Molecular Phylogenetics and Evolution 118 (2018) 330–342

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Assessing the utility of transcriptome data for inferring phylogenetic relationships among coleoid cephalopods

Annie R. Lindgren[a,*], Frank E. Anderson[b]

[a] The Center for Life in Extreme Environments, Department of Biology, Portland State University, 1719 SW 10th Ave, SRTC Rm 246, Portland, OR 97201, USA
[b] Department of Zoology, Southern Illinois University, Carbondale, IL, USA

ARTICLE INFO

Keywords:
Cephalopoda
Decapodiformes
Phylotranscriptomics
Molecular phylogenetics

ABSTRACT

Historically, deep-level relationships within the molluscan class Cephalopoda (squids, cuttlefishes, octopods and their relatives) have remained elusive due in part to the considerable morphological diversity of extant taxa, a limited fossil record for species that lack a calcareous shell and difficulties in sampling open ocean taxa. Many conflicts identified by morphologists in the early 1900s remain unresolved today in spite of advances in morphological, molecular and analytical methods. In this study we assess the utility of transcriptome data for resolving cephalopod phylogeny, with special focus on the orders of Decapodiformes (open-eye squids, bobtail squids, cuttlefishes and relatives). To do so, we took new and previously published transcriptome data and used a unique cephalopod core ortholog set to generate a dataset that was subjected to an array of filtering and analytical methods to assess the impacts of: taxon sampling, ortholog number, compositional and rate heterogeneity and incongruence across loci. Analyses indicated that datasets that maximized taxonomic coverage but included fewer orthologs were less stable than datasets that sacrificed taxon sampling to increase the number of orthologs. Clades recovered irrespective of dataset, filtering or analytical method included Octopodiformes (*Vampyroteuthis infernalis* + octopods), Decapodiformes (squids, cuttlefishes and their relatives), and orders Oegopsida (open-eyed squids) and Myopsida (e.g., loliginid squids). Ordinal-level relationships within Decapodiformes were the most susceptible to dataset perturbation, further emphasizing the challenges associated with uncovering relationships at deep nodes in the cephalopod tree of life.

1. Introduction

The molluscan class Cephalopoda contains some of the most charismatic invertebrates on Earth, and yet, many questions about their evolutionary history remain. The approximately 900 species of extant nautiloids, octopods, bobtail squids, cuttlefishes and squids comprise a group defined by a high degree of morphological diversity, rapid radiation and a poor fossil record for many taxa, all of which make inferring their phylogenetic history challenging. Two major factors may have influenced radiation and diversification of extant cephalopod taxa: the extinction of the ammonites and belemnoids at ~66 mya, which may have opened up new niches, and the radiation of bony fishes, which are direct competitors with, prey of and predators on cephalopods (Aronson, 1991). Several cephalopod clades likely have undergone major Cenozoic radiations, including Oegopsida (~250 sp., most oceanic 'open-eye' squids), Octopodidae (~150 sp., benthic octopods) and Sepiida (~100 sp., cuttlefishes), while other lineages such as *Vampyroteuthis infernalis* appear to have remained relatively unchanged. The timing and tempo of these radiations are difficult to assess due to weak fossil data for many lineages and an uncertain phylogeny for deep nodes (see Allcock et al., 2014 for details), even though significant methodological advances have been made (Rabosky et al., 2013; Stadler, 2011). Internal factors that can affect diversification rates such as genome duplication events have been identified in vertebrates (Jaillon et al., 2004) where duplicated genes likely led to new functions, such as osmoregulation in salmonids (Norman et al., 2012). Although less well studied, evidence for one or more genome duplication events in cephalopods exists (Hallinan and Lindberg, 2011). Lastly, extant cephalopods have undergone several habitat transitions that likely influenced diversification rate and character evolution (e.g., Kröger et al., 2011; Strugnell et al., 2006).

Over the last century, researchers have utilized a variety of approaches to study phylogenetic relationships within Cephalopoda, with limited success. Despite extensive work using morphological data, traditional multi-gene Sanger sequencing techniques or whole mitochondrial genomes, a good understanding of cephalopod ordinal relationships remains elusive (Allcock et al., 2011; Lindgren, 2010; Young and Vecchione, 1996). The largest molecular phylogenetic study in terms of

* Corresponding author.
E-mail addresses: arl3@pdx.edu (A.R. Lindgren), feander@siu.edu (F.E. Anderson).

http://dx.doi.org/10.1016/j.ympev.2017.10.004
Received 8 February 2017; Received in revised form 26 September 2017; Accepted 4 October 2017
Available online 06 October 2017

taxon sampling incorporated publicly available data from six nuclear and four mitochondrial loci for over 400 OTUs to test hypotheses of convergent evolution and for correlation between morphology and habitat, providing new insight and support for some of the major subclades (Lindgren et al., 2012). At present (see Allcock et al., 2014 for a summary), clades that have been largely robust to differences in taxon sampling, data and/or phylogenetic method include Octopodiformes (all octopods and *Vampyroteuthis infernalis*), Incirrata (all octopods lacking fins), Cirrata (finned octopods) and Decapodiformes (open-eye squids, bobtail squids, pygmy squids, loliginids, cuttlefishes and *Spirula spirula*, the ram's horn squid).

The problem of poor support and/or inconsistent resolution is best exemplified in Decapodiformes, the major clade containing the orders Oegopsida (most oceanic "open eye" squids), Bathyteuthoidea (comb-finned squids and their relatives), Idiosepiida (pygmy squids), Sepiida (cuttlefishes), Sepiolida (bobtail squids), Spirulida (*Spirula spirula*) and Myopsida (comprising Loliginidae—a family of mostly large-bodied, muscular, neritic squid, many of major fisheries importance—and Australiteuthidae, a poorly known group of small squid; Lu, 2005). Little progress on resolving relationships among these lineages has been made since the morphological research of Naef (1923). He proposed that extant Decapodiformes should be subdivided into two groups: Sepioidea (containing Idiosepiida, Sepiida, Sepiolida and Spirulida) and Teuthoidea (Myopsida and Oegopsida). However, Naef struggled with the position of Myopsida, due to shared characteristics with both Sepioidea and Oegopsida. Berthold and Engeser (1987) partially supported Naef's hypothesis, but suggested that Spirulida was a sister taxon to sepioids + loliginids, a group they termed "Uniductia." More recently, no molecular study to date has found support for Sepioidea or Teuthoidea *sensu* Naef, and the position of Myopsida varies significantly depending on analytical method, data and taxon sampling (Allcock et al., 2014). In general, decapodiform relationships vary with differences in taxon sampling, type of genetic data used and analytical method employed (e.g., Carlini and Graves, 1999; Lindgren, 2010; Lindgren et al., 2012; Strugnell et al., 2005; Strugnell and Nishiguchi, 2007) and the issue of how the sepioids and loliginids are related to each other and to Oegopsida remains contentious (Allcock et al., 2014).

Next-generation sequencing (NGS) techniques have shown a high degree of success in phylogeny estimation for a variety of taxonomic groups, including mollusks (Kocot et al., 2011; Smith et al., 2011). Some genome/transcriptome-scale studies have included representatives of multiple cephalopod lineages (Kocot et al., 2011; Smith et al., 2011), but these studies were focused on molluscan phylogeny and lacked representatives of major cephalopod lineages (e.g., Oegopsida, Sepiida and Vampyromorpha). Similarly, Albertin et al. (2015) used genome and transcriptome data to infer the phylogenetic position of *Octopus bimaculoides* within Mollusca, but key questions in coleoid cephalopod phylogeny could not be addressed because Nautiloidea, Oegopsida and Vampyromorpha were not sampled. Recently, a study by Strugnell et al. (2017) utilized mitochondrial genome data from two new taxa, *Spirula spirula* and *Sepiadarium austrinum*, to evaluate higher-level relationships, finding support for a close association between Spirulida, Bathyteuthoidea and Oegopsida (a finding also supported by a multigene phylogeny; Lindgren et al., 2012) and provided new hypotheses regarding the placement of Idiosepiida, Sepiida, Sepiolida and Myopsida (Strugnell et al., 2017). Another recent phylogenetic study that included all 39 previously published cephalopod mitochondrial genomes, plus data for four to five mitochondrial protein-coding genes from *Spirula* and four octopods (including the cirrate *Opisthoteuthis massyae*) (Uribe and Zardoya, 2017) was published shortly after Strugnell et al. (2017). Finally, while the present paper was under review, a study by Tanner et al. (2017) was published in which the authors tested hypotheses of cephalopod relationships using a combination of NGS data, including a transcriptome for the cirrate octopod *Grimpoteuthis glacialis* (now *Cirroctopus glacialis*—Collins and Villanueva, 2006; O'Shea, 1999—but we retain the usage of Tanner et al. for clarity) and a small amount of shotgun genome sequence data for *Spirula spirula*.

The present study aims to incorporate new and published transcriptome data to test the sensitivity and utility of large-scale datasets for inferring relationships among cephalopod lineages, particularly within Decapodiformes. Here, we evaluate the utility of NGS data for cephalopod phylogeny by generating datasets using a new cephalopod core ortholog assignment pipeline. Additionally, we employed several filtering steps and analytical approaches to assess the sensitivity of transcriptome data to impacts such as missing data and compositional and rate heterogeneity artifacts.

2. Methods

2.1. Taxon sampling

For our initial analyses, all publicly available cephalopod transcriptome data as of 9 February 2016 (47 total) were downloaded from the Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra/) and the GenBank EST database (https://www.ncbi.nlm.nih.gov/nucest) as fastq or fasta files. Predicted proteins from the *Octopus bimaculoides* genome (Albertin et al., 2015) were also downloaded from https://www.ncbi.nlm.nih.gov/genome/41501.

Novel transcriptome data were also generated for representatives of *Doryteuthis opalescens*, *Octopus bimaculoides*, *Vampyroteuthis infernalis* and *Todarodes pacificus* using 454 pyrosequencing methods (Table 1). For these species, total RNA was extracted using NucleoSpin® RNA XS Isolation kit from RNA*later*® preserved samples. This RNA extraction method was employed specifically to remove any residual pigment from the RNA that could inhibit downstream cDNA synthesis. To reduce the impact of contamination, all RNA was extracted separately for each species at an 'RNA-only' workstation that was cleaned between each extraction. RNA quantity was assessed using a Qubit fluorometer (Invitrogen, Inc.). First strand cDNA was synthesized via the SMARTer® cDNA Library Construction Kit (Clontech Laboratories, Inc.), using the following primer: 5′-AAG CAG TGG TAT CAA CGC AGA GTA CTT TTT TCT TTT TT-3′. We performed second-strand synthesis using the SMARTer cDNA protocol, with 18–22 cycles depending on initial RNA concentration. Successfully amplified cDNA was then cleaned using a phenol:chloroform:isoamyl protocol and quantified with a Qubit fluorometer. Two to 5 ug of cleaned, non-normalized cDNA was shipped to Brigham Young University for subsequent library preparation and titanium pyrosequencing on the Roche 454 platform. Approximately one-sixth of a lane was used for each cDNA, with individual samples barcoded.

2.2. Transcriptome assembly and processing

Transcriptome assemblies were generated from downloaded fastq/fasta files or raw 454 reads using Trinity (April 13, 2014 release) (Grabherr et al., 2011) with default parameters for single read data and *in silico* normalization for particularly large transcriptomes (using the "–normalize reads" option). Assemblies were produced in house at Portland State University (PSU) for the newly generated data, and at Southern Illinois University (SIU) or via the National Center for Genome Assembly Support (NCGAS) using the Mason cluster at Indiana University for the data downloaded from the SRA and EST databases. We used TransDecoder (http://transdecoder.github.io) to find open reading frames and translate nucleotide sequences into amino acid sequences at least 100 amino acids in length.

