


# Major Revisions in Pancrustacean Phylogeny and Evidence of Sensitivity to Taxon Sampling

James P. Bernot <sup>\*,1,2</sup> Christopher L. Owen <sup>3</sup> Joanna M. Wolfe <sup>4</sup> Kenneth Meland,<sup>5</sup> Jørgen Olesen,<sup>6</sup> and Keith A. Crandall <sup>1,7</sup>

<sup>1</sup>Department of Invertebrate Zoology, US National Museum of Natural History, Smithsonian Institution, Washington, DC, USA

<sup>2</sup>Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT, USA

<sup>3</sup>Systematic Entomology Laboratory, USDA-ARS, % National Museum of Natural History, Smithsonian Institution, Washington, DC, USA

<sup>4</sup>Museum of Comparative Zoology and Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA

<sup>5</sup>Department of Biology, University of Bergen, Bergen, Norway

<sup>6</sup>Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark

<sup>7</sup>Department of Biostatistics and Bioinformatics, Milken Institute School of Public Health, George Washington University, Washington, DC, USA

\*Corresponding author: E-mail: james.bernot@uconn.edu.

Associate editor: Tal Pupko

## Abstract

The clade Pancrustacea, comprising crustaceans and hexapods, is the most diverse group of animals on earth, containing over 80% of animal species and half of animal biomass. It has been the subject of several recent phylogenomic analyses, yet relationships within Pancrustacea show a notable lack of stability. Here, the phylogeny is estimated with expanded taxon sampling, particularly of malacostracans. We show small changes in taxon sampling have large impacts on phylogenetic estimation. By analyzing identical orthologs between two slightly different taxon sets, we show that the differences in the resulting topologies are due primarily to the effects of taxon sampling on the phylogenetic reconstruction method. We compare trees resulting from our phylogenomic analyses with those from the literature to explore the large tree space of pancrustacean phylogenetic hypotheses and find that statistical topology tests reject the previously published trees in favor of the maximum likelihood trees produced here. Our results reject several clades including Caridoida, Eucarida, Multicrustacea, Vericrustacea, and Syncarida. Notably, we find Copepoda nested within Allotriocarida with high support and recover a novel relationship between decapods, euphausiids, and syncarids that we refer to as the Syneucarida. With denser taxon sampling, we find Stomatopoda sister to this latter clade, which we collectively name Stomatocarida, dividing Malacostraca into three clades: Leptostraca, Peracarida, and Stomatocarida. A new Bayesian divergence time estimation is conducted using 13 vetted fossils. We review our results in the context of other pancrustacean phylogenetic hypotheses and highlight 15 key taxa to sample in future studies.

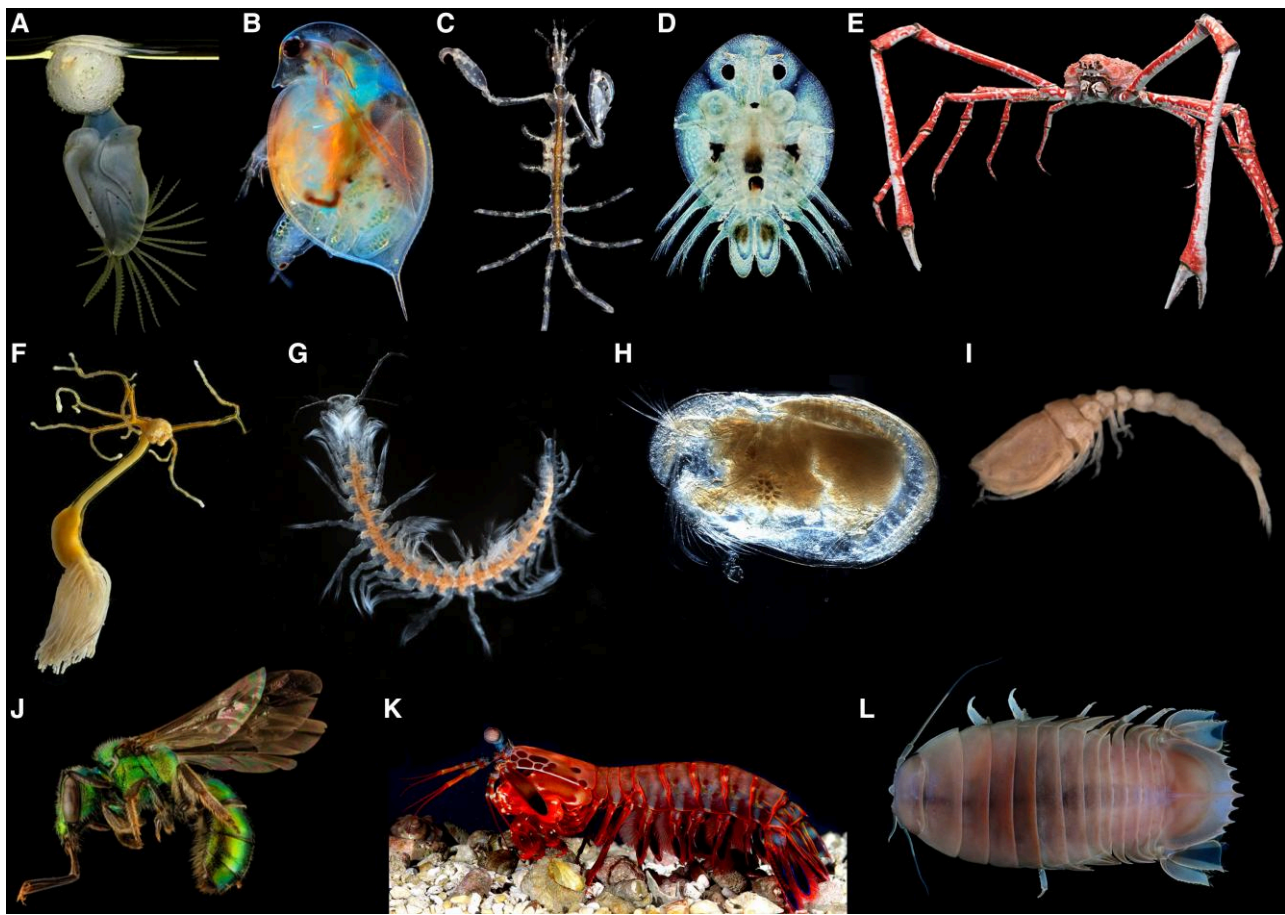
**Key words:** Crustacea, Pancrustacea, phylogeny, evolution, Malacostraca, copepod.

## Introduction

The clade Pancrustacea (“Crustacea” + Hexapoda) is arguably the most successful group of animals on earth. It comprises over 1,236,000 described species, contains more than 80% of extant animal diversity (Roskov et al. 2022), and includes nearly half of all animal biomass on the planet (Bar-On et al. 2018). Pancrustaceans have been a dominant component of earth’s ecosystems for nearly 600 million years (Wolfe et al. 2016). The group includes 57 orders of nonhexapod pancrustaceans (i.e., “crustaceans”) and 31 orders of hexapods (Bracken-Grissom and Wolfe 2020; WoRMS 2023). Many of the most economically important species on earth are pancrustaceans including bees, mosquitos, krill, copepods, and numerous other taxa with key positions in

terrestrial and aquatic food webs. The morphological diversity of body plans in this group is unparalleled among animals (fig. 1), for example, ranging from minute 70  $\mu\text{m}$  tantulocarid larvae (Huys et al. 1993; Petrunina et al. 2018) to Japanese spider crabs with a leg span up to 3.7 m (McClain et al. 2015). Although the species diversity of Pancrustacea is dominated by insects, the  $\sim 64,000$  species of “crustaceans” (WoRMS 2023) make up most of the phylogenetic diversity and morphological disparity of Pancrustacea. “Crustacean” diversity is composed of ten classes: Branchiopoda, Cephalocarida, Copepoda, Ichthyostraca (i.e., Branchiura + Pentastomida), Malacostraca, Mystacocarida, Ostracoda, Remipedia, Tantulocarida, and Thecostraca (WoRMS 2023).

Despite a number of recent pancrustacean phylogenomic studies (Regier et al. 2008, 2010; Andrew 2011;



**Fig. 1.** Morphological diversity of select pancrustaceans: (A) buoy barnacle (Cirripedia), (B) *Daphnia* sp. (Branchiopoda), (C) skeleton shrimp (Amphipoda), (D) *Argulus* sp. (Branchiura), (E) Japanese spider crab (Decapoda), (F) parasitic copepod (Copepoda), (G) *Lasionectes entrichoma* (Remipedia), (H) seed shrimp (Ostracoda), (I) comma shrimp (Cumacea), (J) sweat bee (Hexapoda), (K) mantis shrimp (Stomatopoda), and (L) giant isopod (Isopoda). Image credits: (A) David Fenwick, (B) Marek Miś, (C) David Fenwick, (D) Andrei Savitsky, (E) Michael Wolfe and Hans Hillewaert, (F) Geoff Boxshall, (G) Jørgen Olesen, (H) Anna Syme, (I) Hans Hillewaert, (J) USGS Bee Inventory and Monitoring Lab, (K) Roy L. Caldwell, and (L) Chan T.Y. and Lin C.W.

von Reumont et al. 2012; Oakley et al. 2013; Rota-Stabelli, Lartillot, et al. 2013; Schwentner et al. 2017, 2018; Lozano-Fernandez et al. 2019), the relationships among classes have been particularly challenging to reconstruct for a number of reasons. First, the most recent common ancestor (MRCA) of Pancrustacea is estimated to be >500 million years ago (Ma) (Oakley et al. 2013; Rota-Stabelli, Daley, et al. 2013; Schwentner et al. 2017; Wolfe 2017). The difficulty with confidently estimating deep-time relationships among taxa of this age is further compounded by the suggestion that these relationships are part of a rapid radiation. Ancient rapid radiations are some of the most difficult to resolve due to their age and short internal branch lengths (Fishbein et al. 2001; Rokas and Carroll 2006; Whitfield and Lockhart 2007; One Thousand Plant Transcriptomes Initiative 2019). Generally, to obtain robust nodal supports for rapid radiations, more genes or more taxa are sequenced to add as much information to those short branches as possible, but this is difficult for many pancrustaceans. Nearly half of the 57 “crustacean” orders have not been sampled in

multigene phylogenetic analyses. Many lineages are rare, small bodied, and often have large genomes (Alfsnes et al. 2017; Bracken-Grissom and Wolfe 2020), so maximizing sequence data through whole genome sequencing remains challenging and costly. Over such long time scales (>500 Ma), it is often not whole genes that are conserved, but individual exons. Pancrustaceans are known to have short exons (e.g., Owen et al. 2020), which yield shorter alignments with fewer parsimony informative sites to resolve branches in phylotranscriptomic studies. These exons can also have different evolutionary histories and are vulnerable to weak signal/noise ratios. Both issues can be particularly problematic for multispecies coalescent models (Huang et al. 2010; Bayzid and Warnow 2013; Patel et al. 2013; DeGiorgio and Degnan 2014; Mirarab et al. 2014; Lanier and Knowles 2015; Mirarab and Warnow 2015; Xi et al. 2015; Scornavacca and Galtier 2017). Unfortunately, a robust and comprehensive morphology-based phylogeny is also not available for Pancrustacea due to extreme morphological variation among groups and numerous convergences that make it challenging to

assign homology (Wolfe 2017; Lozano-Fernandez et al. 2019; Bracken-Grissom and Wolfe 2020).

Historically, different analytical approaches have been employed to overcome the difficulties in estimating a robust pancrustacean phylogeny. These include a variety of phylogenetic methods: partitioned analyses to buffer against heterogeneity in evolutionary rates across proteins and across lineages, site-heterogeneous methods (e.g., CAT-GTR and C60) to account for variation in amino acid (AA) frequencies across site, and Dayhoff 6-state recoding (Dayhoff6) to buffer against saturation and AA compositional heterogeneity. Different methods for selecting orthologs have also been used (e.g., sequence similarity vs. tree based). Despite the many different phylogenomic analyses, few have explicitly examined the effects of taxon sampling. Typically, a small number of novel taxa have been added for each study, whereas most data are reused from public databases; the resulting novel topologies are presented, often with an assumption that the additional taxa have led to improved accuracy.

Despite being integral to systematic study design, taxon sampling has received less focus relative to phylogenomic methodology in studies of pancrustacean phylogeny. Prior to the phylogenomics era, there was extensive debate over whether it is more important to sample more taxa or more nucleotide sites (reviewed in Nabhan and Sarkar 2012). In general, those arguing in favor of taxon sampling cite studies that have demonstrated that well-sampled phylogenies lead to more accurate topologies (e.g., Hedtke et al. 2006; Heath et al. 2008) and more accurate branch lengths (e.g., Hugall and Lee 2007) and reduce the number of long branches that may contribute to long branch attraction (LBA) (e.g., Hendy and Penny 1989; Poe 2003). To date, no phylogenomic study has demonstrated the effects of taxon sampling in relation to the pancrustacea phylogeny. This is important because the lack of comprehensive taxonomic coverage at the ordinal rank may impact topological accuracy and studies have shown that taxon sampling does impact tree topology even when thousands of loci are used. For example, Betancur-R et al. (2019) and Branstetter et al. (2017) demonstrated that taxon sampling and density both contribute to accuracy when using genomic data.

In addition to these challenges, many questions remain unanswered regarding the evolutionary relationships within Pancrustacea. Nearly half of the 57 “crustacean” orders have never been included in a multigene phylogeny, and relationships of the orders that have been sampled have often been unstable (Mallatt and Giribet 2006; Regier et al. 2008, 2010; Andrew 2011; von Reumont et al. 2012; Oakley et al. 2013; Rota-Stabelli, Lartillot, et al. 2013; Schwentner et al. 2017, 2018; Lozano-Fernandez et al. 2019). Some key areas of contention include the position of Hexapoda, interrelationships and validity of Multicrustacea (Copepoda, Malacostraca, Thecostraca), and relationships within the most speciose “crustacean” class, Malacostraca. To address these contentions, we estimate pancrustacean relationships with increased taxon sampling of 105 taxa represented by 90 transcriptomes and 15 genomes (supplementary table

S1, Supplementary Material online) and a tree-based approach for ortholog selection, which has been shown to improve phylogenetic reconstruction (Dunn et al. 2013; Yang and Smith 2014; Ballesteros and Hormiga 2016; Smith and Pease 2017). With the resulting phylogeny, we examine deep-level relationships of pancrustaceans and estimate the timing of their divergence using 13 fossils as calibration points. We also compare the results to two different taxon sampling schemes (table 1). Our results demonstrate that, even in the context of hundreds of orthologs, small changes in pancrustacean taxon sampling have major impacts on the resulting topology under all methods used here. We review our results in the context of all other pancrustacean phylogenomic studies to identify areas of conflict, and we suggest potential avenues for improving the resolution of the pancrustacean tree of life, especially by identifying the most crucial lineages to sample in future studies.

## Results

### Data Sets and Taxon Sampling

Our phylogenomic analyses (see supplementary Methods, Supplementary Material online) were conducted on two taxon sets: an earlier analysis (Data set 1) with a more malacostracan focus and a subsequent analysis (Data set 2) that was expanded slightly and better balanced across Pancrustacea (table 1; supplementary Tables S1 and S2, Supplementary Material online). Both data sets used the same methods for ortholog selection and phylogenetic analyses, but orthology inference was completed separately for each taxon set. Data set 2, the final iteration, had the largest and most balanced taxon set and the most robust results, so most of this study focuses on the results of those analyses (figs. 2, 3A–E, and 4; supplementary table S1 and S4, Supplementary Material online). However, through our investigation of these two taxon sets, we identified several reasons for topological differences among analyses and potential sources of error to be wary of, so we briefly summarize these iterations here.

