Systematic Biology Advance Access published September 23, 2016

Running head: Lophotrochozoan phylogenomics

Phylogenomics of Lophotrochozoa with consideration of systematic error

Kevin M. Kocot[1,2*], Torsten H. Struck[3], Julia Merkel[4], Damien S. Waits[1], Christiane Todt[5], Pamela M. Brannock[1], David A. Weese[1,6], Johanna T. Cannon[1,7], Leonid L. Moroz[8], Bernhard Lieb[4], and Kenneth M. Halanych[1*]

[1]*Department of Biological Sciences, 101 Rouse Life Sciences, Auburn University, Auburn, Alabama 36849, USA.*

[2]*Department of Biological Sciences and Alabama Museum of Natural History, 307 Mary Harmon Bryant Hall, The University of Alabama, Tuscaloosa, Alabama 35487 USA.*

[3]*Natural History Museum, Department of Research and Collections, University of Oslo, PO Box 1172 Blindern, N-0318 Oslo, Norway*

[4]*Johannes Gutenberg University, Institute of Zoology, 55099 Mainz, Germany.*

[5]*University Museum of Bergen, The Natural History Collections, University of Bergen, Allégaten 41, 5007 Bergen, Norway.*

[6]*Current address: Department of Biological and Environmental Sciences, Georgia College and State University, Campus Box 81, Milledgeville, Georgia 31061 USA.*

[7]*Current address: Department of Zoology, Naturhistoriska riksmuseet, Box 50007, 104 05 Stockholm, Sweden.*

[8]*The Whitney Laboratory for Marine Bioscience, University of Florida, 9505 Ocean Shore Blvd, St Augustine, Florida 32080, USA.*

*Corresponding author: *kmkocot@ua.edu.*

Current Address: Department of Biological Sciences and Alabama Museum of Natural History, 307 Mary Harmon Bryant Hall, The University of Alabama, Tuscaloosa, Alabama 35487 USA.

ABSTRACT

Phylogenomic studies have improved understanding of deep metazoan phylogeny and show promise for resolving incongruences among analyses based on limited numbers of loci. One region of the animal tree that has been especially difficult to resolve, even with phylogenomic approaches, is relationships within Lophotrochozoa (the animal clade that includes molluscs, annelids, and flatworms among others). Lack of resolution in phylogenomic analyses could be due to insufficient phylogenetic signal, limitations in taxon and/or gene sampling, or systematic error. Here, we investigated why lophotrochozoan phylogeny has been such a difficult question to answer by identifying and reducing sources of systematic error. We supplemented existing data with 32 new transcriptomes spanning the diversity of Lophotrochozoa and constructed a new set of Lophotrochozoa-specific core orthologs. Of these, 638 orthologous groups (OGs) passed strict screening for paralogy using a tree-based approach. In order to reduce possible sources of systematic error, we calculated branch-length heterogeneity, evolutionary rate, percent missing data, compositional bias, and saturation for each OG and analyzed increasingly stricter subsets of only the most stringent (best) OGs for these five variables. Principal component analysis of the values for each factor examined for each OG revealed that compositional heterogeneity and average patristic distance contributed most to the variance observed along the first principal component while branch-length heterogeneity and, to a lesser extent, saturation contributed most to the variance observed along the second. Missing data did not strongly contribute to either. Additional sensitivity analyses examined effects of removing taxa with heterogeneous branch lengths, large amounts of missing data, and compositional heterogeneity. Although our analyses do

not unambiguously resolve lophotrochozoan phylogeny, we advance the field by reducing the list of viable hypotheses. Moreover, our systematic approach for dissection of phylogenomic data can be applied to explore sources of incongruence and poor support in any phylogenomic dataset.

Keywords: Trochozoa, Spiralia, Mollusca, Nemertea, Annelida, Brachiopoda, Phoronida, Entoprocta, Platyzoa, Polyzoa, Bryozoa

Understanding of deep phylogeny has improved with the application of phylogenomic approaches (e.g., Philippe et al. 2004, 2005; Matus et al. 2006; Delsuc et al. 2006; Dunn et al. 2008; Hejnol et al. 2009; Kocot et al. 2011; Smith et al. 2011; Struck et al. 2011; Zhong et al. 2011a; Ryan et al. 2013; Moroz et al. 2014, Torruella et al. 2015, Whelan et al. 2015, etc). Nonetheless, some regions of the tree of life with short internodes, probably due to rapid diversification, still lack resolution. Relationships within Lophotrochozoa (Halanych et al. 1995) are one such example. Lophotrochozoa is a well-supported clade of invertebrates that includes Annelida (including Myzostomida, Pogonophora, Echiura, and Sipuncula), Brachiopoda, Bryozoa (=Ectoprocta), Cycliophora, Dicyemida, Entoprocta (=Kamptozoa), Gastrotricha, Gnathostomulida, Micrognathozoa, Mollusca, Nemertea, Orthonectida, Phoronida, Platyhelminthes, and Rotifera (=Syndermata; including Acanthocephala and Seisonida), (e.g., Eernisse et al. 1992; Halanych et al. 1995; Halanych 2004; Matus et al. 2006; Giribet 2008, 2015; Dunn et al. 2008; Hejnol et al. 2009; Minelli 2009; Kocot et al. 2010; Edgecombe et al. 2011; Nielsen 2011; Dunn et al. 2014; Struck et al. 2014; Kocot 2016). Lophotrochozoa has the distinction of having the greatest diversity of body plans of the three bilaterian 'supergroups' (Lophotrochozoa, Ecdysozoa, and Deuterostomia) including two of the most morphologically variable animal phyla, Mollusca and Annelida.

Briefly, Trochozoa (Roule 1891; as *Trochozoaires* – see Rouse 1999, Peterson and Eernisse

2001, and Kocot 2016 for details on the history of this term) is a subclade of Lophotrochozoa that includes taxa with a trochophore larva (reviewed by Rouse 1999; Henry et al. 2007) or a secondarily modified trochophore larva: Mollusca, Annelida, Nemertea, Brachiopoda, and Phoronida (reviewed by Dunn et al. 2014). Molecular studies based on nuclear ribosomal RNA (rRNA) genes (18S and 28S; e.g., Halanych et al. 1995; Winnepenninckx et al. 1995; Giribet et al. 2000; Peterson and Eernisse 2001; Passamaneck and Halanych 2006; Paps et al. 2009b), sodium potassium ATPase alpha subunit (Anderson et al. 2004), and phylogenomic analyses (e.g., Dunn et al. 2008; Struck et al. 2014; Laumer et al. 2015) have largely supported Trochozoa but relationships among these phyla remain unresolved. Entoprocta, Cycliophora, and Bryozoa, three phyla of small-bodied suspension feeding animals, have also been hypothesized to be nested within Trochozoa by some. Bryozoans were traditionally grouped with Brachiopoda and Phoronida in a clade called Lophophorata (Hyman 1959) while Entoprocta has been hypothesized to be related to Mollusca under the Tetraneuralia hypothesis (Wanninger 2009). However, most molecular studies to date have instead recovered Bryozoa and Entoprocta in a separate lophotrochozoan sub-clade called Polyzoa in which Bryozoa is usually recovered sister to Entoprocta or Entoprocta + Cycliophora when the latter phylum was also sampled (e.g., Hausdorf et al. 2007; Helmkampf et al. 2008; Struck and Fisse 2008; Hausdorf et al. 2010; Witek et al. 2008, 2009; Hejnol et al. 2009; but see Nesnidal et al. 2010, 2013). Platyzoa (Cavalier-Smith 1998; Platyhelminthes, Gastrotrichia, Syndermata, Gnathostomulida, and Micrognathozoa) is a hypothesized grouping of mostly small-bodied animals but no uniting synapomorphy for the group is known. Gnathifera is a well-supported platyzoan clade that includes Rotifera, Gnathostomulida, and Micrognathozoa (Kristensen and Funch 2000). Aside from Gnathifera, support for relationships within Platyzoa and even support for platyzoan monophyly have been weak (Passamaneck and Halanych 2006; Dunn et al. 2008 [Myzostomida was nested within Platyzoa]; Hejnol et al. 2009; Witek et al. 2009) or lacking (e.g.,

Struck et al. 2014; Laumer et al. 2015).