2.3. Dataset construction

Following Garrison et al. (2016), we produced a custom cephalopod core ortholog set using OrthoMCL 2.0 (Li et al., 2003). To generate the cephalopod core ortholog set, we used all-versus-all BLASTP (Altschul

Table 1
SRA project number, number of Trinity contigs and number of HaMStR orthologous groups (recovered using the cephalopod core ortholog set) for all transcriptomes used in this study. ARL = collected by the first author for this study; ^ = used in construction of the cephalopod core ortholog set; * = Tanner et al. (2017) used data from one run in a set of runs generated for different tissue samples in one BioProject, whereas we pooled data from all runs; ** = used only in the "all" and "best2" analyses (where it replaced *Doryteuthis pealei* 1).

Taxon	SRA Run Number	# Contigs	# HaMStR Orthologs (cephalopod core ortholog set)	All	Best1	Tanner et al.	Best1 + *Grimpoteuthis* + *Spirula*	Combined
Abdopus aculeatus	SRR680047	987	60			✓		
Architeuthis dux	Unpublished as of 10/10/2017	–	2062			✓		✓
Bathypolypus arcticus	Unpublished as of 10/10/2017	12559	1581			✓		✓
Chiroteuthis calyx	SRR2102319	11304	470			✓		✓
Doryteuthis opalescens (ARL)	SRR6150357	14895	661	✓				
Doryteuthis pealei ^	SRR1725163, SRR1725164, SRR1725167, SRR1725169, SRR1725171, SRR1725172, SRR1725213, SRR1725235, SRR1725236	431857	2175	✓	✓		✓	✓
Doryteuthis pealei 2**	SRR826777, SRR826778, SRR826780	374601	1967	✓				
Doryteuthis pealei 3	SRR824660, SRR824682,	14065	126	✓				
Doryteuthis pealei 4	LIBEST_027407	22033	479	✓				
Doryteuthis pealei 5	SRR3472304	59820	1960			✓		
Dosidicus gigas	SRR1386212*, SRR1955488	75623	1880	✓	✓	✓*	✓	✓
Euprymna scolopes 1^	SRR871362, SRR871363, SRR871364, SRR871365, SRR871366	69173	2057	✓	✓		✓	✓
Euprymna scolopes 2	SRR1460580	3834	132	✓				
Euprymna scolopes 3	SRR2102326	10154	549	✓				
Euprymna scolopes 4	LIBEST_018865 – LIBEST_018874	35420	662	✓				
Euprymna scolopes 5	SRR3472306		1930			✓		
Galiteuthis armata	SRR2102359	12549	624	✓	✓	✓	✓	✓
Grimpoteuthis glacialis	Unpublished as of 10/10/2017		1930			✓	✓	✓
Hapalochlaena maculosa 2	SRR3105559	82444	2030			✓		✓
Heterololigo bleekeri	DRR018274, DRR018275	223955	1969	✓	✓		✓	✓
Idiosepius notoides	SRR2984343	72961	1887			✓		✓
Idiosepius paradoxus	LIBEST_020620	9079	248	✓	✓		✓	
Lolliguncula brevis	Unpublished as of 10/10/2017	12544	1630			✓		✓
Nautilus pompilius 1^	SRR330442	8682	749	✓	✓		✓	✓
Nautilus pompilius 2	SRR027037, SRR027038, SRR027039, SRR027040, SRR027041, SRR027042, SRR108979*	16149	284	✓		✓*		
Octopus bimaculoides (ARL)	SRR6150355	30678	1163	✓				
Octopus bimaculoides 1^	SRR2045866, SRR2045870, SRR2047107, SRR2047109, SRR2047111, SRR2047114, SRR2047116, SRR2047118, SRR2047120, SRR2047122, SRR2048495, SRR2048496, SRR2048497, SRR2048498, SRR2048521, SRR2048522, SRR2048523, SRR2048524, SRR2048525	494729	2185 (Trinity assembly); 2179 (proteome)	✓	✓		✓	✓
Octopus bimaculoides 2	SRR2102364	11526	665	✓				
Octopus cyanea	SRR725937	1375	83			✓		
Octopus vulgaris 1	SRR1507221, SRR1507224	164502	1854	✓	✓		✓	
Octopus vulgaris 2	SRR331946	121092	1718	✓		✓		✓
Octopus vulgaris 3	SRR026776, SRR026777, SRR026778, SRR026779, SRR026780, SRR026781	26878	494	✓				
Onychoteuthis banksii	Unpublished as of 10/10/2017	21104	1960			✓		✓
Pareledone albimaculata	Unpublished as of 10/10/2017	17293	1901			✓		✓
Sepia esculenta	SRR1281998, SRR128310, SRR1386223	288101	2158	✓	✓	✓	✓	✓
Sepia officinalis 1	LIBEST_027716	43625	1107	✓	✓		✓	
Sepia officinalis 2	SRR1325115	34297	1585			✓		✓
Sepiella japonica^	SRR2889752, SRR2889753 SRR2891123, SRR2891216	276570	2164	✓	✓		✓	✓
Sepioteuthis australis	SRR725780	1090	80			✓		
Sepioteuthis lessoniana	SRR1386192	352345	1488	✓	✓	✓	✓	✓
Spirula spirula	Unpublished as of 10/10/2017	223686	65			✓	✓	
Sthenoteuthis oualaniensis	Unpublished as of 10/10/2017	7846	1058			✓		✓
Todarodes pacificus (ARL)	SRR6150358	13946	631	✓	✓		✓	✓
Uroteuthis edulis 1	SRR2102378	6976	321	✓	✓		✓	
Uroteuthis edulis 2	DRR068682	121764	2024			✓		✓
Uroteuthis noctiluca	SRR725597	264	18			✓		
Vampyroteuthis infernalis (ARL)^	SRR6150356	60004	893	✓	✓		✓	✓

(*continued on next page*)

Table 1 (*continued*)

Taxon	SRA Run Number	# Contigs	# HaMStR Orthologs (cephalopod core ortholog set)	All	Best1	Tanner et al.	Best1 + *Grimpoteuthis* + *Spirula*	Combined
Vampyroteuthis infernalis 2	SRR2102472	824	29			✓		
Watasenia scintillans^	SRR2960126, SRR2960127, SRR2960128, SRR2960129, SRR2960130, SRR2960131	263122	2165	✓	✓		✓	✓

et al., 1990) on the Mason cluster to compare six of our Trinity assemblies—*Doryteuthis pealei* (a.k.a. *D. pealeii*; Aldrich, 1990), *Euprymna scolopes*, *Nautilus pompilius*, *Sepiella japonica, Vampyroteuthis infernalis* and *Watasenia scintillans* (Table 1)—and the predicted transcripts from the *Octopus bimaculoides* genome. The six Trinity assemblies chosen were the largest assemblies we had access to that represented all major extant cephalopod lineages except *Idiosepius* (our initial assembly of publicly available data for this taxon was quite small; Table 1). An e-value cut-off of 10^{-3} was used. Based on the BLASTP results, we conducted Markov clustering using OrthoMCL 2.0 (closely following the protocol described in the OrthoMCL User Guide – http://orthomcl.org/common/downloads/software/v2.0/UserGuide.txt) to yield a set of putatively orthologous groups. Each OG was then processed through a modified version of a pipeline as implemented in Whelan et al. (2015). We eliminated all OGs that were shorter than 100 AA long; each remaining OG was aligned with MAFFT (L-INS-i) (Katoh et al., 2005). We then inferred an approximate maximum likelihood tree (ML) for each OG using FastTreeMP (under the –slow and –gamma settings) (Price et al., 2010) and used PhyloTreePruner (Kocot et al., 2013) to screen each of the resulting trees. In PhyloTreePruner, all nodes with SH (Shimodaira-Hasegawa)-like (Price et al., 2010) local support values < 0.95 on each tree were collapsed into polytomies, and we retained the largest subtree where each taxon was either not represented or was represented by only one sequence, unless *all* sequences for a given taxon formed part of a clade or part of the same polytomy (in which case, all were kept). Any sequence falling outside the maximally inclusive subtree was assumed to be a paralog and was deleted. If multiple in-paralogs were initially retained, only the longest sequence was retained. This returned an alignment for each OG that included (at most) a single, putatively orthologous sequence for each taxon. We used alignments for OGs that included sequences from at least five of the seven reference taxa to build profile hidden Markov models (pHMMs) using hmmbuild and hmmcalibrate in the HMMER package (Eddy et al., 2011). This process yielded a total of 2355 cephalopod core ortholog pHMMs.

All transcriptomes were screened with the pHMMs in HaMStR v.13.2.3 (Ebersberger et al., 2009). We set HaMStR to output all sequences that fulfilled the reciprocity requirement. At this point, we chose to remove several transcriptomes for which HaMStR returned small numbers of orthologs (Table 1). This left us with 30 transcriptome assemblies representing 18 species; no major cephalopod lineages were lost, and five species (*Doryteuthis pealei*, *Euprymna scolopes*, *Nautilus pompilius*, *Octopus bimaculoides*, *Octopus vulgaris*) were still represented by more than one transcriptome. For initial analyses, we created an "all-in" set of orthogroups (OGs) that retained all 30 transcriptomes (hereafter referred to simply as "all"). To better assess the impact multiple transcripts may have on our phylogenetic analyses, we produced two additional sets of OGs in which only the "best" transcriptome was retained for each of the five species listed above. "Best" was determined in two ways—the number of HaMStR core orthologs present (i.e., minimizing missing data at the level of OG) and the total number of amino acids recovered across all OGs (minimizing missing data at the level of individual characters). The same transcriptomes were found to be the "best" in all cases except for *D. pealei* (Table 1), leading us to create a "best1" OG set (based on number of core orthologs present) and a "best2" OG set (based on total number of amino acids recovered). Both "best" datasets included an initial set of 18 transcriptomes representing most major living cephalopod groups—Nautiloidea, Octopodiformes (Octopoda, Vampyromorpha) and Decapodiformes (Idiosepiida, Myopsida, Sepiida and Sepiolida, Oegopsida: families Cranchiidae, Enoploteuthidae and Ommastrephidae, Table 1).

For each initial set of assemblies, we employed a custom script to produce FASTA-formatted files for each OG that included all sequences while removing any duplicated contigs. Each OG was then aligned with MAFFT (L-INS-i) (Katoh et al., 2005). Failure to distinguish orthologs from paralogs can cause errors in phylogenetic inference (Struck, 2013). To filter possible paralogs from our data, we followed a tree-based procedure similar to that described above, in which we inferred an approximate ML tree for each aligned OG with FastTreeMP (Price et al., 2010) (–slow and –gamma settings), and used PhyloTreePruner to screen the resulting trees. In this case, nodes on each ML tree with SH-like local support values < 0.7 were collapsed into polytomies with PhyloTreePruner, with sequences retained as described above. We used PhyloTreePruner to retain only OGs found in at least ~25% (8 taxa for the "all" dataset, 5 for the "best" datasets), ~50% (15 taxa for all, 9 for best), and ~75% (23 taxa for all, 14 for best) of the transcriptomes. Following pruning, all loci were subsequently realigned with MAFFT (L-INS-i), and we used FASconCAT v1.0.pl (Kück and Meusemann, 2010) to concatenate OGs into 25%, 50% and 75% data matrices. Only one locus was retained for the all 75% matrix, so we did not perform downstream analyses on this matrix.