Both data sets included all major pancrustacean clades with chelicerate and myriapod outgroups and collectively comprised 98 and 105 taxa, respectively (see supplementary table S1, Supplementary Material online, and fig. 2 for Data set 2 and supplementary table S2 and fig. S5, Supplementary Material online, for Data set 1). Both data sets were interrogated in detail with a variety of phylogenetic methods using the same parameters (partitioned maximum likelihood [ML], ML with site-heterogeneous C60 models, Bayesian inference [BI] site-heterogeneous CAT-GTR, partitioned ML and CAT-GTR of Dayhoff6 matrices, and coalescent methods). In all analyses of Data set 1, we recovered suspect relationships, particularly that in all analyses we recovered monophyletic Xenocarida (Remipedia + Cephalocarida) as sister to the rest of Allotriocarida (i.e., the clade consisting of Remipedia, Cephalocarida, Branchiopoda, and Hexapoda—here expanded to include Copepoda), and Copepoda as the sister to Hexapoda (fig. 3F–H). Given these surprising findings, we interrogated our taxon selection and added 13 hexapods, 3



**Table 1.** Comparison of the Data Sets Assembled in This Study.

Data Set	Num. Taxa	Num. “Crustacean” Orders	Orthologs	Alignment Length (AA)	Average Orthologs per Species	Average AA per Species	Average Species per Ortholog
Data set 1	98	28	559	80,215	355	48,742	62
Data set 2	105	30	576	121,508	363	64,018	66

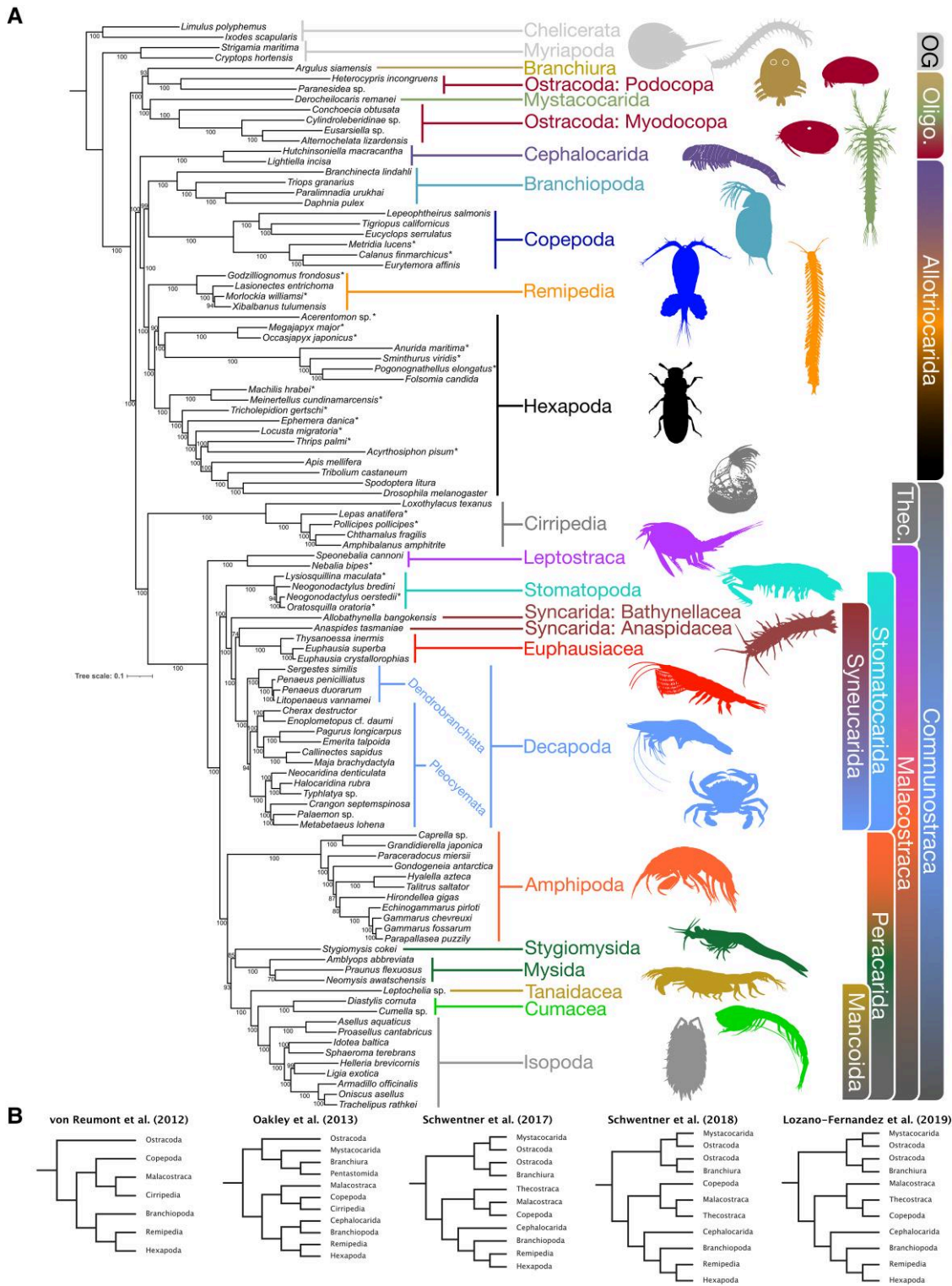
stomatopods, 2 remipedes, 2 copepods, 2 barnacles, and another leptostracan to break long branches (asterisks in [fig. 2](#)). To better balance the taxon sampling, we also removed 16 taxa that represented short branches at the tips of densely sampled clades. The taxa removed for the final data set comprised: 8 amphipods, 4 decapods, 3 isopods, and the second hymenopteran (red taxa in [supplementary fig. S5, Supplementary Material](#) online); this resulted in Data set 2 ([fig. 2A](#) and [tables 1](#) and [2](#); [supplementary table S1, Supplementary Material](#) online).

Data set 2, our final taxon set, is the primary focus of the study. We used genomic and transcriptomic data representing 30 of 57 “crustacean” orders and a phylogenetic diversity of hexapods. In total, 90 transcriptomes and 15 genomes spanning the arthropod tree of life were used ([supplementary table S1, Supplementary Material](#) online). We had particularly high representation of malacostracans not sampled in previous phylogenomic analyses. We identified 576 protein-coding genes with a taxon occupancy >50% with an ortholog occupancy of 84–533 ( $\bar{x} = 363$ , 63%) per species and a taxon occupancy of 53–99 ( $\bar{x} = 66$ , 62%) species per ortholog. The taxon and ortholog occupancy statistics are summarized in [supplementary table S1, Supplementary Material](#) online; statistics for each of the orthologs are given in [supplementary table S3, Supplementary Material](#) online. The final concatenated alignment consisted of 576 orthologs, 121,508 AA positions, and 91,467 parsimony informative sites.

### Taxon Sampling Results

Somewhat surprisingly, with the relatively small changes in taxon sampling between Data sets 1 and 2 (82 taxa shared, ~80% overlapping), we found substantial differences in the topology across all ML, BI, and coalescent-based analyses. Most notably, in Data set 1, Copepoda and Hexapoda were sister taxa in all analyses, and Remipedia and Cephalocarida were sister taxa (e.g., [fig. 3B–E](#) vs. [3F–H](#)). Given that identical methods were used to call orthologs and for phylogenetic reconstruction, we reasoned taxon sampling was driving the differences in topology. Taxon sampling can generally affect two parts of a phylogenomic analysis: 1) the ortholog selection (because clustering algorithms are sensitive to orthogroup structure, which is impacted by phylogenetic relatedness of taxa [[Chen et al. 2007](#); [Altenhoff and Dessimoz 2009](#)]) and 2) the accuracy of the species tree reconstruction. We sought to identify through which of these processes taxon sampling was having its effects. To investigate this, we first controlled for the effect of taxon sampling on ortholog selection by using only those

orthologs that were exclusively shared between the two different taxon data sets; that is, we used the same subset of genes for phylogenetic tree searches of both taxon sets. Approximately half of the orthologs were shared between the two data sets (267 of 559 and 576 orthologs in Data sets 1 and 2, respectively). Using these same genes and the same ML, BI, and coalescent methods, we again recovered very different topologies between these two taxon sets. In fact, the topologies inferred with this shared set of orthologs were nearly identical to the topologies recovered from analysis of the full matrices (i.e., all 559 and 576 orthologs, respectively) ([supplementary fig. S10A and B, Supplementary Material](#) online); within Data sets 1 and 2, the topologies from the full matrix and reduced matrices were completely congruent at the ordinal level and above, except for the position of the mysids, which was different within Data set 2 with low support in the 267 shared ortholog data set. To further explore the effects of taxon sampling, we pruned the additional taxa in Data set 2, removing the 23 taxa that were added relative to Data set 1 (i.e., 13 hexapods, 3 stomatopods, 2 remipedes, 2 copepods, 2 barnacles, and 1 leptostracan). ML analyses of the Data set 2 orthologs without these taxa once again recovered Remipedia and Cephalocarida as sister taxa ([supplementary fig. S11, Supplementary Material](#) online); that group was found in all analyses of Data set 1 ([supplementary figs. S5–S9, Supplementary Material](#) online) but was never recovered in Data set 2 when those additional taxa were included ([fig. 3F–H](#)). We examined long branch (LB) scores to compare the relative lengths of terminal branches in each phylogeny and estimate the degree of taxon-specific LBA ([Struck 2014](#); [Weigert et al. 2014](#)). The LB scores between Data sets 1 and 2 suggest that the longest branches in the phylogeny are attributed to the Cirripedia, Copepoda, Ostracoda: Podocopa, Hexapoda, Branchiopoda, Ostracoda: Myodocopa, and the outgroup. In the Data set 1 phylogeny, the LB scores ranged from –29.3 to 50.3, whereas in the Data set 2 phylogeny, they ranged from –28.3 to 37.1 ([supplementary table S5, Supplementary Material](#) online). In Data set 1, the top 10% of the largest LB scores included taxa within Cirripedia (*Loxothylacus texanus*), Copepoda (*Lepeophtheirus salmonis*, *Tigriopus californicus*, and *Eurytemora affinis*), Hexapoda (*Drosophila melanogaster* and *Folsomia candida*), Branchiopoda (*Branchinecta lindahli*), and Ostracoda: Myodocopa (*Conchoecia obtusata*). In Data set 2, Branchiopoda, Ostracoda: Podocopa, and Ostracoda: Myodocopa were not in the top 10%, whereas the composition of Copepoda and Hexapoda taxa within the top 10% LB scores changed from Data set to 2. The phylogenetic results of the rest of our study focus primarily on the final taxon set, Data set 2.

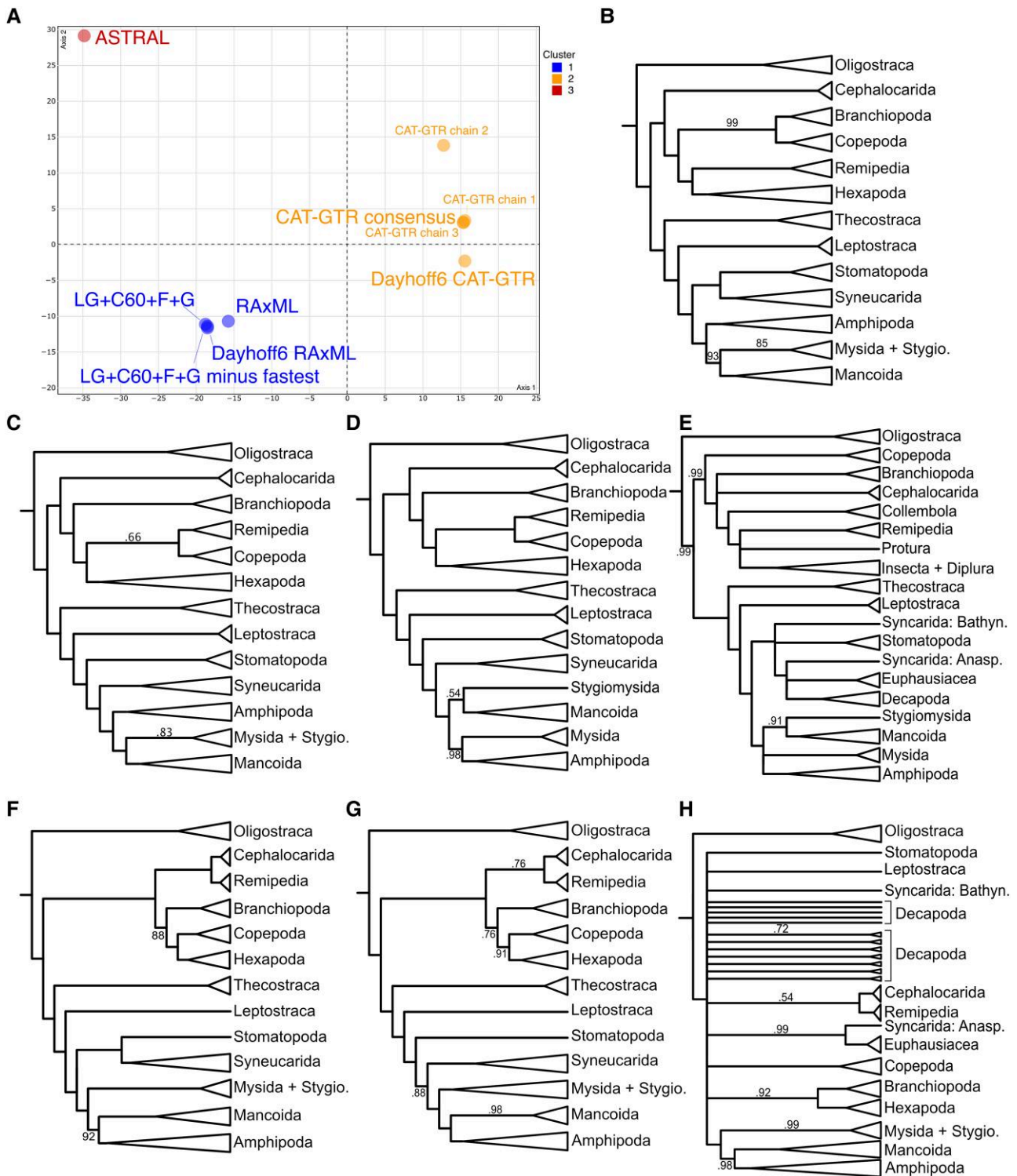


**Fig. 2.** (A) Tree resulting from ML analysis of the Data set 2 AA matrix using LG + C60 + F + G model. Species marked with an asterisk are additional taxa included in Data set 2 relative to Data set 1. (B) Recent phylogenomic hypotheses of pancrustacean relationships.

