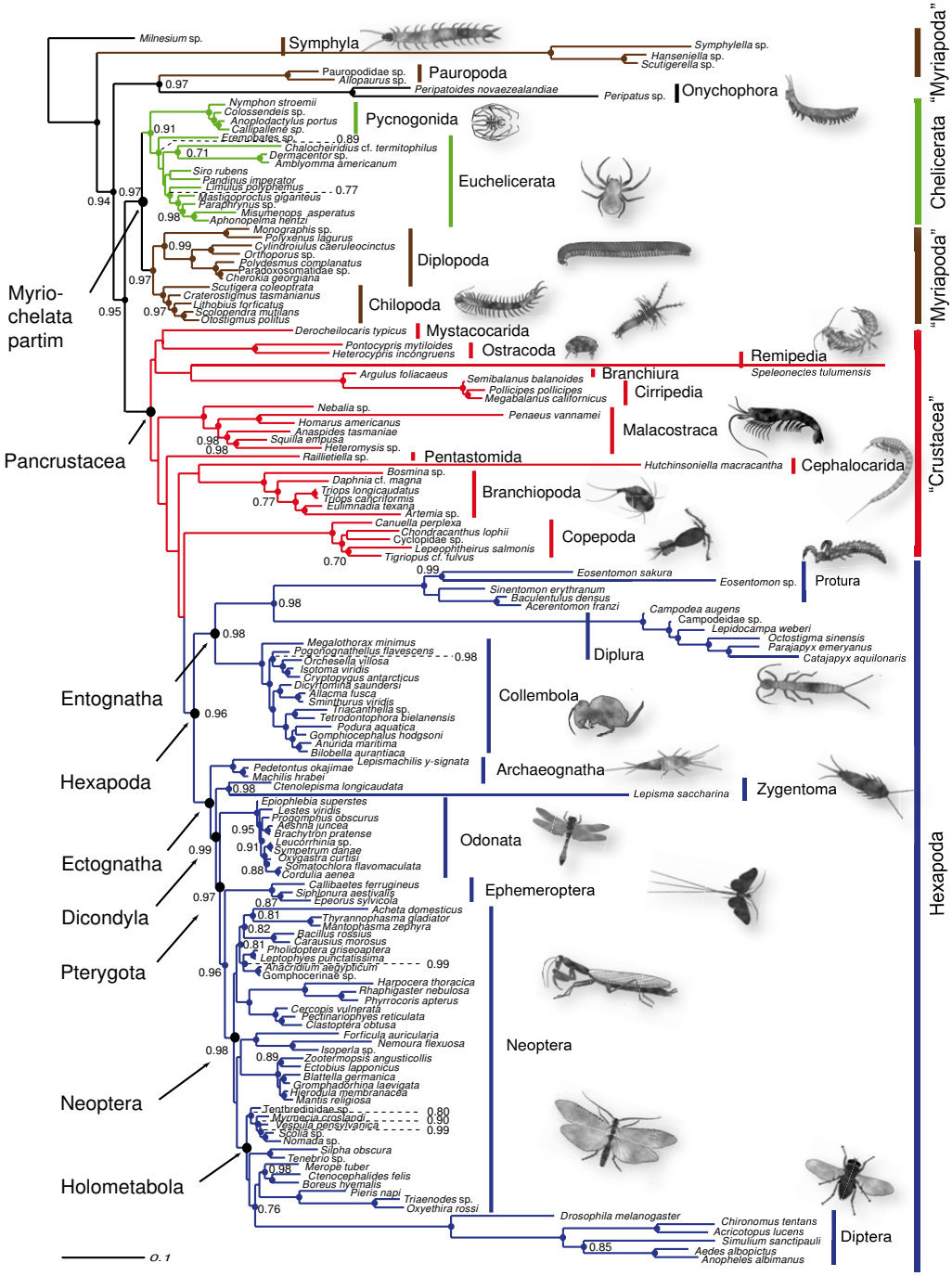


Figure 3

Time-heterogeneous consensus tree. Consensus tree from 56,000 sampled trees of the time-heterogeneous substitution process inferred by *PHASE-2.0*, graphically processed with Adobe Illustrator CS2. Support values below 0.70 are not shown (nodes without dots), nodes with a maximum posterior probability (pP) of 1.0 are represented by dots only. Quotation marks indicate that monophyly is not supported in the given tree.

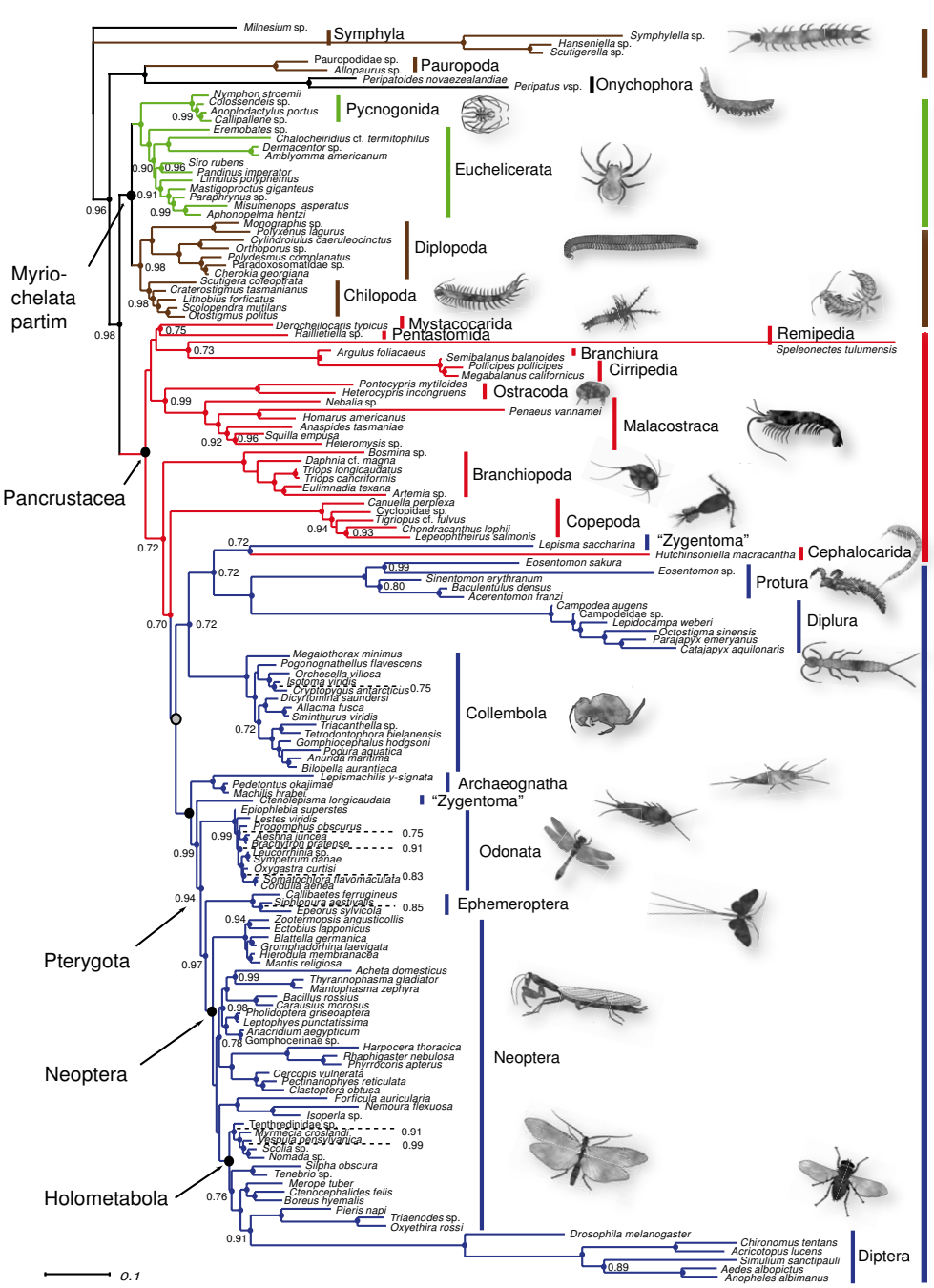


Figure 4

Time-homogeneous consensus tree. Consensus tree from 18,000 sampled trees of the time-homogeneous substitution process inferred by *PHASE-2.0*, graphically processed with Adobe Illustrator CS2. Support values below 0.70 are not shown (nodes without dots), nodes with a maximum posterior probability (pP) of 1.0 are represented by dots only. The grey dot indicates the clade containing all hexapod taxa including *Hutchinsoniella* (Crustacea) + *Lepisma* (Zygentoma); its node value is pP 0.58. Quotation marks indicate that monophyly is not supported in the given tree.

phyletic clade Ephemeroptera + Neoptera (heterogeneous: pP 0.96; homogeneous: pP 0.97), known as the "Chiastomyaria" clade [32,34,35,69]. Blattodea were always paraphyletic with respect to the isopteran representative. This assemblage formed a sister group relationship with Mantodea, thus giving support to a monophyletic Blattopteroidea or Dictyoptera while the position of Dictyoptera among hemimetabolan insects differed. Dermaptera always clustered with Plecoptera. Hemiptera (Heteroptera + Homoptera) in both approaches formed a clade with the remaining orthopterans + ((*Acheta* + Mantophamsmatodea)Phasmatodea) with low statistical support. Caused by *Acheta* orthopteran insects appeared always polyphyletic. Within the monophyletic Holometabola (pP 1.0), Hymenoptera formed the sister group of the remaining taxa.

While the time-heterogeneous and time-homogeneous trees corresponded in overall topologies, they differed in a number of remarkable details.

1) Hexapoda, Entognatha, Ectognatha and Dicondylia were only reconstructed in the time-heterogeneous approach. 2) The cephalocarid *Hutchinsoniella* clustered among crustaceans as sister group to the Branchiopoda only in the heterogeneous approach, this clade formed the sister group to (Copepoda + Hexapoda) although with low support. 3) The time-homogeneous runs revealed highly supported (Malacostraca + Ostracoda) as the sister group to a clade ((Mystacocarida + Pentastomida) + (Branchiura + Cirripedia)). In contrast, in the time-heterogeneous analysis more terminal positioned Malacostraca are the sister group of a clade (Pentastomida((Cephalocarida + Branchiopoda) + (Copepoda + Hexapoda))). The altered postition of Pentastomida was only low supported in this tree. 4) In the homogeneous tree *Hutchinsoniella* emerged as sister taxon to *Lepisma* with low support (pP 0.72), and this cluster was positioned within the remaining hexapods (Figure 4). Hexapoda were monophyletic only in the time-heterogeneous approach, well supported (pP 0.96, Figure 3), with Copepoda as sister group, latter with low support (pP 0.69). 5) In the time-homogeneous tree (Figure 4), Copepoda emerged as sister group, again with a low support value (pP 0.70) of ((*Lepisma* + *Hutchinsoniella*) + "Hexapoda"). 6) Entognatha (pP 0.98), and Ectognatha (pP 1.0) and Dicondylia (pP 0.99) were monophyletic only in the time-heterogeneous tree. 7) The time-heterogeneous tree showed the expected paraphyly of primarily wing-less insects with Archaeognatha as sister group to Zygentoma + Pterygota. 8) Within pterygote insects (Dermaptera + Plecoptera) emerged as sister group of Dictyoptera in the non-stationary tree, contrary as sister group of Holometabola in the stationary tree, both scenarios with negligible support.

Discussion

Among arthropods 18S and 28S rRNA genes have the densest coverage of known sequences. Apart of some exceptions most studies on phylogenetic relationships at least partly rely on rRNA data. Often, however, only one of the genes was used, sometimes even just fragments of a gene [23,32,34,40,42,44,70-72], while only few studies used nearly complete 18S and 28S rRNA sequences [1,11,73]. Despite this wide usage, the reliability of reconstructions based on rRNA markers is still debated (for contradicting views see [34,74,75]. A major cause of concern is the pronounced site heterogeneity of evolutionary rates, the non-stationarity of base composition among taxa and rate variation in time. All three phenomena quickly lead to the erosion of phylogenetic signal [76]. On the one hand, our understanding of the molecular structure of other markers and about taxon-dependent processes of molecular evolution remains poor. On the other hand, our vast background knowledge regarding rRNA molecules offers a unique opportunity to study the effects of selection and application of substitution models in greater detail.

Quality check and character choice in alignments

Phylogenetic signal in sequence data can get noisy due to (i) multiple substitution processes (saturation) and (ii) erroneous homology hypotheses caused by ambiguous sequence alignment. Both effects correspond in that they result in random similarity of alignment regions. Such noisy sections potentially bias tree reconstructions in several ways which have been appreciated for years but only recently been applied, that allow to account for these problems [25,54,77,78]. Exclusion of these ambiguously aligned or saturated regions can help to reduce noise, see e.g. [65]. If this topic is addressed at all, the majority of studies include a manual alignment check for untrustworthy regions [1,4,22,32,34,39,44,71-73]. Only some recent publications addressing arthropod relationships have used automated tools, e.g. [14,79,80].

To identify alignment sections of random similarity prior to tree reconstructions, we used ALISCORE, which, compared to the commonly used Gblocks [81], is not dependent on the specification of an arbitrary threshold [65]. To improve the signal-to-noise ratio we restricted our character choice to alignment sections which contained nucleotide patterns that differ from randomized patterns.

Phylogenetic reconstruction methods

Arthropod phylogenies have been inferred with reconstruction methods like Maximum Parsimony, Maximum Likelihood and Bayesian approaches. We tried to implement knowledge about the evolution of rRNA in two ways: (i) the use of mixed DNA/RNA models is meant to

account for known instances of character dependence due to compensatory mutations in stem regions, (ii) the application of time-heterogeneous models accounts for non-stationary processes that occurred in arthropod lineages. The consensus secondary structure of our dataset, generated with RNAsalsa, can be understood as a model parameter that defines site interactions and thus character dependence due to compensatory mutations [34,82,83]. Neglect of character dependence surely results in unrealistic support values. In single low supported nodes, where the signal-to-noise ratio is at the edge of resolution, such a neglect theoretically can even turn the balance between two competing hypotheses. Additionally a consensus secondary structure is necessary to apply a mixed model approach, since it determines whether the evolution of a given site is modeled by the DNA-model, or as part of a base-pair by the RNA-model. Within the mixed model approach, we opted for DNA-corresponding 16-state RNA models [63]. It can certainly be argued that the choice of 16-state models is problematic because it is difficult to fit these models to real data due to their parameter richness and heavy computational costs. However, even the best choice of a consensus secondary structure can only capture the predominantly conserved structural features among the sequences. This implies that the applied RNA models must be able to cope with mismatches in base-pairing. Less complex RNA models like those of the 6 and 7-state families either ignore mismatches completely or pool these mismatches into a single character state which produces artificial synapomorphies. Additionally, according to Schöninger and v. Haeseler [84], it is more likely that co-variation is a multiple step process which allows for the intermediate existence of instable (non Watson-Crick) pairs. These intermediate states are only described in 16-state RNA models.

Concerning rRNA-genes of arthropods, shifts in base composition are mentioned for Diptera, Diplura, Protura and Symphyla [1,23,34,44,73,85]. Since base compositional heterogeneity within a dataset can mislead phylogenetic reconstruction [61,86,87] and [60], some of these studies discussed observed but not incorporated non-stationary processes as possible explanations for misplacements of some taxa [11,23,24,44,73]. The selective exclusion of these taxa to test for misleading effects on the remaining topology, however, is not appropriate to test whether non-stationarity really fits as the causal explanation of the placement incongruent with other analyses. LogDet methods have been applied to compensate for variations of base frequencies [1,11,44], which leads to some independence of non-stationarity, but among site rate variation (ASRV) cannot be handled efficiently. After detecting compositional base frequency heterogeneity in our data, we chose a non-stationary approach implemented in *PHASE-2.0*. Because no previous study of arthropod phylogeny has used a time-heterogeneous approach including mixed DNA/RNA models, we compared this approach with a "classical" time-homogeneous setup. Our results prove that the time-heterogeneous approach produces improved likelihood values with improved branch lengths estimates and more realistic, though not perfect (see below), topology estimates. Since modeling of general time-heterogeneous processes is in its infancy and since its behavioural effect on real data is relatively unknown [60,61], we favored a set up accounting for the three different "submodels" corresponding to three base frequency categories in our dataset (Additional file 4). The application of the three submodels to individual branches in a tree by the MCMC process was not further constrained. This scheme allowed for a maximum of flexibility without losing the proper mix of parameters.