2.4. Confounding factors

Unfortunately, the immense amount of data generated through high-throughput sequencing is not a panacea for the many issues known to impact phylogenetic analysis, including heterogeneity among lineages in terms of substitution rate and nucleotide/amino acid composition (Felsenstein, 1978; Foster and Hickey, 1999; Hendy and Penny, 1989; Saccone et al., 1990). To explore the impact of these issues on our inferences, we determined best-fitting substitution models for each OG using the ProteinModelSelection.sh script (https://github.com/stamatak/standard RAxML/blob/master/usefulScripts/ProteinModelSelection.pl), then inferred ML trees for each OG under the appropriate model using RAxML version 8.0.23 (Stamatakis, 2014) with 100 rapid bootstrap replicates. We then used TreSpEx.v1.1 (Struck, 2014) to estimate three measures of substitution rate (branch length) heterogeneity for every OG in all of our datasets—(1) the standard deviation of the tip-to-root distance, (2) the average patristic distance (PD) and (3) the LB score (the mean pairwise patristic distance of a taxon to all other taxa in the tree relative to the average pairwise patristic distance across all taxa (Struck, 2014)). These indices were extracted from the LB_scores_summary_perPartition.txt TreSpEx output file; OGs that returned a value equal to or greater than 1.5 times the interquartile range above the median for any of these three indices were removed, and the remaining loci were concatenated with FASconCAT. The TreSpEx-filtered concatenated matrix was then evaluated with BaCoCa v. 1.104r (Kück and Struck, 2014). Data partitions (i.e., OGs) with a p value of less than 0.05 for a chi square test of homogeneity were removed, as were all OGs that were 1.5 times the interquartile range above the

median RCFV value (Zhong et al., 2011). These combined filtering steps were conducted using custom scripts (available in the Mendeley Data package); they should remove all outlier OGs showing very high levels of rate and/or compositional heterogeneity among lineages.

Two other approaches were also used to ameliorate the impact of problematic loci on our inferences. First, as an additional method to reduce the impact of compositional heterogeneity, we recoded amino acids into groups that minimize compositional heterogeneity according to a chi square test of homogeneity using the software package minmaxchisq (http://www.mathstat.dal.ca/tsusko/software.cgi; Susko and Roger, 2007). To retain as much data as possible while minimizing the effect of compositional heterogeneity, we used the maximum number of bins selected by minmax-chisq that yielded a chi square test p-value above 0.05. All recoded data matrices were analyzed on CIPRES (http://www.phylo.org; Miller et al., 2010) under a MULTICAT GTR model, with all free parameters estimated and search parameters as described below (Section 2.6). Second, we employed the MARE v. 0.1.2 software package (Meyer et al., 2011) to reduce the number of OGs in some of our matrices. MARE reduces the size of data matrices by eliminating loci and/or taxa of low information content, a measure based on the "tree likeness" of each OG/taxon estimated using extended geometry mapping (Nieselt-Struwe and von Haeseler, 2001). We used default settings, but employed the –c option to retain all taxa in each matrix, meaning that only OGs were deleted to produce the MARE-reduced matrices.

2.5. "Question-specific" loci

For lineages in which rapid radiation has occurred, reconstructing evolutionary relationships can be challenging, even in the presence of large amounts of data, because insufficient time has passed for large numbers of synapomorphies to accumulate or for genes to coalesce (e.g., Wiens et al., 2008). Furthermore, in cases of ancient rapid radiations, the lengthy time since divergence could increase homoplastic changes that can be erroneously interpreted as phylogenetic "signal" (Whitfield and Lockhart, 2007). For lineages where particularly recalcitrant problems exist, filtering data may only further muddy the phylogenetic waters by supporting conflicting, but often strongly supported, hypotheses (Rokas and Carroll, 2006; Salichos and Rokas, 2013). Recently, other methods to increase accuracy in phylogenomic analyses have been proposed, including limiting analyses to loci that show clock-like behavior or good substitution model fit in posterior predictive simulations (Doyle et al., 2015) or that support uncontroversial clades (Chen et al., 2015). Several uncontroversial coleoid subclades are represented by two or more transcriptomes in our datasets, and restricting analysis to loci that support these groups may shed light on other aspects of cephalopod phylogeny. To find these loci in our sets of OGs, we first deleted any OGs that had one or fewer representatives of Decapodiformes, Octopoda, Myopsida (represented by the family Loliginidae) or Sepiida. For the remaining OGs, we filtered all ML topologies using a topological constraint in PAUP* (Swofford, 2002), only retaining OGs whose ML topologies contained monophyletic Decapodiformes, Octopoda, Loliginidae and Sepiida, as these groups have been previously strongly supported as monophyletic (e.g., Lindgren et al., 2012; Allcock et al., 2014) where represented by multiple species. The OGs whose ML topologies satisfied this constraint were concatenated and analyzed with RAxML as described above; we refer to the resulting matrices as "node control" or "DOLS" matrices (for **D**ecapodiformes, **O**ctopoda, **L**oliginidae and **S**epiida). To maximize the number of loci recovered, we performed this filtering initially only for the largest set of OGs (25%) for each group of transcriptomes (all, best1 and best2).

All data matrices, tree files and scripts, as well as the cephalopod core ortholog set developed for this study, are available via Mendeley Data.

2.6. Maximum likelihood (ML) analyses

We used the ProteinModelSelection.pl script to select best-fitting amino acid substitution models for each OG for each dataset (all, best1 and best2), level of completeness (25%, 50% and 75%) and filtering protocol (TreSpEx/BaCoCa, MARE, etc.). Partitioned maximum likelihood (ML) analyses were conducted with RAxML versions 8.2.8 and 8.2.9 (Stamatakis, 2014) on CIPRES (Miller et al., 2010) with 100 rapid bootstrap pseudoreplicates, using these options: -f a -x < random number seed for rapid bootstrapping; unique for each analysis > -p < random number seed for initial parsimony inferences; unique for each analysis > -# 100 -m PROTGAMMA < amino acid model > -s < inputfile > -n < outputfile > -q < partitionfile > .

2.7. Bayesian analyses

Bayesian analyses were conducted on the all 50% and best1 and best 2 75% unfiltered data matrices under the CAT-GTR model with gamma-distributed rates in PhyloBayes-MPI v. 1.5a (Lartillot et al., 2013, 2009) on CIPRES. Two independent chains were run per analysis, saving points every ten cycles. Analyses were allowed to run for up to 168 hours (the CIPRES limit), constant sites were removed (-dc option), and four categories were used for the discrete gamma distribution (-dgam = 4). Convergence checks were conducted every 1800 seconds and analyses were automatically terminated early by CIPRES if after a burn-in of 500 cycles, the minimum effective sample size exceeded 50, and the "maxdiff" value between chains was less than 0.1 (denoting a "good run", according to the PhyloBayes MPI manual). We assumed that runs that were not automatically terminated before reaching the 168-hour limit did not converge.

2.8. Recent developments

While this manuscript was under review, a phylogenomic study of cephalopod lineages using transcriptome data was published (Tanner et al. 2017). In their study Tanner et al. used some of the same publicly available transcriptome data used in this study, but also used alternate smaller, published transcriptomes (not incorporated in this study) and incorporated new data, most notably from *Spirula spirula*, *Architeuthis dux* (the giant squid) and a cirrate octopus, *Grimpoteuthis glacialis*. Their data processing pipeline also differed from ours—rather than using HaMStR to recover putative orthologs from transcriptome assemblies, they used BLAST searches to extract 197 genes used in a previous study investigating higher-level relationships of Metazoa (Philippe et al., 2011).

To explore the impact of the differences between transcriptome data used and the data processing pipeline methods that were used in these two studies, we constructed and analyzed three additional datasets that incorporated data from Tanner et al. (2017). First, to evaluate whether our pipeline was able to recover the topology published by Tanner et al. we attempted to replicate their dataset following data sources listed in their Supplementary Table 3. For transcriptomes used in Tanner et al. (2017) but not previously incorporated into this study (Table 3), we assembled the raw sequence reads with Trinity under default parameters, matching Tanner et al. (R. Fonseca, *pers. comm.*). Amino acid data for new taxa including *Architeuthis dux*, *Onychoteuthis banksii* and *Sthenoteuthis oualaniensis* and assemblies for *Bathypolypus arcticus*, *Dosidicus gigas*, *Grimpoteuthis glacialis*, *Pareledone* sp. and *Spirula spirula* (the latter a shotgun genome sequence) were provided by R. Fonseca and A. Tanner. Second, to assess the impact that including key missing lineages would have on inferences based on our initial data, we incorporated data for *Grimpoteuthis glacialis* and *Spirula spirula* into our initial dataset, adding these data to the start of our bioinformatics pipeline and processing them as described above. To save computational effort, we only added the *Grimpoteuthis* and *Spirula* data to our best1 dataset. Finally, to estimate relationships using the largest, most data-

rich transcriptomes currently available, we produced a novel dataset in which we combined the best transcriptomes from our initial best1 dataset with the best transcriptomes used by Tanner et al. to produce a "combined" dataset (Table 1). As above, we defined the "best" transcriptomes as those that maximized the number of OGs returned after processing with HaMStR using our cephalopod core ortholog set, retaining only transcriptomes that returned data for at least 400 OGs. Lastly, due to its small size, we also replaced *Idiopsepius paradoxus* with *Idiosepius notoides*—both were not included because the *I. paradoxus* assembly returned a small number of OGs (2 4 8) compared to all others (> 400). All three datasets were processed and analyzed as described above (i.e., our cephalopod core ortholog set was used in HaMStR and 25%, 50% and 75% matrices were produced, followed by filtering of OGs with TreSpEx and BaCoCa, MARE, etc., with subsequent partitioned ML analyses in RAxML).

Finally, to assess whether any of the relationships we recovered were due to use of our cephalopod core ortholog set, and to assess the robustness of our conclusions based on analyses of the combined dataset, we also produced a set of OGs for the combined datasets using the prepackaged lophotrochozoan core ortholog set in HaMStR, followed by partitioned RAxML analyses of unfiltered and filtered (with TreSpEx and BaCoCa) 25%, 50% and 75% data matrices following the protocols described above (see Supplementary Material).

3. Results

3.1. Transcriptomes, datasets and matrices

Novel transcriptomes from four cephalopods (Table 1) were used in this study; these data are available from the NCBI Sequence Read Archive under BioProject accession number SRP119608. The initial 30 transcriptomes incorporated into this study were highly variable in terms of number of contigs assembled by Trinity and the number of orthologs recovered by HaMStR (Table 1). Several transcriptomes contained contigs representing ~2000 orthologs in the cephalopod core ortholog set (e.g., *Doryteuthis pealei*, *Euprymna scolopes*, *Heterololigo bleekeri*, *Octopus bimaculoides*, *Sepia esculenta*, *Sepiella japonica* and *Watasenia scintillans*), while others were much smaller (< 100 orthologs) and were excluded from most subsequent analyses (Table 1). All transcriptomes used in the initial all, best1 and best2 data matrices contained > 600 orthologs in the cephalopod core ortholog set, with two exceptions: *Idiopsepius paradoxus* (248) and *Uroteuthis edulis* (321).

A total of eight datasets were initially produced based on the set of OGs recovered by HaMStR using the cephalopod core ortholog set—25%, 50% and 75% datasets for the best1 and best2 sets of transcriptomes and 25% and 50% datasets for the set of all transcriptomes (only one OG was retained in the all 75% dataset). Datasets were filtered or recoded in various ways to minimize the impact of hidden paralogy as well as among-lineage substitution rate and amino acid compositional heterogeneity. Descriptive statistics for each dataset are presented in Table 2.