### Phylogenetic Results

We completed phylogenetic analysis using the following methods: partitioned ML analyses with RAXML, site-heterogenous ML analyses (C60 family of models) with IQTREE2, BI site-heterogenous CAT-GTR with PhyloBayes, and coalescent analyses with ASTRAL-III. We

also recoded the matrix into Dayhoff6 states to buffer against potential effects of saturation and across lineage compositional heterogeneity; we analyzed the recoded matrix under the CAT-GTR model in PhyloBayes and with a partitioned analysis under the GTR model in RAXML. Although some topological differences occurred

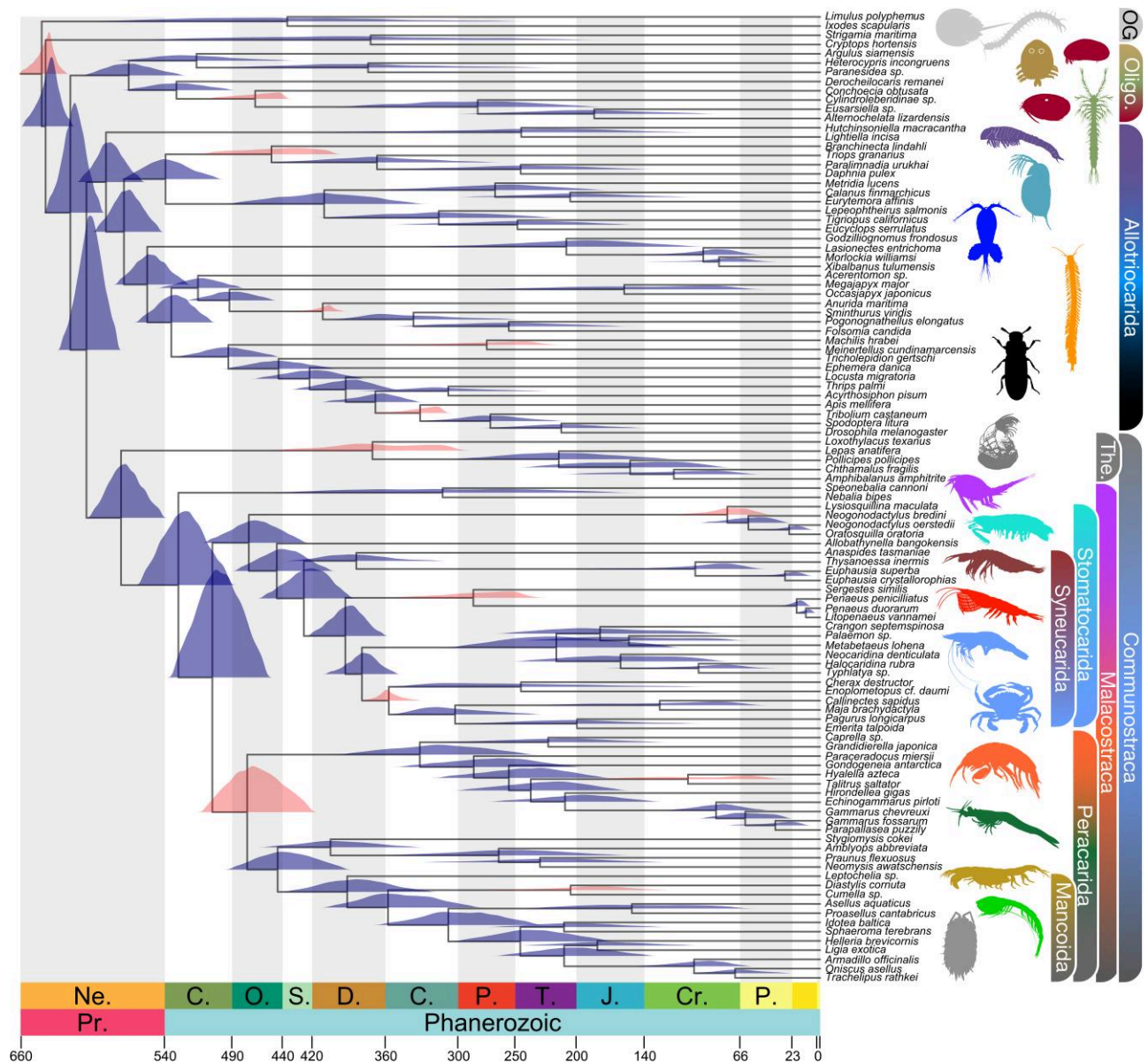


**Fig. 3.** Topological comparisons. Data set 2: (A) MDS of topological tree space of phylogenetic analyses in this study using the [Kendall and Colijn \(2016\)](#) method for defining summary trees, (B) C60 + LG + F + G, (C) CAT-GTR majority-rule consensus, (D) CAT-GTR majority-rule consensus of Dayhoff6 matrix, and (E) ASTRAL resulting from analysis of gene tree nodes with <30% BS collapsed and nodes <0.5 PP collapsed. Data set 1: (F) C60 + LG + F + G, (G) CAT-GTR majority-rule consensus, and (H) CAT-GTR majority-rule consensus of Dayhoff6 matrix with nodes <0.5 PP collapsed (nodes differed between chains after >80,000 generations).

across the BI, ML, and coalescent-based methods, there was general agreement across analyses, especially among ML analyses (fig. 3A; [supplementary figs. S1–S4](#), [Supplementary Material](#) online). All methods supported the

following topological arrangements: Oligostraca is the first group of pancrustaceans to diverge from all the others and contains a polyphyletic Ostracoda; within Altocrustacea (i.e., all pancrustaceans except Oligostraca), Thecostraca is the





**Fig. 4.** Fossil calibrated divergence time estimates for Pancrustacea, based on an MCMCtree analysis of the topology depicted in [figure 2A](#) calibrated with 13 vetted fossils. The fossil calibrated nodes have their posterior age distributions highlighted in light pink.

sister to the Malacostraca (Communostraca hypothesis of [Regier et al. \[2010\]](#)); Allotriocarcia (i.e., that clade consisting of Remipedia, Cephalocarida, Branchiopoda, and Hexapoda) is the sister to Communostraca but is expanded to include Copepoda; Leptostraca is the sister to all other malacostracans; Peracarida is monophyletic; and Decapoda and Euphausiacea form a clade with a paraphyletic Syncarida. The presence of the two clades of syncarids in this latter grouping, with Euphausiacea being closer to one of them (Anaspidacea) than to Decapoda, renders the classically recognized Eucarida polyphyletic, and we propose the name Syncarida for this expanded clade ([figs. 2A](#) and [3B–E](#); [supplementary figs. S1–S4](#), [Supplementary Material](#) online).

Although results from the ML, BI, and coalescent methods showed a degree of congruence, the topologies differed in

some parts of the pancrustacean tree ([fig. 3](#); [supplementary figs. S1–S3](#), [Supplementary Material](#) online). The tree resulting from the ML analysis using the LG + C60 + F + G mixture model ([fig. 2A](#)) is presented as our primary species tree for five reasons: 1) it was the substitution model of best fit by Bayesian Information Criterion (BIC) with the caveat that the CAT-GTR model cannot be tested with standard tests of model fit; 2) the AU-test ([Shimodaira 2002](#)) rejected the other topologies produced by other methods (i.e., ASTRAL and BI) in favor of this one ( $P < 0.001$ ) ([supplementary table S6](#), [Supplementary Material](#) online); 3) the PhyloBayes CAT-GTR analysis with three chains analysis did not fully converge (maxdiff between chains 0.69–1) even after nearly 1 year of continuous run time; 4) site-heterogeneous models like the C60 class account for among site variation in AA propensities

**Table 2.** Comparison of Recent Pancrustacean Phylogenomic Analyses.

Study	Number “Crustacean” Orders	Total Taxa (Number of “Crustaceans”)	Number Orthologs	AA Positions	Gaps per Ortholog	% Gaps in Super Matrix	Orthology Inference
This study (Data set 2)	30	105 (83)	576	121,508	10%	47%	Tree-based
This study (Data set 1)	28	98 (88)	560	80,215	11%	40%	Tree-based
Lozano-Fernandez et al. (2019), Matrix B	23	140 (58)	2,718	53,039	23%	28%	Sequence similarity
Lozano-Fernandez et al. (2019), Matrix A	23	140 (58)	244	57,149	25%	25%	Tree-based
Schwentner et al. (2018)	25	97 (83)	455	112,993	NA	38%	Sequence similarity
Schwentner et al. (2017)	19	40 (26)	1077	301,748	NA	34%	Sequence similarity
Lozano-Fernandez et al. (2016)	6	30 (11)	246	40,657	NA	36%	Tree-based
Oakley et al. (2013)	22	93 (84)	1,002	263,306	NA	80%	Sequence similarity
von Reumont (2012)	14	91 (30)	316	62,638	NA	38%	Sequence similarity

are less prone to artifacts like LBA (Lartillot and Philippe 2004; Lartillot et al. 2007; Le et al. 2008) and were found to converge in this data set in a reasonable time frame; and 5) there was high congruence between this topology and those produced from other methods, including nearly identical topologies (i.e., differing only in a few terminal branches) with the partitioned RAXML analysis with and without Dayhoff6 recoding (fig. 3A; supplementary fig. S1A–C, Supplementary Material online). The resulting LG + C60 + F + G phylogeny is the focus of most of this study, but differences between this topology and those from the other analyses are reviewed in the following paragraphs.

In addition to the relationships noted in the preceding paragraph that were recovered unanimously in all BI, ML, and coalescent-based methods used here, the tree resulting from the LG + C60 + F + G model had other notable findings. First, Stomatopoda was recovered as the sister to Syneucarida with 100% bootstrap (BS) support (this was also recovered in all the other ML analyses). Given the consistency of this more derived position for Stomatopoda in ML and under the multispecies coalescent (fig. 3E; supplementary figs. S1 and S4, Supplementary Material online), we propose the name Stomatocarida for the new clade comprising Stomatopoda and Syneucarida; this divides Malacostraca into three clades: Leptostraca, Stomatocarida, and Peracarida. Second, all ML analyses recovered Amphipoda as the sister to the other peracarids, with a clade comprising Stygiomysida and Mysida as sister to Mancoida (fig. 2A; supplementary fig. S1A–C, Supplementary Material online).

Despite general agreement, the majority-rule posterior consensus trees under the CAT-GTR model differed from the ML topologies in three main respects. First, analyses with the CAT-GTR model of both the AA matrix and Dayhoff6 recoded matrix found Remipedia + Copepoda as the sister to Hexapoda (fig. 3C and D; supplementary figs. S2 and S3, Supplementary Material online). This had maximum support in the Dayhoff6 analysis (supplementary fig. S3, Supplementary Material online) but low support (0.66 PP) with the full AA matrix (supplementary fig. S2A, Supplementary Material online) because one of the three chains recovered copepods as the sister to Remipedia + Hexapoda rather than the sister to Remipedia alone (supplementary fig. S2C, Supplementary Material online).

Second, Stomatopoda was recovered as the sister taxon to all other malacostracans except for the Leptostraca under the CAT-GTR model (fig. 3C and D; supplementary fig. S2A and B, Supplementary Material online). Third, relationships within Peracarida differed. Amphipoda was sister to the other peracarids in all ML analyses (including Dayhoff6 recoding) and in CAT-GTR analyses of the full AA matrix (fig. 3B and C), but CAT-GTR analysis of the Dayhoff6 matrix recovered Amphipoda as the sister to Mysida and Stygiomysida as the sister to Mancoida; this latter relationship was also found in ASTRAL analyses (fig. 3D and E). Variation regarding the Mysida and Stygiomysida is likely due to the fact that the mysids and stygiomysid had some of the lowest ortholog occupancy (*Praunus*, *Amblyops*, and *Stygiomysis* ranked first, third, and fifth in fewest orthologs, respectively) (supplementary table S1, Supplementary Material online). The only other differences among the ML and BI analyses were in shallow nodes within clades, such as between the three mysids (one of the few branches to vary among ML analyses) and the amphipod genera *Gondogenia* and *Hirondellia* (supplementary figs. S1A–C and S2A and B, Supplementary Material online).

ASTRAL topologies were very divergent relative to ML and BI (fig. 3A, B, and E and supplementary figs. S1–S4, Supplementary Material online). Coalescent methods are particularly susceptible to a few sources of error that exist in the evolutionary history of pancrustaceans. Because genes usually consist of multiple exons, which can have different evolutionary trajectories over the >500 Ma history of Pancrustacea, this can violate the nonrecombination assumption of coalescent methods (Scornavacca and Galtier 2017). Furthermore, short orthologs (e.g., average of 211 AA here) may suffer from relatively weak signal-to-noise ratios and high gene tree error. As noted by others, summary methods like ASTRAL can be inappropriate when gene tree estimation error is high (Huang et al. 2010; Bayzid and Warnow 2013; Patel et al. 2013; DeGiorgio and Degnan 2014; Mirarab et al. 2014; Lanier and Knowles 2015; Mirarab and Warnow 2015; Xi et al. 2015). As a result, we interpreted results from ASTRAL, especially those that conflicted with ML or BI, with some skepticism. With that in mind, we summarize the main areas of conflict below. Within Allotriocarida, ASTRAL topologies showed a lack of congruence with other methods in that ASTRAL consistently recovered Copepoda as the sister



to all other Allotriocarida, whereas ML and BI methods consistently had Cephalocarida in this early diverging position (fig. 3E vs. B–D; supplementary fig. S4, Supplementary Material online). Strikingly, ASTRAL also recovered Hexapoda paraphyletic with Protura + Diplura + Insecta more closely related to Remipedia than to Collembola (supplementary fig. S4, Supplementary Material online) albeit with low support. We suspect this is an artifact, but hexapod paraphyly has been suggested before (Nardi et al. 2003). In the full ASTRAL analysis, the node leading to Remipedia and Protura + Diplura + Insecta has 0.96 PP, but support for this node decreased when nodes with low support in gene trees were collapsed to polytomies; when gene tree nodes with <10% and <30% BS support were collapsed prior to ASTRAL, the support values for the Remipedia + Protura + Diplura + Insecta node decreased to 0.91 PP and 0.45 PP, respectively, demonstrating that this node in ASTRAL was supported by gene trees with low support at this node (supplementary fig. S4A–C, Supplementary Material online). Evolutionary relationships of the mysids also differed. Contrary to ML and BI analyses of the full AA matrix, ASTRAL along with analyses of the Dayhoff6 recoded matrix found Mysida was paraphyletic with Stygiomysida as the sister to Mancoidea and the other mysids sister to amphipods but with low support (<0.40 PP in ASTRAL, 50% BS in RAxML with Dayhoff6, and 0.93 PP in CAT-GTR with Dayhoff6) (fig. 3B and D–E). Results regarding mysid and stygiomysid relationships from ASTRAL and under Dayhoff6 recoding may be particularly prone to error due to the low number of orthologs for *Praunus*, *Amblyops*, and *Stygiomysis* (supplementary table S1, Supplementary Material online).

### Divergence Time Estimation Results

We completed divergence time estimates across Pancrustacea using three chains each in MCMCTree and in PhyloBayes with autocorrelated (CIR and lognormal) and uncorrelated (UGAM) clock models. Fossil dates and justifications are given in supplementary table S7, Supplementary Material online. Convergence between chains was assessed by plotting posterior means for each chain against one another for MCMCTree (supplementary fig. S12, Supplementary Material online) and with trace plots for PhyloBayes (supplementary fig. S13, Supplementary Material online). Divergence time estimates from MCMCTree are summarized in figure 4. Unlike the only previous study conducted with similar fossil calibrations (Schwentner et al. 2017), our use of MCMCTree allowed our age estimates to incorporate the full matrix. With these more extensive sequence data in MCMCTree, we retrieved deep root ages for arthropods, pancrustaceans, and the three major clades of pancrustaceans, extending slightly past and into the middle of the Ediacaran period for each. With our PhyloBayes analyses using only 50 loci (supplementary fig. S15, Supplementary Material online), these deeper nodes diverged within the Cambrian. In the MCMCTree analysis, within Allotriocarida, hexapods were estimated to have terrestrialized

in the late Cambrian. In all PhyloBayes models, terrestrial hexapods were estimated as Ordovician. Their sister group, remipedes, has a very wide 95% highest posterior density (HPD) with MCMCTree, reflecting a crown group that may have diverged in the Jurassic (mean age), with a range from Pennsylvanian to Cretaceous. The other large allotriocarid group without an internal calibration, copepods, likely diverged in the Devonian (with a range from Ordovician to end Permian depending on the clock model). Within Multicrustacea, major crown groups showed mean estimates for their divergences in the Ordovician (peracarids and seneu-carids), Devonian (decapods), and Cretaceous (stomatopods). Finally, all sampled higher-level clades within Pancrustacea had diverged prior to the Cenozoic, a result that is consistent using PhyloBayes.