Conflicting phylogenetic hypotheses and non-stationary processes of rRNA evolution

The comparison of our time-homogeneous approach to our time-heterogeneous one was not only meant to show improvements in the application of more realistic models, but also to indicate which incongruities of analyses of rRNA genes may be causally explained by non-stationary processes during the evolution of these genes.

In our time-homogeneous approach, the crustacean *Hutchinsoniella* (Cephalocarida) clustered with *Lepisma* (Zygentoma, Hexapoda) within enthognathans as sister group to Nonoculata (Protura + Diplura), (see Figure 4). This led to the polyphyly or paraphyly of several major groups (e.g. Hexapoda, Entognatha, Ectognatha, Dicondylia). In our time-heterogeneous analysis, Cephalocarida clustered as sister group to Branchiopoda. This result, although marginal supported, is congruent, at least, with some morphological data [88]. Most recent molecular studies have not included Cephalocarida, e.g. [1,11]. Regier et al. [12] reconstructed a sister group relationship of Remipedia and Cephalocarida (likewise represented by *Hutchinsoniella*), but his result also received only moderate bootstrap support. The same clade was presented in Giribet et al. [9] based on morphological and molecular data.

Independent of the sister group relationship of Cephalocarida within crustaceans, the correction of the misplacement of *Hutchinsoniella*, by allowing for non-stationary processes, has a major effect on the heuristic value of our analyses. Not only is the monophyletic status of Hexapoda, Entognatha, Ectognatha, Dicondylia supported after the correction, but likewise a causal explanation is given for the misplacement in the time-homogeneous approach, which cannot be accomplished by alternatively

excluding the taxon. Our time-heterogeneous analyses resulted in a sister group relationship of Diplura and Protura, which lends support to a monophyletic Nonoculata within a monophyletic Entognatha. This result is congruent with trees published by Kjer [32], Luan et al. [44], Mallat and Giribet [1], and Dell'Ampio et al. [23]. Following Luan et al. [44] Dell'Ampio et al. [23] cautioned that Nonoculata may be an artificial cluster caused by a shared nucleotide bias and long branch attraction. Since this node is recovered with high support by our non-stationary approach, Nonoculata cannot be suspected of being an artificial group based on shared compositional biases alone. However, one must keep in mind that Protura and Diplura have longer branches than Ectognatha and Collembola (Figure 3 and 4), and long-branch effects may still be present. Thus monophyly of a clade Nonoculata still awaits support from a data set independent from rRNA sequences.

Clades not affected by non-stationary processes

Symphyla and Pauropoda

Although we tried to break down long branches by a dense taxon sampling, some long-branch problems persisted. We cannot clearly address the reason but, due to the symptoms, assume that saturation by multiple substitution caused signal erosion (class II effect, [25]). To evaluate the impact on the topology of the very likely incorrect positions of Symphyla and Pauropoda, we repeated the time-heterogeneous analysis using a reduced dataset excluding these taxa. We limited the analysis to ten chains with 7, 000, 000 generations each (2, 000, 000 burn-in). Differences occurring in the inferred consensus topology (not shown) of the final three chains (15, 000, 000 generations) show that some nodes are still sensitive to taxon sampling, since e.g. Pycnogonida clustered with (Chilopoda + Diplopoda) after exclusion of pauropod and symphylan sequences. Also the crustacean topology changed, remaining long branch taxa *Hutchinsoniella* and *Speleonectes* clustered together in the reduced dataset, forming a clade with (Branchiura + Cirripedia).

Mandibulata versus Myriochelata

Analyses of rRNA sequences up till now were held to favor Myriochelata (Myriapoda + Chelicerata) over Mandibulata [1,4,11]. Our analyses provide no final conclusion with respect to this conflict, since the position of Pauropoda and Symphyla is unusual, it results in polyphyletic myriapods. The exact reconstruction of the position of myriapods within the Euarthropoda thus demands e.g. the application of new markers and suitable phylogenetic strategies.

Phylogenetic position of Malacostraca and Pentastomida

The position of Malacostraca differs among molecular studies. Often, Malacostraca emerge as nested within the remaining crustacean groups, e.g. [5,89]. Complete mitochondrial genomes place Malacostraca close to insects [90,91]. However, studies of rRNA sequences recover this group as the sister group to all remaining crustaceans [1,11,92]. Since in our stationary tree monophyletic Malacostraca branched off at a more basal split within crustaceans [88,93], forming a sister group relationship to Ostracoda and contrary they branched off at a more terminal split in the non-stationary tree we cannot draw a final conclusion about the placement of Malacostraca. Unfortunately the position of the Pentastomida remains ambiguous in our analyses, we argue that low pP values might be induced by conflicting phylogenetic signal.

Sister group of Hexapoda

The sister group of Hexapoda is still disputed. Most molecular studies support paraphyly of crustaceans with respect to hexapods. A sister group relationship between Branchiopoda and Hexapoda was proposed for the first time by Regier and Shultz [94], yet with low support. Shultz and Regier [5] and Regier et al. [12] corroborated this relationship, which is likewise favored by authors of rRNA-based studies [1,11], despite their result that Cyclopidae (Copepoda) is the sister group of Hexapoda. Our denser taxon sampling further supports Copepoda as the sister group to Hexapoda, but the low support value might indicate conflicting signal. This clade up till now, however, lacks any support from morphological studies.

Ancient splits within pterygote insects

We find that the rRNA data cannot robustly resolve the most ancient splits within Pterygota. Nonetheless, rRNA data, when analyzed under more realistic models favour Chiastomyaria as the most likely hypothesis. Since all three possible arrangements of Odonata, Ephemeroptera and Neoptera likewise receive morphological support, we agree with Whitfield and Kjer [35] that the ambiguity can best be explained by early 'explosive radiation' within Pterygota.

Conclusion

We conclude that the implementation of biologically realistic model parameters, such as site interaction (mixed DNA/RNA models) and compositional heterogeneity of base frequency, is fundamental to robustly reconstruct phylogenies. The most conspicuous examples comparing our tress are a) the position of *Hutchinsoniella* (Crustacea), although a low pP value of 0.59 in the non-stationary tree prohibits conclusions about its internal crustacean relationship and b) the well supported position of *Ctenolepisma* and *Lepisma* (Zygentoma). As a consequence, the monophyly of Hexapoda, Entognatha and Ectognatha and Dicondylia received support only in the time-heterogeneous approach. Sev-

eral artificial clades remain in our analyses which cannot be causally explained unambiguously. However, the examples given here clearly demonstrate that the probability to causally explain some incongruities between different data sets, as well as the correction of certain obvious misplacements, is enhanced by using more complex but realistic models. The present study aimed to incorporate background knowledge on the evolution of molecular sequences in general and ribosomal RNA-genes in special into various steps of data processing. For all steps fully automated methods were used, including an automated secondary structure guided alignment approach, a software that enables to distinguish random similarity from putative phylogenetic signal, mixed models that avoid artefacts due to co-variation among sites, and analyses that account for variation of evolutionary rates among lineages. The resolution of many relationships among arthropods, and the minimization of obvious misplacements demonstrate that the increased computational effort pays off.

Methods

Taxon Sampling

Our taxon sampling was designed to represent a taxonomically even collection of specimens across arthropod groups. In particular, we took care to include taxa which do not differ too widely from the hypothetical morphological ground-pattern of the represented group, when possible [53,78]. In total we included 148 concatenated 18S and 28S rRNA sequences in the analysis (Additional file 1). Of these, we contributed 103 new sequences, 41 for the 18S and 62 for the 28S rRNA gene, respectively. Only sequences which span at least 1500 bp for the 18S and 3000 bp for the 28S were included. For 29 taxa we had to construct chimeran concatenated sequences of 18S and 28S rRNA sequences of different species, marked with an asterisk. Details are listed in Additional file 5, we chose species as closely related as possible depending on it's availability in GenBank. The outgroup included the concatenated 18S and 28S rRNA sequences of *Milnesium* sp. (Tardigrada).

Laboratory work

Collected material was preserved in 94 – 99% ethanol or liquid nitrogen. Samples were stored at temperatures ranging from -20°C to -80°C. DNA extraction of complete specimens or tissue followed different standard protocols. We used phenol-chloroform isoamyl extraction [95], standard column DNA extraction kits DNeasy Blood & Tissue Kit (Qiagen) and NucleoSpin Tissue Kit (Machery-Nagel) following the manual. Single specimens were macerated for extraction, only specimens of *Ctenocephalides felis* were pooled. Manufacturer protocols were modified for all crustaceans, some apterygote hexapods and myriapods (overnight incubation and adding 8 μl RNAse [10 mg/ml] after lysis). Extracted genomic DNA was amplified with the Illustra GenomiPhi V2 DNA Amplification Kit (GE Healthcare) for tiny, rare or hard to collect specimens.

Partly published rRNA primer sets were used, they were designed in part for specific groups (Additional file 6 and 7). The 18S of crustaceans was amplified in one PCR product and sequenced using four primer combinations. The 18S of apterygotes was amplified in three or four fragments (Additional file 8). The 28S of crustaceans and basal hexapods was amplified in nine overlapping fragments starting approximately in the middle of the rRNA 5.8S to the nearly end of the D12 of 28S rRNA (Additional file 9). The 28S of odonats was amplified in seven or eight, the 28S of ephemeropterans and neopterans in eight overlapping fragments (Additional file 10). Primers were ordered from Metabion, Biomers or Sigma-Genosys. PCR products were purified using following kits: NucleoSpin ExtractionII (Machery-Nagel), QIAquick PCR purification kit (Qiagen), peqGOLD Gel Extraction Kit (peqLab Biotechnologie GmbH), MultiScreen PCR Plate (Millipore) and ExoI (Biolabs Inc.)/SAP (Promega). Some samples were purified using a NHAc [4 mol] based ethanol precipitation. In case of multiple bands fragments with the expected size were cut from 1% – 1.5% agarose gel and purified according to manufacturer protocols.

Cycle sequencing and sequence analyses took place on different thermocyclers and sequencers. Cycle sequencing products were purified and sequenced double stranded. Several amplified and purified PCR products were sequenced by Macrogen (Inc.), Korea. Sequencing of the 28S fragment 28V – D10b.PAUR of the Pauropodidae sp. (Myriapoda) was only successful via cloning. Fragments of the 28S rRNA of the diplopod Monographis sp. (Myriapoda) were processed following Mallatt et al. [11] and Luan et al. [44]. Please refer to the electronic supplement (Additional file 11) for detailed information about PCR-conditions, applied temperature profiles (Additional file 12), primer combinations, used chemicals (Additional file 13) and settings to amplify DNA fragments. Sequence electropherograms were analyzed and assembled to consensus sequences applying the software SeqMan (DNASTar Lasergene) or BioEdit 7.0 [96]. All sequences or composed fragments were blasted in NCBI using BLASTN, mega BLAST or "BLAST 2 SEQUENCES" [97] to exclude contaminations.

Alignments and alignment evaluation

Secondary structures of rRNA genes were considered (as advocated in [98-101] to improve sequence alignment. Structural features are the targets of natural selection, thus

the primary sequence may vary, as long as the functional domains are structurally retained. Alignments and their preparation for analyses were executed for each gene separately. We prealigned sequences using MUSCLE v3.6 [102]. Sequences of 24 taxa of Pterygota were additionally added applying a profile-profile alignment [103]. The 28S sequences of *Hutchinsoniella macracantha* (Cephalocarida), *Speleonectes tulumensis* (Remipedia), *Raillietiella* sp. (Pentastomida), *Eosentomon* sp. (Protura) and *Lepisma saccharina* (Zygentoma) were incomplete. Apart from *L. saccharina*, prealignments of these taxa had to be corrected manually. We used the "BLAST 2 SEQUENCES" tool to identify the correct position of sequence fragments in the multiple sequence alignment (MSA) for these incomplete sequences.

The software RNAsalsa [56] is a new approach to align structural RNA sequences based on existing knowledge about structure patterns, adapted constraint directed thermodynamic folding algorithms and comparative evidence methods. It automatically and simultaneously generates both individual secondary structure predictions within a set of homologous RNA genes and a consensus structure for the data set. Successively sequence and structure information is taken into account as part of the alignment's scoring function. Thus, functional properties of the investigated molecule are incorporated and corroborate homology hypotheses for individual sequence positions. The program employs a progressive multiple alignment method which includes dynamic programming and affine gap penalties, a description of the exact algorithm of RNAsalsa will be presented elsewhere.

As a constraint, we used the 28S + 5.8S (U53879) and 18S (V01335) sequences and the corresponding secondary structures of *Saccharomyces cerevisae* extracted from the European Ribosomal Database [104-106]. Structure strings were converted into dot-bracket-format using Perl-scripts. Folding interactions between 28S and 5.8S [74,107,108] required the inclusion of the 5.8S in the constraint to avoid artificial stems. Alignment sections presumably involved in the formation of pseudoknots were locked from folding to avoid artifacts. Pseudoknots in *Saccharomyces cerevisae* are known for the 18S (stem 1 and stem 20, V4-region: stem E23 9, E23 10, E23 11 and E23 13) while they are lacking in the 28S secondary structure. Prealignments and constraints served as input, and RNAsalsa was run with default parameters. We constructed manually chimeran 18S sequences of *Speleonectes tulumensis* (EU370431, present study and L81936) and 28S sequences of *Raillietiella* sp. (EU370448, present study and AY744894). Concerning the 18S of *Speleonectes tulumensis* we combined positions 1–1644 of L81936 and positions 1645–3436 of sequence EU370043. Regarding the 28S of *Raillietiella* we combined positions 1–3331 of AY744894 with positions 3332–7838 of sequence EU370448. Position numbers refer to aligned positions.

RNAsalsa alignments were checked with ALISCORE[65]. ALISCORE generates profiles of randomness using a sliding window approach. Sequences within this window are assumed to be unrelated if the observed score does not exceed 95% of scores of random sequences of similar window size and character composition generated by a Monte Carlo resampling process. ALISCORE generates a list of all putative randomly similar sections. No distinction is made between random similarity caused by mutational saturation and alignment ambiguity. A sliding window size ($w = 6$) was used, and gaps were treated as ambiguities (- N option).

The maximum number of possible random pairwise comparisons (- r: 10,878) was analyzed. After the exclusion of putative random sections and uninformative positions using *PAUP* 4.0b10, alignments were checked for compositional base heterogeneity using the χ^2-test. Additionally, for each sequence the heterogeneity-test was performed for paired and unpaired sites separately. Further heterogeneity-tests were applied to determine the minimal number of base frequency groups.

RNAsalsa generated consensus structure strings for 18S and 28S rRNA sequences, subsequently implemented in the MSA. Randomly similar sections identified by ALISCORE were excluded using a Perl-script. ALISCORE currently ignores base pairings. If ambiguously aligned positions within stems are discarded the corresponding positions will be handled as an unpaired character in the tree reconstruction. The cleaned 18S and 28S alignments were concatenated.

To analyze information content of raw data SplitsTree4 was used to calculate phylogenetic networks (see Huson and Bryant [109] for a review of applications). We compared the network structure based on the neighbornet algorithm [110] and applying the LogDet transformation, e.g. [111,112]. LogDet is a distance transformation that corrects for biases in base composition. The network graph gives a first indication of signal-like patterns and conflict present in the alignments. We used the alignment after filtering of random-like patterns with ALISCORE.