3.2. Phylogenetic relationships

Maximum likelihood analyses of our initial all, best1 and best2 datasets yielded several uncontroversial relationships with high support, including a monophyletic Octopodiformes (cirrate and incirrate octopods plus *Vampyroteuthis infernalis*) and Decapodiformes (represented by all bobtail squids, pygmy squids, cuttlefishes and open-eye squids) (Fig. 1). However, within Decapodiformes, ordinal-level relationships fluctuated with changes in amount of missing data, coding strategy, and filtering method (Fig. 1). The majority of these conflicts involved the phylogenetic placement of representatives of two clades: the bobtail squid *Euprymna scolopes* (Sepiolida) and the pygmy squid *Idiosepius paradoxus* (Idiosepiida). In 19 of 25 topologies, *Euprymna* was sister to all other decapodiforms (Fig. 1, node 1), while in five of the

Table 2
Descriptive statistics for all data matrices used in this study. For each level of data completeness, the number of characters in the matrices produced using Susko and Roger recoding was identical to the original, unfiltered data.

Dataset	# Loci	# Characters	% Missing
All transcriptomes			
25% completeness			
Original data	1377	736,736	73.39
TreSpEx + BaCoCa filtered	1002	481,413	71.84
MARE	323	139,706	63.73
50% completeness			
Original data	281	123,520	63.07
TreSpEx + BaCoCa filtered	212	81,398	60.02
MARE	102	37,357	56.27
Best1 set			
25% completeness			
Original data	1846	979,595	64.14
TreSpEx + BaCoCa filtered	1386	674,371	62.41
MARE	622	297,744	55.67
50% completeness			
Original data	1099	545,717	57.62
TreSpEx + BaCoCa filtered	815	372,422	55.47
MARE	430	198,446	53.07
75% completeness			
Original data	73	28,555	41.77
TreSpEx + BaCoCa filtered	56	18,708	36.05
MARE	48	17,854	39.76
Best2 set			
25% completeness			
Original data	1839	962,001	65.32
TreSpEx + BaCoCa filtered	1371	654,405	63.08
MARE	612	283,892	55.49
50% completeness			
Original data	1061	517,238	58.34
TreSpEx + BaCoCa filtered	784	350,562	55.78
MARE	415	188,271	52.89
75% completeness			
Original data	70	25,980	39.54
TreSpEx + BaCoCa filtered	48	17,872	37.60
MARE	45	16,316	40.02
Tanner et al. (2017)			
25% completeness			
Original data	1754	861,575	61.35
TreSpEx + BaCoCa filtered	1294	619,419	59.25
50% completeness			
Original data	1137	522,504	53.24
TreSpEx + BaCoCa filtered	862	394,758	51.97
75% completeness			
Original data	20	7267	33.85
TreSpEx + BaCoCa filtered	–	–	–
Best1 + *Grimpoteuthis* + *Spirula*			
25% completeness			
Original data	1905	1,021,380	64.89
TreSpEx + BaCoCa filtered	1422	700,560	
50% completeness			
Original data	1052	533,450	57.86
TreSpEx + BaCoCa filtered	779	362,912	
75% completeness			
Original data	70	26,720	41.97
TreSpEx + BaCoCa filtered	53	18,259	38.26
Combined (Cephalopod OGs)			
25% completeness			
Original data	1899	1,031,401	57.75
TreSpEx + BaCoCa filtered	1433	719,955	54.79
50% completeness			
Original data	1315	698,412	49.78
TreSpEx + BaCoCa filtered	977	489,072	47.05
75% completeness			
Original data	557	243,244	36.04
TreSpEx + BaCoCa filtered	415	177,893	34.54

(*continued on next page*)

Table 2 (*continued*)

Dataset	# Loci	# Characters	% Missing
MARE filtered	343	143,227	33.51
Node control ("DOLS")	234	103,266	37.20
Combined (Lophotrochozoan OGs)			
25% completeness			
Original data	868	470,484	58.26
TreSpEx + BaCoCa filtered	675	352,368	56.78
50% completeness			
Original data	416	218,283	43.45
TreSpEx + BaCoCa filtered	319	165,013	42.46
75% completeness			
Original data	55	23,898	28.52
TreSpEx + BaCoCa filtered	41	18,326	28.20

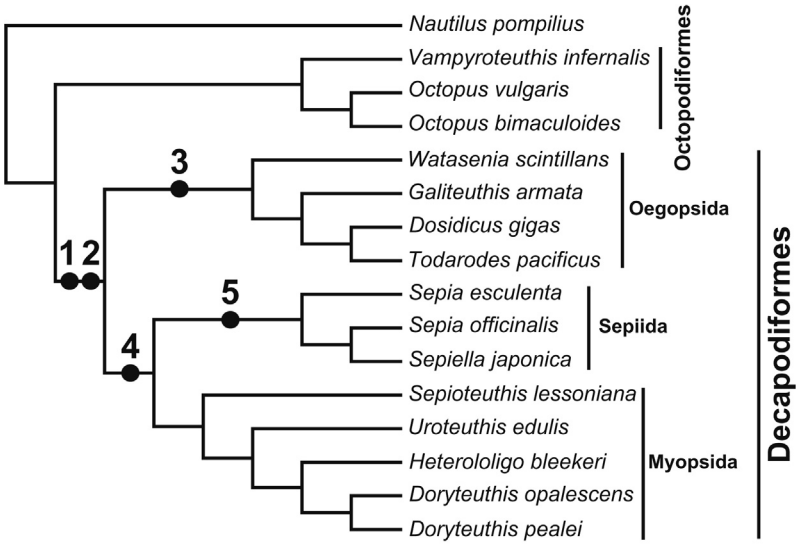
remaining six topologies *Euprymna* was either sister to Sepiida + Myopsida or sister to *Idiosepius*, with this pair sister to Sepiida + Myopsida (Fig. 1, node 4). The position of *Idiosepius* also varied, with it falling as sister to all other decapodiforms (Fig. 1, node 1), sister to all other decapodiforms except *Euprymna* (node 2), sister to Oegopsida (node 3), sister to *Euprymna*, with this pair sister to Sepiida + Myopsida (node 4), sister to Sepiida + Myopsida (node 4) or sister to Sepiida (node 5). In general, bootstrap support for relationships among the decapodiform lineages was low to moderate (< 90%) across all analyses.

Of the three PhyloBayes analyses we attempted for our initial datasets, only two—the analyses of the "best" 75% data matrices—automatically terminated following convergence checks. The analysis of the all 50% matrix did not terminate, and we therefore assume that it failed to converge. In the best1 75% analysis, two chains ran for an average of 13,940 cycles; in the best2 analysis, two chains ran for an average of 10,855 cycles. A 500-cycle burn-in was used in each case, and visual inspection of the trace files in Tracer suggested that the runs had converged, though some ESS values were < 200. Trees produced with the 75% matrices included the same uncontroversial clades as in the maximum likelihood topologies described above. Within Decapodiformes, Myopsida (Loliginidae) was sister to a clade containing all other orders (see Supplementary Material for trees), although support values were low (PP 0.89).

3.3. Additional analyses incorporating recently published data

Several of the transcriptomes we used to construct our initial data matrix were also incorporated in the Tanner et al. (2017) data matrix, but they used six transcriptomes that we excluded from our initial analyses due to their small size (*Abdopus aculeatus*, *Chiroteuthis calyx*, *Octopus cyanea*, *Sepioteuthis australis*, *Uroteuthis noctiluca* and *Vampyroteuthis infernalis*; Table 1). In four cases (*Doryteuthis pealei*, *Euprymna scolopes*, *Nautilus pompilius* and *Octopus vulgaris*), Tanner et al. sampled the same taxa as we did in our initial analyses, but used *different* publicly available transcriptomes. We also included a new larger transcriptome for *V. infernalis* and publicly available data for *Octopus bimaculoides* and *Sepiella japonica* that Tanner et al. (2017) did not use. Finally, five transcriptomes used by Tanner et al. were released to the public after we had begun our initial analyses in 2016 (*Doryteuthis pealei*, *Hapalochlaena maculosa*, *Lolliguncula brevis*, *Sepia officinalis* and *Uroteuthis edulis*; Table 1), though we subsequently included all but the new (smaller) *Doryteuthis pealei* transcriptome in our combined analyses.

Analyses of our version of the Tanner et al. (2017) dataset (hereafter referred to as "tanner17") yielded an array of topologies, several of which were quite different from those published by Tanner et al. (see Supplementary Material for trees). Unfiltered analyses of the 25%, 50% and 75% tanner17 datasets were particularly unstable, while filtering the matrices using TreSpEx and BaCoCa yielded topologies that were more stable across the three levels of missing data. However, none of the topologies we produced using the tanner17 data matrices matched the topologies presented in Tanner et al. (2017) with respect to Decapodiformes. Our tanner17 dataset was highly sensitive to the level of missing data and filtering method, although overall the topologies we recovered with this dataset resembled those generated by the all, best1 and best2 data matrices (e.g., a basal position within Decapodiformes for *Euprymna* and *Idiosepius* and a sister relationship between Sepiida



Analysis	Position of *Euprymna*	Position of *Idiosepius*
2-50F, 2-50S, 2-25F, 2-25M*	1	2
2-25T	2	1
1-75F, 1-75T, 1-75S, 1-50S, 1-50M 2-75F, 2-75S, 2-75M, 2-50M	1	3
1-25D	4	5
1-50F, 1-50T, 1-25F, 1-25T, 1-25M*	1	5
1-25S, 2-75T, 2-25S, 2-25D	4	4
2-50T	1	4

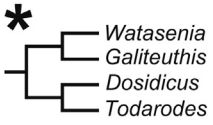


Fig. 1. Backbone tree of all unique maximum likelihood topologies inferred for all best1 and best2 data matrices. Numbers on branches and in the table depict inferred positions of *Euprymna scolopes* and *Idiosepius paradoxus* across analyses. Analyses are coded as follows: 1 = best1, 2 = best2; 25 = 25%, 50 = 50%, 75 = 75%; D = data matrix filtered to retain only loci that return monophyly of four well-supported clades (DOLS); F = full (unfiltered) data matrix, partitioned RAxML; M = MARE-filtered data matrix; S = data matrix recoded using the Susko and Roger (2007) method, MULTIGAMMA RAxML; T = TreSpEx-and-BaCoCa-filtered data matrix, partitioned RAxML (see text for details). For example, "1-25M" is the MARE-filtered best1 25% dataset. * = alternative topology recovered within Oegopsida. ML analysis of the 1-75M matrix returned a unique topology: (*Euprymna*(*Idiosepius*(Sepiida(Oegopsida(Myopsida))))).

and Myopsida; Supplementary Material).

The addition of key lineages Spirulida (*Spirula spirula*) and Cirrata (*Grimpoteuthis glacialis*) to our best1 dataset caused significant destabilization, primarily due to the variable placement of *Spirula spirula* across analyses (see Supplementary Material for trees). Analyses of the dataset with the highest number of OGs and the highest amount of missing data (25%) recovered *Spirula* as sister to Octopodiformes (which is clearly incorrect), whereas analyses of the 50% dataset recovered *Spirula* as sister to Sepiidae. The instability of *Spirula* was likely due to the very small number (65) of OGs recovered using the cephalopod core ortholog set.

By contrast, analyses of the combined dataset (Table 1), in which we integrated the best transcriptomes from our initial data matrix with the best transcriptomes used by Tanner et al. (2017), yielded topologies that were fairly robust to changes in the amount of missing data, coding strategy and which core ortholog set was used in HaMStR (Fig. 2). However, it is important to note that the *Spirula spirula* data were excluded from the combined dataset. Traditional relationships recovered in our initial analyses and Tanner et al. (2017), such as monophyly of Decapodiformes and Octopodiformes, were also recovered with high support values in the combined analyses. Additionally, within Octopodiformes, Incirrata was monophyletic, with incirrate octopods comprising two clades, the first including the two *Octopus* spp. and *Hapalochlaena,* with this clade sister to a *Bathypolypus* + *Pareledone* clade. Each of the major decapodiform groups for which multiple transcriptomes were available—Myopsida (Loliginidae), Oegopsida and Sepiida—were also well supported as monophyletic (100% bootstrap support; Fig. 2).