## Discussion

### Impacts of Taxon Sampling

We analyzed two similar taxon sets using parallel methods for ortholog selection and phylogenetic analysis. After recovering some unusual relationships in Data set 1 (e.g., Copepoda sister to Hexapoda and Stomatopoda sister to syncarids and eucarids), we added additional hexapods, copepods, stomatopods, and a second leptostracan. We also removed some closely related amphipods, decapods, and isopods at the tips of the tree (red taxa in supplementary fig. S5, Supplementary Material online) to better balance the taxon sampling such that malacostracans made up 50% of the data set rather than 70%. Because we recovered substantially different topologies between Data sets 1 and 2, which differed only in taxon sampling with 75% of taxa shared, we further investigated the effects of taxon sampling on the pancrustacean phylogeny.

We reasoned taxon sampling could be affecting two components of phylogenomic analysis: ortholog selection and tree topology accuracy. To disentangle these effects, we controlled for ortholog selection by using the same genes: only the 267 orthologs that were shared between the two data sets (roughly 50% of the orthologs). Using identical orthologs and only slightly different taxa, we repeated the same phylogenetic analyses and once again recovered incongruent topologies—topologies that were nearly identical to the originals recovered from analysis of the full ortholog alignments of each data set (supplementary fig. S10A and B, Supplementary Material online). So although taxon sampling did impact ortholog identification (in that only half of orthologs were shared), by controlling for ortholog selection, our results demonstrate that the effect of taxon sampling differences on the species tree reconstruction alone (rather than differences in the loci) was enough to drive the topological differences in the resulting species tree, an impact we found surprisingly large given the small change in taxon coverage. We further explored sensitivity to taxon sampling by removing the additional 23 taxa (i.e., 13 hexapods, 3 stomatopods, 2 remipedes, 2 copepods, 2 barnacles, and 1

leptostracan) that were sampled in Data set 2 relative to Data set 1. After removing these taxa from the Data set 2 matrix, the topology changed in two major ways: Cephalocarida + Remipedia was recovered with high support (97% BS) as the sister to the rest of Allotriocarida (100% BS), and Branchiopoda was the sister to Hexapoda (97% BS) (supplementary fig. S11, Supplementary Material online). Although Cephalocarida + Remipedia was supported in all analyses of Data set 1 (fig. 3F and G), usually as the sister to the rest of Allotriocarida, these relationships were never recovered in Data set 2. Yet using the Data set 2 orthologs and 20% fewer taxa (removing 23 of 105 taxa), these relationships were recovered with high support, further highlighting the importance of taxon sampling in the pancrustacean phylogeny, especially within Allotriocarida.

Although the results of the controlled ortholog experiments were surprising, we do find evidence that supports decades of literature suggesting taxon sampling and density increases phylogenetic accuracy. The biggest differences between Data set 1 and Data set 2 are the relationships estimated within Allotriocarida. In Data set 1, we estimated a sister relationship between copepods and hexapods (supplementary fig. S5, Supplementary Material online), whereas in Data set 2, we usually recovered the typical sister relationship between remipedes and hexapods (fig. 2) (but see discussion of Copepoda under the CAT-GTR model below). The taxon sampling differences between these data sets for these lineages are an increase in the number of noninsect hexapods (i.e., Entognatha: Diplura, Collembola, and Protura) from one representative (Data set 1) to seven (Data set 2), an increase in remipede taxa from two representatives (Data set 1) to four (Data set 2), and an increase in copepods from four (Data set 1) to six (Data set 2). We believe that the reduced taxon sampling caused the spurious relationships in Data set 1, due in part to the relative branch lengths of these groups. Generally speaking, the branches near the base of Allotriocarida are short, but nearly all of the branches are relatively long toward the tips, a particularly challenging pattern for phylogenetic reconstruction. Despite using a substitution model that accounts for nonstationarity and has been shown to be more robust against LBA (Lartillot et al. 2007), we estimated copepods were the sister to hexapods in Data set 1 with a single taxon representative of noninsect hexapods that had a branch length of 0.84 AA substitutions per site (supplementary table S5, Supplementary Material online). In Data set 2, although there are still relatively long branches in the noninsect hexapods, the additional taxa decreased the average terminal branch length in this lineage from 0.84 to 0.27 AA substitutions per site. When we increased the taxon density of early diverging hexapods, reducing the effects of LBA, we recovered a closer relationship with remipedes. Yet, when we removed the additional hexapods and remipedes from Data set 2, we once again recovered “Xenocarida” at the base of Allotriocarida and Branchiopoda sister to Hexapoda (supplementary fig. S11, Supplementary

Material online). These results confirm that taxon sampling continues to be a dominant factor in phylogenomics, even in the context of hundreds of carefully selected orthologs. Other phylogenomic studies are also finding that subsampling deep lineages may cause topological inaccuracies (e.g., Sharma et al. 2014; Branstetter et al. 2017; Betancur-R. et al. 2019) and the suspect relationships recovered with the more limited sampling in Data set 1 further support this.

### Systematics and Targets for Future Sampling

In general, most pancrustacean phylogenomic studies have shared a high degree of overlap in taxon selection and methodology. All have focused on single-copy protein-coding genes under one of three phases of data generation: Sanger sequencing (Regier et al. 2005, 2008; Regier et al. 2010; Rota-Stabelli, Lartillot, et al. 2013), expressed sequence tags (ESTs) from 454 sequencing (von Reumont et al. 2012; Oakley et al. 2013), and Illumina-based RNA-Seq (Schwentner et al. 2017, 2018; Lozano-Fernandez et al. 2019). Besides the Sanger sequence studies, most have had roughly similar alignment sizes, once the number of orthologs, AA positions, and gaps is accounted for (table 2). Despite these similarities, the topologies recovered have been surprisingly variable.

Here, we sampled more “crustacean” orders (i.e., 30 vs. <26) and sampled them more densely relative to previous studies (table 2). This was especially true within Malacostraca where we sampled more syncarids, peracarids (especially isopods and amphipods), decapods, and stomatopods than prior studies (e.g., 54 malacostracans vs. 4–26). We assembled our final matrix (Data set 2) with an emphasis on maintaining balanced taxon sampling to the greatest extent possible. Taxon sampling was balanced by 1) selecting taxa representing deep splits in clades with known phylogenetic relationships and 2), when phylogenetic relationships were unclear, sampling as many taxa as possible followed by iteratively building species trees and subsampling clades to prune closely related species at the tips of densely sampled clades, retaining primarily the deep splits within clades. This was done because uneven taxon sampling is known to affect homolog clustering and therefore ortholog selection (Chen et al. 2007; Altenhoff and Dessimoz 2009) and because less balanced taxon sampling had a large impact on topology in our analyses of Data set 1, even when controlling for ortholog selection (supplementary fig. S10A, Supplementary Material online).

Beyond taxon sampling, our study differed from most others in several ways. First, we used a tree-based approach to ortholog identification, which has been shown to improve phylogenetic reconstructions in simulation studies (Smith and Pease 2017) and published data sets (Dunn et al. 2013; Yang and Smith 2014; Ballesteros and Hormiga 2016). A few other pancrustacean studies have used a tree-based approach to ortholog identification, but most have used the sequence similarity approach in

OMA (Altenhoff et al. 2011) (table 2). Second, we place less emphasis on results recovered only with Dayhoff6 recoding, given that it may remove phylogenetic signal more than it ameliorates saturation and compositional heterogeneity (Hernandez and Ryan 2021 but see Foster et al. 2022, Giacomelli et al. 2022) (still, we present a CAT-GTR and an ML analysis of a Dayhoff6 matrix [fig. 3A, D, and H; supplementary figs. S1B, S3, S6, and S8, Supplementary Material online]); instead, we favored accounting for saturation and compositional heterogeneity with site-heterogeneous models in addition to partitioned and coalescent-based analyses. Third, we did not rely primarily on results under the CAT-GTR model. In pancrustacean phylogenomic analyses, CAT-GTR chains frequently do not fully converge, as was the case here despite nearly a year of run time, and without convergence, the results are statistically invalid (Gelman and Rubin 1992; Huelsenbeck et al. 2002; Whelan and Halanaych 2017). Furthermore, Li et al. (2021) showed that CAT-GTR analyses of the metazoan tree of life often have hundreds of additional rate categories yet fail to fit better than site-heterogeneous models with many fewer categories. Given the issues with convergence under CAT-GTR, we emphasized the phylogeny resulting from the LG + C60 + F + G mixture model, which has not been used in previous pancrustacean phylogenomic analyses. Site-heterogeneous models like the C60 class still account for among site variation in AA propensities, are less prone to artifacts like LBA (Lartillot and Philippe 2004; Lartillot et al. 2007; Le et al. 2008), and were found to converge in this data set in a reasonable time frame (five independent tree searches produced identical topologies). It was also the model of best fit, the highest likelihood tree, and was not rejected by the AU-test (unlike CAT-GTR) (supplementary table S6, Supplementary Material online). Finally, the tree resulting from the C60 analysis was robust; that is, it was nearly identical to those produced from the partitioned RAxML analyses with and without Dayhoff6 recoding; all of these topologies were completely congruent with respect to clades at the ordinal level and above (fig. 3A–E; supplementary fig. S1, Supplementary Material online).

Across the different pancrustacean phylogenomic studies, some clades have been consistently recovered. The position and composition of Oligostraca has been relatively constant compared with the other major clades (Regier et al. 2010; von Reumont et al. 2012; Oakley et al. 2013; Rota-Stabelli, Lartillot, et al. 2013; Schwentner et al. 2017, 2018; Lozano-Fernandez et al. 2019). Although a few analyses have found Oligostraca to be more closely related to Malacostraca + Thecostraca (see von Reumont et al. [2012], fig. 2, and Rota-Stabelli, Lartillot et al. [2013]), the vast majority have recovered Oligostraca as the sister to all other pancrustaceans (Regier et al. 2010; von Reumont et al. 2012; Oakley et al. 2013; Schwentner et al. 2017, 2018; Lozano-Fernandez et al. 2019). A major question that remains is the monophyly of ostracods. Oakley et al. (2013) recovered a monophyletic Ostracoda

with relatively dense sampling of ostracods, but more recent studies that have relied primarily on ostracod transcriptomes rather than ESTs, and have included fewer ostracods as a result, have often found a nonmonophyletic Ostracoda as we have here. Clearly, expanded taxon sampling of the Ostracoda is needed, which will also enable the incorporation of the rich fossil record of ostracods for divergence time analyses. Ostracod nonmonophyly is a particularly interesting in the light of carapace evolution within Oligostraca. The topology recovered here in, and in most recent phylogenomic studies, suggests that either Branchiura or Mystacocarida independently lost an ostracod-like bivalved carapace or that it evolved independently in the Podocopa and Mydocopa. The question of ostracod monophyly will be best explored not just in the context of expanded sampling of ostracods (Ellis et al. 2023) but also by expansion of the poorly sampled Mystacocarida and Branchiura. In most recent phylogenomic studies, Mystacocarida is represented only by *Derocheilocaris remanei* and Branchiura only by *Argulus siamensis* (von Reumont et al. 2012; Rota-Stabelli, Lartillot, et al. 2013; Schwentner et al. 2017, 2018; Lozano-Fernandez et al. 2019). Therefore, the phylogeny of Oligostraca would benefit most from expanded taxon sampling of the early diverging ostracods *Manawa staceyi* and *Platycopa*, the other genus of Mystacocarida (*Ctenocheilocaris*), and other ichthyostracans, especially Pentastomida and the early diverging branchiuran *Dolops* (Møller et al. 2008).

Most pancrustacean phylogenomic studies have recovered a clade comprising some combination of Branchiopoda, Cephalocarida, Hexapoda, and Remipedia, in a group termed Allotriocarida. In terms of the interrelationships, the sister group to hexapods has received the most attention. Regier et al. (2010) found Remipedia + Cephalocarida (which they named Xenocarida) as the sister to Hexapoda, but subsequent studies have usually found Remipedia alone sister to Hexapoda, with Cephalocarida diverging earliest from the rest of Allotriocarida (Oakley et al. 2013; Schwentner et al. 2017, 2018; Lozano-Fernandez et al. 2019). Multiple studies have noted the Cephalocarida + Remipedia pairing may be an artifact of LBA (Rota-Stabelli, Daley, et al. 2013; Schwentner et al. 2017; Lozano-Fernandez et al. 2019). We recovered Cephalocarida as the sister to all other allotriocaridans with maximum support in all concatenation-based analyses of Data set 2. However, the pancrustacean phylogeny seems to be particularly prone to a “Xenocarida” artifact. In all concatenated analyses of Data set 1, including CAT-GTR of the AA and Dayhoff6 matrices, we recovered “Xenocarida” with high support, usually as an early diverging branch in Allotriocarida (fig. 3F and G; supplementary figs. S5–S8, Supplementary Material online). In a striking example, when using the Data set 2 orthologs, which never supported “Xenocarida,” once we removed the additional remipedes and early diverging hexapods from the alignment, we again recovered Cephalocarida + Remipedia, further indicating a



LBA artifact (supplementary fig. S11, Supplementary Material online). Sampling members of the three remaining cephalocarid genera may help stabilize their relationship (WoRMS 2023).

Notably, we recovered Allotriocarida with the marked addition of Copepoda in all analyses with high support (fig. 3B–E). Although we find robust support for an Allotriocarida expanded to include Copepoda, the relationship between constituent members of Allotriocarida is less clear. All ML analyses (partitioned, site-heterogenous, Dayhoff6, and subsampled matrices) consistently found copepods sister to branchiopods with high support (95–100% BS) (supplementary fig. S1A–C, Supplementary Material online). Analyses of the AA matrix and the Dayhoff6 matrix with the CAT-GTR model recovered Copepoda + Remipedia as the sister to Hexapoda (fig. 3C and D). Meanwhile all ASTRAL analyses estimated Copepoda as the sister to all other allotriocaridans (supplementary fig. S4, Supplementary Material online) but with low support. These conflicting results highlight a lack of resolution within Allotriocarida. Two prior studies occasionally recovered Copepoda in Allotriocarida in a subset of their analyses, but the position of copepods was variable: Lozano-Fernandez et al. (2019) found Copepoda sister to Remipedia (their figure 1B) or Remipedia + Hexapoda (their figure 1C), whereas Rota-Stabelli, Daley, et al. (2013) found Copepoda sister to Branchiopoda (their figure 1C and D). Still, the exact position of copepods continues to be one of the least resolved parts of the pancrustacean tree of life (Rota-Stabelli, Daley, et al. 2013; Lozano-Fernandez et al. 2019). We think a more precise position of Copepoda rests in sampling Platycopepoda, the copepod order sister to all others (Huys and Boxshall 1991), with genome-scale data. Species of Platycopepoda have scarcely been sequenced at all, but this taxon should shorten the LB leading to Copepoda, which can ameliorate phylogenetic error (Hendy and Penny 1989).

Copepoda being positioned within Allotriocarida is not straightforward to explain from a morphological perspective. Some morphological support linking Copepoda to other allotriocaridan taxa has been mentioned previously, but this was before Allotriocarida was recognized as a clade, so a re-evaluation is needed. A relationship between copepods and remipedes has been suggested given that both possess a series of six cephalic anterior limbs, including the maxillipeds, fused into the cephalosome (Boxshall 1983). Posterior to the cephalosome, both copepods and remipedes also possess biramous, flattened, paddle-like swimming legs (Yager 1981). Superficially, the remipede body plan resembles a copepod that serially added leg-bearing segments, or alternatively, the copepod body plan resembles a remipede with truncated development of leg-bearing segments. Analyses using the CAT-GTR model here lend support to these putative homologies (fig. 3C and D). Itô (1989) hypothesized an evolutionary relationship between Copepoda, Remipedia, and Cephalocarida. He proposed that the three-segmented

endopod of the copepod trunk limb was derived from an ancestral remipede-like four-segmented endopod, which is supported by the fact that some remipedes even possess three-segmented endopods on their posterior trunk limbs. Itô (1989) further noted a potential relationship between copepods, remipedes, and cephalocarids based on the hypothesis that the endopods of copepods and remipedes are derived from an ancestral five-segmented endopod like that seen in cephalocarid trunk limbs.