Poor support and incongruence in phylogenomic analyses could be due to insufficient phylogenetic signal (e.g., due to closely spaced branching events), limitations in taxon and/or gene sampling, or systematic error (Philippe et al. 2011). Here, we focus on identifying and reducing sources of systematic error in phylogenomic datasets. One source of systematic error is compositional heterogeneity (Nesnidal et al. 2010, 2013). Biases in amino acid composition can result in erroneous phylogenetic reconstructions in which unrelated taxa with deviant amino acid usage are artificially grouped (Jermiin et al. 2004, Desluc et al. 2005, Rodríguez-Ezpeleta et al. 2007, Nesnidal et al. 2010). Recently, Nesnidal et al. (2013) examined lophotrochozoan relationships using a phylogenomic approach, paying particular attention to compositional heterogeneity. Although support varied among analyses, they recovered Phoronida + Bryozoa sister to Brachiopoda (i.e., Lophophorata) contrary to molecular studies supporting Polyzoa. Examination of compositional heterogeneity revealed that Polyzoa, Brachiopoda + Phoronida, and Kryptrochozoa (Giribet et al. 2009; a hypothesized grouping of Brachiopoda, Phoronida, and Nemertea) were supported by characters with apparently deviant amino acid compositions, whereas there was no indication for compositional heterogeneity in characters supporting Lophophorata. Thus, the authors concluded that support for Polyzoa and Kryptrochozoa in previous phylogenomic studies may have been an artifact due to compositional bias. Excluding taxa with exceptionally biased amino acid usage may ameliorate effects of compositional heterogeneity. In cases where such taxa are central to the question being addressed, the next best approach appears to be excluding the most compositionally heterogeneous genes and retaining more conserved, homogeneous genes (e.g., Nesnidal et al. 2013).

Another potential source of systematic error in phylogenomic analyses is missing data (Philippe et al. 2004; Wiens et al. 2006; Wiens and Moen 2008; Lemmon et al. 2009). Roure et al.

(2013) examined effects of missing data on deep metazoan phylogeny by progressively deleting data from an initially complete supermatrix. They showed that realistic patterns of missing data negatively influenced phylogenetic inference beyond the expected decrease in resolving power by reducing the number of species available for the detection of multiple substitutions at a given site. Thus, they argued that smaller (i.e., with fewer genes) but more complete datasets might be advantageous relative to larger (i.e., with more genes) but sparser datasets. Their results also support previous studies (e.g., Sanderson et al. 2011) indicating that inclusion of incomplete but short-branched, slowly evolving taxa helps to ameliorate artifacts due to missing data.

Struck et al. (2014) examined lophotrochozoan phylogeny paying special attention to compositional bias and missing data as well as long-branch attraction. A "brute force" approach by Struck et al. (2014) including all taxa and genes selected by their pipeline recovered small-sized and simply organized platyzoans as a clade. However, platyzoans exhibit considerable branch-length heterogeneity with most (but not all) sampled platyzoans having much longer branches than other lophotrochozoans. Struck et al. (2014) calculated pairwise patristic distances and a novel measure called LB score, which represents a taxon's percentage deviation from the average pairwise distance between taxa. When they excluded taxa and genes most likely to be susceptible to long-branch attraction, Platyzoa was recovered as a paraphyletic assemblage, consistent with the hypothesis that this group is an artifact of long-branch attraction (Dunn et al. 2008). Effects of saturation, including long-branch attraction, can be ameliorated by analyzing amino acids rather than nucleotide datasets, excluding genes with very high levels of saturation in favor of less saturated genes, and using best-fitting models of sequence evolution (Struck et al. 2008; Philippe et al. 2011; Dordel et al. 2010; Nosenko et al. 2013).

In addition to systematic error, poor support and incongruence for relationships within

Lophotrochozoa in previous phylogenomic studies could also stem from problems with orthology inference. Recently, Struck (2013) showed that even a small number of overlooked paralogs could have dramatic effects on phylogenomic analyses. HaMStR (Ebersberger et al. 2009) is a program that identifies sequences that are orthologs to a pre-defined set of 'core orthologs' using profile hidden Markov models (HMMs) and BLAST (Altschul et al. 1990). However, this software is dependent on taxon sampling of the core orthologs used. Given the paucity or absence of lophotrochozoans in available core ortholog sets, re-evaluation of lophotrochozoan phylogeny with a more suitable core ortholog set and/or confirmation of orthology using phylogenetic tree-based approaches (Kocot and Citarella et al. 2013; Dunn et al. 2013; Yang and Smith 2014) is desirable.

In order to improve understanding of lophotrochozoan phylogeny and explore the impact of potential sources of systematic error in phylogenomic datasets, we performed analyses on datasets with up to 74 taxa and 653 genes. To assess the impact of several factors that may cause systematic errors, we calculated amino acid composition bias, percent missing data, branch-length heterogeneity, average patristic distance, and saturation for each orthologous group (OG) and analyzed increasingly strict subsets of only the most stringent or 'best' OGs (i.e., those least likely to cause systematic error) according to each of these factors. We also examined the effects of removing taxa with high amounts of missing data, biased amino acid composition, and high LB scores.

MATERIALS AND METHODS

Taxon Sampling

Taxa were chosen to span the extant diversity of Lophotrochozoa while minimizing potentially deleterious effects of missing data (Roure et al. 2013). Our taxon sampling is biased

towards Trochozoa because 1) we intentionally avoided sampling platyzoans shown to have exceptionally long branch lengths (e.g., some of the gastrotrichs sampled by Struck et al. 2014 and some of the flatworms sampled by Laumer et al. 2015) and 2) a secondary goal of this study was to reexamine relationships within Mollusca in light of new data for key groups. Predicted transcripts from publicly available genomes were employed whenever possible. However, given the paucity of high-quality genomes from lophotrochozoans, the majority of our dataset consisted of Illumina transcriptomes. Taxon sampling and details on data used are presented in Supplementary Table 1 and details on specimen collection, tissues used, and RNA extraction for 32 newly sequenced taxa (including eleven molluscs, four brachiopods, two phoronids, three nemerteans, six annelids, four entoprocts, one cycliophoran, one chaetognath, and one priapulid) are presented in Supplementary Table 2. Some of transcriptomes employed were published in our studies addressing Toll-like receptors in Lophotrochozoa (Halanych and Kocot 2014) and nemertean toxin genes (Whelan et al. 2014) but have not yet been brought to bear on lophotrochozoan phylogeny.

Notably, some lophotrochozoan taxa that we were unable to sample were not included in this study. These include micrognathozoans (which are known only from remote freshwater habitats in Greenland and the Subantarctic), dicyemids (obligate endoparasites of cephalopods thought to be lophotrochozoans), and orthonectids (a rarely collected parasitic group thought to be lophotrochozoans). Transcriptome data collected from the bryozoan *Pectinatella magnifica* were found to contain annelid contamination and were excluded.