Within Decapodiformes, the majority of analyses recovered consistent relationships at the ordinal level with 100% bootstrap support for many nodes: Idiosepiida (*Idiosepius notoides*) fell as sister to the remaining decapodiforms, followed by Sepiolida (*Euprymna scolopes*), with cuttlefishes (Sepiidae) and Loliginidae forming a monophyletic group that was sister to Oegopsida (Fig. 2), with 100% bootstrap support values in all cases except for 75% DOLS, where it dropped to 88%. The majority of relationships within individual decapodiform clades were also consistent across the combined analyses, including monophyly for all families where multiple representatives were present (Ommastrephidae, Sepiidae, Loliginidae), although paraphyly was recovered for the cuttlefish genus *Sepia* (with *Sepia officinalis* more closely related to *Sepiella japonica* than to *Sepia esculenta*) as seen in recent mitochondrial genome studies (Strugnell et al., 2017; Uribe and Zardoya, 2017). Relationships among genera within Loliginidae (all supported with 100% bootstrap support across analyses) are concordant with previous Sanger-sequencing studies based on mitochondrial data (Anderson, 2000a; Anderson et al., 2014), combined mitochondrial and morphological data (Anderson, 2000b), and combined mitochondrial and nuclear data (Sales et al., 2013). Although the order Oegopsida was monophyletic across all analyses, family-level relationships were somewhat variable (Fig. 2). The most common pattern was a close

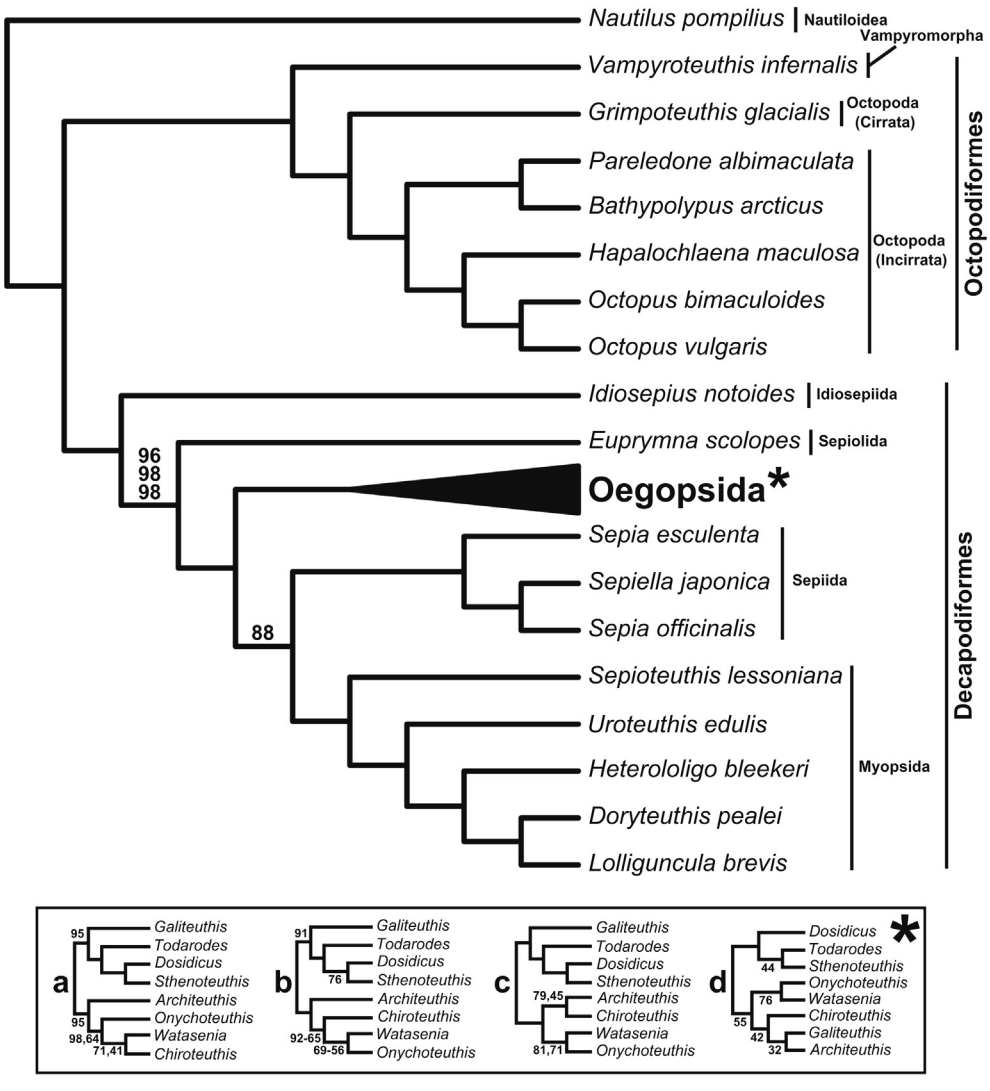


Fig. 2. Phylogenetic hypothesis reflecting all partitioned RAxML analyses of all combined data matrices (25%, 50% and 75% matrices across all filtering and recoding schemes; see text for details) generated using the cephalopod core ortholog set in HaMStR; all nodes outside Oegopsida have 100% ML bootstrap support except where noted (clade comprising all decapodiforms except *Idiosepius*: 96% BS support for 75% Susko and Roger recoded matrix, 98% BS support in 75% MARE-filtered and 75% DOLS-filtered matrix; Sepiida + Myopsida clade: 88% BS support for 75% DOLS-filtered matrix). Inset shows alternative resolutions and bootstrap values/ranges (all nodes without values have 100% bootstrap support across all analyses that returned that topology) within Oegopsida across analyses; a = 75% Susko and Roger recoded, 50% TreSpEx + BaCoCa filtered; b = 25% unfiltered, 25% TreSpEx + BaCoCa filtered, 50% unfiltered, 75% TreSpEx + BaCoCa filtered; c = 75% unfiltered, 75% DOLS filtered; d = 75% MARE filtered.

association between families Cranchiidae and Ommastrephidae, with this clade sister to a second clade that included representatives from families Architeuthidae, Chiroteuthidae, Enoploteuthidae and Onychoteuthidae. Overall, relationships recovered within Oegopsida are fairly stable for the larger data matrices; the greatest degree of variation was observed within the Architeuthidae-Chiroteuthidae-Onychoteuthidae-Enoploteuthidae clade, where three different sets of relationships were recovered across all analyses (Fig. 2).

With regard to inter-ordinal relationships, analyses of combined data matrices constructed using the lophotrochozoan core ortholog set corroborated those based on the cephalopod core ortholog set described above, though often with lower bootstrap support (Supplementary Material). There were a few exceptions to this general pattern. In two cases (the 75% unfiltered and 75% TreSpEx-and-BaCoCa filtered matrices), Idiosepiida was recovered as sister to a clade comprising Oegopsida, Sepiida and Myopsida (BS = 68% unfiltered, 48% filtered), with Sepiolida sister to the rest of Decapodiformes (i.e., *Idiosepius* and *Euprymna* swapped positions). The 50% unfiltered topology was nonsensical, placing *Vampyroteuthis* as sister to *Todarodes pacificus* (BS = 98%) and supporting a sister group relationship between Myopsida and Oegopsida (including *Vampyroteuthis*) (BS = 81%). This pattern was not seen in any other analysis in this study. Examination of single-gene trees suggested that a few OGs in this data set included sequences that were nearly identical for *Todarodes* and *Vampyroteuthis*, possibly due to bleed through during pyrosequencing. However, filtering the 50% data set with TreSpEx and BaCoCa removed most of these OGs and returned a topology identical to that seen in Fig. 2 (with relationships among oegopsids as shown in 2b).

4. Discussion

4.1. The utility of phylotranscriptomics for cephalopod phylogenetics

Regardless of the dataset construction method, the degree of missing data or filtering method, several uncontroversial relationships were consistently recovered in this study: Decapodiformes, Octopodiformes and all families for which multiple members were included all formed well-supported clades. Surprisingly, when we took several of the transcriptomes used in our best1 dataset and incorporated data for several additional species used by Tanner et al. and not previously available to us (*Grimpoteuthis glacialis, Pareledone albimaculata, Bathypolypus arcticus, Hapalochlaena maculosa, Sthenoteuthis oualaniensis, Architeuthis dux, Chiroteuthis calyx* and *Onychoteuthis banksii*) into a combined dataset, our topology became largely stable when subjected to various filtering strategies (Fig. 2), but differed from that presented by Tanner et al. (2017). Transcriptome data are clearly useful for phylogenetic inference within Cephalopoda, but the shift in stability for trees generated from our initial to our combined datasets indicates that large datasets produced by transcriptome data remain susceptible to issues related to missing data and taxon sampling.

4.2. (In)stability of ordinal relationships within Decapodiformes

Historically, hypotheses of evolutionary relationships among the decapodiform orders have varied greatly, depending on the type of analysis and the data used, from morphology (e.g., Naef, 1923; Young and Vecchione, 1996), to fossil data (e.g., Berthold and Engeser, 1987), to combined morphology and molecular data (Lindgren et al., 2004) and finally to different types of molecular data alone (see Allcock et al., 2014 for a review). Our initial all, best1 and best2 datasets comprising 18 taxa yielded variable ordinal-level relationships within Decapodiformes, particularly with respect to the relative positions of the sepioid orders Idiosepiida and Sepiolida (Fig. 1). By contrast, topologies generated using the combined dataset (26 taxa) were relatively stable with respect to ordinal-level relationships across filtering/coding strategy (Fig. 2). This observation was most evident for Idiosepiida; *Idiosepius paradoxus* was one of the smallest samples included in our initial datasets, returning only 248 OGs from the cephalopod core ortholog set (Table 1). For the combined dataset, we chose to use a newer, larger transcriptome for a different representative (*Idiosepius notoides*) that yielded more OGs (1887; Table 1). In the initial datasets, the position of Idiosepiida was highly variable, while in analyses of the combined dataset, the placement of Idiosepiida largely stabilized regardless of filtering strategy, amount of missing data, etc., falling as sister to all other decapodiforms, followed by Sepiolida. Idiosepiids are small animals (1–3 cm long); so one plausible explanation for the instability of *Idiosepius paradoxus* in our initial analyses is contamination with prey (e.g., fish or crustacean) or parasite (e.g., nematode or flatworm) sequences. To explore this possibility, we subjected all 473 *Idiosepius paradoxus* sequences (representing 248 OGs) to a BLAST search against the NCBI non-redundant protein sequence database. Of these 473 sequences, 471 returned significant hits. Of those 471, the top hit was a cephalopod sequence in 380 cases (*Octopus bimaculoides* in 365 cases). Of the remaining 91 top hits, nearly half (44) were non-cephalopod mollusks, followed by one nematode, two crustacean and fourteen vertebrate top hits (one human, 80% identity), along with several mostly low-similarity (< 80% identity) hits to several other taxa including fungi, annelids, insects and cyanobacteria. The four sequences that had high-similarity hits that seemed like the most plausible contaminants—*Pristionchus pacificus* (a free-living terrestrial nematode, 98% identity), *Labrus bergylta* (a northeastern Atlantic perciform fish, 90% identity) and two *Daphnia pulex* (79% identity)—were removed by PhyloTreePruner prior to phylogenetic analyses. In short, there is little evidence that contaminant sequences affected the placement of *Idiosepius paradoxus* in our initial analyses. While the placement of Idiosepiida as sister to all other Decapodiformes has been found elsewhere (e.g., Lindgren et al., 2012), this is the first topology to recover this relationship with any degree of support. In the combined dataset, cuttlefishes (Sepiida) and loliginids (Myopsida) formed a monophyletic group sister to the open-eyed squid order Oegopsida. It is worth noting that none of our analyses (Figs. 1 and 2) found support for Uniductia *sensu* Berthold and Engeser (1987), Sepioidea *sensu* Naef (1923), or Teuthoidea *sensu* Naef (1923).