However, trunk limb evolution of Copepoda and Remipedia needs evaluation in the broader context of Allotriocarida. Differing from the mentioned copepod and remipede biramous limbs, two other allotriocaridan taxa, cephalocarids and branchiopods, have phyllopodous (flattened) trunk limbs with a double function as they are involved in locomotion and feeding simultaneously. In both cephalocarids and branchiopods, multiple endites along the median edge trunk limbs play a role in collecting food and transporting it forwardly to the mouth region (Sanders 1963; Fryer 1983; Olesen 2007). Interestingly, the cephalocarid/branchiopod type of trunk limbs bears much resemblance to that seen in many Cambrian microfossils such as *Rehbachella kinnekullensis* and *Dala peilertae*, a notion that has been used to argue for a feeding apparatus involving trunk limbs being ancestral to “Crustacea” (Walossek 1993; Olesen et al. 2011), although the fossils may represent larval stages (Boxshall 2007; Wolfe and Hegna 2014) that typically use trunk limbs for feeding. Given current phylogenomic results, these similarities may be interpreted as ancestral, perhaps even a novelty, to the Allotriocarida lineage. Consequently, because of the phylogenetic position of Copepoda, Remipedia, and Hexapoda deeply nested within Allotriocarida, “trunk limb-based feeding” could have been lost in all these taxa, perhaps independently. For the Hexapoda, a transition into a purely cephalic-based feeding system may have been an exaptation for conquering terrestrial habitats. Hexapoda and Remipedia are likely sister groups (Data set 2; figs. 2A and 3B–D), so a cephalic raptorial feeding apparatus made up of uniramous maxilla 1 and maxilla 2 may have been present in their common ancestor. In the recent Remipedia, a morphologically similar appendage (maxilliped) was added to the cephalic feeding apparatus, whereas the uniramous maxilla 1 and 2 was modified into the palp-bearing “maxillae” and “labium” of insect, the evolutionary details of which needs exploration.

Nonetheless, the presence of Copepoda within Allotriocarida does have implications for the evolution of the other constituent clades. First, given that copepods were almost certainly ancestrally marine and hyperbenthic (Huys and Boxshall 1991), their close relationship with branchiopods, hexapods, and remipedes provides additional support for the hypothesis that these clades were also ancestrally marine (von Reumont et al. 2012; Lozano-Fernandez et al. 2016; Schwentner et al. 2017). A marine origin is also supported by the Cambrian or Ordovician divergence estimated for hexapods here (fig.

4; supplementary fig. S15, Supplementary Material online). Schwentner et al. (2017) noted that the loss of the mandibular palp in adults (though present in juveniles of Branchiopoda, Cephalocarida, and Remipedia) might be an apomorphy for Allotriocarida. However, the position of Copepoda recovered here suggests the evolutionary history regarding the loss of the mandibular palp in adults is homoplasious. Either the mandibular palp was lost separately in the adults of Branchiopoda, Cephalocarida, and Remipedia or Copepoda is unique in Allotriocarida for retaining the mandibular palp in adulthood (Olesen et al. 2014); the latter scenario is more parsimonious and is supported by other neotenic features proposed for Copepoda (Gurney 1942).

Malacostraca and Thecostraca have a consistent phylogenetic affinity in all recent phylogenomic studies, but relationships among Multicrustacea (Copepoda, Malacostraca, and Thecostraca) have been one of the least stable areas of the pancrustacean tree of life, mostly due to variability in position of Copepoda (Regier et al. 2010; von Reumont et al. 2012; Oakley et al. 2013; Rota-Stabelli, Lartillot, et al. 2013; Schwentner et al. 2017, 2018; Lozano-Fernandez et al. 2019). Previous studies have typically found Copepoda as the sister to Malacostraca + Thecostraca or sister to one of those taxa individually. However, as noted above, some analyses from previous studies and all those in this study recover Copepoda within Allotriocarida, rejecting the traditional Multicrustacea (Copepoda + Malacostraca + Thecostraca). We recovered Communostraca (Malacostraca + Thecostraca) in all analyses. Although Lozano-Fernandez et al. (2019) did not find consistent support for Communostraca across all analyses, our results support their suggestion that the location of male and female gonopores on different body somites is a synapomorphy for Communostraca. To further resolve the communostracan phylogeny and more comprehensively test the validity of “Multicrustacea,” it is important to sample Platycoptoidea and the early diverging thecostracan lineages Acrothoracica, Ascothoracida, Facetotecta, and Tantulocarida (Petrunina et al. 2014; Chan et al. 2021).

With over 43,500 extant described species (WoRMS 2023), Malacostraca is the most speciose “crustacean” class, but it has received relatively little attention in pancrustacean phylogenomic studies. Most did not comment on malacostracan interrelationships because they sampled only 4–26 species (Regier et al. 2010; von Reumont et al. 2012; Oakley et al. 2013; Schwentner et al. 2017). One constant is Leptostraca as the sister to all other malacostracans. Schwentner et al. (2018) included the most malacostracans prior to this study and examined interrelationships among 26 species, most of which were decapods (60%). Here, we examined malacostracan relationships with expanded taxon sampling, doubling the number of species sampled (52), especially peracarids and stomatopods. An earlier version of this analysis (Data set 1) included 16 additional amphipods, isopods, and decapods, but these branches comprised shallow splits, and these taxa were removed to maintain a more balanced taxon set (see taxa labeled red in supplementary fig. S5,

Supplementary Material online). With the inclusion of Bathynellacea for the first time, we were able to test the monophyly of Syncarida (Anaspidacea + Bathynellacea) and it was polyphyletic in all analyses, with Anaspidacea as sister to Euphausiacea resulting in a paraphyletic Eucarida (i.e., Euphausiacea + Decapoda). Our results support the hypothesis of Serban (1972, 1973) that the Syncarida is polyphyletic with Bathynellacea in an earlier diverging position. Notwithstanding the nonmonophyly of Syncarida and Eucarida respectively, a clade with all their subtaxa (Anaspidacea, Bathynellacea, Eucarida, and Decapoda) (figs. 2 and 3B–D) was recovered with maximum support in all ML and BI analyses, and we propose the name Syneucarida for this clade (figs. 2A and 3B–D).

The position of Stomatopoda has been variable between studies and within our analyses here. Although the CAT-GTR analysis in Schwentner et al. (2018) found Stomatopoda in a more classical, early diverging position in Malacostraca, all of our ML and ASTRAL analyses recovered Stomatopoda as sister to Syneucarida. We propose the name Stomatocarida for Stomatopoda + Syneucarida, dividing Malacostraca into three clades: Leptostraca, Stomatocarida, and Peracarida. Notably, CAT-GTR analyses here did recover Stomatopoda in the more basal position (fig. 3C and D) found by Schwentner et al. (2018). In the CAT-GTR analysis of the Dayhoff6 matrix in Lozano-Fernandez et al. (2019), Stomatopoda was recovered as the sister to Mysida; we suspect this to be an artifact, perhaps due to LBA, given that we never recovered this grouping in any analysis under our expanded sampling of mysids and stomatopods. Several improvements can be made to better resolve relationships among stomatocaridan taxa. Sampling more stomatopods may address the variability in their phylogenetic relationships; we attempted to include *Hemisquilla californiensis* (SRR2103462–3) and *Pseudosquilla ciliata* (SRR2103518, SRR2103524) in our study but removed them from phylogenetic analyses because we were never able to recover more than 13% of orthologs from these samples. Species of *Hemisquilla* would be particularly valuable to include given a recent study suggested they are sister to the other stomatopods (Koga and Rouse 2021). Since both syncarid orders are represented by a single species each, sampling additional syncarids would be beneficial.

Our expanded taxon sampling of peracarids enabled us to examine relationships in this clade in greater detail than previous phylogenomic studies. In all ML and BI analyses of the AA matrix, we found a monophyletic Peracarida with amphipods sister to all other peracarids and Mysida sister to Mancoida (Isopoda + Cumacea + Tanaidacea) (figs. 2 and 3B, 3C). Those results contrast Schwentner et al. (2018), where results varied and Mysida was frequently sister to the other peracarids. With expanded taxon sampling here comprising two additional mysids, a stygiomysid, a second cumacean, eight and nine select isopods, and amphipods, respectively, we consistently recovered mysids sister to Mancoida in all concatenated analyses of the AA matrix. However, in CAT-GTR analysis of the Dayhoff6 matrix here, Mysida and Stygiomysida were each sister to

Amphipoda and Mancoida, respectively (fig. 3D), although this result should be treated with caution given the low percentage of orthologs retrieved from the mysid and stygiomysid taxa and the information loss associated with Dayhoff6 recoding. Within Mancoida, all our analyses grouped Cumacea sister to Isopoda, contrary to the Tanaidacea + Cumacea relationship recovered in most analyses in Schwentner et al. (2018). These results also disagree with the morphological phylogenetic hypothesis of Richter and Scholtz (2001) that linked Isopoda + Tanaidacea. Interestingly, in ML analyses of a Data set 1, which included eight additional amphipods and three isopods (all shallow branches that were pruned to create a more balanced taxon set here), we did find Mysida sister to all other peracarids, similar to Schwentner et al. (2018). This result, however, was not robust within that data set: ASTRAL and CAT-GTR analyses of the same matrix consistently found mysids sister to Mancoida, just as in all concatenation methods of our Data set 2 matrix here. Taken together, these differences among our analyses, as well as those of Schwentner et al. (2018) and Höpel et al. (2022), suggest a surprising amount of instability at the base of Peracarida.

Peracarida is one of the pancrustacean taxa most in need of sampling effort. There are 12 extant orders of peracarids, and half of them have yet to be sampled in phylogenomic analyses (WoRMS 2023). Resolving the backbone of the peracarid phylogeny requires sampling the six remaining orders: Bochsacea, Ingolfiellida, Lophogastrida, Mictacea, Spelaeogriffacea, and Thermosbaenacea. These orders are crucial not just for the peracarid tree of life, but also for the larger Malacostraca given that some have questioned whether Lophogastrida and Thermosbaenacea belong in Peracarida at all (Siewing 1956; Schram and Hof 1998). A recent mitochondrial genome study by Höpel et al. (2022) included Lophogastrida and recovered a monophyletic Peracarida with Lophogastrida sister to Mysida and Stygiomysida, a hypothesis that would be interesting to test with nuclear loci. Sampling these taxa would enable a robust test of the monophyly of Peracarida and might provide more resolution for the position of Stomatopoda given the short branches found separating Peracarida, Stomatopoda, and Syneucarida here.

### Divergence Time Estimation

Our MCMCtree divergence time estimates retrieved deep splits of arthropods and the three main pancrustacean clades (Oligostraca, Allotriocarida, and Communostraca) earlier than the Cambrian, preceding the oldest known crown group arthropod fossils (first appearing around 521 Ma; Daley et al. 2018) and pancrustacean fossils (stem and crown groups simultaneously appearing about 514 Ma; Zhai et al. 2019; Hegna et al. 2020). It has been proposed that molecular clock models may overestimate the time of divergence of the crown group MRCA and that their stem groups may go extinct quickly after the MRCA, together suggesting a “long fuse” divergence estimate is unlikely (Budd and Mann 2020a). It is possible

that our MCMCtree results, using the uncorrelated independent rates clock model, represent such an example of overestimation of the crown group age, as older root ages have been observed before with this software and clock model (Barba-Montoya et al. 2017). To further investigate, we compared MCMCtree to divergence times estimated under three different clock models in PhyloBayes (supplementary fig. S15, Supplementary Material online). We found that the arthropod root was within the Cambrian using autocorrelated clock models (CIR and log-normal) and in the uncorrelated (UGAM) analysis crown group Pancrustacea diverged in the Cambrian. These chronograms all estimated shorter branch lengths for the presumed early, rapid divergences, whereas our results with MCMCtree appear to “smooth” the rate of early evolution at deep nodes. Another hint comes from the relatively narrow posterior age distributions at these deep nodes in all analyses (compared with the marginal priors; supplementary figs. S14 and S15, Supplementary Material online), with wide distributions at many shallow nodes, which suggest a decrease in rates over time that may challenge clock models (dos Reis, Thawornwattana, et al. 2015). It is not so simple as to assume that uncorrelated clock models overestimate divergence times in our data set, as the PhyloBayes UGAM analysis resulted in the youngest ages for many internal nodes. It is therefore unclear what drives the differences among clock models for internal node age estimates, although perhaps autocorrelated models are better able to cope with putative rapid divergences in the Cambrian (Lee et al. 2013; Daley et al. 2018; Budd and Mann 2020b) and a subsequent slowdown, similar to that proposed for one possible topology of chelicerates (Lozano-Fernandez et al. 2020).

Broadly, most major pancrustacean clades (Oligostraca, Allotriocarida, Communostraca, and most classes) were established in the early Paleozoic. The Late Cambrian origin for terrestrial Hexapoda estimated by MCMCtree is consistent with some recent studies (e.g., Lozano-Fernandez et al. 2016; Schwentner et al. 2017), whereas the much younger age (at least 100 Ma younger, up to 330 Ma) of the sister group, crown Remipedia, presents further challenges to the quest for identifying a stem group of either clade in the fossil record, as a genuine ghost lineage indicates a long time to pioneer different habitats and many unknown morphological changes.

Most shallower posterior age estimates were roughly consistent with their fossil records (e.g., Wolfe et al. 2016; Hegna et al. 2020), highlighting the importance of appropriately vetted age priors in divergence time studies. Our new topological result supporting Syneucarida may prompt re-evaluation of Paleozoic fossils previously assigned to the extinct “syncarid” group Palaeocaridacea (Schram 1984; Hegna et al. 2020). Although we were not able to include many peracarid fossil calibrations due to their lack of phylogenetic framework (Hegna et al. 2020), ages within the group were consistent in several cases, including isopods with a mean age range from the Carboniferous to Permian, depending on the clock model.



Other clades which do not have fossil calibrations available (e.g., remipedes) or that were not appropriate to use with our particular molecular taxon sampling (e.g., copepods, most of amphipods) exemplified wider posterior ages under all clock models, often varying between models. In one standout case, penaeid shrimp, there are known crown group fossils from the Triassic (Wolfe et al. 2019), but they could not be used as priors, and we retrieved impossibly young posterior ages with all clock models.

## Conclusions

Given the apparent sensitivity to taxon sampling in the pancrustacean phylogeny and the fact that the clade contains >1,000,000 described species, taxon sampling should be expanded strategically for improved resolution. The primary objective should be to sample the 27 “crustacean” orders that have not yet been sampled with transcriptomic data. Of utmost priority, we identify just 15 crucial taxa that should break many of the longest branches in the phylogeny: *Manawa* and *Platycopa* (Ostracoda); Pentastomida and *Dolops* (Ichthyostraca); Platycopioidea (Copepoda); Acrothoracica, Ascothoracida, Facetotecta, and Tantulocarida (Thecostraca); and Bochsacea, Ingolfiellida, Lophogastrida, Mictacea, Spelaeogriffacea, and Thermosbaenacea (Peracarida). These taxa should be prioritized in genome sequencing efforts (Lewin et al. 2022). Until we have full genome sequences for the major branches of the tree of life (Lewin et al. 2018), we believe adding carefully curated sequences for taxa that have yet to be sampled is the most promising avenue for resolving the pancrustacean phylogeny.