Molecular Techniques

Different methods were used by the Halanych, Lieb, and Moroz labs to generate transcriptome data (Supplementary Table 2). For the Halanych lab taxa, total RNA was usually

extracted from frozen or RNAlater-fixed tissue using TRIzol (Invitrogen) and purified using the RNeasy kit (Qiagen) with on-column DNase digestion. In cases where only a small amount of tissue was available and low RNA yield was expected, RNA extraction and purification were performed using the RNeasy Micro kit (Qiagen) with on-column DNAse digestion or the RNAqueous Micro kit (Ambion) without DNAse digestion. RNA concentration was measured using a Nanodrop (Thermo) and RNA quality was evaluated on a 1% SB agarose gel. For most libraries, first-strand cDNA was synthesized from 1 µg of total RNA. If much less than 1 µg of total RNA was available, 1 µl of RNase-OUT (Invitrogen) was mixed with all of the remaining eluted RNA, this mixture was vacuum centrifuged to a volume of 3 µl, and all 3 µl were used to make cDNA. First-strand cDNA synthesis was performed using the SMART cDNA library construction kit (Clontech) as per the manufacturer's instructions except that the 3' primer was replaced with the CapTrsa-CV oligo (5'-AAGCAGTGGTATCAACGCAGAGT CGCAGTCGGTACTTTTTTCTTTTTTV-3') as per Meyer et al. (2009). Full-length cDNA was then amplified using the Advantage 2 PCR system (Clontech) using the minimum number of PCR cycles necessary (usually 15 to 19) and sent to The Hudson Alpha Institute for Biotechnology (Huntsville, AL, USA) for sequencing library preparation and sequencing. Each library was sequenced using approximately one-sixth of an Illumina HiSeq 2000 lane with 2 X 100 bp paired-end chemistry.

For the Lieb lab taxa, total RNA was extracted from RNAlater-fixed tissue using Exiqon miRCURY RNA Isolation Kit for animal tissue and sent to Genterprise (Germany) for library preparation and sequencing. Total RNA quality and quantity were evaluated using an Agilent Bioanalyzer 2100 and a Nanodrop spectrophotometer. Illumina RNASeq libraries were prepared using the TruSeq RNA v2 protocol with minor modifications. Briefly, poly A+ RNA was isolated and fragmented followed by first-strand cDNA synthesis, second strand synthesis,

and purification of double-stranded cDNA (ds cDNA) with the SPRI-TE Nucleic Acid Extractor using the SPRIworks fragment library system I (Beckman Coulter). Size selection was performed to isolate fragments approximately 200-400 bp in length. Fragments were then end-repaired, end-adenylated, adaptor-ligated, and PCR-amplified with 14 cycles. Each library was sequenced using one-sixth of an Illumina HiSeq 2000 or 2500 lane with 2 X 100 bp paired-end chemistry.

The chaetognath, *Sagitta* sp., was sequenced by the Moroz lab. Animals were collected from Friday Harbor Laboratories in spring-summer. RNA isolation, quantification, sequencing library constriction, and Ion Proton (ThermoFisher) sequencing were performed according to protocols described in Kohn et al. (2013).

Sequence Assembly and Processing

We improved upon previous versions of our bioinformatic pipelines (Kocot et al. 2011, Kocot 2013, Kocot et al. 2013). For the Halanych and Lieb lab taxa, raw PE Illumina reads were digitally normalized using khmer (normalize-by-median.py -C 30 -k 20 -N 4 -x 2.5e9; Brown et al. 2012) and assembled using the October 5, 2012 release of Trinity (Grabherr et al. 2011). The Sanger *Brachionus plicatilis* expressed sequence tag (EST) data were processed and assembled using the EST2uni pipeline (Forment et al. 2008). This software removes low-quality regions with lucy (Chou and Holmes 2001), removes vector sequences with lucy and SeqClean (http://compbio.dfci.harvard.edu/tgi/software), masks low complexity regions with RepeatMasker (www.repeatmasker.org), and assembles contigs with CAP3 (Huang and Madan 1999). For the Struck et al. (2014) taxa, assembly was conducted using CLC Genomics Work Bench using the default settings with scaffolding, and expected insert size of 200-400 bp, keeping only contigs larger than 200 bp. For *Sagitta* sp. (Moroz lab), Ion Proton transcriptome assembly was performed

as described in Moroz et al. (2014). Publicly available data were downloaded as assemblies when possible (see Supplementary Table 1). In cases where assemblies were not available, publicly available Illumina data were digitally normalized using khmer and assembled using the October 5, 2012 version of Trinity as described above or both normalization and assembly were conducted using the April 13, 2014 release of Trinity. All contigs were translated with TransDecoder (https://sourceforge.net/p/transdecoder/) and amino acid sequences shorter than 100 amino acids were deleted.

Because preliminary analyses indicated that the *Symbion americana* transcriptome was contaminated with transcripts from its lobster host, we used a BLAST-based filter to remove this contamination. Translated *Symbion* transcripts were compared to a database containing translated transcripts from our four entoproct transcriptomes and translated predicted transcripts from the genome of *Daphnia pulex* using BLASTP. A sequence was kept if it satisfied one of the following criteria: 1) had hits to only entoproct transcriptomes, or 2) had a hit to an entoproct transcriptome with an e-value two orders of magnitude greater than its best hit to a *Daphnia* transcript.

Development of a Custom Core Ortholog Set

In order to improve on the orthology inference approaches used in previous studies, we employed HaMStR version 13 (Ebersberger et al. 2009) with a specifically curated core-ortholog set based on a broadly sampled set of lophotrochozoans. This "Lophotrochozoa-Kocot" core ortholog set was generated by first conducting an all-versus-all BLASTP (Altschul et al., 1990) comparison of the transcripts of *Brachionus plicatilis* (Rotifera), *Capitella teleta* (Annelida), *Crassostrea gigas* (Mollusca), *Hemithiris psittacea* (Brachiopoda), *Lottia gigantea* (Mollusca), *Loxosoma pectinaricola* (Entoprocta), *Malacobdella grossa* (Nemertea), *Phoronis psammophila* (Phoronida), and *Schmidtea mediterranea* (Platyhelminthes) with an e-value cut-off of 10^{-5}.

Capitella, *Crassostrea*, and *Lottia* were represented by predicted transcripts from those genomes (Zhang et al. 2012, Simakov et al. 2013) while other taxa were represented by our Illumina transcriptomes and the publicly available *Brachionus* EST data. Next, based on the BLASTP results, Markov clustering was conducted in OrthoMCL 2.0 (Li et al. 2003) with an inflation parameter of 2.1 following Hejnol et al. (2009) and preliminary analyses of an earlier version of this dataset (Kocot 2013).

Resulting putatively orthologous groups (55,556 in total) were processed with a modified version of the bioinformatic pipeline employed by Kocot et al. (2013). First, any sequences shorter than 200 amino acids in length were discarded. Next, each candidate core ortholog group was aligned with MAFFT (Katoh et al. 2005) using the automatic alignment strategy with a "maxiterate" value of 1,000. To screen candidate core ortholog groups for evidence of paralogy, an "approximately maximum likelihood tree" was inferred for each remaining alignment using FastTree 2 (Price et al. 2010) using the "slow" and "gamma" options. PhyloTreePruner (Kocot and Citarella et al. 2013) was then employed to use a tree-based approach to screen each candidate OG for evidence of paralogy. First, nodes with support values below 0.90 were collapsed into polytomies. Next, the maximally inclusive subtree was selected where each taxon was represented by no more than one sequence or, in cases where more than one sequence was present for any taxon, all sequences from that taxon formed a clade or were part of the same polytomy. Putative paralogs (sequences falling outside of this maximally inclusive subtree) were then deleted from the input alignment. In cases where multiple sequences from the same taxon formed a clade or were part of the same polytomy, all sequences except the longest were deleted. Lastly, in order to eliminate orthology groups with poor taxon sampling, all groups sampled for fewer than seven of the nine taxa were discarded (resulting in 2,630 OGs). *Lottia gigantea* (Gastropoda) was selected as the HaMStR primer taxon because it was the best represented taxon

in terms of number of genes sampled. Because HaMStR requires a primer taxon sequence for all OGs, those not sampled for *Lottia* (371) were discarded. The 2,259 remaining alignments were used to build pHMMs for HaMStR with hmmbuild and hmmcalibrate from the HMMER package (Eddy 2011).