Phylogenetic reconstruction

Mixed DNA/RNA substitution models were chosen, in which sequence partitions corresponding to loop regions were governed by DNA models and partitions corresponding to stem regions by RNA models that consider co-variation. Among site rate variation [113] was imple-

mented in both types of substitution models. Base frequency tests indicated that base composition was inhomogeneous among taxa (see results), suggesting non-stationary processes of sequence evolution. To take such processes into account the analyses were performed in *PHASE-2.0* [63] to accommodate this compositional heterogeneity to minimize bias in tree reconstruction. Base compositional heterogeneity is implemented in *PHASE-2.0* according to the ideas developed by Foster [87].

We limited the number of candidate models to the REV + Γ, TN93 + Γ and the HKY85 + Γ models for loop regions and the corresponding RNA16I + Γ, RNA16J + Γ and RNA16K + Γ models for stem regions. Site heterogeneity was modeled by a discrete gamma distribution [114] with six categories. The extent of invariant characters was not estimated since it was shown to correlate strongly with the estimation of the shape parameter of the gamma distribution [113,115-117]. The data was partitioned into four units representing loop and stem regions of 18S rRNA and loop and stem regions of 28S rRNA. DNA and RNA substitution model parameters were independently estimated for each partition. Substitution models were selected based on results of time-homogeneous setups. We tested three different combinations of substitution models, REV + Γ & RNA16I + Γ, TN93 + Γ & RNA16J + Γ and HKY85 + Γ & RNA16K + Γ. We used Dirichlet distribution for priors, proposal distribution and Dirichlet priors and proposals for a set of exchangeability parameters (Additional file 14) described in Gowri-Shankar and Rattray [60].

Appropriate visiting of the parameter space according to the posterior density function [118] was checked by plotting values of each parameter and monitoring their convergence. This was calculated for all combinations after 500,000 generations (sampling period: 150 generations). We discarded models in which values of several parameters did not converge. For models which displayed convergence of nearly all parameter values, we re-run MCMC processes with 3,000,000 generations and a sampling period of 150 generations. Prior to comparison of the harmonic means of *lnL* values, 299,999 generations were discarded as burn-in. After a second check for convergence the model with the best fitness was selected applying a Bayes Factor Test (BFT) to the positive values of the harmonic means calculated from *lnL* values [67,68]. The favored model ($2lnB_{10} > 10$) was used for final phylogenetic reconstructions.

To compare results of time-homogeneous and time-heterogeneous models, 14 independent chains of 7,000,000 generations and two chains of 10 million generations for both setups were run on Linux clusters (Pentium 4, 3.0 GHz, 2 Gb RAM, and AMD Opteron Dual Core, 64 bit systems, 32 Gb RAM). For each chain the first two million generations were discarded as burn-in (sampling period of 1000). The setup for the time-homogeneous approach was identical to the pre-run except for number of generations, sampling period and burn-in. The setting for the time-heterogeneous approach differed (Additional file 4). We followed the method of Foster [87] and Gowri-Shankar and Rattray [60] in the non-homogeneous setup whereby only a limited number of composition vectors can be shared by different branches in the tree. Exchangeability parameters (average substitution rate ratio values, rate ratios and alpha shape parameter) were fixed as input values. Values for these parameters were computed from results of the preliminary time-homogeneous pre-run (3,000,000 generations). A consensus tree was inferred in *PHASE mcmcsummarize* using the output of the pre-run. This consensus tree topology and the model file of this run served as input for a ML estimation of parameters in *PHASE optimizer*. Estimated values of exchangeability parameters from the resulting *optimizer* output file and estimated start values for base frequencies were fed into *mcmcphase* for the time-heterogeneous analysis. Values of exchangeability parameters remained fixed during the analysis. The number of allowed base frequency categories (models) along the tree was also fixed. The number of base frequency groups was set to three "submodels"), reflecting base frequency heterogeneity.

Harmonic means of *lnL* values of these 16 independent chains were again compared with a BFT to identify possible local optima in which a single chain might have been trapped. We only merged sample data of chains with a $2lnB_{10}$-value < 10 [67] using a Perl-script to construct a "metachain" [119]. Finally we included ten time-heterogeneous chains and three time-homogeneous chains. The assembled meta-chains included 56 million generations for the non-stationary approach (Additional file 15) and 18 million generations for the time-homogeneous approach (Additional file 16), burn-ins were discarded. Consensus trees and posterior probability values were inferred using *mcmcsummarize*. Branch lengths of the time-homogeneous and time-heterogeneous consensus tree were estimated using three *mcmcphase* chains (4 million generations, sampling period 500, topology changes turned off, starting tree = consensus tree, burn-in: 1 million generations) from different initial states with a Gowri-Shankar modified *PHASE* version. To infer mean branch lengths we combined data with the described branch lengths and *mcmcsummarize*. These mean branch lengths were used to redraw the consensus tree (Additional file 4).

Localities of sampled specimen used for amplication are listed in Additional file 17.

List of abbreviations

rRNA: ribosomal RNA; PCR: polymerase chain reaction; RNA: ribonucleic acid; DNA: deoxyribonucleic acid; df: degree of freedom; P: probability; pP: posterior probability; sp.: species epithet not known; *ln*: natural logarithm or log_e; BFT: Bayes Factor Test.

Authors' contributions

BMvR, KM and BM conceived the study, designed the setup and performed all analyses. VG complemented *PHASE-2.0* and contributed to *PHASE-2.0* analyses setup. RRS, HOL, BM provided RNAsalsa and software support. JWW allocated the neighbornet-analysis. BMvR, KM, ED, SS, HOL, DB and YL contributed sequence data and designed primers. BMvR, KM, BM, NUS and JWW wrote the paper with comments and revisions from ED, VG, RRS, DB, SS, GP, HH and YL. All authors read and approved the final manuscript.

Additional material

Additional file 1

Taxa list. *Taxa list of sampled sequences. * indicates concatenated 18S and 28S rRNA sequences from different species. For combinations of genes to construct concatenated sequences of chimeran taxa, see Table S1. ** contributed sequences in the present study (author of sequences).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-119-S1.xls]

Additional file 2

LogDet corrected network of concatenated 18S and 28S rRNA alignment. *LogDet corrected network plus invariant site models (30.79% invariant sites) using SplitsTree4 based on the concatenated 18S and 28S rRNA alignment after exclusion of randomly similar sections evaluated with* ALISCORE.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-119-S2.pdf]

Additional file 3

Bayesian support values for selected clades. *List of Baysian support values (posterior probability, pP) for selected clades of the time-heterogeneous and time-homogeneous tree.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-119-S3.xls]

Additional file 4

Detailed flow of the analysis procedure in the software package PHASE-2.0. *Options used in* PHASE-2.0 *are italicized above the arrows and are followed by input files. Black arrows represent general flows of the analysis procedure, green arrows show that results or parameter values after single steps were inserted or accessed in a further process. Red block-arrows mark the final run of the time-heterogeneous and time-homogeneous approach with 16 chains each (2 × 118,000,000 generations).* ***First row****: I.) We prepared 3 control files (control.mcmc) for* mcmcphase *using three different mixed models. This "pre-run" was used for a first model selection (500,000 generations for each setting). We excluded model (C) based on non-convergence of parameter values. II.) We repeated step one (I.) with 3,000,000 generations using similar control files (different number of generations and random seeds) of the two remaining model settings. Calculated ln likelihoods values of both chains were compared in a BFT resulting in the exclusion of mixed model (A). Parameter values of the remaining model (B) were implemented in the time-heterogeneous setting. III.) We started the final analysis (final run) using sixteen chains for both the time-homogeneous and the time-heterogeneous approach. In the final time-homogeneous approach, the control files were similar to step II.) except for a different number of generations and random seeds.* ***Second row****: Additional steps were necessary prior to the computation of the final time-heterogeneous chains. We applied* mcmcsummarize *for the selected mixed model (B) to calculate a consensus tree.* Optimizer *was executed to conduct a ML estimation for each parameter value (opt.mod) based on the inferred consensus tree and optimized parameter-values (mcmc-best.mod), a data file delivered by* mcmcphase*. Estimated values were implemented in an initial.mod file. The initial.mod file and its parameter values were accessed by the control files of the final time-heterogeneous chains (only topology and base frequencies estimated).* ***Third row****: Trees were reconstructed separately for the time-homogeneous and time-heterogeneous setting. All chains of each approach were tested in a BFT against the chain with the best* lnL*. We only included chains with a* $2\ln B_{10}$*-value > 10. From these chains we constructed a metachain for each setting using Perl and applied* mcmcsummarize *to infer the consensus topology. To estimate branch lengths properly we ran* mcmcphase*, resulting branch lengths were implemented in the consensus trees. Finally, both trees were optimized using graphic programs (Dendroscope, Adobe Illustrator CS II).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-119-S4.pdf]

Additional file 5

List of chimeran species for concatenated 18S and 28S rRNA sequences
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-119-S5.xls]

Additional file 6

Primer list 18S rRNA
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-119-S6.xls]

Additional file 7

Primer list 28S rRNA
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-119-S7.xls]

Additional file 8

Primercard of the 18S rRNA gene for hexapods, myriapods and crustaceans. *Primers used for hexapods or myriapods are shown in the upper part, primers for crustaceans in the lower part. Positions of forward primers are marked with green arrows, those of reverse primers with red arrows. When different primers with identical position were used, all primer labels are given at the single arrow for the specific position. Primers and their combinations are given in Additional file 6 and 11.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-119-S8.pdf]

Additional file 9

Primercard of the 28S rRNA gene for crustaceans, hexapods and myriapods. *Positions of forward primers are tagged with green arrows, those of reverse primers with red arrows. When different primers with identical position were used, all primer labels are given at the single arrow for the specific position. Primers and their combinations are given in Additional file 7 and 11.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-119-S9.pdf]

Additional file 10

Primercard of the 28S rRNA gene for pterygots. *Positions of forward primers are assigned by green arrows, those of reverse primers with red arrows. When different primers with identical position were used, all primer labels are given at the single arrow for the specific position. Primers and their combinations are given in Additional file 7 and 11.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-119-S10.pdf]

Additional file 11

Supplementary Information. *Supplementary information for lab work (amplificaion, purification and sequencing of PCR products).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-119-S11.pdf]

Additional file 12

PCR temperature-profiles
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-119-S12.xls]

Additional file 13

PCR chemicals
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-119-S13.xls]

Additional file 14

Setting of exchangeability parameters used in pre-runs. *Listed settings of exchangeability parameters used in pre-runs in* PHASE-2.0.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-119-S14.xls]

Additional file 15

Included chains to infer the time-heterogeneous consensus tree. *Number of chains, generations per chain, harmonic means (*lnL*) and* $2\ln B_{10}$*-values included to infer the time-heterogeneous consensus tree.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-119-S15.xls]

Additional file 16

Included chains to infer the time-homogeneous consensus tree. *Number of chains, generations per chain, harmonic means (*lnL*) and* $2\ln B_{10}$*-values included to infer the time-homogeneous consensus tree.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-119-S16.xls]

Additional file 17

Localities of sampled taxa
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-119-S17.xls]

Acknowledgements

We thank Matty Berg, Anke Braband, Antonio Carapelli, Erhard Christian, Romano Dallai, Johannes Dambach, Wolfram Dunger, Erich Eder, Christian Epe, Pietro Paolo Fanciulli, Makiko Fikui, Francesco Frati, Peter Frenzel, Yan Gao, Max Hable, Bernhard Huber, Herbert Kliebhan, Stefan Koenemann, Franz Krabb, Ryuichiro Machida, Albert Melber, Wolfgang Moser, Reinhard Predel, Michael Raupach, Sven Sagasser, Kaoru Sekiya, Marc Sztatecsny, Dieter Waloßek, Manfred Walzl and Yi-ming Yang for help in collecting specimens, for providing tissue or other DNA-samples or for laboratory help. Thanks also go to Andreas Wißkirchen, Theory Department, Physikalisches Institut, University of Bonn for using their computational power. We thank Berit Ullrich, Oliver Niehuis and Patrick Kück for providing Perl-Scripts and Thomas Stamm for suggestions on the discussion structure. Special thanks go to John Plant for linguistic help. This work was supported by the German Science Foundation (DFG) in the priority program SPP 1174 "Deep Metazoan Phylogeny" http://www.deep-phylogeny.org. Work by JWW, BMvR is supported by the DFG grant WA 530/34; BM, KM are funded by the DFG grant MI 649/6, HH and SS are supported by the DFG grant HA 1947/5. NUS, ED, DB and GP are funded by the Austrian Science Foundation (FWF) grant P 20497-B17.

References

1. 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* 2006, **40(3):**772-794.
2. Zrzavý J, Štys P: **The basic body plan of arthropods: insights from evolutionary morphology and developmental biology.** *J Evol Biol* 1997, **10(3):**653-367.
3. Dohle W: **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)* 2001, **37(3):**85-103.
4. Friedrich M, Tautz D: **Ribosomal DNA phylogeny of the major extant arthropod classes and the evolution of myriapods.** *Nature* 1995, **376(6536):**165-167.
5. Shultz JW, Regier JC: **Phylogenetic analysis of arthropods using two nuclear protein-encoding gene ssupports a crustacean + hexapod clade.** *Proc Biol Sci* 2000, **267(1447):**1011-1019.