Within Oegopsida, relationships fluctuated across datasets, but this is not surprising, for two reasons. First, *Chiroteuthis* was the smallest transcriptome that we included in the combined analyses, with only 470 OGs in the cephalopod core ortholog dataset (Table 1). Second, only six of 24 oegopsid families were included here, which represents only a fraction of oegopsid diversity, making it difficult to accurately infer relationships among these families. However, Sanger-sequencing based studies (e.g., Lindgren, 2010; Lindgren et al., 2012) did hypothesize that Cranchiidae and Ommastrephidae fell near the base of the oegopsid clade, a finding shown in the majority of our analyses.

The only combined topologies that showed any degree of instability with respect to ordinal relationships were those generated with the lophotrochozoan core ortholog set (Supplementary Material), where many bootstrap support values were lower and the relative positions of *Idiosepius* and *Euprymna* switched places depending on amount of missing data. The lophotrochozoan core ortholog set, which was developed using genomic data from representatives of five phyla (Annelida, Arthropoda, Mollusca, Platyhelminthes and Nematoda), comprises roughly 900 fewer OGs than our cephalopod core ortholog set. This means that combined data matrices generated using the lophotrochozoan core ortholog set were consistently smaller (both in terms of number of OGs and overall number of characters), and in some cases more than an order of magnitude smaller, than were the equivalent combined data matrices generated using the cephalopod core ortholog set (Table 2), emphasizing the benefit to creating lineage-specific ortholog sets.

4.3. (In)congruence and –omic based datasets

This study represents the fourth '-omics' based analysis of cephalopod relationships published in 2017. While all provide exciting and important information with regard to cephalopod phylogenetics, the topologies generated differ markedly in the type of data included, analytical methods and inferred relationships, particularly among orders of Decapodiformes (Fig. 3). Strugnell et al. (2017) utilized a smaller amino acid sequence data matrix of 13 mitochondrial protein-coding genes for 36 taxa, but was able to include *Spirula spirula* and *Bathyteuthis abyssicola*, while Uribe and Zardoya (2017) used a combination of 39 complete mitochondrial genomes and partial mitochondrial gene sequences from key lineages including *Spirula*. Tanner et al. (2017) utilized transcriptomic data from 26 cephalopod taxa for 180 genes, though several of those transcriptomes included significant amounts of missing data. Our combined dataset (26 taxa) was restricted to relatively large transcriptomes that contained many OGs, but this resulted in exclusion of the key order Spirulida (Figs. 2 and 3).

Topologies in all four studies are consistent in monophyly for Octopodiformes and Decapodiformes (Fig. 3, 100% support). Within Octopodiformes, topologies from this study, Uribe and Zardoya and Tanner et al. show monophyly for Octopoda (100%), a finding not tested in Strugnell et al. Within Decapodiformes, the relative position of the orders Sepiida, Sepiolida, Idiosepiida, Spirulida and Loliginidae differ significantly across these four studies, with varying degrees of support (Fig. 3). Strugnell et al., Uribe and Zardoya and Tanner et al. all hypothesize that Sepiida is sister all other decapodiforms (100% support), while our analyses suggest Idiosepiida as sister to all other decapodiforms, followed by Sepiolida, with a Sepiida + Loliginidae clade sister to Oegopsida. A close relationship between Sepiolida and Idiosepiida was hypothesized by Tanner (85% BS support), but not by Strugnell et al., where Idiosepiida and Loliginidae fell in a monophyletic group, or by Uribe and Zardoya, where Idiosepiida was sister to all other decapodiform lineages, except Sepiida. Furthermore, the position of Idiosepiida was not well resolved in Strugnell et al., and findings from their gene order data suggested a possible closer affinity to other sepioid orders. The placement of Spirulida has been contentious in both morphological and molecular studies where it has been suggested to be part of the sepioid complex (e.g., Naef, 1923), sister to a sepioid + myopsid clade (Berthold and Engeser, 1987), or sister to Bathyteuthoidea + Oegopsida (Lindgren et al., 2012). The recovery of a clade comprising Spirulida and Oegopsida by Uribe and Zardoya, Strugnell et al. and Tanner et al. is suggestive of a close association. Unfortunately, our pipeline (using either the lophotrochozoan or cephalopod core ortholog set) was unable to recover sufficient OGs to include data for *Spirula spirula* in the combined analyses. To fully test these hypotheses regarding Spirulida, Oegopsida and Bathyteuthoidea, a larger transcriptome or genome dataset is needed.

Given the variation across these four studies with respect to data type (mitochondrial genome vs. transcriptome/nuclear genome), taxon sampling, and the degree to which the datasets incorporated missing data, it is not entirely surprising that different topologies are recovered. Recovery of strong support for conflicting patterns of relationships—both for topologies based on mitochondrial data (Strugnell et al., 2017; Uribe and Zardoya, 2017) and for those based on transcriptome data (Tanner et al., 2017; this study)—is disturbing, but far from unprecedented in phylogenomic studies. Several prominent examples of such conflict have come to light recently, most notably within deep metazoan (Pisani et al., 2015; Shen et al., 2017; Simion et al., 2017; Whelan et al., 2015) and angiosperm phylogeny (Drew et al., 2014; Goremykin et al., 2015; Xi et al., 2014). The apparent conflict among these three studies, coupled with relatively high support values, only reinforces the notion that much work remains to be done to generate more robust hypotheses of ordinal-level relationships, particularly with respect to the sepioid groups.

Our initial and combined datasets also illustrate that large datasets remain susceptible to the impacts of missing data. In our initial analyses, the instability of ordinal-level relationships was most evident for matrices that included smaller transcriptomes; the resulting topologies were highly variable and in some cases, the relationships observed were nonsensical (e.g., *Spirula* sister to Octopodiformes). The controversies surrounding missing data are well known (e.g., de Queiroz and Gatesy, 2007; Lemmon et al., 2009; Wiens, 2003) and missing data can lead to erroneous placement of incomplete taxa or an increase in tree reconstruction artifacts. However, in some cases, only a limited amount of material is available for key taxonomic groups and limited taxon sampling may lead to potential long-branch attraction artifacts (Anderson and Swofford, 2004; Hillis et al., 2003). Furthermore, inclusion of smaller transcriptomes (i.e., more missing data) can be beneficial in phylogenomic studies (Roure et al., 2013). As with all previous phylogenetic studies of Cephalopoda (as summarized in Allcock et al., 2014), the greatest areas of conflict identified here were those discussed by Naef (1923): whether the sepioid orders form a clade and whether myopsids are more closely related to oegopsids or sepioids.

5. Conclusion

While genomic data are known to be useful for inferring deep-level relationships in many cases, such as for Mollusca (e.g., Kocot et al., 2011; Smith et al., 2011), a strong phylogenetic signal must be present. Extant cephalopods (particularly Decapodiformes) seem to be the product of ancient rapid radiations, which may confound our ability to

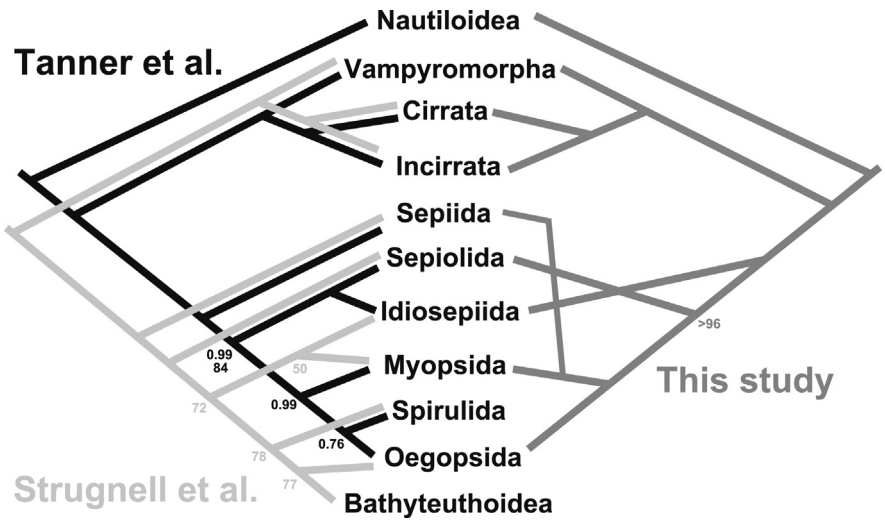


Fig. 3. Comparison of topologies recovered by Tanner et al. (2017) (black; transcriptome data), Strugnell et al. (2017) (red; mitochondrial genomes) and in the present study (blue; combined analyses based on the cephalopod core ortholog set; identical topology to Fig. 2). All nodes have 100% ML bootstrap support (posterior probabilities for Tanner et al.), except where noted. The phylogenies published by Uribe and Zardoya (2017) are similar to Strugnell et al.'s tree, but Uribe and Zardoya recover Idiosepiida as sister to all decapodiforms except Sepiida and Spirulida as sister to Oegopsida, with Bathyteuthoidea sister to Spirulida + Oegopsida (based on four mitochondrial genes for Spirulida). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

resolve deep-node relationships, even in the presence of a large number of loci. In these cases, taxon sampling (e.g., Pyron, 2015), orthology inference and data filtering methods are all likely to have significant impacts. Based on the data generated here, several key conclusions can be drawn with regard to the utility of transcriptomes for inferring cephalopod phylogeny. First, the size of the NGS data matrix matters with respect to topological stability: our analyses indicate that inclusion of larger transcriptomes (and in some cases, exclusion of small transcriptomes) yielded datasets with more orthologs that were much less sensitive to analytic perturbation. For example, our combined datasets produced more consistent topologies with higher levels of branch support than recovered by more traditional multigene datasets as well as previously published "-omics" datasets (including our own initial analyses). Additionally, in our combined analyses, topologies generated using our cephalopod core ortholog set were more consistent across levels of missing data, filtering method, etc., than those generated based on the lophotrochozoan core ortholog set. Second, to fully test the feasibility of recovering stable ordinal relationships, additional taxon sampling is sorely needed for several important cephalopod groups where transcriptome data are either very limited (e.g., Spirulida, Sepiolida, Idiosepiida and Cirrata) or lacking altogether (order Bathyteuthoidea, sepioid family Sepiadariidae and the majority of pelagic octopods).

The sensitivity of deep-level relationships to the ratio of phylogenetic signal to noise as represented by our initial and combined datasets coupled with the ancient rapid radiations that impacted extant decapodiforms in particular likely play a major role in our (in)ability to recover deep relationships consistently. Whether or not hundreds or thousands of loci will resolve relationships previously unrecoverable with only a few loci (e.g., Sharma et al., 2014) for cephalopods has yet to be seen. However, with basic biological data available for only 25% of known cephalopods (Tittensor et al., 2010) and only a single genome sequenced (Albertin et al., 2015), there is a substantial amount of potential to expand our understanding of phylogenetic relationships among cephalopods.

Declarations

Availability of data and material

The datasets generated during and analyzed in this study, custom scripts used in the phylogenomics pipeline, the cephalopod core ortholog set for use in HaMStR and phylogenetic trees are available in the Dryad repository (Supplementary Material).

Competing interests

We have no competing interests.

Authors' contributions

ARL and FEA designed the study; ARL collected the new transcriptome data; FEA implemented the bioinformatics pipelines; ARL and FEA performed the phylogenetic analyses; ARL and FEA wrote and edited the manuscript.