Dataset Construction

Translated transcripts for all 74 taxa were then searched against the 2,259 Lophotrochozoa-Kocot pHMMs in HaMStR 13 using the default options. The "-representative" option was not used because it is not compatible with PhyloTreePruner, and the "-strict" option could not be used because not all taxa in the core OG set were sampled for all genes (only the primer taxon *Lottia* was guaranteed to be sampled for all genes). Sequences matching an OG's pHMM were compared to the proteome of *Lottia* using BLASTP. If the *Lottia* amino acid sequence contributing to the pHMM was the best BLASTP hit in each of these back-BLASTs, the sequence was then assigned to that OG.

In order to reduce missing data, sequences shorter than 50 amino acids were deleted and OGs (323) sampled for fewer than 50 of the 74 taxa were discarded. Redundant sequences that were identical (at least where they overlapped) were then removed with UniqHaplo (http://raven.iab.alaska.edu/~ntakebay/), leaving only unique sequences for each taxon. In theory, this approach could result in the unnecessary deletion of sequences if two or more different taxa had identical sequences. Spot-checking the number of sequences sampled for the two closely related *Tubulanus polymorphus* OTUs revealed no such problem in practice. Next, in cases where one of the first or last 20 characters of an amino acid sequence was an X (corresponding to a codon with an ambiguity, gap, or missing data), all characters between the X and that end of the sequence were deleted and treated as missing data. This step was retained from an earlier version

of our pipeline where it was important because 454 contig ends containing Xs are often obviously mistranslated. Each OG was then aligned with MAFFT (Katoh et al. 2005) using the automatic alignment strategy with a "maxiterate" value of 1,000. Alignments were then trimmed with Aliscore (Misof and Misof 2009) and Alicut (Kück 2009) with the default options to remove ambiguously aligned regions. Next, a consensus sequence was inferred for each alignment using the EMBOSS program infoalign (Rice et al. 2000). For each sequence in each single-gene amino acid alignment, the percentage of positions of that sequence that differed from the consensus of the alignment were calculated using the infoalign's "change" calculation. Any sequence with a "change" value greater than 75 was deleted. This step helped exclude incorrectly aligned sequences. Subsequently, a custom script was used to delete any putatively mistranslated sequence regions; these regions contained 20 or fewer amino acids in length surrounded by ten or more gaps on either side. At this point, OGs with alignments shorter than 50 amino acids in length (248 OGs) were discarded. Lastly, we deleted sequences that did not overlap with all other sequences in the alignment by at least 20 amino acids, starting with the shortest sequences not meeting this criterion. This step was necessary for downstream single-gene phylogenetic tree reconstruction. Finally, OGs sampled for fewer than 50 taxa (653 OGs) were discarded.

In some cases, a taxon was represented in an OG by two or more sequences (splice variants, lineage-specific gene duplications [=inparalogs], overlooked paralogs, or exogenous contamination). To select the best sequence for each taxon and help exclude overlooked paralogs or exogenous contamination, we built approximate maximum likelihood trees in FastTree 2 and used PhyloTreePruner to select the best sequence for each taxon as described above. Only OGs sampled for at least 50 taxa after pruning with PhyloTreePruner were retained. In addition to reducing paralogs, this approach should also help exclude contamination such as foreign sequences coming from gut contents, epibionts, or "bleed-through" during Illumina sequencing.

Further screening for paralogs and exogenous contamination was implemented using TreSpEx 1.0 (Struck 2014). First, single-gene trees were constructed for each OG in RAxML 7.7.6 (Stamatakis, 2014). Next, the TreSpEx *a priori* paralogy screening function based on bootstrap support was used (TreSpEx.v1.pl -fun a -gts Y -lowbs 95 -upbs 100 -possc 1 -poslb 2 -lowbl 4 -upbl 4 -possb 3 -maxtaxa 3 -blt 0.00001). As strong bootstrap support of 95 or higher for a clade in a single-gene tree might also stem from true phylogenetic signal, groups with *a priori* evidence of monophyly (i.e., Mollusca, Brachiopoda, Phoronida, Annelida, Nemertea, Platyhelminthes, Syndermata, Gastrotricha, Entoprocta, and Ecdysozoa) were masked with the "-gts" option for the further analyses. This screening revealed no cross-contaminating sequences in the OGs according to the short branch criterion (-possb; see Struck 2014 and TreSpEx manual for details). Second, sequences flagged as possible paralogs were then screened using a BLAST-based approach (TreSpEx.v1.pl -fun c -ppf Pruned_PotentialParalogsBootstrap.txt -ipt trees.txt -ipa alignments.txt -db1 Tribolium_castaneum -db2 Apis_mellifera -ediff 5 -ltp 0.1 -utp 0.85 -evalue 1e-20). Sequences indicated as "certain paralogs" after this BLAST search were excluded. Furthermore, all flagged sequences, for which this BLAST search did not return a hit allowing certain assessment of paralogy, were subjected to a second round of three BLAST searches with different databases each consisting of two paired species (i.e., *Drosophila melanogaster*/*Caenorhabditis elegans*, *Mus musculus*/*Bos taurus*, and *Capitella teleta*/*Lottia gigantea*). This was done to increase the likelihood to retrieve a hit allowing certain assessment of paralogy. Again, all sequences indicated as certain paralogs at this round in any of the three searches were excluded. Finally, to be conservative, all flagged sequences that still did not return a hit allowing certain assessment of paralogy after the second round were also excluded. Pruning of excluded sequences from the OGs was done with the aid of TreSpEx. After screening for and excluding paralogs with TreSpEx, all remaining 638 OGs were concatenated using FASconCAT

(Kück and Meusemann 2010) to make the “complete dataset” (Supplementary Table 3).

Sensitivity Analyses

We sought to assess factors that could cause systematic error. Therefore, for each OG we calculated five indices: 1) standard deviation of branch-length heterogeneity (LB; Struck 2014); 2) average patristic distance (PD); 3) percent missing data; 4) amino acid composition bias as measured by relative composition frequency variability (RCFV; Zhong et al. 2011b, Nesnidal et al. 2013); and 5) saturation as measured by the slope of patristic distance versus uncorrected p-distance (Nosenko et al. 2013). These factors will henceforth be referred to as LB, PD, missing data, RCFV, and saturation, respectively. Put simply, larger values for LB, PD, missing data, and RCFV are 'worse' (more likely to cause systematic error) than smaller values, in general, whereas larger values for our measure of saturation are 'better' than smaller values. These factors have been shown by recent studies (e.g., Philippe et al. 2011; Roure et al. 2013: Nesnidal et al. 2013: Nosenko et al. 2013: Struck et al. 2014) to be those that pose the most risk to contributing systematic error to phylogenomic analyses. LB, PD, and saturation were calculated using TreSpEx (Struck 2014) and RCFV and missing data were calculated using BaCoCa (Kück and Struck 2014).

For each of the five factors examined (Supplementary Fig. 1), we sorted OGs from 'best' to 'worst' by sextiles and constructed a series of increasingly smaller (more stringent) data matrices with the best 5/6, 4/6, 3/6, 2/6, and 1/6 OGs according to each factor. Our naming convention (Table 1) for these matrices indicates the factor being examined and the number of OGs remaining after some 'bad' OGs were deleted. For example, the dataset comprising the best 532 out of all 638 OGs with respect to LB is named LB_532; this indicates that the worst 1/6 OGs according to LB have been deleted. Because differences in the number of OGs within matrices

may confound comparisons, we also conducted analyses of each sextile so that differences in topologies resulting from any two sextiles (e.g. the sextile containing the best 1/6 OGs versus the sextile containing the second-best 1/6 OGs) could be directly compared. Our naming convention for these matrices indicates the factor being examined and the ranked sextile being examined. For example, LB_6of6 contains the 106 OGs in the sixth-best (i.e., worst) sextile of OGs based on LB whereas LB_2of6 contains the 107 OGs in the second-best sextile. Further, in an attempt to simultaneously reduce systematic error introduced by all five of the examined factors, we assembled datasets containing 1) only OGs that were ranked in at least the best 5 sextiles (i.e., the top 532 OGs of each category) for all five categories (Best_296_all_cat; 296 OGs) and 2) only OGs that were ranked in at least the best 4/6 in all five categories (Best_135_all_cat; 135 OGs). Attempts to assemble stricter subsets (e.g., OGs ranked in the best 3/6 or better for all 5 categories) resulted in small datasets with fewer than 100 genes, which were not considered further. In order to examine the relative influence of each of these factors on each other, a principal component analysis (PCA) was conducted on the data for LB, PD, missing data, RCFV, and saturation presented in Supplementary Table 3 using R (R Core Team 2013).