Additional data types could help resolve difficult nodes. Although identifying more orthologs is worthwhile, it should be done carefully and not just for the sake of more genes given that the inclusion of a small number of paralogs can introduce strong erroneous signal that can mislead phylogenetic reconstructions (Shen et al. 2017; Smith and Hahn 2021). It is noteworthy that the number of orthologs identified has been relatively consistent despite a variety of ortholog identification methods and taxon sets (table 2), which may indicate that there are not many additional conserved protein-coding orthologs across pancrustaceans to be added. Particular areas of the pancrustacean phylogeny may be better resolved by more clade-specific analyses, which often yield more orthologs and more complete matrices (Schwentner et al. 2018; Laumer et al. 2019; Wolfe et al. 2019). New methods that do not rely solely on orthologs but also incorporate phylogenetic signal in paralogs, thus leverage substantially more data, may help resolve challenging nodes in the tree of life (Hellmuth et al. 2015; Smith and Hahn 2021; Smith et al. 2022). Phylogenetic signal can also be mined from synteny, which is more conserved and has clearer homology than coding sequence; these analyses are showing great promise at other challenging nodes in metazoan phylogeny (Moret and Warnow 2005; Hu et al. 2014; Simakov et al. 2022, Schultz et al. 2023).

Unfortunately, given the small number of chromosome-scale genomes for noninsect pancrustaceans (only 75 species have genome assemblies available in the National Center for Biotechnology Information [NCBI]), synteny analyses are dependent on expanded “crustacean” genome sequencing efforts (Bernot et al. 2022). Finally, in light of the recent discoveries of a number of incredibly preserved pancrustacean fossils (Zhang et al. 2007; Wolfe et al. 2016; Luque and Gerken 2019; Zhai et al. 2019; Robin et al. 2021), it is exciting to consider that new fossil discoveries may inform our understanding of pancrustacean evolution.

## Materials and Methods

### RNA Extraction and Sequencing

We collected fresh specimens of *Amblyops abbreviata*, *Diastylis cornuta*, and *Echinogammarus pirloti* and stored them in RNAlater (Invitrogen, Waltham, MA) or in no preservative at  $-80^{\circ}\text{C}$ . Total RNA was extracted using TRIzol (Invitrogen, Waltham, MA) according to the manufacturer’s instructions. Following extraction, total RNA was cleaned using the Nucleospin RNA Clean-up (Macherey-Nagel, Düren, Germany) silica-based column to further purify the RNA. Ribosomal RNA was removed using Ribo-zero (Illumina, San Diego, CA), and the quality of the RNA was checked on a Bioanalyzer. Both *D. cornuta* and *E. pirloti* were sequenced on an Illumina HiSeq 2500 with 125 bp paired-end reads at Duke University, whereas *A. abbreviata* was sequenced on an Ion Torrent at the University of Bergen. Sample information and sequence data are available in NCBI BioProject PRJNA997050.

### Data Set and Transcriptome Assembly

In total, 149 transcriptomes and 16 genome assemblies spanning the arthropod tree of life were examined in this study (supplementary tables S1 and S2, Supplementary Material online). The genomes of two chelicerates, *Limulus polyphemus* and *Ixodes scapularis*, and two myriapods, *Cryptops hortensis* and *Strigamia maritima*, were selected as outgroup taxa based on previous phylogenetic studies (Regier et al. 2010; Schwentner et al. 2017, 2018; Lozano-Fernandez et al. 2019). Following analysis of Data set 1, we expanded the taxon sampling in Data set 2 to include an additional 13 hexapods, 3 stomatopods, 2 remipedes, 2 copepods, 2 barnacles, and a second leptostracan (taxa marked with asterisks in fig. 2A). Additionally, 14 transcriptomes were shown in be low quality in initial analysis of Data set 1 because they contained <20% of orthologs (marked with asterisks in supplementary table S2, Supplementary Material online); these taxa were excluded from the Data set 1 matrix prior to species tree inference and were excluded from all downstream analyses in Data set 2. To better balance the taxon sampling in Data set 2, we also removed 16 taxa from densely sampled clades that were shown to be closely related to other species in Data set 1 (see red taxon labels in supplementary fig. S5, Supplementary Material

online); the taxa removed for the final data set comprised 8 amphipods, 4 decapods, 3 isopods, and the second hymenopteran. The final taxon set, Data set 2 (supplementary table S1, Supplementary Material online), comprised 90 transcriptomes and 15 genomes. Three taxa that represented important branches in the phylogeny but had only 14–20% of orthologs were retained for downstream phylogenetic analyses: *A. abbreviata* (Mysida), *Praunus flexuosus* (Mysida), and *Neogonodactylus oerstedii* (Stomatopoda). All computational analyses were carried out on the high-performance computing cluster at George Washington University.

Raw reads for all transcriptomes were assembled de novo as follows. Raw read quality was assessed using FastQC v0.11.8 (Andrews 2018), reads were subjected to quality and adapter trimming using Trimmomatic v0.33 (ILLUMINACLIP: TruSeq3-PE-2.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:50) (Bolger et al. 2014), and quality trimming and adapter removal was confirmed using FastQC again after trimming. Trimmed reads were error-corrected using Rcorrector (Song and Florea 2015) with default settings. Error-corrected reads were assembled using Trinity (Grabherr et al. 2011; Haas et al. 2013) under default parameters. Assembled contigs were translated to AA sequences using TransDecoder v5.2.0 (Haas et al. 2013) with open reading frames identified using default parameters. Redundancy in AA sequences resulting from Transdecoder was reduced using CDHIT v4.6 (Li and Godzik 2006; Fu et al. 2012) with a 99% similarity threshold.

### Ortholog Identification

Orthologs were identified using an explicit phylogenetic approach following Yang and Smith (Yang and Smith 2014) (unless otherwise noted, named scripts are from [https://bitbucket.org/yangya/phylogenomic\\_dataset\\_construction/src/master](https://bitbucket.org/yangya/phylogenomic_dataset_construction/src/master)). The predicted proteins from the transcriptomes and genomes were subjected to an all-by-all BlastP v2.9.0 (Altschul et al. 1990, 1997; Camacho et al. 2009) search (-max\_target\_seqs 1,000 -evalue 10), and the resulting Blast output was filtered for the hit fraction being at least 0.4 (Chiu et al. 2006). Filtered Blast hits were further clustered using MCL v12.068 (Van Dongen 2000, 2008) with a -log E-value cutoff set to 5 and an I-value of 1.4 to identify homologous protein sequences. Fasta files were written from the MCL output using write\_fasta\_files\_from\_mcl.py.

Each cluster of homologs was then aligned individually with MAFFT v7.13 (-genafpair-maxiterate 1,000 if <1,000 sequences; -auto if >1,000 sequences) (Katoh and Standley 2013) and trimmed using phyutility (minimum column occupancy = 0.1) (Smith and Dunn 2008), and trees were built using either RAXML v8.2.9 (Stamatakis 2014) under the model "PROTGAMMALG" for clusters with <1,000 sequences or FastTree v2.1.8 (Price et al. 2010) under the model "-lg" for clusters >1,000 sequences since the LG matrix was the model of best fit for the majority of orthogroups. The resulting trees may contain

branches representing paralogs or misassembled contigs. These were identified and filtered using the following three methods. First, divergent sequences were removed from clusters if a terminal branch was longer than 1.0 or more than 15× longer than its sister using trim\_tips.py. Next, if monophyletic or paraphyletic tips from the same taxa were present in a tree, only the sequence with the highest number of nonambiguous characters in the trimmed alignment was kept and the rest removed following previously published methods (Smith et al. 2011; Dunn et al. 2013; Yang and Smith 2014). Lastly, potential deep paralogs were removed using cut\_long\_internal\_branches.py with an internal branch length cutoff of 1.8 and a minimum number of taxa of 15. Cutoff files were written from the trimmed trees and alignments, and the entire process of aligning, trimming alignments, building trees, and removing paralogs and long branches was repeated. After the second round of refinement, the trees were called homolog trees and were further pruned to infer orthologs.

Orthologs were called using the maximum inclusion method (Dunn et al. 2008, 2013; Yang and Smith 2014; Ballesteros and Hormiga 2016). After pruning the homolog trees to identify maximum inclusion orthologs, the remaining subtrees may contain terminal taxa subtended by long branches as a result of the subtree trimming method (Yang and Smith 2014). To account for this, the trees were trimmed once more using a range of permissive-to-strict branch length trimming parameters, referred to from here on as permissive, medium, and strict branch trimming, with relative branch lengths of 10×, 12×, or 15× and absolute branch lengths of 0.75, 0.85, or 1.0 at the permissive, medium, and strict levels, respectively. Previous analyses showed that the more strict branch length trimming parameters only resulted in the loss of ~40 orthologs, so the strict branch length trimming was used here. The resulting orthologs were aligned with MAFFT and trimmed with Gblocks v0.91b using less strict parameters (Talavera and Castresana 2007), and the final matrix was produced using concatenate\_matrices.py (Yang and Smith 2014) with a minimum number of sites set at 50 AA and a minimum taxon cutoff of 53/105 taxa (50.5%). The resulting matrix was composed of 576 orthologs.

### Phylogenetic Analyses

Phylogenetic analyses were completed using concatenation and coalescent methods. Concatenation analyses were done in both ML partitioned analysis and ML and BI site-heterogenous models. In the ML framework, the partitioned analysis was performed by using a clustering algorithm to group orthologs based on sequence properties and comparing alternative clusters and evolutionary models using the BIC (Lanfear et al. 2014, 2017) (v2.1.1) (-rclusterf), which identified 90 partitions. In 2 instances, a partition did not contain all 20 AA states, which can cause problems with phylogenetic parameter estimation, so for both of these cases, the 2 partitions were combined

with a partition using a similar model of evolution (ortholog 61 combined with ortholog 62 [both JTT], and ortholog 71 combined with ortholog 52 [JTT and JTTDCMUT]). We estimated the concatenated ML phylogeny with these partitions and best fitting models of evolution using RAXML (v8.2.12); to ensure a thorough exploration of tree space and support value estimation, we completed 500 BS replicates with every fifth BS tree used as a starting tree for the ML tree search.

Mixture models were also used for ML and BI tree searches because they have been shown to account for among site variation in AA propensities and thus are less prone to artifacts like LBA (Lartillot and Philippe 2004; Lartillot et al. 2007; Le et al. 2008), without the information loss inherent in recoding strategies such as Dayhoff6 (Hernandez and Ryan 2021). In the ML mixture model framework, the model of best fit was LG + C60 + F + G as tested in IQTREE (v1.6.11) (Nguyen et al. 2015). We built an initial tree using the LG + C60 + F + G model, and the resulting tree was used as a guide tree for a posterior mean site frequency model (PMSF) (Wang et al. 2018) (-m LG + C60 + F + G -ft) with 100 BS replicates. To test for consistency in the C60 analysis, the LG + C60 + F + G model was run from a parsimony starting tree and the tree search was repeated five times from random starting trees; all resulting trees were identical, suggesting the ML mixture model tree search was not stuck on a local optimum. To exclude rapidly evolving genes that may exhibit mutational saturation, we calculated average branch lengths as a proxy for rate (Oakley et al. 2013) using ETE3 and divided by the number of taxa; the 10% ( $n = 58$ ) fastest evolving genes (those with the highest average branch lengths) were removed from the concatenated ortholog matrix to produce another matrix (LG + C60 + F + G minus fastest).

BI analyses were completed using the CAT-GTR model of PhyloBayes-MPI (v1.8) (Lartillot et al. 2013) with at least two independent chains. Each chain was run for at least 18,000 generations. Convergence between the three chains in the BI analysis was assessed using the PhyloBayes bpcomp module sampling every ten trees with the first 25% of trees excluded as burn-in. Support values were obtained by calculating the posterior probability at each node. The results of the CAT-GTR analyses of the AA matrices are as follows. Data set 2: 3 chains, 18,000 cycles, maxdiff = 1, meandiff = 0.007, and minimum effective size = 548. Data set 1: 4 chains, 100,000 cycles, maxdiff = 1, meandiff = 0.01, and minimum effective size = 129.

To reduce potential effects of saturation and AA usage bias (Susko and Roger 2007), which have been shown to exist in crustaceans (Rota-Stabelli, Lartillot, et al. 2013), the concatenated matrix was recoded into Dayhoff6 states (Susko and Roger 2007; Giacomelli et al. 2022) (but see Hernandez and Ryan 2021; Foster et al. 2022). Subsequent phylogenetic analyses were carried out with RAXML using a GTR substitution model on the same 88 partitions as used in the analysis of the full matrix with an automated BS convergence criterion (autoMRE). A BI analysis was completed using the

CAT-GTR model on the recoded matrix with 2 independent chains run for at least 80,000 generations each; the consensus tree of both chains was made by sampling 10 trees with the first 25% of trees excluded as burn-in; convergence between chains was assessed as above. The results of the CAT-GTR analyses of the Dayhoff6 matrices are as follows. Data set 2: 2 chains, 80,000 cycles, maxdiff = 1, meandiff = 0.006, and minimum effective size = 3,643. Data set 1: 2 chains, 80,000 cycles, maxdiff = 1, meandiff = 0.09, and minimum effective size = 3,533.

For the multispecies coalescent phylogeny, individual gene trees were built for each ortholog using IQTREE with the AA substitution model of best fit by BIC score and 200 BS replicates. The species tree was estimated by using the ML gene trees as input in ASTRAL-III (v5.6.3) (Mirarab et al. 2016; Zhang et al. 2018). The sensitivity of the ASTRAL-III analyses to a number of ortholog features was explored. Ortholog features were based on Shen et al. (2016), which identified gene properties most strongly associated with phylogenetic signals. These properties were measured using the python package ETE3 (Huerta-Cepas et al. 2016) unless otherwise noted. The following were measured for each ortholog: number of taxa, alignment length, total AA in alignment, number of gaps in alignment, percent of gaps, number of variable sites, proportion of variable sites, number of parsimony informative sites, proportion of parsimony informative sites, tree length, average branch length (tree length/number of taxa), and compositional homogeneity (supplementary table S4, Supplementary Material online). ML ortholog branches with <10% and 30% BS support in the individual ortholog trees were also collapsed prior to running ASTRAL-III following Zhang et al. (2018). Branch support for the ASTRAL-III analyses was assessed using local posterior probabilities (Sayyari and Mirarab 2016).

We identified shared orthogroups between the data sets using BlastP. Specifically, a reciprocal BlastP analysis was performed between data sets with the AA sequences that had not been trimmed for the phylogenetic analyses to avoid missing AA sites. Because we predicted orthologs using maximum inclusion and the different data sets include different taxa, we only considered orthogroups that share the same sequences (i.e., 100% identity across their overlapping length) and the same taxa. The different taxa in each data set ultimately affect the MCL homolog clustering and ortholog prediction; therefore, we only chose to compare orthologs from different data sets with the same taxa and sequences.

Topological differences from the phylogenetic analyses were assessed using Robinson-Foulds (RF) symmetric distances calculated in ETE3 (Huerta-Cepas et al. 2016) and the information metric of Kendall and Colijn (2016) using the R package TreeSpace (Jombart et al. 2017). A multidimensional scaling (MDS) plot showing topological variation between analyses based on the Kendall and Colijn metric was also calculated with TreeSpace (fig. 3). Topologies from the different phylogenetic analyses were compared by AU-test (Shimodaira 2002) in IQTREE



(v1.6.11) along with published pancrustacean phylogenies (Regier et al. 2010; von Reumont et al. 2012; Oakley et al. 2013; Rota-Stabelli, Lartillot, et al. 2013; Schwentner et al. 2017, 2018; Lozano-Fernandez et al. 2019) using the full ortholog matrix and an LG + G + F + I model with 100,000 BS with the RELL method.

We also estimated LB scores with Phykit v1.11.10 (Steenwyk et al. 2021) to compare the relative lengths of terminal branches in each phylogeny and measure the degree of taxon-specific LBA. The LB score for each taxon is the mean patristic distance between itself and all other taxa directly proportional to the mean of all patristic distances for all taxa in the phylogeny (Struck 2014; Weigert et al. 2014). The larger the LB score, the longer the terminal branch with respect to all other taxa in the tree. Although this metric serves as a measure to compare terminal branches in a phylogeny, it cannot be used here to compare terminal branches in different phylogenies because branch lengths are parameter estimates and our model changed in each phylogeny due to different taxon and gene sampling.