6. Friedrich M, Tautz D: **Arthropod rDNA phylogeny revisited: A consistency analysis using Monte Carlo simulation.** *Ann Soc Entomol Fr (New Series)* 2001, **37(1–2):**21-40.
7. Hwang UW, Friedrich M, Tautz D, Park CJ, Kim W: **Mitochondrial protein phylogeny joins myriapods with chelicerates.** *Nature* 2001, **413(6852):**154-157.
8. Regier JC, Shultz JW: **Elongation factor-2: A useful gene for arthropod phylogenetics.** *Mol Phylogenet Evol* 2001, **20:**136-148.
9. Giribet G, Edgecombe GD, Wheeler WC: **Arthropod phylogeny based on eight molecular loci and morphology.** *Nature* 2001, **413(6852):**157-161.
10. Pisani D, Poling L, Lyons-Weiler M, Hedges SB: **The colonization of land by animals: molecular phylogeny and divergence times among arthropods.** *BMC Biol* 2004, **2:**1.
11. Mallatt JM, Garey JR, Shultz JW: **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* 2004, **31:**178-191.
12. Regier JC, Shultz JW, Kambic RE: **Pancrustacean phylogeny: hexapods are terrestrial crustaceans and maxillopods are not monophyletic.** *Proc Biol Sci* 2005, **272(1561):**395-401.
13. Hassanin A: **Phylogeny of Arthropoda inferred from mitochondrial sequences: strategies for limiting the misleading effects of multiple changes in pattern and rates of substitution.** *Mol Phylogenet Evol* 2006, **38:**100-16.
14. Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, Sorensen MV, Haddock SHD, Schmidt-Rhaesa A, Okusu A, Kristensen RM, Wheeler WC, Martindale MQ, Giribet G: **Broad phylogenomic sampling improves resolution of the animal tree of life.** *Nature* 2008, **452(7188):**745-749.
15. Fanenbruck M, Harzsch S, Wägele JW: **The brain of the Remipedia (Crustacea) and an alternative hypothesis on their phylogenetic relationships.** *Proc Natl Acad Sci USA* 2004, **101(11):**3868-3873.
16. Harzsch S, Müller CHG, Wolf H: **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* 2005, **215(2):**53-68.
17. Harzsch S: **Neurophylogeny: Architecture of the nervous system and a fresh view on arthropod phyologeny.** *Integr Comp Biol* 2006, **46(2):**162-194.
18. Ungerer P, Scholtz G: **Filling the gap between identified neuroblasts and neurons in crustaceans adds new support for Tetraconata.** *Proc Biol Sci* 2008, **275(1633):**369-376.
19. Nardi F, Spinsanti G, Boore JL, Carapelli A, Dallai R, Frati F: **Hexapod origins: monophyletic or paraphyletic?** *Science* 2003, **299(5614):**1887-1889.
20. Cameron SL, Miller KB, D'Haese CA, Whiting MF, Barker SC: **Mitochondrial genome data alone are not enough to unambiguously resolve the relationships of Entognatha, Insecta and Crustacea *sensu lato* (Arthropoda).** *Cladistics* 2004, **20(6):**534-557.
21. Cook CE, Yue Q, Akam M: **Mitochondrial genomes suggest that hexapods and crustaceans are mutually paraphyletic.** *Proc Biol Sci* 2005, **272(1569):**1295-1304.
22. Carapelli A, Liò P, Nardi F, Wath E van der, Frati F: **Phylogenetic analysis of mitochondrial protein coding genes confirms the reciprocal paraphyly of Hexapoda and Crustacea.** *BMC Evol Biol* 2007:S8.
23. Dell'Ampio E, Szucsich NU, Carapelli A, Frati F, Steiner G, Steinacher A, Pass G: **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 Scr* 2009, **38(2):**155-170.
24. Hassanin A, Léger N, Deutsch J: **Evidence for multiple reversals of asymmetric mutational constraints during the evolution of the mitochondrial genome of Metazoa, and consequences for phylogenetic inferences.** *Syst Biol* 2005, **54(2):**277-298.
25. Wägele JW, Mayer C: **Visualizing differences in phylogenetic information content of alignments and distinction of three classes of long-branch effects.** *BMC Evol Biol* 2007, **7:**147.
26. Rota-Stabelli O, Telford ML: **A multi criterion approach for the selection of optimal outgroups in phylogeny: Recovering some support for Mandibulata over Myriochelata using mitogenomics.** *Mol Phylogenet Evol* 2008, **48:**103-111.
27. Bäcker H, Fanenbruck M, Wägele JW: **A forgotten homology supporting the monophyly of Tracheata: The subcoxa of insects and myriapods re-visited.** *Zool Anz* 2008, **247(3):**185-207.
28. Kadner D, Stollewerk A: **Neurogenesis in the chilopod *Lithobius forficatus* suggests more similarities to chelicerates than to insects.** *Dev Genes Evol* 2004, **214(8):**367-379.
29. Stollewerk A, Simpson P: **Evolution of early development of the nervous system: a comparison between arthropods.** *Bioessays* 2005, **27(9):**874-883.
30. Stollewerk A, Chipman AD: **Neurogenesis in myriapods and chelicerates and its importance for understanding arthropod relationships.** *Integr Comp Biol* 2006, **46(2):**195-206.
31. Ogden TH, Whiting MF: **The problem with "the Paleoptera Problem": sense and sensitivity.** *Cladistics* 2003, **19(5):**432-442.
32. Kjer KM: **Aligned 18S and insect phylogeny.** *Syst Biol* 2004, **53(3):**506-514.
33. Kukalová-Peck J, Lawrence JF: **Relationships among coleopteran suborders and major endoneopteran lineages: Evidence from hind wing characters.** *Eur J Entomol* 2004, **101:**95-144.
34. Misof B, Niehuis O, Bischoff I, Rickert A, Erpenbeck D, Staniczek A: **Towards an 18S phylogeny of hexapods: Accounting for group-specific character covariance in optimized mixed nucleotide/doublet models.** *Zoology (Jena)* 2007, **110(5):**409-429.
35. Whitfield JB, Kjer KM: **Ancient rapid radiations of insects: Challenges for phylogenetic analysis.** *Annu Rev Entomol* 2008, **53:**449-472.
36. Koch M: **Monophyly and phylogenetic position of the Diplura (Hexapoda).** *Pedobiologia (Jena)* 1997, **41(9):**9-12.
37. Kristensen NP: **The groundplan and basal diversification of the hexapods.** In *Arthropod Relationships* London: Chapman and Hall:281-293.
38. Carapelli A, Frati F, Nardi F, Dallai R, Simon C: **Molecular phylogeny of the apterygotan insects based on nuclear and mitochondrial genes.** *Pedobiologia (Jena)* 2000, **44(3–4):**361-373.
39. Carapelli A, Nardi F, Dallai R, Boore JL, Liò P, Frati F: **Relationships between hexapods and crustaceans based on four mitochondrial genes.** In *Crustacean and Arthropod Relationships, Volume 16 of Crustacean Issues* Edited by: Koenemann S, Jenner RA. CRC Press; 2005:295-306.
40. D'Haese CA: **Were the first springtails semi-aquatic? A phylogenetic approach by means of 28S rDNA and optimization alignment.** *Proc Biol Sci* 2002, **269(1496):**1143-1151.
41. Luan Yx, Zhang Y, Qiaoyun Y, Pang J, Xie R, Yin W: **Ribosomal DNA gene and phylogenetic relationships of Diplura and lower hexapods.** *Sci China C Life Sci* 2003, **46:**67-76.
42. Giribet G, Edgecombe GD, Carpenter JM, D'Haese CA, Wheeler WC: **Is Ellipura monophyletic? A combined analysis of basal hexapod relationships with emphasis on the origin of insects.** *Org Divers Evol* 2004, **4(4):**319-340.
43. Regier JC, Shultz JW, Kambic RE: **Phylogeny of basal hexapod lineages and estimates of divergence times.** *Ann Entomol Soc Am* 2004, **97(9):**411-419.
44. Luan Yx, Mallatt JM, Xie Rd, Yang Ym, Yin Wy: **The phylogenetic positions of three basal-hexapod groups (Protura, Diplura, and Collembola) based on on ribosomal RNA gene sequences.** *Mol Biol Evol* 2005, **22(7):**1579-1592.
45. Szucsich NU, Pass G: **Incongruent phylogenetic hypotheses and character conflicts in morphology: The root and early branches of the hexapodan tree.** *Mitt Dtsch Ges Allg Angew Entomol* 2008, **16:**415-429.
46. Jow H, Hudelot C, Rattray M, Higgs PG: **Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution.** *Mol Biol Evol* 2002, **19(9):**1591-1601.
47. Galtier N: **Sampling properties of the bootstrap support in molecular phylogeny: Influence of nonindependence among sites.** *Syst Biol* 2004, **53:**38-46.
48. Fox GE, Woese CR: **The architecture of 5S rRNA and its relation to function.** *J Mol Evol* 1975, **6:**61-76.
49. Wuyts J, De Rijk P, Peer Y Van de, Pison G, Rousseeuw P, De Wachter R: **Comparative analysis of more than 3000 sequences reveals the existence of two pseudoknots in area V4 of eukaryotic small subunit ribosomal RNA.** *Nucleic Acids Res* 2000, **28(23):**4698-4708.

50. Gutell JC, Robin R, Lee , Cannone JJ: **The accuracy of ribosomal RNA comparative structure models.** *Curr Opin Struct Biol* 2002, **12(3):**301-310.
51. Ban N, Nissen P, Hansen J, Moore PB, Steitz TA: **The complete atomic structure of the large ribosomal subunit at 2.4 Å Resolution.** *Science* 2000, **289(5481):**905-920.
52. Noller HF: **RNA structure: reading the ribosome.** *Science* 2005, **309(5740):**1508-1514.
53. Lartillot N, Philippe H: **Improvement of molecular phylogenetic inference and the phylogeny of Bilateria.** *Philos Trans R Soc Lond B Biol Sci* 2008, **363(1496):**1463-1472.
54. Rodríguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H: **Detecting and overcoming systematic errors in genome-scale phylogenies.** *Syst Biol* 2007, **56(3):**389-399.
55. Philippe H, Germot A, Moreira D: **The new phylogeny of eukaryotes.** *Curr Opin Genet Dev* 2000, **10(6):**596-601.
56. Stocsits RR, Letsch H, Hertel J, Misof B, Stadler PF: *RNAsalsa. Version 0.7.3, current versions* 2008 [http://rnasalsa.zfmk.de]. Zoologisches Forschungsmuseum A. Koenig, Bonn
57. Brown JM, Lemmon AR: **The importance of data partitioning and the utility of bayes factors in bayesian phylogenetics.** *Syst Biol* 2007, **56(4):**643-655.
58. Galtier N, Gouy M: **Inferring phylogenies from DNA sequences of unequal base compositions.** *Proc Natl Acad Sci USA* 1995, **92(24):**11317-11321.
59. Tarrío R, Rodríguez-Trelles F, Ayala FJ: **Shared nucleotide composition biases among species and their impact on phylogenetic reconstructions of the Drosophilidae.** *Mol Biol Evol* 2001, **18(8):**1464-1473.
60. Gowri-Shankar V, Rattray M: **A reversible jump method for bayesian phylogenetic inference with a nonhomogeneous substitution model.** *Mol Biol Evol* 2007, **24(6):**1286-1299.
61. Blanquart S, Lartillot N: **A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution.** *Mol Biol Evol* 2006, **23(11):**2058-2071.
62. Gowri-Shankar V, Rattray M: **On the correlation between composition and site-specific evolutionary rate: Implications for phylogenetic inference.** *Mol Biol Evol* 2006, **23(2):**352-364.
63. Gowri-Shankar V, Jow H: *PHASE: a software package for Phylogenetics And Sequence Evolution. 2.0* University of Manchester; 2006.
64. Telford MJ, Wise MJ, Gowri-Shankar V: **Consideration of RNA secondary structure significantly improves likelihood-based estimates of phylogeny: Examples from the Bilateria.** *Mol Biol Evol* 2005, **22(4):**1129-1136.
65. Misof B, Misof K: **A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion.** *Syst Biol* 2009, **58:**syp006.
66. Swofford DL: *PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods). Version 4* Sinauer Associates, Sunderland, Massachusetts; 2003.
67. Kaas RE, Raftery AE: **Bayes Factors.** *Journal of the American Statistical Association* 1995, **90(430):**773-795.
68. Nylander JAA, Ronquist F, Huelsenbeck JP, Nieves-Aldrey JL: **Bayesian phylogenetic analysis of combined data.** *Syst Biol* 2004, **53(21):**47-67.
69. Boudreaux BH: *Arthropod phylogeny: with special reference to insects* John Wiley & Sons Inc; 1979.
70. Edgecombe GD, Giribet G: **Myriapod phylogeny and the relationships of Chilopoda.** *Biodiversidad, Taxonomía y Biogeografía de Artrópodos de México: Hacia una Síntesis de su Conocimiento* 2002, **III:**143-168.
71. Kjer KM, Carle FL, Litman J, Ware J: **A Molecular Phylogeny of Hexapoda.** *Arthropod Systematics & Phylogeny* 2006, **64:**35-44.
72. Yamaguchi S, Endo K: **Molecular phylogeny of Ostracoda (Crustacea) inferred from 18S ribosomal DNA sequences: implication for its origin and diversification.** *Mar Biol* 2003, **143:**23-38.
73. Gai YH, Song DX, Sun HY, Zhou KY: **Myriapod monophyly and relationships among myriapod classes based on nearly complete 28S and 18S rDNA sequences.** *Zool Sci* 2006, **23(12):**1101-1108.
74. Gillespie JJ, Johnston JS, Cannone JJ, Gutell RR: **Characteristics of the nuclear (18S, 5.8S, 28S and 5S) and mitochondrial (12S and 16S) rRNA genes of Apis mellifera (Insecta: Hymenoptera): structure, organization, and retrotransposable elements.** *Insect Mol Biol* 2005, **15(5):**657-686.
75. Jordal B, Gillespie JJ, Cognato AI: **Secondary structure alignment and direct optimization of 28S rDNA sequences provide limited phylogenetic resolution in bark and ambrosia beetles (Curculionidae: Scolytinae).** *Zool Scr* 2008, **37:**43-56.
76. Simon C, Buckley TR, Frati F, Stewart JB, Beckenbach AT: **Incorporating Molecular Evolution into Phylogenetic Analysis, and a New Compilation of Conserved Polymerase Chain Reaction Primers for Animal Mitochondrial DNA.** *Annu Rev Ecol Evol Syst* 2006, **37:**545-579.
77. Susko E, Spencer M, Roger AJ: **Biases in phylogenetic estimation can be caused by random sequence segments.** *J Mol Evol* 2005, **61(3):**351-359.
78. Philippe H, Delsuc F, Brinkmann H, Lartillot N: **Phylogenomics.** *Annual Review of Ecology, Evolution, and Systematics* 2005, **36:**541-562.
79. Roeding F, Hagner-Holler S, Ruhberg H, Ebersberger I, von Haeseler Arndt, Kube M, Reinhardt R, Burmester T: **EST sequencing of Onychophora and phylogenomic analysis of Metazoa.** *Mol Phylogenet Evol* 2007, **45(3):**942-951.
80. Podsiadlowski L, Kohlhagen H, Koch M: **The complete mitochondrial genome of *Scutigerella causeyae* (Myriapoda: Symphyla) and the phylogenetic position of Symphyla.** *Mol Phylogenet Evol* 2007, **45:**251-260.
81. Castresana J: **Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis.** *Mol Biol Evol* 2000, **17(4):**540-552.
82. Hancock JM, Tautz D, Dover GA: **Evolution of the secondary structures and compensatory mutations of the ribosomal RNAs of *Drosophila melanogaster* .** *Mol Biol Evol* 1988, **5(4):**393-414.
83. Stephan W: **The rate of compensatory evolution.** *Genetics* 1996, **144:**419-426.
84. Schöninger M, von Haeseler A: **A stochastic model for the evolution of autocorrelated DNA sequences.** *Mol Phylogenet Evol* 1994, **3(3):**240-247.
85. Friedrich M, Tautz D: **An episodic change of rDNA nucleotide substitution rate has occurred during the emergence of the insect order Diptera.** *Mol Biol Evol* 1997, **14(6):**644-653.
86. Jermiin LS, Ho SYW, Ababneh F, Robinson J, Larkum AWD: **The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated.** *Syst Biol* 2004, **53(4):**638-643.
87. Foster PG: **Modeling compositional heterogeneity.** *Syst Biol* 2004, **53(3):**485-495.
88. Walossek D: **On the Cambrian diversity of Crustacea.** In *Crustaceans and the Biodiversity Crisis, Proceedings of the Fourth International Crustacean Congress, Amsterdam, The Netherlands, July 20–24, 1998 Volume 1*. Edited by: von Vaupel Klein FRSJC. Brill Academic Publishers, Leiden; 1998:3-27.
89. Edgecombe GD, Wilson GDF, Colgan DJ, Gray MR, Cassis G: **Arthropod Cladistics: Combined analysis of histone H3 and U2 snRNA sequences and morphology.** *Cladistics* 2000, **16(2):**155-203.
90. Lim JT, Hwang UW: **The complete mitochondrial genome of *Pollicipes mitella* (Crustacea, Maxillopoda, Cirripedia): non-monophylies of Maxillopoda and Crustacea.** *Mol Cells* 2006, **22(3):**314-322.
91. Wilson K, Cahill V, Ballment E, Benzie J: **The complete sequence of the mitochondrial genome of the crustacean *Penaeus monodon*: Are malacostracan crustaceans more closely related to insects than to branchiopods?** *Mol Biol Evol* 2000, **17(6):**863-874.
92. Glenner H, Thomsen PF, Hebsgaard MB, Sørensen MV, Willerslev E: **Evolution: The origin of insects.** *Science* 2006, **314(5807):**1883-1884.
93. Zhang Xg, Siveter DJ, Waloszek D, Maas A: **An epipodite-bearing crown-group crustacean from the Lower Cambrian.** *Nature* 2007, **449(7162):**595-598.
94. Regier JC, Shultz JW: **Molecular phylogeny of the major arthropod groups indicates polyphyly of crustaceans and a new hypothesis for the origin of hexapods.** *Mol Biol Evol* 1997, **14(9):**902-913.
95. Schierwater B, Hadrys H: **Environmental factors and metagenesis in the hydroid *Eleutheria dichotoma* .** *Invertebr Reprod Dev* 1998, **34(2–3):**139-148.