We also conducted taxon-based sensitivity analyses to examine the effects of removing taxa. Based on the distribution of LB scores among sampled taxa, we identified three natural breaks in our dataset (Supplementary Fig. 1a). First we excluded taxa (6) with an LB score at or above 39.21. Next we excluded taxa (13) with an LB score at or above 15.90. In order to examine placement of *Bugula* (Bryozoa) and *Symbion* (Entoprocta), who both had LB scores above 15.90, we conducted three additional analyses where we systematically restored *Bugula*, *Symbion*, or both to determine if removal of other taxa with long branches would improve support for their placement. Note that per-OG and per-taxon branch-length heterogeneity are calculated differently (Struck 2014), but the same abbreviation is used herein to refer to both indices for simplicity.

Similarly, based on the distribution of missing data in the sampled taxa, we identified two breaks in our dataset (Supplementary Fig. 1c). First, taxa (3) with missing data values greater than 80.0% were excluded. Next, taxa (25) with missing data values greater than 37.8% were excluded. Likewise we identified two natural breaks in the taxa in our dataset with respect to RCFV (Supplementary Fig. 1d). First taxa (2) with RCFV above 0.00107 were excluded. Next, taxa (13) with RCFV above 0.00063 were excluded. Leaf stability was calculated for each taxon in Roguenarok (http://rnr.h-its.org/) based on the RAxML bootstrap file from the analysis of the "complete dataset" (see below).

Hierarchical Clustering Analysis of Missing Data

Hierarchical clustering was conducted to determine if missing data exhibited a pattern of non-randomness, which could have an effect on phylogenetic reconstruction (Lemmon et al. 2009 Roure et al. 2013). We used BaCoCa (Kück and Struck 2013) to calculate the degree of missing data in common for each taxon pair in the complete dataset. From this taxon vs. taxon matrix, BaCoCa uses R (R Core Team 2013) to generate hierarchical clustering diagrams and heatmaps. If any taxa group together in both this hierarchical clustering analysis and the phylogenetic tree, this grouping is possibly due to shared missing data and should be treated with caution.

Phylogenetic Analyses

Phylogenetic analyses of all datasets were conducted using ML with the MPI version of RAxML 7.7.6 on the Auburn University CASIC HPC system with up to 100 CPUs used per analysis. Matrices were partitioned by gene and the PROTGAMMALGF model was used for all partitions. Spot-checks on haphazardly selected datasets using ProtTest revealed that LG was the best-fitting model for the vast majority of OGs, but preliminary ML analyses of these

datasets using the best-fitting model for each OG consistently recovered trees with identical branching patterns and comparable branch lengths and support values (data not shown). Thus, this step was omitted to reduce the computational and organizational complexity of this project. For each ML analysis, the tree with the best likelihood score after 10 random addition sequence replicates was retained and topological robustness (i.e., nodal support) was assessed with 100 replicates of rapid bootstrapping (the -f a command line option was used). For discussion purposes, support values below 70 are considered weakly supported, values between 70 and 90 are considered to have moderate support, and those above 90 are considered strongly supported.

Bayesian inference (BI) analyses were attempted using Phylobayes MPI 1.5a (Lartillot et al. 2013) on the Auburn University CASIC HPC system with 8 CPUs per chain. The CAT-GTR model was used to account for site-specific rate heterogeneity. Because of their computational intensity (Whelan and Halanych accepted manuscript), BI analyses were only run on datasets corresponding the 'best' 1/6 of the OGs according to each of the five factors examined. All BI analyses were conducted with four parallel chains run for around 15,000-20,000 cycles each (nearly six months of run time using 8 CPU cores per chain). Manual examination of '.trace' files produced by Phylobayes with Tracer (http://tree.bio.ed.ac.uk/software/tracer/) revealed that the vast majority of the variation in likelihood score occurred within the first 3,000-5,000 sampled trees (roughly 25% to 33% of the sampled) for all analyses. We discarded the first 5,000 trees from each chain as burn-in and calculated a 50% majority rule consensus tree from the remaining trees from each chain. Phylobayes bpcomp values of >0.3 for all five analyses (1.0 for most) indicated that the chains had not converged according to this strict measure. To further assess convergence of Bayesian Inference analyses, we calculated average standard deviation of spits frequencies (ASDSF; Ronquist et al. 2012). For each best 1/6 dataset, PhyloBayes tree files for the four independent chains were imported into MrBayes version 3.2.5

(Ronquist et al. 2012). ASDSF was calculated using the sumt command with 25% trees discarded as burn-in.

Correlation of Bootstrap Support and Number of Positions

Sensitivity analyses based on a reduced subset of OGs may have reduced support for a node simply because the analysis is based on a smaller data matrix. Thus, for the complete dataset and all subsets with complete taxon sampling based on the 'best' genes according to each of the five factors examined (e.g., LB_532 to LB_106 but not LB_6of6 to LB_2of6), we plotted bootstrap support versus the number of positions in the data matrix for a set of key phylogenetic hypotheses and determined if there was a significant correlation using linear regression.

Single-Gene Tree Congruence with Complete Dataset

Following the approach of Sharma et al. (2014), we used parse_gene_trees.py (https://github.com/claumer) to examine the number of genes that supported certain nodes of interest in the tree recovered by ML analysis of the complete dataset (Fig. 1). We examined all 638 single-gene trees and identified the number of OGs potentially decisive for a given node (those that sample at least one member of each descendant lineage of the investigated node plus at least two outgroups) and identified the number of genes within that set that are congruent with that node. For hypotheses that were not recovered in the analysis of the complete dataset, we enforced constraints on tree topology and examined the number of OGs congruent with these hypotheses. Constraint trees were generated in RAxML as described above except a constraint topology was provided via the -g flag.

RESULTS AND DISCUSSION

Data Matrices

Our bioinformatic pipeline retained only OGs inferred to be orthologous among the nine taxa used to construct the Lophotrochozoa-Kocot core ortholog set (2,259 OGs total). This resulted in a final matrix ("complete dataset") of 638 OGs totaling 121,980 amino acid positions in length. After trimming with Aliscore and Alicut, the average OG length was 191 amino acids and the longest was 415 amino acids. All OGs were sampled for at least 50 taxa but some were sampled for as many as 71 (of 74) taxa with an average of 57 taxa sampled per OG. Missing data in the complete dataset was 28.88% (71.12% matrix occupancy). Annotations and characteristics of each OG including length, number of taxa sampled, and values for each of the five factors examined are presented in Supplementary Table 3. In addition to the complete dataset, we assembled 66 other data matrices for sensitivity analyses examining the effects of removing genes or taxa with poor scores for LB, PD, missing data, RCFV, and slope (Table 1).

Phylogenetic Analysis of Complete Dataset

ML analysis of the complete dataset (Fig. 1) strongly supported a clade comprising all lophotrochozoans (not including the chaetognath *Sagitta*; bootstrap support, bs = 100). Likewise, Trochozoa including Mollusca, Brachiopoda, Phoronida, Nemertea, and Annelida was also strongly supported (bs = 99). Mollusca was sister to a weakly supported clade including Brachiopoda, Phoronida, Nemertea, and Annelida. Brachiopoda and Phoronida were recovered as sister taxa with strong support (bs = 99) and sister to Annelida + Nemertea, whose relationship was weakly supported. Sister to Trochozoa was a clade (bs = 64) composed of a strongly supported Platyzoa (bs = 99) and a weakly supported Polyzoa. Within Platyzoa, Gastrotricha was

recovered sister to a paraphyletic Gnathifera with Gnathostomulida sister to Platyhelminthes (bs = 90). Within Polyzoa, the bryozoan *Bugula* was sister to a strongly supported (bs = 100) clade of Cycliophora+Entoprocta. With exception of relationships among some conchiferan molluscs, all bootstrap support values within phyla were ≥92.