Acknowledgement

We are grateful to Dick Young, Michael Vecchione and two anonymous reviewers for providing feedback on this manuscript. Thanks also to Alistair Tanner and Rute de Fonseca who provided us with access to their assemblies as well as guidance on how their dataset was constructed. All new transcriptome sequence data were obtained in the laboratory of Todd Oakley at UCSB. Thanks also to Sabrina Pankey who collected the *Vampyroteuthis infernalis* and *Todarodes pacificus* specimens used for this study. We would like to thank Carrie Ganote at the National Center for Genome Assembly Support for her tireless efforts to help FEA with OrthoMCL and MySQL. This work would not have been possible without CIPRES, NCGAS, Indiana University, and Portland State University who provided the computational facilities for the initial data processing steps and construction of the cephalopod core ortholog set. Kevin Horn (SIU) provided valuable bioinformatics assistance that greatly streamlined several analyses, and Bryce Corbett (SIU) assisted FEA with phylogenetic analyses. All photos used in the graphical abstract were obtained via Wikimedia Commons with two exceptions: we thank Dr. Michael Vecchione (National Museum of Natural History, USA) for allowing us to use his photograph of *Cirroctopus glacialis* and Dr. Tom Davis (Southern Cross University, Australia) for granting us permission to use his *Idiosepius notoides* photograph. This work was supported by the U.S. National Science Foundation DEB-1036516 to FEA (WormNet II Assembling the Annelid Tree of Life) and DEB-090633 to ARL.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2017.10.004.

References

Albertin, C.B., Simakov, O., Mitros, T., Wang, Z.Y., Pungor, J.R., Edsinger-Gonzales, E., Brenner, S., Ragsdale, C.W., Rokhsar, D.S., 2015. The octopus genome and the evolution of cephalopod neural and morphological novelties. Nature 524, 220–224. http://dx.doi.org/10.1038/nature14668.

Aldrich, F.A., 1990. Lol-i-go and far away: a consideration of the establishment of the species designation *Loligo pealei*. In: Gilbert, D.L., Adelman, W.J., Arnold, J.M. (Eds.), Squid as Experimental Animals. Plenum Publishing Corporation, New York, New York, pp. 27–34.

Allcock, A.L., Cooke, I.R., Strugnell, J.M., 2011. What can the mitochondrial genome reveal about higher-level phylogeny of the molluscan class Cephalopoda? Zool. J. Linn. Soc. 161, 573–586. http://dx.doi.org/10.1111/j.1096-3642.2010.00656.x.

Allcock, A.L., Lindgren, A., Strugnell, J.M., 2014. The contribution of molecular data to our understanding of cephalopod evolution and systematics: a review. J. Nat. Hist. 49, 1373–1421.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410. http://dx.doi.org/10.1016/S0022-2836(05)80360-2.

Anderson, F.E., 2000a. Phylogeny and historical biogeography of the loliginid squids (Mollusca: cephalopoda) based on mitochondrial DNA sequence data. Mol. Phylogenet. Evol. 15, 191–214. http://dx.doi.org/10.1006/mpev.1999.0753.

Anderson, F.E., 2000b. Phylogenetic relationships among loliginid squids (Cephalopoda: Myopsida) based on analyses of multiple data sets. Zool. J. Linn. Soc. 130, 603–633. http://dx.doi.org/10.1111/j.1096-3642.2000.tb02203.x.

Anderson, F.E., Bergman, A., Cheng, S.H., Pankey, M.S., Valinassab, T., 2014. Lights out: the evolution of bacterial bioluminescence in Loliginidae. Hydrobiologia 725, 189–203. http://dx.doi.org/10.1007/s10750-013-1599-1.

Anderson, F.E., Swofford, D.L., 2004. Should we be worried about long-branch attraction in real data sets? Investigations using metazoan 18S rDNA. Mol. Phylogenet. Evol. 33, 440–451.

Aronson, R.B., 1991. Ecology, paleobiology and evolutionary constraint in the octopus. Bull. Mar. Sci. 49, 1–2.

Berthold, T., Engeser, T., 1987. Phylogenetic analysis and systematization of the Cephalopoda (Mollusca). Verhandlungen Naturwissenschaftlichen Vereins Hambg.

Carlini, D.B., Graves, J.E., 1999. Phylogenetic analysis of cytochrome c oxidase I sequences to determine higher-level relationships within the coleoid cephalopods. Bull. Mar. Sci. 64, 57–76.

Chen, M.-Y., Liang, D., Zhang, P., 2015. Selecting question-specific genes to reduce incongruence in phylogenomics: a case study of jawed vertebrate backbone phylogeny. Syst. Biol. 64, 1104–1120. http://dx.doi.org/10.1093/sysbio/syv059.

Collins, M.A., Villanueva, R., 2006. Taxonomy, ecology and behaviour of the cirrate octopods. In: Gibson, R.N., Atkinson, R.J.A., Gordon, J.D.M. (Eds.), Oceanography & Marine Biology: An Annual Review. CRC Press, Boca Raton, Florida, pp. 277–322.

de Queiroz, A., Gatesy, J., 2007. The supermatrix approach to systematics. Trends Ecol. Evol. 22, 34–41. http://dx.doi.org/10.1016/j.tree.2006.10.002.

Doyle, V.P., Young, R.E., Naylor, G.J.P., Brown, J.M., 2015. Can we identify genes with increased phylogenetic reliability? Syst. Biol. 64, 824–837. http://dx.doi.org/10.1093/sysbio/syv041.

Drew, B.T., Ruhfel, B.R., Smith, S.A., Moore, M.J., Briggs, B.G., Gitzendanner, M.A., Soltis, P.S., Soltis, D.E., 2014. Another look at the root of the angiosperms reveals a familiar tale. Syst. Biol. 63, 368–382. http://dx.doi.org/10.1093/sysbio/syt108.

Ebersberger, I., Strauss, S., von Haeseler, A., 2009. HaMStR: profile hidden markov model based search for orthologs in ESTs. BMC Evol. Biol. 9, 157. http://dx.doi.org/10.1186/1471-2148-9-157.

Eddy, S.R., Altschul, S., Gish, W., Miller, W., Myers, E., Lipman, D., Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Krogh, A., Brown, M., Mian, I., Sjölander, K., Haussler, D., Durbin, R., Eddy, S., Krogh, A., Mitchison, G., Altschul, S., Bundschuh, R., Olsen, R., Hwa, T., Schäffer, A., Aravind, L., Madden, T., Shavirin, S., Spouge, J., Altschul, S., Wootton, J., Gertz, E., Agarwala, R., Morgulis, A., Yu, Y., Gertz, E., Agarwala, R., Schäffer, A., Altschul, S., Hunter, S., Apweiler, R., Attwood, T., Bairoch, A., Bateman, A., Finn, R., Mistry, J., Tate, J., Coggill, P., Heger, A., Chaudhary, V., Liu, F., Matta, V., Meng, X., Nambiar, A., Walters, J., Qudah, B., Chaudhary, V., Landman, J., Ray, J., Walters, J., Sachdeva, V., Kistler, M., Speight, E., Tzeng, T., Maddimsetty, R., Derrien, S., Quinton, P., Jacob, A., Lancaster, J., Buhler, J., Chamberlain, R., Oliver, T., Yeow, L., Schmidt, B., Horn, D., Houston, M., Hanrahan, P., Walters, J., Balu, V., Kompalli, S., Chaudhary, V., Chukkapalli, G., Guda, C., Subramaniam, S., Rekapalli, B., Halloy, C., Zhulin, I., Sun, Y., Buhler, J., Sun, Y., Buhler, J., Johnson, L., Eddy, S., Portugaly, E., Smith, T., Waterman, M., Madera, M., Gough, J., Johnson, S., Freyhult, E., Bollback, J., Gardner, P., Eddy, S., Rognes, T., Farrar, M., Wozniak, A., Rognes, T., Seeberg, E., Milosavljević, A., Jurka, J., Karplus, K., Barrett, C., Hughey, R., Rabiner, L., Melnikoff, S., Quigley, S., Pearson, W., Grundy, W., Gertz, E., Yu, Y., Agarwala, R., Schäffer, A., Altschul, S., Price, G., Crooks, G., Green, R., Brenner, S., 2011. Accelerated profile HMM searches. PLoS Comput. Biol. 7, e1002195. http://dx.doi.org/10.1371/journal.pcbi.1002195.

Felsenstein, J., 1978. Cases in which parsimony and compatibility methods will be positively misleading. Syst. Zool. 27, 401–410.

Foster, P.G., Hickey, D.A., 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J. Mol. Evol. 48, 284–290. http://dx.doi.org/10.1007/PL00006471.

Garrison, N.L., Rodriguez, J., Agnarsson, I., Coddington, J.A., Griswold, C.E., Hamilton, C.A., Hedin, M., Kocot, K.M., Ledford, J.M., Bond, J.E., 2016. Spider phylogenomics: untangling the Spider Tree of Life. PeerJ 4, e1719. http://dx.doi.org/10.7717/peerj.1719.

Goremykin, V.V., Nikiforova, S.V., Cavalieri, D., Pindo, D.M., Lockhart, P., 2015. The root of flowering plants and total evidence. Syst. Biol. 64, 879–891. http://dx.doi.org/10.1093/sysbio/syv028.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29, 644–652. http://dx.doi.org/10.1038/nbt.1883.

Hallinan, N.M., Lindberg, D.R., 2011. Comparative analysis of chromosome counts infers three paleopolyploidies in the Mollusca. Genome Biol. Evol. 3, 1150–1163.

Hendy, M.D., Penny, D., 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. 38, 297–309.

Hillis, D.M., Pollock, D.D., McGuire, J.A., Zwickl, D.J., Thorne, J., 2003. Is sparse taxon sampling a problem for phylogenetic inference? Syst. Biol. 52, 124–126.

Jaillon, O., Aury, J.-M., Brunet, F., Petit, J.-L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature 431, 946–957.

Katoh, K., Kuma, K., Toh, H., Miyata, T., 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33, 511–518. http://dx.doi.org/10.1093/nar/gki198.

Kocot, K.M., Cannon, J.T., Todt, C., Citarella, M.R., Kohn, A.B., Meyer, A., Santos, S.R., Schander, C., Moroz, L.L., Lieb, B., Halanych, K.M., 2011. Phylogenomics reveals deep molluscan relationships. Nature 477, 452–456. http://dx.doi.org/10.1038/nature10382.

Kocot, K.M., Citarella, M.R., Moroz, L.L., Halanych, K.M., 2013. PhyloTreePruner: a phylogenetic tree-based approach for selection of orthologous sequences for phylogenomics. Evol. Bioinform. Online 9, 429–435. http://dx.doi.org/10.4137/EBO.S12813.

Kröger, B., Vinther, J., Fuchs, D., 2011. Cephalopod origin and evolution: a congruent picture emerging from fossils, development and molecules: Extant cephalopods are younger than previously realised and were under major selection to become agile, shell-less predators. BioEssays 33, 602–613. http://dx.doi.org/10.1002/bies.201100001.

Kück, P., Meusemann, K., 2010. FASconCAT: convenient handling of data matrices. Mol. Phylogenet. Evol. 56, 1115–1118. http://dx.doi.org/10.1016/j.ympev.2010.04.024.

Kück, P., Struck, T.H., 2014. BaCoCa–a heuristic software tool for the parallel assessment of sequence biases in hundreds of gene and taxon partitions. Mol. Phylogenet. Evol. 70, 94–98. http://dx.doi.org/10.1016/j.ympev.2013.09.011.

Lartillot, N., Lepage, T., Blanquart, S., 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25, 2286–2288. http://dx.doi.org/10.1093/bioinformatics/btp368.