96. Hall TA: **BioEdit: a user-friendly biological alignment sequence EDITOR and analysis program for Windows95/98/NT.** *Nucleic Acids Symp Ser* 1999, **41(2–3):**95-98.
97. Tatusova TA, Madden TL: **BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences.** *FEMS Microbiol Lett* 1999, **174(2):**247-250.
98. Kjer KM: **Use of rRNA secondary structure in phylogenetic studies to identify homologous positions: An example of alignment and data presentation from the frogs.** *Mol Phylogenet Evol* 1995, **4(3):**314-330.
99. Hickson RE, Simon C, Perrey SW: **The performance of several multiple-sequence alignment programs in relation to secondary-structure features for an rRNA sequence.** *Mol Biol Evol* 2000, **17(4):**530-539.
100. Buckley TR, Simon C, Chambers GK: **Exploring among-site rate variation models in a maximum likelihood framework using empirical data: Effects of model assumptions on estimates of topology, branch lengths, and bootstrap support.** *Syst Biol* 2001, **50:**67-86.
101. Misof B, Niehuis O, Bischoff I, Rickert A, Erpenbeck D, Staniczek A: **A hexapod nuclear SSU rRNA secondary-structure model and catalog of taxon-specific structural variation.** *J Exp Zool B Mol Dev Evol* 2006, **306B:**70-88.
102. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32(5):**1792-1797.
103. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity.** *BMC Bioinformatics* 2004, **5:**113.
104. Peer Y Van de, De Rijk P, Wuyts J, Winkelmans T, De Wachter R: **The European Small Subunit Ribosomal RNA database.** *Nucleic Acids Res* 2000, **28:**175-176.
105. Wuyts J, Peer Y Van de, Winkelmans T, De Wachter R: **The European database on small subunit ribosomal RNA.** *Nucleic Acids Res* 2002, **30:**183-185.
106. Wuyts J, Perrière G, Peer Y Van de: **The European ribosomal RNA database.** *Nucleic Acids Res* 2004, **32(Suppl 1, Database):**D101-103.
107. Michot B, Bachellerie JP, Raynal F: **Structure of mouse rRNA precursors. Complete sequence and potential folding of the spacer regions between 18S and 28S rRNA.** *Nucleic Acids Res* 1983, **11(10):**3375-3391.
108. Gillespie JJ, Munro JB, Heraty JM, Yoder MJ, Owen AK, Carmichael AE: **A secondary structural model of the 28S rRNA expansion segments D2 and D3 for chalcidoid wasps (Hymenoptera: Chalcidoidea).** *Mol Biol Evol* 2005, **22(7):**1593-1608.
109. Huson DH, Bryant D: **Application of phylogenetic networks in evolutionary studies.** *Mol Biol Evol* 2006, **23(2):**254-267.
110. Bryant D, Moulton V: **Neighbor-Net: An agglomerative method for the construction of phylogenetic networks.** *Mol Biol Evol* 2004, **21(2):**255-265.
111. Penny D, Lockhart PJ, Steel MA, Hendy MD: **The role of models in reconstructing evolutionary trees.** In *Models in phylogeny reconstruction, The Systematics Association Special Volume Series* Edited by: Scotland RW, Diebert DJ, Williams DM. Oxford University Press; 1994:211-230.
112. Steel M, Huson D, Lockhart PJ: **Invariable sites models and their use in phylogeny reconstruction.** *Syst Biol* 2000, **49(2):**225-232.
113. Yang Z: **Among-site rate variation and its impact on phylogenetic analyses.** *Trends Ecol Evol (Amst.)* 1996, **11(9):**367-372.
114. Yang Z: **Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods.** *J Mol Evol* 1994, **39(3):**306-314.
115. Kelchner SA, Thomas MA: **Model use in phylogenetics: nine key questions.** *Trends Ecol Evol (Amst.)* 2007, **22(2):**87-94.
116. Sullivan J, Swofford DL: **Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated?** *Syst Biol* 2001, **50(5):**723-729.
117. Waddell PJ, Penny D, Moore T: **Hadamard conjugations and modeling sequence evolution with unequal rates across sites.** *Mol Phylogenet Evol* 1997, **8:**33-50.
118. Zwickl DJ, Holder MT: **Model parameterization, prior distributions, and the general time-reversible model in bayesian phylogenetics.** *Syst Biol* 2004, **53(6):**877-888.
119. Beiko RG, Keith JM, Harlow TJ, Ragan MA: **Searching for convergence in phylogenetic Markov Chain Monte Carlo.** *Syst Biol* 2006, **55(4):**553-565.

SUPPLEMENT - MANUSCRIPT II

MBE Advance Access published August 19, 2009

Hemocyanin suggests a close relationship of Remipedia and Hexapoda

Research Article

Beyhan Ertas[1], Björn M. von Reumont[2], Johann-Wolfgang Wägele[2], Bernhard Misof[1], Thorsten Burmester[1]

[1]Biozentrum Grindel und Zoologisches Museum, Universität Hamburg, Martin-Luther-King-Platz 3, D-20146 Hamburg, Germany

[2]Zoologisches Forschungsmuseum Alexander Koenig, Adenauerallee 160, D-53113 Bonn, Germany

Corresponding author: Thorsten Burmester
Biozentrum Grindel und Zoologisches Museum
Universität Hamburg
Martin-Luther-King-Platz 3
D-20146 Hamburg
Germany
Phone: (+49) 40-42838-3913
Fax: (+49) 40-42838-3937
E-mail: thorsten.burmester@uni-hamburg.de

Key words: Crustacea, Hemocyanin, Hexapoda, Insecta, Pancrustacea, Remipedia

Running head: Remipede hemocyanin

Abbreviations: Hc, hemocyanin; PCR, polymerase chain reaction; SDS-PAGE, sodium dodecyl sulfate-polyacrylamide gel electrophoresis; see Supplementary Table 1 for abbreviations of hemocyanins.

Abstract

The remipedia are enigmatic crustaceans from anchialine cave systems, first described only 30 years ago, whose phylogenetic affinities are as yet unresolved. Here we report the sequence of hemocyanin from *Speleonectes tulumensis* Yager, 1987 (Remipedia, Speleonectidae). This is the first proof of the presence of this type of respiratory protein in a crustacean taxon other than Malacostraca. *S. tulumensis* hemocyanin consists of multiple distinct (at least three) subunits (StuHc1 to 3). Surprisingly, the sequences are most similar to hexapod hemocyanins. Phylogenetic analyses showed that the *S. tulumensis* hemocyanin subunits StuHc1 and StuHc3 associate with the type 1 hexapod hemocyanin subunits, whereas StuHc2 associates with the type 2 subunits of hexapods. Together, remipede and hexapod hemocyanins are in the sister-group position to the hemocyanins of malacostracan crustaceans. Hemocyanins provide no indication of a close relationship of Myriapoda and Hexapoda, but tend to support Pancrustacea (Crustacea + Hexapoda). Our results also indicate that Crustacea are paraphyletic and that Hexapoda may have evolved from a Remipedia-like ancestor. Thus, Remipedia occupy a key position for the understanding of the evolution of hexapods, which are and have been one of the world's most speciose lineage of animals.

Introduction

In 1979 Yager discovered in an anchialine cave in the Bahamas the first specimens of Remipedia, which represented a new class of crustaceans with currently about 20 described extant species (Yager 1981; Koenemann et al. 2007b). Living remipedes harbor a number of unique features, which include the loss of eyes, biramous antennulae, three pairs of postmandibular mouthparts adapted to a predatory feeding mode and to grooming, and the lack of tagmatization of the trunk (Schram and Lewis 1989; Koenemann et al. 2007c; van der Ham and Felgenhauer 2008). The phylogenetic affinities of Remipedia are still controversial. On the basis of the homonomous segmentation of their trunk, which supposedly represents the ancestral ground pattern in the evolution of Arthropoda, it has been suggested that Remipedia occupy a basal position within Crustacea or even Mandibulata (Schram 1986; Wills 1997; Giribet, Edgecombe, and Wheeler 2001). Molecular phylogenetic analyses of ribosomal RNA and mitochondrial genomes, as well as limb morphology have indicated an association of Remipedia with various crustacean classes assigned to the "Maxillopoda" (Ito 1989; Spears and Abele 1997; Lim and Hwang 2006). However, these results may be due to long branch attraction phenomena (Spears and Abele 1997; Telford et al. 2008; von Reumont et al. 2009). In fact, the pattern of mitochondrial gene arrangement convincingly excluded Remipedia from the maxillopod assemblage, which in that study comprised Branchiura, Cephalocarida, Cirripedia, and Pentastomida (Lavrov, Brown, and Boore 2004). Other molecular phylogenetic analyses have found Remipedia at various positions within Crustacea, albeit with poor support (Regier and Shultz 2001; Regier, Shultz, and Kambic 2005; Hassanin 2006; Carapelli et al. 2007). A recent study of arthropod brain morphology suggested that Remipedia belong to a monophylum of Malacostraca and Hexapoda within paraphyletic crustaceans (Fanenbruck, Harzsch, and Wagele 2004; Fanenbruck and Harzsch 2005). The discovery of free-living lecithotrophic remipede larvae and analysis of early larval

development of Remipedia tentatively support their assumed relationship to malacostracan crustaceans (Koenemann et al. 2007a; Koenemann et al. 2009).

Oxygen transport in the hemolymph of various arthropod and mollusk taxa is facilitated by copper-proteins referred to as hemocyanins (Markl and Decker 1992; van Holde and Miller 1995; Burmester 2002). Arthropod and mollusk hemocyanins are not related, but evolved independently from distinct copper-containing enzymes (Burmester 2001; Burmester 2002). Hemocyanins of Arthropoda are large hexameric or oligohexameric proteins composed of similar or identical subunits in the size range of 75 kDa. Each subunit can bind to an O_2 molecule by the virtue of two copper ions that are coordinated by six histidine residues. Hemocyanins have been thoroughly studied in Chelicerata and Crustacea, occur in Myriapoda (Jaenicke et al. 1999; Kusche and Burmester 2001), and have recently been identified in Onychophora (Kusche, Ruhberg, and Burmester 2002) and Hexapoda (Hagner-Holler et al. 2004; Pick, Hagner-Holler, and Burmester 2008; Pick, Schneuer, and Burmester 2009). Within Crustacea, hemocyanins have been thought to be confined to Malacostraca (Mangum 1985; Markl and Decker 1992). Hemocyanin sequences have been shown to be informative for the inference of phylogenies within Arthropoda (Burmester 2001; Kusche and Burmester 2001; Kusche, Ruhberg, and Burmester 2002; Kusche et al. 2003). Thus, lineage-specific presence and phylogenetic analyses could be useful in assessing the position of Remipedia.

Here we report the identification and molecular cloning of hemocyanin cDNA from the remipede *Speleonectes tulumensis* and show that molecular phylogenetic analyses of these hemocyanin sequences provide evidence for a close relationship of Remipedia and Hexapoda.

Materials and Methods

Sample preparation

Speleonectes tulumensis was collected on Yucatan Peninsula, Mexico. Animals were cut in small pieces and the tissue was preserved in RNAlater (Qiagen, Hilden, Germany). Total RNA and protein was extracted using TriFast^{TM} (Peqlab, Erlangen, Germany) according to the manufacturer's instructions. The protein and RNA samples were immediately used or kept frozen at -20 °C until use.

Cloning of hemocyanin cDNA

Three µg total RNA were converted into cDNA by Superscript III reverse transcriptase employing an oligo-dT primer according the manufacturer's instructions (Invitrogen, Carlsbad, USA). The resulting cDNA was used for standard PCR, using degenerate primers (forward primer: 5'-ATGGAYTTYCCNTTYTGGTGGA-3'; reverse primer 5'-GTNGCGGTYTCRAARTGYTCCAT-3') that had been derived from conserved coding regions of arthropod hemocyanins (amino acid sequences: MDFPFWW and MEHFETAT) (Burmester 2001; Hagner-Holler et al. 2004; Pick, Schneuer, and Burmester 2009). Fragments of the expected size were cloned into the pGem-T Easy/JM109 *E. coli* system (Promega, Mannheim, Germany) and 14 independent clones were sequenced by a commercial service (Genterprise, Mainz, Germany). 5' and 3' RACE experiments were carried out by a RNA ligase-mediated rapid amplification method employing the GeneRacer Kit with SuperScript III reverse transcriptase. Sets of gene-specific primers were constructed according to the partial sequences (Supplementary Table 1). The cDNA fragments were cloned and sequenced as described. Sequences were

assembled using ContigExpress (Vector NTI Advance 10.3; Invitrogen) and GeneDoc 2.7 (Nicholas and Nicholas 1997).

Western Blotting

Protein extracts were denatured in sample buffer (31.25 mM Tris-HCl, pH 6.8, 1% SDS, 2.5% β-mercaptoethanol, 5% glycerol) at 95 °C for 5 min and loaded onto a 10% polyacrylamide gel. SDS-PAGE was carried out according to standard procedures. Semi-dry electro-transfer of proteins onto nitrocellulose membranes (Hartenstein, Würzburg, Germany) was carried out for 2 h at 0.8 mA/cm^2. Non-specific binding sites were blocked for 1 h with 2% non-fat dry milk in TBS (10 mM Tris-HCl, pH 7.4; 140 mM NaCl). The membranes were incubated for 2 h at room temperature with polyclonal antibodies that had been raised against various insect and crustacean hemocyanins (anti-*Homarus americanus*, anti-*Panulirus interruptus*, anti-*Cancer pagurus*, and anti-*Thermobia domestica* hemocyanin antibodies) diluted 1:5.000 in 2 % milk/TBS. The nitrocellulose filters were washed three times with TBS for 15 min and incubated for 1 h with a goat anti-rabbit antibody coupled with alkaline phosphatase (Dianova, Hamburg, Germany), diluted 1:10.000 in TBS. After a final washing step, detection was carried out with nitro-blue-tetrazolium-chloride and 5-bromo-4-chloro-3-indolyl-phosphate as substrates.

Sequence and phylogenetic analyses

Tools provided with the ExPASy Molecular Biology Server of the Swiss Institute of Bioinformatics (http://www.expasy.org) were used for the analyses of DNA and amino acid sequences. Signal peptides were predicted using SignalP 1.1 (Nielsen et al. 1997). Deduced amino acid sequences of *S. tulumensis* hemocyanin were included in a previously published multiple alignment of arthropod hemocyanins, insect hexamerins and crustacean

pseudohemocyanins (Pick, Schneuer, and Burmester 2009) employing MAFFT with the L-INS-i method and the BLOSUM 62 matrix (Katoh et al. 2005). A list of sequences used in this study is provided in Supplementary Table 1.

After the exclusion of N- and C-terminal extensions, the final multiple sequence alignment contained 800 positions and 96 sequences. The appropriate model of amino acid sequence evolution (WAG + Gamma model; Whelan and Goldman 2001) was selected by ProtTest (Abascal, Zardoya, and Posada 2005) using the Akaike Information Criterion (AIC). Bayesian phylogenetic analysis was performed using MrBayes 3.1.2 (Huelsenbeck and Ronquist 2001). We assumed the WAG model with a gamma distribution of substitution rates. We used uninformative priors on all trees. Metropolis-coupled Markov chain Monte Carlo (MCMCMC) sampling was performed with one cold and three heated chains that were run for 8,000,000 generations. Starting trees were random and trees were sampled every 100^{th} generation. Two independent runs were performed in parallel and were continued until runs had converged (average standard deviations of split frequencies were stationary and lower than 0.005). The program Tracer 1.4 (http://tree.bio.ed.ac.uk/software/tracer/) was used to examine log-likelihood plots and MCMC summaries for all parameters. Posterior probabilities were estimated on the final 60,000 trees (burnin = 20,000). RAxML 7.0.4, assuming the WAG evolutionary model with gamma distributions, was used for maximum likelihood analyses, and the resulting tree was tested by bootstrapping with 1000 replicates. TREE-PUZZLE 5.2 (Schmidt et al. 2002) was used to test alternative tree topologies. Hypothesis testing was performed by four methods: A Shimodaira-Hasegawa (SH) test (Shimodaira and Hasegawa 1999), a two sided Kishino-Hasegawa (2sKH) test (Kishino and Hasegawa 1989), a one sided KH test based on pairwise SH tests (Goldman, Anderson, and Rodrigo 2000), and an Expected Likelihood Weight (ELW) test (Strimmer and Rambaut 2002). SH, 1sKH, and ELW tests performed 1,000 resamplings using the RELL method.