Principal Component Analysis

We used a principal component analysis (PCA) to understand covariation of five factors previously suggested to contribute to systematic error (saturation, branch length heterogeneity [LB], percent missing data, average patristic distance [PD], and compositional heterogeneity [RCFV]) across the 638 orthologous groups (OGs) considered in this study. This analysis showed that the first principal component (PC1) explains 34.7% of the variance present in different measurements across all OGs and the second principal component (PC2) accounted for another 22.6% of the variance (Fig. 2, Supplementary Table 4). RCFV and PD strongly contribute to PC1 (PC values of 0.61 and 0.60, respectively) and their load vectors are longest along the x-axis. The load vector for RCFV is nearly a horizontal line indicating that it makes virtually no contribution to PC2. LB contributes the strongest to PC2 (0.68) with missing data being the second strongest contributor (0.54). Hence, of all measurements, RCFV and PD contributed most to the variance observed in the dataset across the different OGs.

Amino acid compositional heterogeneity shows strong correlation with PD but not with missing data, branch-length heterogeneity, or saturation in this dataset. Not surprisingly, the contribution of both RCFV and PD to PC1 is equally strong. This is different for PC2 where LB contributes more strongly than missing data. This might be explained by the fact that LB negatively contributes to PC1 indicating that it is not positively correlated with RCFV and PD. In contrast, missing data is slightly positively correlated with the two, but also with LB. Thus,

similar results in the sensitivity analyses based on either amino acid compositional heterogeneity (RCFV) or evolutionary rate (PD) might not be as independent as assumed *a priori* as they, at least partially, measure the same variation across the OGs. Stronger independence seems to be present for LB, missing data and saturation, although saturation and LB exhibit similar load vectors in the first two principal components, which may also be indicative of partial covariance.

Our results indicate that amino acid compositional heterogeneity (RCFV) and evolutionary rate (PD) are correlated, which is not surprising even though they are usually treated as independent variables. Evolutionary rate depends on the number of substitutions in an OG. The more substitutions per sequence, the higher the rate of evolution. Because higher rates of evolution will cause substitution biases to accumulate faster, individual sequences are likely to deviate more strongly from the average amino acid composition in the dataset, and hence measurements of compositional heterogeneity like RCFV also increase.

On the other hand, measures of overall substitutional rate, like average patristic distance (PD), are often used as a proxy to detect long-branch attraction assuming that long-branch attraction is primarily caused by fast-evolving genes. Hence, evolutionary rate and long-branch attraction are not treated independently. In contrast, our PCA showed that PD was not strongly correlated with LB. Whereas LB score is a proxy of actual branch-length heterogeneity (i.e., difference in substitutions across taxa within an OG), PD gauges the overall substitutional rate of an OG. OGs can have similar average pairwise PDs, but variation in individual rate measurements among taxa may show considerable, or limited, heterogeneity. Likewise, OGs can have similar branch-length heterogeneity, but the overall average PD can differ. Consider two OGs with all else being equal including their gene genealogy, but one OG has twice the overall evolutionary rate and hence each individual branch in the tree is twice as long. In this case the branch-length heterogeneity would be identical, but the average PD would differ by a factor of 2.

Based on the PCA results, saturation and LB were more correlated than saturation and overall substitution rate (PD). Model-based approaches like ML or BI performed worse at modelling evolutionary rates accurately in heterotachous datasets (Kolaczkowski and Thornton 2004; Gadagkar and Kumar 2005; Roure and Philippe 2011), that is datasets with strong branch-length heterogeneity. Hence, correction for saturation in such datasets may be less effective. The stronger correlation of saturation and branch-length heterogeneity could reflect these problems associated with such heterogeneous datasets. Given these considerations and the PCA results, amino acid compositional heterogeneity and evolutionary rate should not be treated as independent variables in sensitivity analyses *a priori*. The same is true for hand branch-length heterogeneity and saturation. The common use of measurements of evolutionary rate, like PD, is not a good predictor of long-branch attraction and instead we advocate that direct measurements of branch-length heterogeneity, like LB, should be used.

Sensitivity Analysis (i) Branch-Length Heterogeneity

We examined the standard deviation of branch-length heterogeneity (LB) on a per OG basis. We note that branch-length heterogeneity did not necessarily result in long-branch attraction, but it is a prerequisite for this phenomenon to occur. ML analysis of the LB_532 dataset, in which the "worst" 1/6 OGs according to LB were deleted (dataset naming convention in Table 1 and Methods), resulted in the same branching pattern (Supplementary Fig. 3) as the complete dataset with comparable support at key nodes. However, ML analyses of all other trimmed LB datasets (LB_425, LB_319, LB_213, LB_106) resulted in a different topology within Trochozoa (Supplementary Figs. 4-6, Fig. 3). Support for Annelida sister to Brachiopoda + Phoronida showed a trend of increasing support as OGs with high LB were excluded (Table 2), even though the overall number of OGs analyzed decreased. A Student's t-test was significant (α-level of 0.05;

Supplementary Table 6) for observed negative correlations between LB and bootstrap support for Annelida sister to Brachiopoda + Phoronida and this grouping sister to Mollusca. Moreover, bootstrap support values for two alternative hypotheses (Annelida + Nemertea and Annelida + Nemertea sister to Brachiopoda + Phoronida) were significantly positively correlated with LB. Thus, bootstrap support for these hypotheses, which was not correlated with number of positions, increased with increasing LB scores. ML analyses of the dataset with the most reduced LB (LB_106) yielded the strongest support of all ML analyses for relationships among trochozoan phyla (bs = 85 for Annelida sister to Brachiopoda + Phoronida and bs = 95 for Mollusca sister to that clade; Fig. 3). ML analysis of the same matrix with all non-trochozoans excluded (LB_106_no_outgroup; Supplementary Fig. 7) also resulted in a topology consistent with Brachiopoda + Phoronida sister to Annelida. Support for a clade including Entoprocta + Cycliophora and Trochozoa was also significantly negatively correlated with LB, suggesting that topologies placing Entoprocta + Cycliophora in a clade with Bryozoa and/or Platyzoa might be due to LBA.

Struck et al. (2014) examined lophotrochozoan phylogeny excluding taxa and OGs most likely susceptible to long-branch attraction and recovered Platyzoa as a paraphyletic assemblage. ML analyses of our datasets, in contrast, recovered Platyzoa monophyletic. However, support for Platyzoa decreased as OGs with high values for LB were excluded (Table 2; Supplementary Table 6). Most notably, bootstrap support for Platyzoa drops to a mere 17 in the ML analysis of LB_107. Student's t-test showed that LB score and bootstrap support for Platyzoa were significantly positively correlated, while this is not the case for bootstrap support and number of positions (Supplementary Table 6). Similar significant positive correlation of bootstrap support was observed with LB score, but not with number of positions, for Polyzoa, Platyzoa+Polyzoa, and Bryozoa + Platyzoa, all independent of the number of positions employed.