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. http://dx.doi.org/10.1093/sysbio/syt022.

Lemmon, A.R., Brown, J.M., Stanger-Hall, K., Lemmon, E.M., 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Syst. Biol. 58, 130–145. http://dx.doi.org/10.1093/sysbio/syp017.

Li, L., Stoeckert, C.J., Roos, D.S., 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13, 2178–2189. http://dx.doi.org/10.1101/gr.1224503.

Lindgren, A.R., Giribet, G., Nishiguchi, M.K., 2004. A combined approach to the phylogeny of Cephalopoda, (Mollusca). Cladistics 20, 454–486. http://dx.doi.org/10.1111/j.1096-0031.2004.00032.x.

Lindgren, A.R., 2010. Molecular inference of phylogenetic relationships among Decapodiformes (Mollusca: Cephalopoda) with special focus on the squid order Oegopsida. Mol. Phylogenet. Evol. 56, 77–90. http://dx.doi.org/10.1016/j.ympev.2010.03.025.

Lindgren, A.R., Pankey, M.S., Hochberg, F.G., Oakley, T.H., 2012. A multi-gene phylogeny of Cephalopoda supports convergent morphological evolution in association with multiple habitat shifts in the marine environment. BMC Evol. Biol. 12, 129. http://dx.doi.org/10.1186/1471-2148-12-129.

Lu, C.-C., 2005. A new family of myopsid squid from Australasian waters (Cephalopoda:Teuthida). Phuket Mar. Biol. Cent. Res. Bull. 66, 71–82.

Meyer, B., Meusemann, K., Misof, B., 2011. MARE: MAtrix REduction—a tool to select optimized data subsets from supermatrices for phylogenetic inference. Bonn Zent. fuur Mol. Biodiversitätsforsch. am ZFMK.

Miller, M., Pfeiffer, W., Schwartz, T., 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees, in: Gateway Computing Environments Workshop (GCE), 2010. IEEE, pp. 1–8.

Naef, A., 1923. Cephalopoda. Fauna e Flora del Golfo di Napoli. Monograph No. 35. Teil I, Band 1, Lfg. 2, Fauna e Flora del Folgo di Napoli. Monograph No. 35. English translation: A Mercado (1972). Israel Program for Scientific Translations Ltd.

Nieselt-Struwe, K., von Haeseler, A., 2001. Quartet-mapping, a generalization of the likelihood-mapping procedure. Mol. Biol. Evol. 18, 1204–1219.

Norman, J.D., Robinson, M., Glebe, B., Ferguson, M.M., Danzmann, R.G., 2012. Genomic arrangement of salinity tolerance QTLs in salmonids: a comparative analysis of Atlantic salmon (*Salmo salar*) with Arctic charr (*Salvelinus alpinus*) and rainbow trout (*Oncorhynchus mykiss*). BMC Genomics 13, 420. http://dx.doi.org/10.1186/1471-2164-13-420.

O'Shea, S., 1999. The marine fauna of New Zealnd: Octopoda (Mollusca: Cephalopoda), NIWA Biodiversity Memoir. NIWA Research, Wellington.

Philippe, H., Brinkmann, H., Copley, R.R., Moroz, L.L., Nakano, H., Poustka, A.J., Wallberg, A., Peterson, K.J., Telford, M.J., 2011. Acoelomorph flatworms are deuterostomes related to Xenoturbella. Nature 470, 255–258.

Pisani, D., Pett, W., Dohrmann, M., Feuda, R., Rota-Stabelli, O., Philippe, H., Lartillot, N., Wörheide, G., 2015. Genomic data do not support comb jellies as the sister group to all other animals. Proc. Natl. Acad. Sci. 112, 15402–15407. http://dx.doi.org/10.1073/pnas.1518127112.

Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS ONE 5, e9490. http://dx.doi.org/10.1371/journal.pone.0009490.

Pyron, R.A., 2015. Post-molecular systematics and the future of phylogenetics. Trends Ecol. Evol. 30, 384–389. http://dx.doi.org/10.1016/j.tree.2015.04.016.

Rabosky, D.L., Santini, F., Eastman, J., Smith, S.A., Sidlauskas, B., Chang, J., Alfaro, M.E., 2013. Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. Nat. Commun. 4, 1958. http://dx.doi.org/10.1038/ncomms2958.

Rokas, A., Carroll, S.B., 2006. Bushes in the tree of life. PLoS Biol. 4, e352. http://dx.doi.org/10.1371/journal.pbio.0040352.

Roure, B., Baurain, D., Philippe, H., 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Mol. Biol. Evol. 30, 197–214. http://dx.doi.org/10.1093/molbev/mss208.

Saccone, C., Lanave, C., Pesole, G., Preparata, G., 1990. Influence of base composition on quantitative estimates of gene evolution. Methods Enzymol. 180, 570–583.

Sales, J.B. de L., Shaw, P.W., Haimovici, M., Markaida, U., Cunha, D.B., Ready, J., Figueiredo-Ready, W.M.B., Schneider, H., Sampaio, I., 2013. New molecular phylogeny of the squids of the family Loliginidae with emphasis on the genus *Doryteuthis* Naef, 1912: Mitochondrial and nuclear sequences indicate the presence of cryptic species in the southern Atlantic Ocean. Mol. Phylogenet. Evol. 68, 293–299. http://dx.doi.org/10.1016/j.ympev.2013.03.027.

Salichos, L., Rokas, A., 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497, 327–331. http://dx.doi.org/10.1038/nature12130.

Sharma, P.P., Kaluziak, S.T., Perez-Porro, A.R., Gonzalez, V.L., Hormiga, G., Wheeler, W.C., Giribet, G., 2014. Phylogenomic interrogation of Arachnida reveals systemic conflicts in phylogenetic signal. Mol. Biol. Evol. 31, 2963–2984. http://dx.doi.org/10.1093/molbev/msu235.

Shen, X.-X., Hittinger, C.T., Rokas, A., 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nat. Ecol. Evol. 1, 126.

Simion, P., Philippe, H., Baurain, D., Jager, M., Richter, D.J., Di Franco, A., Roure, B., Satoh, N., Quéinnec, É., Ereskovsky, A., Lapébie, P., Corre, E., Delsuc, F., King, N., Wörheide, G., Manuel, M., 2017. A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. Curr. Biol. 27, 958–967. http://dx.doi.org/10.1016/j.cub.2017.02.031.

Smith, S.A., Wilson, N.G., Goetz, F.E., Feehery, C., Andrade, S.C.S., Rouse, G.W., Giribet, G., Dunn, C.W., 2011. Resolving the evolutionary relationships of molluscs with phylogenomic tools. Nature 480, 364–367. http://dx.doi.org/10.1038/nature10526.

Stadler, T., 2011. Mammalian phylogeny reveals recent diversification rate shifts. Proc. Natl. Acad. Sci. U.S.A. 108, 6187–6192. http://dx.doi.org/10.1073/pnas.1016876108.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313. http://dx.doi.org/10.1093/bioinformatics/btu033.

Struck, T.H., 2014. TreSpEx—detection of misleading signal in phylogenetic reconstructions based on tree information. Evol. Bioinforma. 10, 51. http://dx.doi.org/10.4137/EBO.S14239.

Struck, T.H., 2013. The impact of paralogy on phylogenomic studies - a case study on annelid relationships. PLoS ONE 8, e62892. http://dx.doi.org/10.1371/journal.pone.0062892.

Strugnell, J., Jackson, J., Drummond, A.J., Cooper, A., 2006. Divergence time estimates

for major cephalopod groups: evidence from multiple genes. Cladistics 22, 89–96.
Strugnell, J., Nishiguchi, M.K., 2007. Molecular phylogeny of coleoid cephalopods (Mollusca: Cephalopoda) inferred from three mitochondrial and six nuclear loci: a comparison of alignment, implied alignment and analysis methods. J. Molluscan Stud. 73, 399–410. http://dx.doi.org/10.1093/mollus/eym038.
Strugnell, J., Norman, M., Jackson, J., Drummond, A.J., Cooper, A., 2005. Molecular phylogeny of coleoid cephalopods (Mollusca: Cephalopoda) using a multigene approach; the effect of data partitioning on resolving phylogenies in a Bayesian framework. Mol. Phylogenet. Evol. 37, 426–441.
Strugnell, J.M., Hall, N.E., Vecchione, M., Fuchs, D., Allcock, A.L., 2017. Whole mitochondrial genome of the Ram's Horn Squid shines light on the phylogenetic position of the monotypic order Spirulida (Haeckel, 1896). Mol. Phylogenet. Evol. 109, 296–301. http://dx.doi.org/10.1016/j.ympev.2017.01.011.
Susko, E., Roger, A.J., 2007. On reduced amino acid alphabets for phylogenetic inference. Mol. Biol. Evol. 24, 2139–2150. http://dx.doi.org/10.1093/molbev/msm144.
Swofford, D.L., 2002. PAUP*. Phylogenetic Analysis Using Parsimony (* and Other Methods).
Tanner, A.R., Fuchs, D., Winkelmann, I.E., Gilbert, M.T.P., Pankey, M.S., Ribeiro, Â.M., Kocot, K.M., Halanych, K.M., Oakley, T.H., da Fonseca, R.R., Pisani, D., Vinther, J., 2017. Molecular clocks indicate turnover and diversification of modern coleoid cephalopods during the Mesozoic Marine Revolution. Proc. R. Soc. London B Biol. Sci. 284.
Tittensor, D.P., Mora, C., Jetz, W., Lotze, H.K., Ricard, D., Berghe, E. Vanden, Worm, B., 2010. Global patterns and predictors of marine biodiversity across taxa. Nature 466, 1098–1101. http://dx.doi.org/10.1038/nature09329.
Uribe, J.E., Zardoya, R., 2017. Revisiting the phylogeny of Cephalopoda using complete mitochondrial genomes. J. Molluscan Stud. 83, 133–144. http://dx.doi.org/10.1093/mollus/eyw052.
Whelan, N.V., Kocot, K.M., Moroz, L.L., Halanych, K.M., 2015. Error, signal, and the placement of Ctenophora sister to all other animals. Proc. Natl. Acad. Sci. U.S.A. 112, 5773–5778. http://dx.doi.org/10.1073/pnas.1503453112.
Whitfield, J.B., Lockhart, P.J., 2007. Deciphering ancient rapid radiations. Trends Ecol. Evol. 22, 258–265. http://dx.doi.org/10.1016/j.tree.2007.01.012.
Wiens, J., Kuczynski, C., Smith, S., Mulcahy, D., Sites, J., Townsend, T., Reeder, T., Zamudio, K., 2008. Branch lengths, support, and congruence: testing the phylogenomic approach with 20 nuclear loci in snakes. Syst. Biol. 57, 420–431. http://dx.doi.org/10.1080/10635150802166053.
Wiens, J.J., 2003. Missing data, incomplete taxa, and phylogenetic accuracy. Syst. Biol. 52, 528–538. http://dx.doi.org/10.1080/10635150390218330.
Xi, Z., Liu, L., Rest, J.S., Davis, C.C., 2014. Coalescent versus concatenation methods and the placement of *Amborella* as sister to water lilies. Syst. Biol. 63, 919–932. http://dx.doi.org/10.1093/sysbio/syu055.
Young, R.E., Vecchione, M., 1996. Analysis of morphology to determine primary sister-taxon relationships within coleoid cephalopods. Am. Malacol. Bull. 12, 91–112.
Zhong, M., Hansen, B., Nesnidal, M., Golombek, A., Halanych, K.M., Struck, T.H., 2011. Detecting the symplesiomorphy trap: a multigene phylogenetic analysis of terebelliform annelids. BMC Evol. Biol. 11, 369. http://dx.doi.org/10.1186/1471-2148-11-369.