Results and Discussion

Identification of hemocyanin in *S. tulumensis*

A set of degenerate oligonucleotide primers were designed according to conserved regions arthropod hemocyanin sequences (Burmester 2001; Pick, Schneuer, and Burmester 2009). RT-PCR using total RNA extracted from an adult *S. tulumensis* resulted in the amplification of fragments in the range of 550 bp. After cloning and sequencing, three distinct hemocyanin cDNA sequences were identified, which were termed StuHc1, StuHc2 and StuHc3, respectively. The missing 5' and 3' regions of the hemocyanin cDNAs were obtained by 5' and 3'-RACE. The full length cDNA sequences measure 2422, 2531, and 2506 bp (Supplementary Fig. 1). The deduced amino acid sequences comprise 686, 672, and 671 amino acids, of 51.6 to 55.6% identity (Fig. 1).

Sequence comparisons with other hemocyanins showed that the residues required for reversible oxygen binding and subunit cooperativity (Hazes et al. 1993; Burmester 2002) are present in all three sequences of *S. tulumensis*, suggesting that these proteins are able to carry out respiratory functions. The conserved residues include the six copper-binding site histidines in the central second domain, as well as three phenylalanines that stabilize the binding of O_2 in the first and second domain (Fig. 1). The phenylalanine in the first domain of the subunits is assumed to be a key residue in the regulation of oxygen binding in most hemocyanins (Hazes et al. 1993).

All *S. tulumensis* hemocyanins harbor the typical N-terminal leader sequences required for trans-membrane export via the endoplasmic reticulum (cf. Fig. 1 and Supplementary Fig. 1), resulting in native polypeptides of 656 to 667 amino acids with inferred molecular masses of 75.7 to 77.1 kDa. The occurrence of hemocyanin proteins in *S. tulumensis* was further established by Western blotting employing various polyclonal antibodies and antisera that had been raised against insect or crustacean hemocyanins. Two distinct bands in the range of 75 to 80 kDa were recognized by

most antibodies (Fig. 2), which are in excellent agreement with the molecular masses predicted from the translated cDNA sequences: StuHc1 is interpreted to form the upper band, and StuHc2 and StuHc3 are together the lower band. In summary, Remipedia do have functional hemocyanin sequences that occur in three subunits and are expressed in the adult organism.

Respiratory proteins in Remipedia and other Crustacea

Among the crustacean groups, hemocyanins have only been identified in Malacostraca (Mangum 1985; Markl and Decker 1992; Burmester 2002). In various non-malacostracan crustaceans (Branchiopoda, Ostracoda, Copepoda and Cirripedia), hemoglobins usually serve as oxygen carriers (Mangum 1985). Sequence data corroborate these findings because extensive BLAST searches on genomic sequences or expressed sequence tags (ESTs) in EMBL/GenBank and in the database of the "Deep Metazoan Phylogeny" project did not recover any hemocyanin from a non-malacostracan crustacean. There have been only two reports of the possible presence of hemocyanins in Cirripedia and Remipedia. It is, however, uncertain whether the putative hemocyanin of the parasitic cirriped *Sacculina carcini* is an endogenous protein or actually derives from the decapod host, *Carcinus maenas* (Herberts and de Frescheville 1981). Yager (1991) observed large crystal structures in remipede hemocytes that resemble those of the *Limulus polyphemus* hemocyanin. In a more recent report, van der Ham and Felgenhauer (2007) also speculated about the presence of a hemocyanin-like protein in *Speleonectes* sp. Here we demonstrate that hemocyanin actually occurs in a remipede species, which is in addition the first unambiguous report of such respiratory proteins in a non-malacostracan crustacean.

Remipedia dwell in a high-saline and oxygen-poor environment (usually <1 mg/ml O_2; (Koenemann, Schram, and Iliffe 2006) below the halocline interface between the seawater and the overlying well-oxygenated freshwater that is typical for anchialine cave systems (Pohlmann, Iliffe, and Cifuentes 1997). Remipedia have adapted to this hypoxic environment, and a

respiratory protein that augments the delivery of oxygen within the circulatory system is certainly advantageous. Hemocyanin has been identified in Onychophora and thus was present in the last common ancestor of Arthropoda (Burmester 2002; Kusche, Ruhberg, and Burmester 2002). It is extremely unlikely that hemocyanin was re-invented in Remipedia. It must be assumed that this respiratory protein was present in the last common ancestor of Malacostraca and Remipedia, and was the principal oxygen carrier of stemline crustaceans. Thus, the occurrence of hemocyanin is a plesiomorphic character of Crustacea. As mentioned above, most non-malacostracan crustaceans use hemoglobin not hemocyanin for oxygen transport (Mangum 1985). However, this hemoglobin is a secondary invention (apomorphic character), which most likely derived from an intracellular globin of unknown function. It is currently uncertain whether hemoglobin emerged only once in the Crustacea, which would provide evidence that at least part of the "Entomostraca" *sensu* Walossek (1999), *i.e.* Branchiopoda and "Maxillopoda", may represent a monophylum.

Hexapod affinities of remipede hemocyanins

Surprisingly, BLAST searches and pairwise sequence comparisons revealed that the hemocyanin subunits of *S. tulumensis* display the highest sequence identity with insect hemocyanin subunits. For example, StuHc1 and StuHc3 share 60.4% and 53.4%, respectively, of the amino acids with the hemocyanin 1 of the firebrat *Thermobia domestica*, whereas StuHc2 and the hemocyanin 2 of the cockroach *Blaptica dubia* are 57.1% identical. At least 10% lower identity scores were observed when the *S. tulumensis* hemocyanins were compared to those of malacostracan crustaceans or other arthropod hemocyanins. This evidence is a first rough indicator of a close relationship of hexapod and remipede hemocyanins.

The amino acid sequences of Remipedia hemocyanins were included in an alignment of arthropod hemocyanins and related proteins (Burmester and Scheller 1996; Burmester 2002; Hagner-Holler et al. 2004; Pick, Schneuer, and Burmester 2009) (Supplementary Fig. 2).

Phylogenetic analyses employing maximum likelihood and Bayesian methods resulted in essentially identical trees (Fig. 3; Supplementary Figs. 3 and 4). *S. tulumensis* hemocyanins and hexapod proteins form a monophyletic clade that excludes other crustacean (malacostracan) sequences (ML bootstrap support: 48%; Bayesian posterior probability: 0.96). The other crustacean hemocyanins and pseudohemocyanins were monophyletic (100%; 1.0), but in none of the trees did the hemocyanins of the Remipedia join this clade. The relative position of the myriapod hemocyanins remains uncertain and received poor support. These proteins were either associated with the chelicerate hemocyanins (thereby supporting the "Paradoxopoda" hypothesis; Mallatt, Garey, and Shultz 2004) or are in sister-group position to the crustacean plus hexapod proteins (supporting the traditional "Mandibulata"). In none of the trees did the myriapod hemocyanins join hexapod hemocyanins and hexamerins.

S. tulumensis hemocyanin subunits 1 and 3 (StuHc1 and StuHc3) form a common branch. This is reasonably well supported in Bayesian analyses (0.95) but only in 46% of the ML bootstrapped trees (Fig. 3). A common clade of StuHc1 and StuHc3 and hexapod hemocyanin type 1 subunits plus hexamerins was recovered with 1.00 posterior probability in the Bayesian tree and supported by an ML bootstrap value of only 54%. The association of StuHc2 with the hexapod hemocyanin subunits 2 receives high support throughout (97%; 1.00) and is further corroborated by a unique and conserved insertion of nine amino acids in beta-sheet 3A (amino acid position 416-420 in StuHc2; Fig. 1). Such insertion is not present in any other hemocyanin.

Tree topologies were further evaluated by hypothesis testing (Table 1). The multiple ratio tests recovered the tree topology presented in Fig. 3 as best result and significantly reject StuHc1, StuHc2 and StuHc3 from the clade of other crustacean (malacostracan) hemocyanins. Hence, there is substantial molecular phylogenetic and structural evidence that the remipede hemocyanin subunits are orthologs of hexapod subunits. Thus, the lineage leading to the remipede and hexapod hemocyanins split into two distinct subunit types before these taxa diverged.

Are Remipedia and Hexapoda sister-groups?

The origin of Hexapoda is notoriously disputed. Based on a large number of morphological characters it has long been assumed that Hexapoda evolved from a myriapod-like ancestor, and that these taxa form the subphylum "Tracheata" or "Atelocerata" (e.g., Brusca and Brusca 2003). However, this view has been challenged by molecular phylogenetic approaches that have provided evidence that Hexapoda are somehow, in fact, allied with Crustacea (Friedrich and Tautz 1995; Boore, Lavrov, and Brown 1998; Hwang et al. 2001; Kusche and Burmester 2001; Mallatt, Garey, and Shultz 2004). Such topology was recovered in our analyses as well (Fig. 3). A relationship of Hexapoda and Crustacea also received support in some comparative morphological and developmental biology studies (Giribet and Ribera 2000; Richter 2002), and thus Hexapoda and Crustacea have been joined in a common taxon named either "Tetraconata" (Richter 2002) or "Pancrustacea" (Zrzavý and Štys 1997).

In many analyses, Crustacea form a paraphyletic assemblage with respect to Hexapoda, but it has remained uncertain which crustacean taxon is the sister-group of the Hexapoda. In most studies only two crustacean classes were considered: Malacostraca and Branchiopoda. Analyses employing mitochondrial genomes suggest a close relationship of Malacostraca and Hexapoda with exclusion of Branchiopoda (Garcia-Machado et al. 1999; Wilson et al. 2000; Hwang et al. 2001). Such topology is also supported by some morphological evidence such as brain structure (Harzsch 2002), and by comparative studies of embryonic development (Averof and Akam 1995). On the other hand, a clade consisting of Branchipoda and Hexapoda tends to be supported by rRNA genes and multigene analyses (Regier et al. 2005, 2008; Mallatt and Giribet 2006; Roeding 2007; Dunn et al. 2008).

Here we provide molecular evidence that actually the Remipedia may be the closest living relatives of the Hexapoda. This relationship is suggested in our study 1) by the presence of both hexapod-type hemocyanin subunits in Remipedia (but absence of any other subunit type), 2) by a unique sequence motif insertion that is shared StuHc2 and hexapod type 2 subunits, and 3) by the orthology of remipede and hexapod hemocyanin subunits (Fig. 3), which is also supported by statistical tests of alternative tree topologies. A sister-group relationship of Remipedia and Hexapoda was tentatively retrieved in another molecular tree using elongation factor 2 sequences (Regier and Shultz 2001) and was also suggested by Telford et al. (2008) on the basis of unpublished material. In a large-scale phylogenomic study by Regier et al. (2008), a clade consisting of Remipedia and Hexapoda was consistently recovered, which in that study also included the Cephalocarida (which are not considered here).

Interestingly, a recent morphological study supports a close relationship of Remipedia and Hexapoda: Based on the structure of the arthropod brain, Fanenbruck and co-workers (Fanenbruck, Harzsch, and Wagele 2004) proposed a common clade of Malacostraca, Remipedia and Hexapoda (although they did not include Myriapoda in their analyses) with the exclusion of Branchiopoda and "Maxillopoda". It was noted that the arrangement of nerves, axonal tracts, neuropil compartments and cell clusters in the brains of these taxa are similar, but distinct from that of any other crustacean class. This hypothesis is now reinforced by the fact that all three taxa employ hemocyanin as respiratory protein, whereas other crustacean taxa use hemoglobins (Mangum 1985; Markl and Decker 1992; Burmester 2002). Therefore, Pancrustacea may be divided into two clades consisting of the possibly monophyletic Entomostraca (which have lost hemocyanins) on one hand (Walossek 1999), and Malacostraca, Remipedia and Hexapoda (taxon N.N.) on the other (Fanenbruck, Harzsch, and Wagele 2004). However, it should be noted that we cannot reject the hypothesis of multiple independent losses of hemocyanin and independent evolution of hemoglobin in various "entomostracan" lineages. If this is the case and the cited morphological evidence is ignored, the minimal topology recovered in our analyses is

(Malacostraca, (Remipedia, Hexapoda)). Then an "entomostracan", possibly branchiopod, ancestor of Hexapoda cannot be excluded (Regier et al. 2005; Mallatt and Giribet 2006; Roeding et al. 2007; Dunn et al. 2008).

Conclusions: Remipedia-like ancestor of Hexapoda?

Although recent hypotheses focused on a branchiopod-like crustacean as possible ancestor of hexapods (Glenner et al. 2006), our findings suggest that alternatively a remipede ancestor could be proposed. Extant and fossil Remipedia may provide evidence for the *Bauplan* of the hexapod stem-lineage. The oldest known hexapod fossil is the collembolan *Rhyniella praecursor* from the Devonian Rhynie *Lagerstätte* in Scotland. This species harbors various derived features, thus offering little information on early morphology of Hexapoda. Interestingly, Haas, Waloszek, and Hartenberger (2003) noted that *Tesnusocaris goldichi*, a remipedian fossil from the Carboniferous (Emerson and Schram 1991) may be a stem-lineage hexapod. This proposal was based on the unbranched structure of the first antenna and the long and filiform caudal appendages. It may be further speculated that the enigmatic fossil *Devonohexapodus bocksbergensis* (Haas, Waloszek, and Hartenberger 2003), which was described as a marine hexapod from the lower Devonian, may represent a transitional form between Remipedia and Hexapoda. Hexapod-like structures include leg-like palps of maxillae, absence of a second pair of antennae, and three pairs of longer uniramous thoracopods with six podomeres, but the homonomous trunk with its 38 trunk segments and the “abdominal” leglets actually provide an overall Remipedia-like appearance. In summary, it is possible that both fossil and living Remipedia occupy a key position for unraveling the evolution of Pancrustacea and thus for the understanding of morphological and functional innovations eventually resulting in the emergence of Hexapoda. However, additional molecular and morphological studies are required to unravel the position of Remipedia and to stably resolve their relationship to Hexapoda.

Supplementary Material

Supplementary Table 1. List of sequences used for phylogenetic analysis.

Supplementary Fig. 1. Nucleotide and amino acid sequences of *S. tulumensis* hemocyanin subunits.

Supplementary Fig. 2. Multiple sequence alignment of selected arthropod hemocyanins, crustacean pseudo-hemocyanins and insect hexamerins.

Supplementary Fig. 3. Maximum likelihood tree.

Supplementary Fig. 4. Bayesian phylogenetic tree.

Acknowledgements

We wish to thank Christian Pick for providing arthropod hemocyanin sequences and advice. We are grateful to Tom Iliffe, Brett Gonzalez and Jesse Henson for their help with collecting the *S. tulumensis* specimens. This work is supported by the Deutsche Forschungsgemeinschaft (Bu956/8, Bu956/9, Mi649/6, Wa530/34) within the priority project 1174 "Deep Metazoan Phylogeny".

Literature Cited

Abascal, F., R. Zardoya, and D. Posada. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics **21**:2104-2105.