Datasets most stringently trimmed according to LB have strongest support in ML analyses for relationships among trochozoan phyla, suggesting that removing OGs with high branch-length heterogeneity reduces conflict in our data, at least for this region of the tree. Notably, examination of the density distribution plot for LB (Supplementary Fig. 1a) reveals a dramatic tail of around 100 OGs with very high values for LB (>51.0) relative to the majority of OGs (most OGs had LB scores around 13-51). However, removal of these apparent outliers (LB_532; Supplementary Fig. 3) recovered the same branching order as analysis of the complete dataset. Manual examination of some of the 100 single-gene alignments with the highest LB scores revealed a small number (usually 1-5) of incorrectly aligned or partially mistranslated sequences in an otherwise high-quality alignment. This explains why removal of these apparently "very bad" OGs in terms of LB did not affect overall tree shape. Further reduction of datasets by removing the worst OGs in terms of LB score shifted support in favor of Annelida sister to Brachiopoda + Phoronida relative to analysis of the complete dataset. Given our observation of OGs in the tail of the LB score density plot with a small number of problematic sequences, we suggest that future studies might benefit from calculating single-OG LB scores and manually evaluating alignments with exceptionally high LB scores prior to concatenation and supermatrix analysis. This would be a much faster way to implement some manual dataset refinement without the arduous task of manually examining every OG as performed by Kocot et al. (2011). Notably, employing this step prior to screening OGs for paralogy with PhyloTreePruner (Kocot and Citarella et al. 2013) may have prevented some OGs from being discarded at that step.

Analyses were also conducted on datasets corresponding to the remaining five sextiles of the dataset with respect to LB (LB_6of6 - LB_2of6, LB = 29.54-19.19; Supplementary Figs. 8-12) to examine effect of LB without the influence of differences in matrix size. There was variability in tree topology and support among analyses, but no clear patterns were apparent.

We also examined branch-length heterogeneity score on a per-taxon basis (LB ≥15.90 and LB ≥39.21) and inclusion of *Bugula* and *Symbion*. For all five branch-length heterogeneity analyses, resulting tree topologies and support values were comparable to the analysis of the complete dataset (Supplementary Figs. 13-17). If Polyzoa is an artifact (see below), removal of the long-branched taxon *Symbion* does not affect support for this node (bs = 100, Supplementary Fig. 13). By excluding taxa with LB scores above the natural breaks identified in our dataset, we excluded all platyzoans except Gastrotricha and are thus unable to compare placement of most platyzoan phyla in this analysis with our other analyses and those of Struck et al. (2014) or Laumer et al. (2015).

Sensitivity Analysis (ii) Average Patristic Distance

To assess potential influences of fast- and slow-evolving genes on the reconstruction, we also conducted sensitivity analyses examining PD. ML analyses excluding the most quickly evolving 1/6 to 4/6 OGs according to PD (PD_532 - PD_213) recovered the same branching pattern and comparable support as the analysis of the complete dataset (Supplementary Figs. 18-21). However, ML analysis of the dataset composed of only the most slowly evolving 1/6 OGs (PD_106) according to PD yielded different relationships within Trochozoa, but support for this topology was weak (Fig. 3). Support for Platyzoa decreased as groups with high values for PD were excluded (Table 1) and analysis of the most slowly evolving 1/6 OGs in terms of PD (PD_106; Fig. 4) even found weak support for platyzoan paraphyly, consistent with Struck et al. (2014) and Laumer et al. (2015). According to a Student's t-test, the positive correlation of bootstrap support for Platyzoa and PD was significant, but so was the correlation of bootstrap support and the number of positions analyzed (Supplementary Table 6).

Analyses of the remaining five sextiles of the dataset with respect to PD (PD_6of6 -

PD_2of6) recovered various topologies and support within Trochozoa was weak in all of these single-sextile analyses (Supplementary Figs. 22-26). However, support for relationships among trochozoan phyla was strong but consistent with the analysis of the complete dataset when outgroups were removed from PD_106 (Supplementary Fig. 27). In summary, excluding OGs with high average patristic distance (i.e., fast-evolving OGs) appears to favor a close relationship of Mollusca, Brachiopoda+Phoronida, and Annelida as well as platyzoan paraphyly, but the possibility of these results simply being due to a decrease in the number of positions analyzed cannot be excluded.

Sensitivity Analysis (iii) Missing Data

Missing data can negatively influence phylogenetic inference by reducing the number of positions available for the detection of multiple substitutions (Roure et al. 2013). ML analyses of datasets with less missing data but fewer OGs (Missing_532 - Missing_106; Supplementary Figs. 28-31; Fig. 5) recovered the same branching pattern among phyla as the analysis of the complete dataset. For most nodes, bootstrap support had a tendency to decrease as the number of OGs decreased, even though the percentage of missing data also decreased. This is perhaps not surprising as even our complete dataset has less than 30% missing data. However, for Brachiopoda + Phoronida, Trochozoa, Polyzoa, and Platyzoa the degree of missing data and bootstrap support were significantly positively correlated, while in contrast, these factors were significantly negatively correlated for Lophophorata (Bryozoa + Phoronida + Brachiopoda) at a very low level of bs support (Table 2; Supplementary Table 6). In these cases where bootstrap support was correlated with the amount of missing data, the number of positions analyzed did not matter (no correlation to bootstrap support). Hence, decreasing the proportion of missing data even in an already well-covered dataset can have some influence, but will most likely affect the few taxa with

poor coverage. Interestingly, the bryozoan *Bugula* is among the taxa with worst coverage (84.93% missing data; Supplementary Table 1). Whereas reducing missing data strengthened support for a close relationship of Bryozoa to the other lophophorate taxa (Brachiopoda and Phoronida) in Nesnidal et al. (2013), such a relationship was not found in our analyses. In contrast, support for Annelida + Nemertea and a clade including Annelida, Nemertea, Brachiopoda, and Phoronida increased as missing data decreased, but this was also true with respect to the number of amino acid positions employed (Table 2, Supplementary Table 6). Interestingly, the PCA showed that missing data is not specifically strongly positively correlated with any of the other biases. However, with the exception of RCFV, other measurements of biases (LB, PD and saturation) tended to improve as the degree of missing data decreased (Table 1). Examination of the relationship of missing data and these other biases in other empirical datasets would be of great interest.

Analyses conducted on datasets corresponding to the remaining five sextiles of the dataset (Missing_6of6 - Missing_2of6; Supplementary Figs. 32-36) yielded variable relationships among trochozoan phyla in different analyses but support for inter-phylum relationships (aside from Brachiopoda + Phoronida) were generally weak in all analyses. Notably, the analysis of Missing_4of6 (31.61% missing data, Supplementary Fig. 34) recovered a monophyletic Lophophorata with Bryozoa sister to Phoronida as reported by Nesnidal et al. (2013), but support for this node was weak. Support for relationships among trochozoan phyla was strong and consistent with the analysis of the complete dataset when non-trochozoan taxa were removed from Missing_106 (Supplementary Fig. 37).

ML analysis of a dataset in which the three taxa with >80.0% missing data were excluded (Supplementary Fig. 1b; Missing_<_0.8; Supplementary Fig. 38) recovered Mollusca sister to Brachiopoda + Phoronida with this clade sister to Annelida although support was weak for both

nodes. Analysis of a more strictly reduced dataset in which taxa with >37.8% missing data were excluded (Missing_<_0.378, 49 remaining taxa; Supplementary Fig. 39) resulted in the same general branching pattern and level of support for trochozoan relationships as observed in the analysis of the complete dataset. Roure et al. (2013) and Sanderson et al. (2011) showed that the inclusion of incomplete but short-branched, slowly evolving taxa helps to ameliorate artifacts due to missing data. Although our removal of taxa with >37.8% missing data excluded some short-branched taxa, it excluded many more fast-evolving taxa including three of the four longest-branched taxa in the analysis of the complete dataset. Visualization of the distribution of missing data with hierarchical clustering (Supplementary Fig. 40) showed no correlation between shared missing data and tree topology. Taken together, these results indicate that missing data had little direct influence on our topology with the possible exception of the bryozoan *Bugula*.