### Divergence Time Estimation

Divergence time estimation was based on 12 vetted internal fossil calibrations (6 from Wolfe et al. (2016), 2 from Wolfe et al. (2019), and 4 new) and the root prior was defined based on the Euarthropoda node (Wolfe et al. 2016, node 4) with a gamma distribution with mean 575 Ma and sd 61 Ma (fossil ages and justifications in [supplementary table S7, Supplementary Material](#) online). Divergence times for the main analysis were estimated using MCMCTree (dos Reis and Yang 2011; dos Reis, Donoghue, et al. 2015) and the full AA matrix of Data set 2 using the fixed topology of the highest likelihood tree, which resulted from the LG + C60 + F + G analysis. Divergence times were calculated using an independent, lognormal model (clock = 2) and approximate likelihood calculation; the Hessian calculation for approximate likelihood was done using an LG + G4 matrix (LG was the model of best fit by Akaike information criterion (AIC) and BIC). We ran 3 chains for 20 million generations each, treating the first 25% as burn-in and sampling every 1,000 trees, and a fourth chain with the same settings without data to sample from the time prior. Convergence of the three chains was assessed visually by plotting the distributions of the three chains against each other. The distributions of the chains were nearly perfectly linear at 2 million generations. To further ensure convergence, we ran each chain for 20 million generations and assessed convergence visually in the same way ([supplementary fig. S7, Supplementary Material](#) online). Results in [figure 4](#) and [supplementary figure S9, Supplementary Material](#) online, were plotted using the *MCMCTreeR* package (Puttick 2019).

Divergence times were also estimated in PhyloBayes v1.8 (Lartillot et al. 2013) using a fixed topology from the C60 + LG + F + G analysis. Due to the size of our data matrices and time to convergence, we assembled an AA alignment consisting of the 50 loci with the highest normalized RF distance compared with the species tree resulting from the LG

+ C60 + F + G analysis of the full matrix. This subsampled matrix was then used for divergence time estimation in PhyloBayes (Mongiardino Koch 2021). The inability to use the entire alignment is why we included these as supplementary analyses. We used the C20 + LG substitution model and compared the uncorrelated gamma multipliers (UGM) and lognormal (LN) relaxed clock models (Drummond et al. 2006) and the autocorrelated CIR clock model (Lepage et al. 2007) with three chains per run. Although the topology was fixed, we used a birth–death tree model, with soft bounds allowing 5% of the probability distribution outside the input fossil ages. An automatic stopping rule was implemented, with tests of convergence every 100 cycles, until the default criteria of effective sample sizes and parameter discrepancies between chains were met (50 and 0.3, respectively). Although many values did converge to <0.3, as is commonly the case with PhyloBayes, not all values fully converged even after months of runtime (sigma, mu, scale, and p2 were consistently >0.3, whereas all other parameters were <0.1). To further assess convergence, we visualized trace plots of logL values for each of the 3 chains running for each model in R ([supplementary table S8, Supplementary Material](#) online). Trees and their posterior distributions were generated from completed chains after the initial 20% of sampled generations were discarded. We compared estimated posterior age distributions to the marginal prior by removing sequence data using the -prior flag (Warnock et al. 2012; Brown and Smith 2017).

### Supplementary Material

[Supplementary data](#) are available at *Molecular Biology and Evolution* online.

### Acknowledgments

We thank Nicolás Mongiardino Koch for helpful discussions on divergence time estimation and for a script for generating logL plots of PhyloBayes chains. We are grateful to David Fenwick and Aphotomarine.com for the use of their buoy barnacle photograph. We are thankful to three reviewers whose comments improved this manuscript. This material is based in part upon work supported by the NSF Postdoctoral Research Fellowships in Biology Program under Grant No. 2010898 to J.P.B. J.M.W. is supported by NSF Division of Environmental Biology Grant No. 1856679.

### Data Availability

All alignments and tree files from this study are available on Dryad: <https://doi.org/10.5061/dryad.dr7sqvb2h>. Sample information and sequence data for de novo transcriptomes are available in NCBI BioProject PRJNA997050.

**Conflict of interest statement.** Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not

imply recommendation or endorsement by the USDA; USDA is an equal opportunity provider and employer.

## References

- Alfsnes K, Leinaas HP, Hessen DO. 2017. Genome size in arthropods; different roles of phylogeny, habitat and life history in insects and crustaceans. *Ecol Evol*. **7**:5939–5947.
- Altenhoff AM, Dessimoz C. 2009. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol*. **5**:e1000262.
- Altenhoff AM, Schneider A, Gonnet GH, Dessimoz C. 2011. OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Res*. **39**:D289–D294.
- Altschul S, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol*. **215**:403–410.
- Altschul S, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. **25**:3389–3402.
- Andrew DR. 2011. A new view of insect–crustacean relationships II. Inferences from expressed sequence tags and comparisons with neural cladistics. *Arthropod Struct Dev*. **40**:289–302.
- Andrews S. 2018. *FastQC: a quality control tool for high throughput sequence data*. Available from: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- Ballesteros JA, Hormiga G. 2016. A new orthology assessment method for phylogenomic data: unrooted phylogenetic orthology. *Mol Biol Evol*. **33**:2117–2134.
- Bar-On YM, Phillips R, Milo R. 2018. The biomass distribution on earth. *Proc Natl Acad Sci*. **115**:6506–6511.
- Barba-Montoya J, dos Reis M, Yang Z. 2017. Comparison of different strategies for using fossil calibrations to generate the time prior in Bayesian molecular clock dating. *Mol Phylogenet Evol*. **114**:386–400.
- Bayzid MS, Warnow T. 2013. Naive binning improves phylogenomic analyses. *Bioinformatics* **29**:2277–2284.
- Bernot JP, Avdeyev P, Zamyatin A, Dreyer N, Alexeev N, Pérez-Losada M, Crandall KA. 2022. Chromosome-level genome assembly, annotation, and phylogenomics of the gooseneck barnacle *Pollicipes pollicipes*. *GigaScience* **11**:giac021.
- Betancur-RR, Arcila D, Vari RP, Hughes LC, Oliveira C, Sabaj MH, Ortí G. 2019. Phylogenomic incongruence, hypothesis testing, and taxonomic sampling: the monophyly of characiform fishes\*. *Evolution* **73**:329–345.
- Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**:2114–2120.
- Boxshall GA. 1983. A comparative functional analysis of the major maxillopodan groups. In: Schram FR, editor. *Crustacean phylogeny*. Rotterdam: A. A. Balkema. p. 121–143.
- Boxshall GA. 2007. Crustacean classification: on-going controversies and unresolved problems. *Zootaxa* **1668**(1):313–325.
- Bracken-Grissom H, Wolfe JM. 2020. The pancrustacean conundrum: a conflicted phylogeny with emphasis on Crustacea. In: Poore GCB, Thiel M, editors. *Evolution and biogeography. Vol. 8. Natural history of the Crustacea*. Oxford: Oxford University Press. p. 80–104.
- Branstetter MG, Danforth BN, Pitts JP, Faircloth BC, Ward PS, Buffington ML, Gates MW, Kula RR, Brady SG. 2017. Phylogenomic insights into the evolution of stinging wasps and the origins of ants and bees. *Curr Biol*. **27**:1019–1025.
- Brown J, Smith S. 2017. The past sure is tense: on interpreting phylogenetic divergence time estimates. *Syst Biol*. **67**:340–353.
- Budd GE, Mann RP. 2020a. The dynamics of stem and crown groups. *Sci Adv*. **6**:eaaz1626.
- Budd GE, Mann RP. 2020b. Survival and selection biases in early animal evolution and a source of systematic overestimation in molecular clocks. *Interface Focus*. **10**:20190110.
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics*. **10**:421.
- Chan BK, Dreyer N, Gale AS, Glenner H, Ewers-Saucedo C, Pérez-Losada M, Kolbasov GA, Crandall KA, Høeg JT. 2021. The evolutionary diversity of barnacles, with an updated classification of fossil and living forms. *Zool J Linn Soc*. **193**(3):789–846.
- Chen F, Mackey AJ, Vermunt JK, Roos DS. 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS One* **2**:e383.
- Chiu JC, Lee EK, Egan MG, Sarkar IN, Coruzzi GM, DeSalle R. 2006. OrthologID: automation of genome-scale ortholog identification within a parsimony framework. *Bioinformatics* **22**:699–707.
- Daley AC, Antcliff JB, Drage HB, Pates S. 2018. Early fossil record of Euarthropoda and the Cambrian explosion. *Proc Natl Acad Sci U S A*. **115**:5323–5331.
- DeGiorgio M, Degnan JH. 2014. Robustness to divergence time underestimation when inferring species trees from estimated gene trees. *Syst Biol*. **63**:66–82.
- dos Reis M, Donoghue PCJ, Yang Z. 2015. Bayesian molecular clock dating of species divergences in the genomics era. *Nat Rev Genet*. **17**:71–80.
- dos Reis M, Thawornwattana Y, Angelis K, Telford MJ, Donoghue PCJ, Yang Z. 2015. Uncertainty in the timing of origin of animals and the limits of precision in molecular timescales. *Curr Biol*. **25**:2939–2950.
- dos Reis M, Yang Z. 2011. Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. *Mol Biol Evol*. **28**:2161–2172.
- Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol*. **4**:e88.
- Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, et al. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* **452**:745–749.
- Dunn CW, Howison M, Zapata F. 2013. Agalma: an automated phylogenomics workflow. *BMC Bioinformatics*. **14**:330.
- Ellis EA, Goodheart JA, Hensley NM, González VL, Reda NJ, Rivers TJ, Morin JG, Torres E, Gerrish GA, Oakley TH. 2023. Sexual signals persist over deep time: ancient co-option of bioluminescence for courtship displays in cypridinid ostracods. *Syst Biol* **72**:264–274.
- Fishbein M, Hibsich-Jetter C, Soltis DE, Hufford L. 2001. Phylogeny of Saxifragales (angiosperms, eudicots): analysis of a rapid, ancient radiation. *Syst Biol*. **50**:817–847.
- Foster PG, Schrepf D, Szöllösi GJ, Williams TA, Cox CJ, Embley TM. 2022. Recoding amino acids to a reduced alphabet may increase or decrease phylogenetic accuracy. *Syst Biol*. **2022**:syac042.
- Fryer G. 1983. Functional ontogenetic changes in *Branchinecta ferox* (Milne-Edwards) (Crustacea: Anostraca). *Philos Trans R Soc Lond B*. **303**:229–343.
- Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**:3150–3152.
- Gelman A, Rubin DB. 1992. Inference from iterative simulation using multiple sequences. *Statist Sci*. **7**:457–472.
- Giacomelli M, Rossi ME, Lozano-Fernandez J, Feuda R, Pisani D. 2022. Resolving tricky nodes in the tree of life through amino acid recoding. *Iscience*. **2022**(12):105594.
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. **29**:644–652.
- Gurney R. 1942. Larvae of decapod Crustacea. *Ray Soc*. **129**:1–306.
- Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. 2013. De novo transcript sequence reconstruction from RNA-Seq using the Trinity platform for reference generation and analysis. *Nat Protoc*. **8**:1494–1512.
- Heath TA, Hedtke SM, Hillis DM. 2008. Taxon sampling and the accuracy of phylogenetic analyses. *J Syst Evol*. **46**:239–257.

- Hedtke SM, Townsend TM, Hillis DM. 2006. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. *Syst Biol*. **55**:522–529.
- Hegna TA, Luque J, Wolfe JM. 2020. The fossil record of the pancrustacea. In: Poore GCB, Thiel M, editors. *Evolution and biogeography*. Vol. 8. *Nat hist Crustacea*. Oxford: Oxford Univ Press. p. 21–52.
- Hellmuth M, Wieseke N, Lechner M, Lenhof H-P, Middendorf M, Stadler PF. 2015. Phylogenomics with paralogs. *Proc Natl Acad Sci*. **112**:2058–2063.
- Hendy MD, Penny D. 1989. A framework for the quantitative study of evolutionary trees. *Syst Zool*. **38**:297.
- Hernandez AM, Ryan JF. 2021. Six-state amino acid recoding is not an effective strategy to offset compositional heterogeneity and saturation in phylogenetic analyses. *Syst Biol*. **70**:1200–1212.
- Höpel CG, Yeo D, Grams M, Meier R, Richter S. 2022. Mitogenomics supports the monophyly of Mysidacea and Peracarida (Malacostraca). *Zool Scripta*. **51**:603–613.
- Hu F, Lin Y, Tang J. 2014. MLGO: phylogeny reconstruction and ancestral inference from gene-order data. *BMC Bioinformatics*. **15**:354.
- Huang H, He Q, Kubatko LS, Knowles LL. 2010. Sources of error inherent in species-tree estimation: impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. *Syst Biol*. **59**:573–583.
- Huelsenbeck JP, Larget B, Miller RE, Ronquist F. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. *Syst Biol*. **51**:673–688.
- Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol*. **33**:1635–1638.
- Huggall AF, Lee MSY. 2007. The likelihood node density effect and consequences for evolutionary studies of molecular rates. *Evolution*. **61**:2293–2307.
- Huys R, Boxshall GA. 1991. *Copepod evolution*. London: Ray Soc.
- Huys R, Boxshall GA, Lincoln RJ. 1993. The tantulocaridan life cycle: the circle closed? *J Crustacean Biol*. **13**:432–442.
- Itô T. 1989. Origin of the basis in copepod limbs, with reference to remipedian and cephalocarid limbs. *J Crustacean Biol*. **9**:85–103.
- Jombart T, Kendall M, Almagro-Garcia J, Colijn C. 2017. treespace: statistical exploration of landscapes of phylogenetic trees. *Mol Ecol Resour*. **17**:1385–1392.
- Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. **30**:772–780.
- Kendall M, Colijn C. 2016. Mapping phylogenetic trees to reveal distinct patterns of evolution. *Mol Biol Evol*. **33**:2735–2743.
- Koga C, Rouse GW. 2021. Mitogenomics and the phylogeny of mantis shrimps (Crustacea: Stomatopoda). *Diversity (Basel)*. **13**:647.
- Lanfear R, Calcott B, Kainer D, Mayer C, Stamatakis A. 2014. Selecting optimal partitioning schemes for phylogenomic datasets. *BMC Evol Biol*. **14**:82.
- Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B. 2017. Partitionfinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol Biol Evol*. **34**:772–773.
- Lanier HC, Knowles LL. 2015. Applying species-tree analyses to deep phylogenetic histories: challenges and potential suggested from a survey of empirical phylogenetic studies. *Mol Phylogenet Evol*. **83**:191–199.
- Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol Biol*. **7**:S4.
- Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol*. **21**:1095–1109.
- Lartillot N, Rodrigue N, Stubbs D, Richer J. 2013. Phylobayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst Biol*. **62**:611–615.
- Laumer CE, Fernández R, Lemer S, Combosch D, Kocot KM, Riesgo A, Andrade SCS, Sterrer W, Sørensen MV, Giribet G. 2019. Revisiting metazoan phylogeny with genomic sampling of all phyla. *Proc R Soc B Biol Sci*. **286**:20190831.
- Le SQ, Gascuel O, Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics*. **24**:2317–2323.
- Lee MSY, Soubrier J, Edgecombe GD. 2013. Rates of phenotypic and genomic evolution during the Cambrian explosion. *Curr Biol*. **23**:1889–1895.
- Lepage T, Bryant D, Philippe H, Lartillot N. 2007. A general comparison of relaxed molecular clock models. *Mol Biol Evol*. **24**:2669–2680.
- Lewin HA, Richards S, Lieberman Aiden E, Allende ML, Archibald JM, Bálint M, Barker KB, Baumgartner B, Belov K, Bertorelle G, et al. 2022. The Earth BioGenome Project 2020: starting the clock. *Proc Natl Acad Sci*. **119**:e2115635118.
- Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, Durbin R, Edwards SV, Forest F, Gilbert MTP, et al. 2018. Earth BioGenome Project: sequencing life for the future of life. *Proc Natl Acad Sci*. **115**:4325–4333.
- Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. **22**:1658–1659.
- Li Y, Shen X-X, Evans B, Dunn CW, Rokas A. 2021. Rooting the animal tree of life. *Mol Biol Evol*. **38**:4322–4333.
- Lozano-Fernandez J, Carton R, Tanner AR, Puttick MN, Blaxter M, Vinther J, Olesen J, Giribet G, Edgecombe GD, Pisani D. 2016. A molecular palaeobiological exploration of arthropod terrestrialization. *Philos Trans R Soc B Biol Sci*. **371**:20150133.
- Lozano-Fernandez J, Giacomelli M, Fleming J, Chen A, Vinther J, Thomsen PF, Glenner H, Palero F, Legg DA, Iliffe TM, et al. 2019. Pancrustacean evolution illuminated by taxon-rich genomic-scale data sets with an expanded remipede sampling. *Genome Biol Evol*. **11**:2055–2070.
- Lozano-Fernandez J, Tanner AR, Puttick MN, Vinther J, Edgecombe GD, Pisani D. 2020. A Cambrian–Ordovician terrestrialization of arachnids. *Front Genet*. **11**:182.
- Luque J, Gerken S. 2019. Exceptional preservation of comma shrimp from a mid-Cretaceous Lagerstätte of Colombia, and the origins of crown Cumacea. *Proc R Soc B Biol Sci*. **286**:20191863.
- Mallatt J, Giribet G. 2006. Further use of nearly complete 28S and 18S rRNA genes to classify Ecdysozoa: 37 more arthropods and a kinorhynch. *Mol Phylogenet Evol*. **40**:772–794.
- McClain CR, Balk MA, Benfield MC, Branch TA, Chen C, Cosgrove J, Dove ADM, Gaskins L, Helm RR, Hochberg FG, et al. 2015. Sizing ocean giants: patterns of intraspecific size variation in marine megafauna. *PeerJ*. **3**:e715.
- Mirarab S, Bayzid MS, Boussau B, Warnow T. 2014. Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science*. **346**:1250463.
- Mirarab S, Bayzid MS, Warnow T. 2016. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. *Syst Biol*. **65**:366–380.
- Mirarab S, Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*. **31**:i44–i52.
- Møller OS, Olesen J, Avenant-Oldewage A, Thomsen PF, 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 Struct Dev*. **37**:333–346.
- Mongiardino Koch NM. 2021. Phylogenomic subsampling and the search for phylogenetically reliable loci. *Mol Biol Evol*. **38**:4025–4038.
- Moret BME, Warnow T. 2005. Advances in phylogeny reconstruction from gene order and content data. *Methods Enzymol*. **395**:673–700.
- Nabhan AR, Sarker IN. 2012. The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. *Brief Bioinform*. **13**:122–134.
- Nardi F, Spinsanti G, Boore JL, Carapelli A, Dallai R, Frati F. 2003. Hexapod origins: monophyletic or paraphyletic? *Science*. **299**(5614):1887–1889.



- Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol.* **32**:268–274.
- Oakley TH, Wolfe JM, Lindgren AR, Zaharoff AK. 2013. Phylotranscriptomics to bring the understudied into the fold: monophyletic ostracoda, fossil placement, and pancrustacean phylogeny. *Mol Biol Evol.* **30**:215–233.
- Olesen J. 2007. Monophyly and phylogeny of Branchiopoda, with focus on morphology and homologies of branchiopod phyllopodous limbs. *J Crustacean Biol.* **27**:165–183.
- Olesen J, Haug JT, Maas A, Waloszek D. 2011. External morphology of *Lightiella monniotae* (Crustacea, Cephalocarida) in the light of Cambrian “Orsten” crustaceans. *Arthropod Struct Dev.* **40**(5): 449–478.
- Olesen J, Martinsen SV, Iliffe TM, Koenemann S. 2014. Chapter 15. Remipedia. In: Martin JW Olesen J, Høeg JT, editors. *Atlas of crustacean larvae*. Balt: Johns Hopkins Univ Press. p. 84–89.
- One Thousand Plant Transcriptomes Initiative. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**:679–685.
- Owen CL, Stern DB, Hilton SK, Crandall KA. 2020. Hemiptera phylogenomic resources: tree-based orthology prediction and conserved exon identification. *Mol Ecol Resour.* **20**:1346–1360.
- Patel S, Kimball RT, Braun EL. 2013. Error in phylogenetic estimation for bushes in the tree of life. *J Phylogenetics Evol Biol.* **01**:110.
- Petrunina AS, Høeg JT, Kolbasov GA. 2018. Anatomy of the Tantulocarida: first results obtained using TEM and CLSM. Part I: tantulus larva. *Org Divers Evol.* **18**:459–477.
- Petrunina AS, Neretina TV, Mugue NS, Kolbasov GA. 2014. Tantulocarida versus Thecostraca: inside or outside? First attempts to resolve phylogenetic position of Tantulocarida using gene sequences. *J Zool Syst Evol Res.* **52**:100–108.
- Poe S. 2003. Evaluation of the strategy of long-branch subdivision to improve the accuracy of phylogenetic methods. *Syst Biol.* **52**: 423–428.
- Price MN, Dehal PS, Arkin AP. 2010. Fasttree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* **5**:e9490.
- Puttick MN. 2019. MCMCtree: functions to prepare MCMCtree analyses and visualise posterior ages on trees. *Bioinformatics* **35**:5321–5322.
- Regier J, Shultz J, Ganley A, Hussey A, Shi D, Ball B, Zwick A, Stajich J, Cummings M, Martin J, et al. 2008. Resolving arthropod phylogeny: exploring phylogenetic signal within 41 kb of protein-coding nuclear gene sequence. *Syst Biol.* **57**:920–938.
- Regier JC, Shultz JW, Kambic RE. 2005. Pancrustacean phylogeny: hexapods are terrestrial crustaceans and maxillopods are not monophyletic. *Proc R Soc B Biol Sci.* **272**:395–401.
- Regier JC, Shultz JW, Zwick A, Hussey A, Ball B, Wetzer R, Martin JW, Cunningham CW. 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature* **463**:1079–1083.
- Richter S, Scholtz G. 2001. Phylogenetic analysis of the Malacostraca (Crustacea). *J Zool Syst Evol Res.* **39**:113–136.
- Robin N, Gueriau P, Luque J, Jarvis D, Daley AC, Vonk R. 2021. The oldest peracarid crustacean reveals a Late Devonian freshwater colonization by isopod relatives. *Biol Lett.* **17**:20210226.
- Rokas A, Carroll SB. 2006. Bushes in the tree of life. *PLoS Biol.* **4**: e352.
- Roskov Y, Ower G, Orrell T, Nicolson D, Bailly N, Kirk PM, Bourgoin T, DeWalt RE, Decock W, van Nieukerken E, et al. editors 2022. *Species 2000 & ITIS catalogue of life*, 10th February 2022. Leiden, the Netherlands: Species 2000: Naturalis. ISSN 2405-8858. Digital resource at [www.catalogueoflife.org/col](http://www.catalogueoflife.org/col).
- Rota-Stabelli O, Daley AC, Pisani D. 2013. Molecular timetrees reveal a Cambrian colonization of land and a new scenario for ecdysozoan evolution. *Curr Biol.* **23**:392–398.
- Rota-Stabelli O, Lartillot N, Philippe H, Pisani D. 2013. Serine codon-usage bias in deep phylogenomics: Pancrustacean relationships as a case study. *Syst Biol.* **62**:121–133.
- Sanders HL. 1963. The Cephalocarida. Functional morphology, larval development, comparative external anatomy. *Mem Conn Acad Arts Sci.* **15**:1–80.
- Sayyari E, Mirarab S. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. *Mol Biol Evol.* **33**: 1654–1668.
- Schram FR. 1984. Fossil syncarida. *Trans San Diego Soc Nat Hist.* **20**: 189–246.
- Schram FR, Hof CH. 1998. Fossils and the interrelationships of major crustacean groups. In: Edgecombe GD, editors. *Arthropod fossils and phylogeny*. New York: Columbia University Press. p. 233–302.
- Schultz DT, Haddock SH, Bredeson JV, Green RE, Simakov O, Rokhsar DS. 2023. Ancient gene linkages support ctenophores as sister to other animals. *Nature* **618**:110–117.
- Schwentner M, Combosch DJ, Pakes Nelson J, Giribet G. 2017. A phylogenomic solution to the origin of insects by resolving crustacean-hexapod relationships. *Curr Biol.* **27**:1818–1824.e5.
- Schwentner M, Richter S, Rogers DC, Giribet G. 2018. Tetraconatan phylogeny with special focus on Malacostraca and Branchiopoda: highlighting the strength of taxon-specific matrices in phylogenomics. *Proc R Soc B Biol Sci.* **285**:20181524.
- Scornavacca C, Galtier N. 2017. Incomplete lineage sorting in mammalian phylogenomics. *Sys Bio.* **66**(1):112–120.
- Serban E. 1972. Bathynella (Podophallocarida, Bathynellacea). *Travaux de l'Institut de Spéologie 'Émile Racovitza'*. **11**:11–225.
- Serban E. 1973. Sur le processus de la pléonisation du péreion dans l'ordre des bathynellacea (Crustacea, Malacostraca, Podophallocarida). *Bijdragen tot de Dierkunde.* **43**:173–201.
- Sharma PP, Kaluziak ST, Pérez-Porro AR, González VL, Hormiga G, Wheeler WC, Giribet G. 2014. Phylogenomic interrogation of Arachnida reveals systemic conflicts in phylogenetic signal. *Mol Bio Evol.* **31**(11):2963–2984..
- Shen X-X, Hittinger CT, Rokas A. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nat Ecol Evol.* **1**:0126.
- Shen X-X, Salichos L, Rokas A. 2016. A genome-scale investigation of how sequence-, function-, and tree-based gene properties influence phylogenetic inference. *Genome Biol Evol.* **8**:2565–2580.
- Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. *Syst Biol.* **51**:492–508.
- Siewing R. 1956. Untersuchungen zur Morphologie der Malacostraca (Crustacea). *Zool Jahrb Abt Anat Ontog Tiere.* **75**:39–176.
- Simakov O, Bredeson J, Berkoff K, Marletaz F, Mitros T, Schultz DT, O'Connell BL, Dear P, Martinez DE, Steele RE, et al. 2022. Deeply conserved synteny and the evolution of metazoan chromosomes. *Sci Adv.* **8**:eabi5884.
- Smith ML, Hahn MW. 2021. New approaches for inferring phylogenies in the presence of paralogs. *Trends Genet.* **37**:174–187.
- Smith ML, Vanderpool D, Hahn MW. 2022. Using all gene families vastly expands data available for phylogenomic inference. *Mol Biol Evol.* **39**:msac112.
- Smith SA, Dunn CW. 2008. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics* **24**: 715–716.
- Smith SA, Pease JB. 2017. Heterogeneous molecular processes among the causes of how sequence similarity scores can fail to recapitulate phylogeny. *Brief Bioinform.* **18**:451–457.
- Smith SA, Wilson NG, Goetz FE, Feehely C, Andrade SCS, Rouse GW, Giribet G, Dunn CW. 2011. Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature* **480**:364–367.
- Song L, Florea L. 2015. Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. *GigaScience* **4**:48.
- Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.
- Steenwyk JL, Buida TJ, Labella AL, Li Y, Shen X-X, Rokas A. 2021. PhyKIT: a broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data. *Bioinformatics* **37**:2325–2331.

- Struck T. 2014. Trespex--detection of misleading signal in phylogenetic reconstructions based on tree information. *Evol Bioinform Online*. **10**:51–67.
- Susko E, Roger AJ. 2007. On reduced amino acid alphabets for phylogenetic inference. *Mol Biol Evol*. **24**:2139–2150.
- Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol*. **56**:564–577.
- Van Dongen S. 2000. Graph clustering by flow simulation [PhD thesis]. University of Utrecht.
- Van Dongen S. 2008. Graph clustering via a discrete uncoupling process. *SIAM J Matrix Anal Appl*. **30**:121–141.
- von Reumont BM, Jenner RA, Wills MA, Dell’Ampio E, Pass G, Ebersberger I, Meyer B, Koenemann S, Iliffe TM, Stamatakis A, et al. 2012. Pancrustacean phylogeny in the light of new phylogenomic data: support for Remipedia as the possible sister group of hexapoda. *Mol Biol Evol*. **29**:1031–1045.
- Walossek D. 1993. The Upper Cambrian Rehbachiella and the phylogeny of Branchiopoda and Crustacea. *Fossils and Strata*. **32**:1–202.
- Wang H-C, Minh BQ, Susko E, Roger AJ. 2018. Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. *Syst Biol*. **67**:216–235.
- Warnock RCM, Yang Z, Donoghue PCJ. 2012. Exploring uncertainty in the calibration of the molecular clock. *Biol Lett*. **8**:156–159.
- Weigert A, Helm C, Meyer M, Nickel B, Arendt D, Hausdorf B, Santos SR, Halanych KM, Purschke G, Bleidorn C, et al. 2014. Illuminating the base of the annelid tree using transcriptomics. *Mol Biol Evol*. **31**:1391–1401.
- Whelan NV, Halanych KM. 2017. Who let the CAT out of the bag? Accurately dealing with substitutional heterogeneity in phylogenomic analyses. *Syst Biol*. **66**:232–255.
- Whitfield JB, Lockhart PJ. 2007. Deciphering ancient rapid radiations. *Trends Ecol Evol*. **22**:258–265.
- Wolfe JM. 2017. Metamorphosis is ancestral for crown Euarthropods, and evolved in the Cambrian or earlier. *Integr Comp Biol*. **57**:499–509.
- Wolfe JM, Breinholt JW, Crandall KA, Lemmon AR, Moriarty Lemmon E, Timm LE, Siddall ME, Bracken-Grissom HD. 2019. A phylogenomic framework, evolutionary timeline, and genomic resources for comparative studies of decapod crustaceans. *Proc R Soc B Biol Sci*. **286**:20190079.
- Wolfe JM, Daley AC, Legg DA, Edgecombe GD. 2016. Fossil calibrations for the arthropod tree of life. *Earth-Sci Rev*. **160**:43–110.
- Wolfe JM, Hegna TA. 2014. Testing the phylogenetic position of Cambrian pancrustacean larval fossils by coding ontogenetic stages. *Cladistics* **30**(4):366–390.
- WoRMS (2023). *Crustacea*. [cited 2023 March 11]. Available from: <https://www.marinespecies.org/aphia.php?p=taxdetails&id=1066>
- Xi Z, Liu L, Davis CC. 2015. Genes with minimal phylogenetic information are problematic for coalescent analyses when gene tree estimation is biased. *Mol Phylogenet Evol*. **92**:63–71.
- Yager J. 1981. Remipedia, a new class of Crustacea from a marine cave in the Bahamas. *J Crustacean Biol*. **1**:328–333.
- Yang Y, Smith SA. 2014. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. *Mol Biol Evol*. **31**:3081–3092.
- Zhai D, Ortega-Hernández J, Wolfe JM, Hou X, Cao C, Liu Y. 2019. Three-dimensionally preserved appendages in an early Cambrian stem-group pancrustacean. *Curr Biol*. **29**:171–177.
- Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*. **19**:153.
- Zhang X, Siveter DJ, Waloszek D, Maas A. 2007. An epipodite-bearing crown-group crustacean from the Lower Cambrian. *Nature* **449**:595–598.