Averof, M., and M. Akam. 1995. Hox genes and the diversification of insect and crustacean body plans. Nature **376**:420-423.

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

Brusca, R. C., and G. J. Brusca. 2003. Invertebrates. Sinauer Associates, Sunderland Mass.

Burmester, T. 2001. Molecular evolution of the arthropod hemocyanin superfamily. Mol. Biol. Evol. **18**:184-195.

Burmester, T. 2002. Origin and evolution of arthropod hemocyanins and related proteins. J. Comp. Physiol. [B] **172**:95-107.

Burmester, T., and K. Scheller. 1996. Common origin of arthropod tyrosinase, arthropod hemocyanin, insect hexamerin, and dipteran arylphorin receptor. J. Mol. Evol. **42**:713-728.

Carapelli, A., P. Lio, F. Nardi, E. van der Wath, and F. Frati. 2007. Phylogenetic analysis of mitochondrial protein coding genes confirms the reciprocal paraphyly of Hexapoda and Crustacea. BMC Evol. Biol. **7 Suppl 2**:S8.

Dunn, C. W., A. Hejnol, D. Q. Matus, K. Pang, W. E. Browne, S. A. Smith, E. Seaver, G. W. Rouse, M. Obst, G. D. Edgecombe, M. V. Sorensen, S. H. D. Haddock, A. Schmidt-Rhaesa, A. Okusu, R. M. Kristensen, W. C. Wheeler, M. Q. Martindale, and G. Giribet. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature **452**:745-749.

Emerson, M. J., and F. R. Schram. 1991. Remipedia, part 2: paleontology. Proc. San Diego Soc. Nat. Hist. **7**:1-52.

Fanenbruck, M., and S. Harzsch. 2005. A brain atlas of *Godzilliognomus frondosus* Yager, 1989 (Remipedia, Godzilliidae) and comparison with the brain of *Speleonectes tulumensis* Yager,

1987 (Remipedia, Speleonectidae): implications for arthropod relationships. Arthropod Struc. Dev. **34** 343–378.

Fanenbruck, M., S. Harzsch, and J. W. Wagele. 2004. The brain of the Remipedia (Crustacea) and an alternative hypothesis on their phylogenetic relationships. Proc. Natl. Acad. Sci. U S A **101**:3868-3873.

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

Garcia-Machado, E., M. Pempera, N. Dennebouy, M. Oliva-Suarez, J. C. Mounolou, and M. Monnerot. 1999. Mitochondrial genes collectively suggest the paraphyly of Crustacea with respect to Insecta. J. Mol. Evol. **49**:142-149.

Gaykema, W. P. J., W. G. J. Hol, J. M. Vereijken, N. M. Soeter, H. J. Bak, and J. J. Beintema. 1984. 3.2 Å structure of the copper-containing, oxygen-carrying protein *Panulirus interruptus* haemocyanin. Nature **309**:23-29.

Giribet, G., G. D. Edgecombe, and W. C. Wheeler. 2001. Arthropod phylogeny based on eight molecular loci and morphology. Nature **413**:157-161.

Giribet, G., and C. Ribera. 2000. A review of arthropod phylogeny: New data based on ribosomal DNA sequences and direct character optimization. Cladistics **16**:204–231.

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

Goldman, N., J. P. Anderson, and A. G. Rodrigo. 2000. Likelihood-based tests of topologies in phylogenetics. Syst. Biol. **49**:652-670.

Haas, F., D. Waloszek, and R. Hartenberger. 2003. *Devonohexapodus bocksbergensis*, a new marine hexapod from the Lower Devonian Hunsrück Slates, and the origin of Atelocerata and Hexapoda Org. Divers. Evol. **3**:39-54.

Hagner-Holler, S., A. Schoen, W. Erker, J. H. Marden, R. Rupprecht, H. Decker, and T. Burmester. 2004. A respiratory hemocyanin from an insect. Proc. Natl. Acad. Sci. U S A **101**:871-874.

Harzsch, S. 2002. The phylogenetic significance of crustacean optic neuropils and chiasmata: a re-examination. J Comp Neurol **453**:10-21.

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. Phylogenet. Evol. **38**:100-116.

Hazes, B., K. A. Magnus, C. Bonaventura, J. Bonaventura, Z. Dauter, K. H. Kalk, and W. G. Hol. 1993. Crystal structure of deoxygenated *Limulus polyphemus* subunit II hemocyanin at 2.18 A resolution: clues for a mechanism for allosteric regulation. Protein Sci. **2**:597-619.

Herberts, C., and J. de Frescheville. 1981. Occurrence of hemocyanin in. the rhizocephalan crustacea *Sacculina carcini* Thompson. Comp. Biochem. Physiol. B **70**:657-659

Huelsenbeck, J. P., and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics **17**:754-755.

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

Ito, T. 1989. Origin of the basis in copepod limbs, with reference to remipedian and cephalocarid limbs. J. Crust. Biol. **9**:85–103.

Jaenicke, E., H. Decker, W. Gebauer, J. Markl, and T. Burmester. 1999. Identification, structure, and properties of hemocyanins from diplopod Myriapoda. J. Biol. Chem. **274**:29071-29074.

Katoh, K., K. Kuma, H. Toh, and T. Miyata. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. **33**:511-518.

Kishino, H., and M. Hasegawa. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J. Mol. Evol. **29**:170-179.

Koenemann, S., J. Olesen, F. Alwes, T. M. Iliffe, M. Hoenemann, P. Ungerer, C. Wolff, and G. Scholtz. 2009. The post-embryonic development of Remipedia (Crustacea) - additional results and new insights. Dev. Genes Evol. **219**:131-145.

Koenemann, S., F. R. Schram, A. Bloechl, T. M. Iliffe, M. Hoenemann, and C. Held. 2007a. Post-embryonic development of remipede crustaceans. Evol. Dev. **9**:117-121.

Koenemann, S., F. R. Schram, M. Hönemann, and T. M. Iliffe. 2007b. Phylogenetic analysis of Remipedia (Crustacea). Org. Divers. Evol. **7**:33-51.

Koenemann, S., F. R. Schram, and T. M. Iliffe. 2006. Trunk segmentation in Remipedia. Crustaceana **79** 607-631.

Koenemann, S., F. R. Schram, T. M. Iliffe, L. M. Hinderstein, and A. Bloechl. 2007c. The behavior of Remipedia (Crustacea), with supporting field observations. J. Crust. Biol. **27**:534-542.

Kusche, K., and T. Burmester. 2001. Diplopod hemocyanin sequence and the phylogenetic position of the Myriapoda. Mol. Biol. Evol. **18**:1566-1573.

Kusche, K., A. Hembach, S. Hagner-Holler, W. Gebauer, and T. Burmester. 2003. Complete subunit sequences, structure and evolution of the 6 x 6-mer hemocyanin from the common house centipede, *Scutigera coleoptrata*. Eur. J. Biochem. **270**:2860-2868.

Kusche, K., H. Ruhberg, and T. Burmester. 2002. A hemocyanin from the Onychophora and the emergence of respiratory proteins. Proc. Natl. Acad. Sci. U S A **99**:10545-10548.

Lavrov, D. V., W. M. Brown, and J. L. Boore. 2004. Phylogenetic position of the Pentastomida and (pan)crustacean relationships. Proc. Biol. Sci. **271**:537-544.

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

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

Mallatt, J. M., and G. Giribet. 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

Mangum, C. P. 1985. Oxygen transport in invertebrates. Am. J. Physiol. **248**:505-514.

Markl, J., and H. Decker. 1992. Molecular structure of the arthropod hemocyanins. Adv. Comp. Environm. Physiol. **13**:325-376.

Nicholas, K. B., and H. B. J. Nicholas. 1997. GeneDoc: Analysis and Visualization of Genetic Variation.

Nielsen, H., J. Engelbrecht, S. Brunak, and G. von Heijne. 1997. A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Int. J. Neural Syst. **8**:581-599.

Pick, C., S. Hagner-Holler, and T. Burmester. 2008. Molecular characterization of hemocyanin and hexamerin from the firebrat *Thermobia domestica* (Zygentoma). Insect Biochem. Mol. Biol. **38**:977-983.

Pick, C., M. Schneuer, and T. Burmester. 2009. The occurrence of hemocyanin in Hexapoda. FEBS J. **276**:1930-1941.

Pohlmann, J. W., T. M. Iliffe, and L. A. Cifuentes. 1997. A stable isotope study of organic cycling and the ecology of an anchialine cave ecosystem. Marine Ecology Progress Series **155** 17-27.

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

Regier, J. C., J. W. Shultz, and R. E. Kambic. 2005. Pancrustacean phylogeny: hexapods are terrestrial crustaceans and maxillopods are not monophyletic. Proc. Biol. Sci. **272**:395-401.

Regier, J. C., J. W. Shultz, A. R. Ganley, A. Hussey, D. Shi, B. Ball, A. Zwick, J. E. Stajich, M. P. Cummings, J. W. Martin, and C. W. Cunningham. 2008. Resolving arthropod phylogeny: exploring phylogenetic signal within 41 kb of protein-coding nuclear gene sequence. Syst. Biol. **57**:920-938.

Richter, S. 2002. The Tetraconata concept: hexapod-crustacean relationships and the phylogeny of Crustacea. Organ. Divers. Evol. **2**:217–237.

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.

Schmidt, H. A., K. Strimmer, M. Vingron, and A. von Haeseler. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics **18**:502-504.

Schram, F. R. 1986. Crustacea. Oxford University Press, Oxford, New York.

Schram, F. R., and C. A. Lewis. 1989. Functional morphology of feeding in Nectiopoda. Pp. 115–122 *in* B. E. Felgenhauer, L. Watling, and A. B. Thistle, eds. Crustacean Issues 6: Functional Morphology of Feeding and Grooming in Crustacea. A. A. Balkema Press, Rotterdam.

Shimodaira, H., and M. Hasegawa. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. **16**:1114-1116.

Spears, T., and L. G. Abele. 1997. Crustacean phylogeny inferred from 18S rDNA. Pp. 169-187 *in* R. A. Fortey, and R. H. Thomas, eds. Arthropod Relationships. Chapman and Hall, London.

Strimmer, K., and A. Rambaut. 2002. Inferring confidence sets of possibly misspecified gene trees. Proc. Biol. Sci. **269**:137-142.

Telford, M. J., S. J. Bourlat, A. Economou, D. Papillon, and O. Rota-Stabelli. 2008. The evolution of the Ecdysozoa. Philos. Trans. R. Soc. Lond. B Biol. Sci. **363**:1529-1537.

van der Ham, J. L., and B. E. Felgenhauer. 2007. The functional morphology of the putative injecting apparatus of *Speleonectes tanumekes* (Remipedia). J. Crust. Biol. **27**:1–9.

van der Ham, J. L., and B. E. Felgenhauer. 2008. Ultrastructure and functional morphology of glandular setae and distal claws of cephalic appendages of *Speleonectes tanumekes* (Crustacea: Remipedia). Arthropod Struct. Dev. **37**:235-247.

van Holde, K. E., and K. I. Miller. 1995. Hemocyanins. Adv. Protein Chem. **47**:1-81.

von Reumont, B. M., K. Meusemann, N. U. Szucsich, E. Dell' Ampio, V. Gowri-Shankar, D. Bartel, S. Simon, H. O. Letsch, R. R. Stocsits, Y. X. Luan, J. W. Waegele, G. Pass, H. Hadrys, and B. Misof. 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. 1999. On the Cambrian diversity of Crustacea. Pp. 3–27 *in* F. R. Schram, and J. C. von Vaupel Klein, eds. Crustaceans and the Biodiversity Crisis, Proceedings of the Fourth International Crustacean Congress. Brill, Leiden.

Whelan, S., and N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. **18**:691-699.

Wills, M. 1997. A phylogeny of recent and fossil Crustacea derived from morphological characters. Pp. 189–209 *in* R. A. Fortey, and R. H. Thomas, eds. Arthropod Relationships. Chapman & Hall, London.

Wilson, K., V. Cahill, E. Ballment, and J. Benzie. 2000. The complete sequence of the mitochondrial genome of the crustacean *Penaeus monodon*: are malacostracan crustaceans more closely related to insects than to branchiopods? Mol. Biol. Evol. **17**:863-874.

Yager, J. 1981. Remipedia, a new class of Crustacea from a marine cave in the Bahamas. J. Crust. Biol. **1**:328-333.

Yager, J. 1991. The Remipedia (Crustacea): Recent investigations of their biology and phylogeny. Verh. Dtsch. Zool. Ges. **84**:261-269.

Zrzavý, J., and P. Štys. 1997. The basic body plan of arthropods: insights from evolutionary morphology and developmental biology. J. Evol. Biol. **10**:353–367.

TABLE 1. –Statistical hypothesis tests.

Tree[1]	log L	difference	S.E.	SH[2]	1sKH[2]	2sKH[2]	ELW[2]
1	-74424.60	0.00	best	1.0000 +	1.0000 +	best	0.9983 +
2	-74509.58	84.98	21.5520	0.0000 -	0.0000 -	-	0.0000 -
3	-74469.85	45.25	12.0780	0.0410 -	0.0020 -	-	0.0017 -
4	-74460.23	35.63	12.6676	0.0120 -	0.0000 -	-	0.0000 -

[1]Tree #1: Optimal tree as presented in Fig. 3: StuHc1 and 3 associated with type 1 hexapod hemocyanins, and StuHc2 associated with type 2 hexapod hemocyanins. Tree #2: StuHc1, 2, and 3 in sister-group position to the other crustacean (malacostracan) hemocyanins. Tree #3: StuHc1 and 3 in sister-group position to the other crustacean (malacostracan) hemocyanins. Tree #4: StuHc2 in sister-group position to the other crustacean (malacostracan) hemocyanins. [2]Test methods: SH, Shimodaira-Hasegawa test; 2sKH, two sided Kishino-Hasegawa test; 1sKH, one sided KH test based on pairwise SH tests; ELW, Expected Likelihood Weight test. Plus signs (+) denote the confidence sets, minus signs (-) denote rejection of a tree. All tests used 5% significance level.

Figure Legends

FIG. 1.–Comparison of Remipedia (*S. tulumensis*) hemocyanin subunits (StuHc1-3) with the sequences of the hemocyanin subunits of the stonefly *P. marginata* (PmaHc1, PmaHc2) and the hemocyanin subunit a from the spiny lobster *Panulirus interruptus* (PinHcA). The copper-binding histidines are shaded in black, other conserved residues in light grey. The asterisks in the upper row denote the phenylananines that stabilize the O_2-binding. In the lower row, the secondary structure of PinHcA is given (Gaykema et al. 1984). The conserved insertion of nine amino acids in beta-sheet 3A of StuHc2 and hexapod hemocyanin subunit 2 are shaded in dark grey.

FIG. 2. –Identification of hemocyanin protein in *S. tulumensis*. About 20 µg of total protein extract from *S. tulumensis* were applied per lane and separated by SDS-PAGE. Hemocyanin subunits were detected by polyclonal antibodies raised against various insect and crustacean hemocyanins, as indicated in the upper row. The antibodies were: α-HamHc, anti-*Homarus americanus* hemocyanin; α-PalHc, anti-*Panulirus interruptus* hemocyanin; α-CpaHc, anti-*Cancer pagurus* hemocyanin; α-TdoHc1, anti-*Thermobia domestica* hemocyanin 1; α-TdoHc2, anti-*Thermobia domestica* hemocyanin 2. The positions of the molecular mass standards are given on the left side.

FIG. 3. – Simplified Bayesian phylogenetic tree of arthropod hemocyanins and hexamerins. The numbers at the nodes show maximum likelihood non-parametric bootstrap support (above) and Bayesian posterior probabilities (below). The complete trees are shown in Supplementary Figs. 3 and 4.

Fig. 1

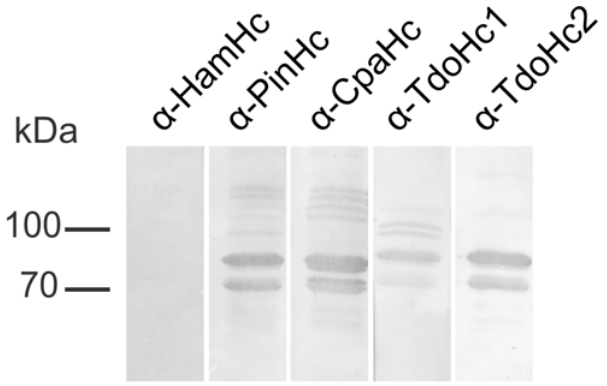


Fig. 2

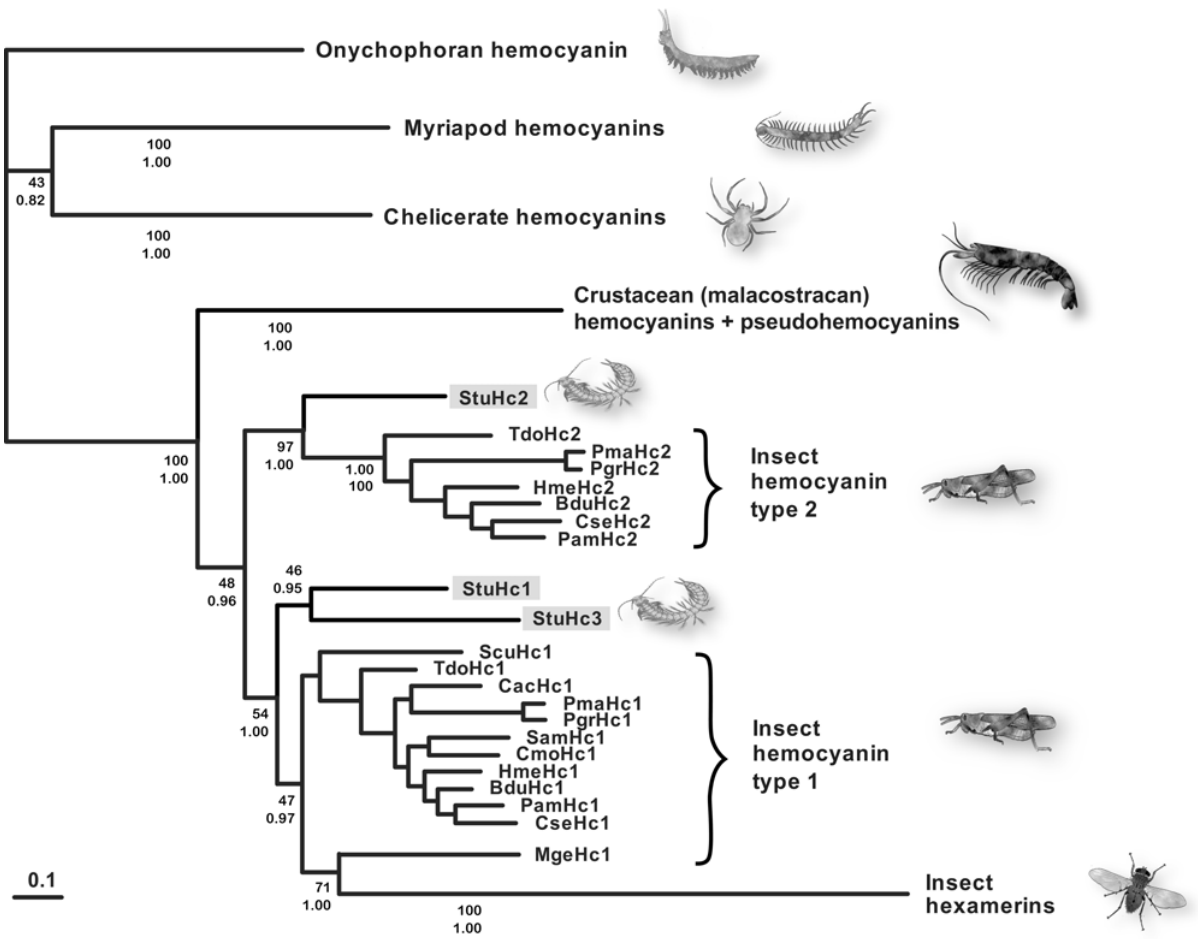

Onychophoran hemocyanin
Myriapod hemocyanins
Chelicerate hemocyanins
Crustacean (malacostracan)
hemocyanins + pseudohemocyanins
StuHc2
TdoHc2
PmaHc2
PgrHc2
HmeHc2
BduHc2
CseHc2
PamHc2
Insect
hemocyanin
type 2
StuHc1
StuHc3
ScuHc1
TdoHc1
CacHc1
PmaHc1
PgrHc1
SamHc1
CmoHc1
HmeHc1
BduHc1
PamHc1
CseHc1
Insect
hemocyanin
type 1
MgeHc1
Insect
hexamerins
0.1
100
1.00
43
0.82
100
1.00
100
1.00
97
1.00
1.00
100
100
1.00
48
0.96
46
0.95
54
1.00
47
0.97
71
1.00
100
1.00

Fig. 3

CURRICULUM VITAE - ERKLÄRUNG

CURRICULUM VITAE

Björn Marcus von Reumont

bmvr@arcor.de

UNIVERSITY

Since Oct	2005	PhD study: "Molecular insights to crustacean phylogeny" within the priority programm: "Deep Metazoan Phylgeny" of the DFG
Mar	2004-2005	Master thesis: „Phylogeography of the *Zygaena angelicae/transalpina* complex based on phylogenetic and Nested Clade Analysis using mt-DNA"
Feb	2004	Master exam: Zoology. Cell Biology and Chemistry

TEACHING EXPERIENCE

July	2008	Teaching assistent: "Marine Invertebrate Zoology"
May	2007	Teaching assistent: "Marine Invertabrate Zoology"
Dec	2006	Teaching assistent: "Phylogeny and Systematics"
Dec	2005	Teaching assistent: "Phylogeny and Systematics"
regularly		Introducing students to molecular systematics and laboratory work

FIELDWORK

Sep	2009	Fieldtrip to Ferrol. Spain (collection of crustaceans, focus: Mystacocarida)
Jan	2009	Fieldtrip to Ferrol. Spain (collection of crustaceans, focus: Leptostraca)
Nov	2008	Fieldtrip to the Yucatan Peninsula. Mexico (cave diving expedition for the collection of Remipedia)
Sep	2008	Fieldtrip to Sardinia. Italy (collection of crustaceans by diving)
Aug	2008	Fieldtrip to Isla di Giglio. Italy (collection of crustaceans by diving)
Jul	2007	Course "Tropical Marine Ecology" (MARB 656). Texas A&M Galveston (TAMUG)
		Collection of Remipedia in Mexican caves (Cavern diving)
May	2007	Fieldtrip to Isla di Giglio. Italy (collection of crustaceans by diving)
May	2006	Stay at the DZMB Wilhelmshaven. Germany. "Meionord-Expedition" on RV Heincke (sampling of benthic and pelagic meiofauna and other crustaceans in the "Waddensea")
Mar	2006	Fieldtrip to Ferrol. Spain (Biological Station Ferrol. collection of Mystacocarida and other crustaceans)
Jul	2004	Fieldtrip to Greece (collection of butterflies)
Jun	2004	Field trips in Germany. Austria and Czech Republic (butterflies collection)

WORKSHOPS (ORGANIZED)

Feb	2008	Co-organiser of the 5th meeting "European Basal Hexapod Workgroup" and the 1st "Arthropod Primer Toolbox Meeting", ZFMK Bonn
Nov	2007	Organiser of the workshop "Alignment masking and analyses of secondary structures", ZFMK Bonn, with AG Prof. Koenemann,
Nov	2006	Co-organiser of the third meeting "European Basal Hexapod Workgroup", ZFMK Bonn
Feb	2006	Co-organiser of the second meeting "European Basal Hexapod Workgroup", ZFMK Bonn

WORKSHOPS (PARTICIPATED)

Apr	2007	Phylogeny workshop. Vienna. Austria: "tree alignment and databases"
May	2006	"cDNA libraries and cloning". Max Planck Institute for Molecular Genetics. Berlin. Germany
Dec	2005	RNA extraction and cloning. Gutenberg University. Mainz. Germany

NON UNIVERSITARY ACTIVITIES

Since Sep	2008	technical diving, cave diving education
Since May	2007	Scuba diving
Since Oct	2005	Reviewer for MPE (Molecular Phylogenetics and Evolution)
Jun – Sep	2005	Scientific assistant for network and computational administration
Aug	1999 –	Student assistant of the "Informatikzentrum der Sparkassenorganisation GmbH
Apr	2005	(SIZ)"
Language skills		English fluently. Frensh basically

FUNDING

1998 – 2002	Molinari Stiftung
2004	Alexander Koenig Stiftung (travel allowance)
2005 – 2007	DFG (project funding)
2008	Alexander Koenig Stiftung (travel allowance)
2007 – 2009	DFG (project funding)
2009	Raiffeisenbank Rhein Sieg, Bonn (travel allowance)
2009 – 2011	DFG (project funding)
Equipment	Poseidon Tauchprodukte GmbH (2008) (equipment allowance)

PUBLICATIONS

Boye C. Boye P. Eiden K. Meusemann K. Meyer-Cords C. **von Reumont B.** Schweitzer U **(2002)**: Die Ofenkaulen im Siebengebirge als Fledermausquartier: Die aktuell vorkommenden Arten. Bestände und Gefährdungen. *Decheniana* (Bonn) 155. 81-103

Boye C. Meusemann K. Meyer-Cords C. **von Reumont B (2002)**: Fledermauskundler fangen viele Fledermäuse – Bericht von einer Netzfangaktion im Siebengebirge bei Bonn. *Nyctalus* (N.F.) 8. 231-239

Schmidt C; **von Reumont BM:** Myriapoda. *Brockhaus - Faszination Natur*. Wirbellose Tiere I. **(2006)**

Pirow R. Buchen I. Richter M. Allmer C. Nunes F. Günsel A. Heikens W. Lamkemeyer T. **v. Reumont BM**. Hetz SK **(2009)**: Cationic composition and acid-base state of the extracellular fluid and specific buffer value of hemoglobin from the branchiopod crustacean *Triops cancriformis. J. Comp. Phys. B, 179(3): 369-81.*

v. Reumont BM, Meusemann K, Szucsich NU, Dell'Ampio E, Gowri-Shankar V, Bartel D, Simon S, Letsch HO, Stocsits R, Luan YX, Wägele JW, Pass G, Hadrys H, Misof B **(2009)**: Can comprehensive background knowledge be incorporated into substitution models to improve phylogenetic analyses? A case study on major arthropod relationships. *BMC Evolutionary Biology 9:119.*

Ertas B. **v. Reumont BM**. Wägele. JW. Misof B. Burmester T **(2009)**: Hemocyanin suggests a close relationship of Remipedia and Hexapoda. *Molecular Biology and Evolution, Advanced access published, doi:10.1093/molbev/msp186.*

PRESENTATIONS

17.11.2006 (talk)	*DFG DMP Meeting Göttingen. Germany*; **v. Reumont**. Waegele. Waloßek: Phylogeny of the Entomostraca – a status report
15.03.2007 (talk)	*Deutsche Crustaceologen Tagung . Senckenberg Museum Frankfurt. Germany*; **v. Reumont**. Waegele. Waloßek: Die Crustacea im Deep Metazoan Phylogeny Programm
02.05.2007 (talk)	*DFG DMP Meeting Bonn. Germany*; **v. Reumont**. Waegele. Waloßek: *Phylogeny of the Crustacea formerly focused on the Entomostraca*
15.01.2008 (talk)	*DFG DMP Meeting Bonn. Germany*; **v. Reumont**. Waegele. Waloßek: Phylogenie der Crustacea
09.04.2008 (talk)	*10. GfBS Meeting. Göttingen. Germany*; Wägele. **v. Reumont**. Meusemann. Misof*:* Progress in Deep Metazoan phylogeny? On the relevance of data quality
06.2008 (talk)	*Dresdener Zoologische Kolloquien. Germany*; Misof. Meusemann. **v. Reumont**. Waegele: Deep phylogeny of arthropods
22.06.2008 (talk)	*Evolution 2008. Minnesota. USA*; **v. Reumont**. Meusemann. Misof. Waegele: A new approach to test the Mandibulata concept and to infer the phylogeny of Crustacea
22.06.2008 (talk)	*Evolution 2008. Minnesota USA ;* Meusemann. Misof. **v. Reumont**. Wägele. Roeding. Burmester: Insights form EST data for phylogenetic reconstruction of arthropod

relationships.

27.08.2008 (talk)	*XXth International congress of Zoology. Paris. France:* Wägele. **v. Reumont**. Meusemann. Misof: Metazoan phylogenetics: surprising new results and the depeptive phylogenetic signal
09.10.2008 (talk)	*Advances of crustacean phylogeny. Rostock. Germany*; Wägele. **v. Reumont**. Meusemann. Misof: Phylogenetic signal and noise in alignments.
30.01.2009 (talk)	*DMP Meeting Bonn, Germany*: **v. Reumont**, Waegele*:* Molecular phylogeny of crustaceans
04.03.2009 (talk)	*Celebrating Darwin: from the origin of species to Deep Metazoan phylogeny (2009). Berlin, Germany;* **v. Reumont**, Meusemann, Misof, Wägele: Molecular insights to crustacean phylogeny within arthropods – present, promises and prospective (home)work
04.03.2009 (talk)	*Celebrating Darwin: from the origin of species to Deep Metazoan phylogeny (2009). Berlin, German;* Meusemann, **v. Reumont**, Wägele, Misof: Inferring arthropod phylogeny from EST data: improvements, difficulties, prospects.
22.03.2009 (talk)	*Kuratorium Alexander Koenig Stiftung, ZFMK Bonn;* **v. Reumont**: Tauchen in das Erdaltertum – Auf der Suche nach den rätselhaften Remipdia
12.08.2009	*GfBS meeting 2009, Leiden, Netherlands;* Meusemann, Misof, Wägele, **v. Reumont**: Inferring arthropods phylogeny from EST data: Improvements, difficulties, prospects
20-23.02.2007 (poster)	*9. GfBS Meeting. Vienna. Austria*; **v. Reumont**. Waegele. Waloßek: Molecular phylogeny of the Entomostraca within the arthropods
07-11.04.2008 (poster)	*10. GfBS Meeting. Göttingen. Germany*; **v. Reumont**. Meusemann. Misof. Waegele*:* A new approach to test the Mandibulata concept and to infer the phylogeny of Crustacea
07-11.04.2008 (poster)	*10. GfBS Meeting. Göttingen. Germany*; Meusemann. **v. Reumont**. Misof. Waegele: The Atelocerata – a vanishing hypothesis? Molecular phylogeny of „basal" hexapods within arthropods – insights from EST data
19-22.09.2008 (poster)	*101. DZG Meeting. Jena. Germany*; Ertas. **v. Reumont**. Misof. Waegele. Burmester*:*A hemocyanin from *Speleonectes tulumensis* (Crustacea. Remipdia)
07-11.10.2008 (poster)	*Advances of Crustacean Phylogeny. Rostock. Germany*; **v. Reumont**. Meusemann. Misof. Wägele et al.: Improved realism of substitution models in rRNA data of Crustacea within Arthropoda.

ERKLÄRUNG

Hiermit erkläre ich, dass diese Arbeit selbstständig verfasst wurde, und keine anderen Quellen und Hilfsmittel als die angegebenen benutzt wurden. Die Stellen der Arbeit, die anderen Werken dem Wortlaut oder Sinn nach entnommen sind, wurden entsprechend gekennzeichnet.

Diese Arbeit liegt oder lag keiner anderen Prüfungsbehörde in ähnlicher oder anderer Form vor.

Ort, Datum Unterschrift