Sensitivity Analysis (iv) Compositional Heterogeneity

Compositional heterogeneity has also been implicated as an important source of systematic error in analyses of Lophotrochozoa (Nesnidal et al. 2010, 2013; Zhong et al. 2011b). ML analyses on datasets with less compositional heterogeneity but fewer OGs (RCFV_532-RCFV_107; Fig. 6, Supplementary Figs. 41-44) generally recovered the same branching pattern as the analysis of the complete dataset with moderate to weak support for most inter-phylum relationships. In some analyses, Polyzoa was monophyletic and sister to either Platyzoa (RCFV_213 and RCFV 319; Supplementary Figs. 44 and 43, respectively) or Trochozoa (RCFV_423 and RCFV_532; Supplementary Figs. 42 and 41, respectively), although in the ML analysis of OGs with the least compositional heterogeneity (RCFV_107; Fig. 6), Bryozoa was sister to Platyzoa, and Entoprocta + Cycliophora was sister to Trochozoa (again with weak support). Analyses conducted on datasets corresponding to the remaining five sextiles of the dataset (RCFV_6of6 - RCFV_2of6;

Supplementary Figs. 45-49) showed no clear trends. When non-trochozoans were removed from RCFV_107, relationships among trochozoan phyla were consistent with the analysis of the complete dataset but support for Annelida + Nemertea dropped to 72 (Supplementary Fig. 50).

Details on amino acid composition for each of the sampled taxa are provided in Supplementary Table 5. When we excluded taxa with the highest RCFV scores (the flatworm *Schmidtea* and the rotifer *Brachionus*; RCFV_<_0.00107), relationships within Trochozoa were different but poorly supported (Supplementary Fig. 51). Specifically, deletion of just these two taxa resulted in a tree with Mollusca sister to Brachiopoda + Phoronida. Although support for the relative placement of Mollusca and Brachiopoda + Phoronida in both the analysis of the complete dataset and the analysis of RCFV_<_0.00107 was weak, this result is surprising because the deleted taxa are not trochozoans, but relatively distantly related platyzoans. These results further support previous assertions that inclusion of taxa with high RCFV values (deviant amino acid compositions) can influence placement of distantly related taxa during phylogenetic reconstruction (Zhong et al. 2011b). A topology with Mollusca sister to Brachiopoda + Phoronida was also recovered (but weakly supported) when we excluded the thirteen taxa with RCFV values above 0.00063 (RCFV_<_0.00063; Supplementary Fig. 52). Overall, our results were not sensitive to excluding OGs or taxa with high RCFV (at least among the remaining taxa).

Sensitivity Analysis (v) Saturation

ML analysis of a dataset where the most saturated 1/6 OGs were excluded (Slope_532; Supplementary Fig. 53) recovered Annelida sister to Brachiopoda + Phoronida with this clade sister to Mollusca and Nemertea sister to all other trochozoans (as seen in analyses of the best OGs in terms of LB), but these relationships were weakly supported. ML analyses of datasets with an intermediate amount of saturation (Slope_425 - Slope_214; Supplementary Figs. 54-56) resulted in

the same topology for Trochozoa as in the analysis of the complete dataset with weak support among phyla. ML analysis of just the least saturated 1/6 OGs (Slope_106; Fig. 7) recovered Mollusca sister to Brachiopoda + Phoronida with Annelida sister to this clade and Nemertea sister to all other trochozoans (again with weak support throughout). Interestingly, bootstrap support for a clade comprising Mollusca, Annelida, Brachiopoda, and Phoronida (all trochozoan taxa except Nemertea) was significantly positively correlated with the slope of patristic distance versus uncorrected p-distance (our measure of saturation; Supplementary Table 6). In contrast to the other measurements, a positive correlation here means that as saturation in the dataset is reduced, bootstrap support increases. Analyses conducted on datasets corresponding to the remaining five sextiles of the dataset (Slope_6of6 - Slope_2of6; Supplementary Figs. 57-61) yielded variable trees with weak support for most interphylum relationships and no clear trends in terms of support values. When non-trochozoans were removed from Slope_106, relationships among trochozoan phyla were consistent with the analysis of the complete dataset with moderate support for Annelida + Nemertea (bs = 89; Supplementary Fig. 62). Taken together, these results suggest that saturation may be an important factor influencing trochozoan relationships, particularly placement of Nemertea.

Sensitivity Analysis (vi) Most Stringent Selection of OGs According to All Five Factors

In addition to examining each of the five factors separately, two additional datasets were constructed based on 296 OGs ranked in the most stringent 5/6 for all five categories (Best_296_all_cat; Supplementary Fig. 63) and 135 OGs ranked in the best 4/6 for all five factors (Best_135_all_cat; Supplementary Fig. 64). ML analysis of Best_296_all_cat yielded the same branching pattern as analysis of the complete dataset with comparable support throughout the tree.

Analysis of further reduced Best_135_all_cat dataset yielded the same general branching pattern within Trochozoa, but relationships outside of this clade were markedly different. Aside from moderate support for Trochozoa (bs = 71), all lophotrochozoan inter-phylum relationships were weakly supported. Notably, the cycliophoran *Symbio*n was recovered within Ecdysozoa, possibly suggesting that our attempt to exclude all lobster contamination from this taxon may have failed. However, *Symbion* is a rather long-branched taxon so it is possible that this topology is due to long-branch attraction or simply inadequate signal for correct placement of this taxon.

Bayesian Inference Analyses

Given their computationally intensive nature, BI analyses (Supplementary Figs. 65-69) with the CAT-GTR model were only attempted on datasets with the most stringent 1/6 OGs for each factor examined. Despite running for nearly six months and ~15-20 thousand generations, Phylobayes bpcomp values indicated that the chains did not converge for any analysis (see Whelan and Halanych accepted manuscript). However, bpcomp >0.3 is a strict cutoff and examination of ASDSF values indicated that the BI analysis of Slope_106 analysis had converged (ASDSF = 0.0411). The resulting topology (Supplementary Fig. 68) had generally poor support for interphylum relationships although a clade of all lophotrochozoans except Gnathifera (and Chaetognatha) and Entoprocta + Cycliophora was strongly supported (posterior probability, pp = 1.0).

Summary of Conflict and Consensus Among Analyses

There is some consistency across our analyses, but also considerable incongruence at key nodes. Our ML analyses generally recovered the following groupings with strong support: Lophotrochozoa, Trochozoa (as Annelida, Brachiopoda, Mollusca, Nemertea, and Phoronida), Brachiopoda + Phoronida, and Entoprocta + Cycliophora. Platyzoa was also recovered in most ML

analyses, but support for this grouping was more variable and often decreased as putative sources of systematic error were reduced. Here we discuss several phylogenetic hypotheses that warrant further discussion.

H1: (Mollusca,((Nemertea,Annelida),(Phoronida,Brachiopoda))) (Fig. 8a). ML analyses of large datasets with all taxa and datasets based on the most stringent OGs in terms of alignment-based scores (missing data and RCFV) recover H1, but most nodes are rarely strongly supported.

H2: (Nemertea,(Mollusca,Annelida,(Phoronida,Brachiopoda))) (Fig. 8b). When genes with high values for LB are excluded, ML analyses recover Annelida sister to Brachiopoda + Phoronida with the strongest support among trochozoan phyla of any ML analyses conducted herein. Excluding taxa with high values for RCFV or very high amounts of missing data in ML analyses favors Mollusca sister to Brachiopoda + Phoronida, which was also recovered in analyses where we analyzed only the most stringent OGs with respect to the tree-based measurements PD and slope. We note, however that Mollusca sister to Brachiopoda + Phoronida is not strongly supported in any analysis.

H3: Platyzoa (Fig. 8c) Most ML analyses conducted herein recovered a monophyletic Platyzoa, sometimes with strong support. However, as most sources of systematic error were reduced, support for Platyzoa decreased and ML analyses of some datasets (LB_3of6, PD_106, Missing_213, Missing_2of6, RCFV_4of6, Slope_106, Slope_6of6, Best_135_all_cat) even recovered platyzoan paraphyly, albeit with weak support. Moreover, the one BI analysis that successfully converged as judged by ASDSF, Slope_106, supported platyzoan paraphyly with a posterior probability of 1.0.

H4: Mollusca sister to Entoprocta + Cycliophora (Fig. 8d), partially consistent with the Tetraneuralia hypothesis, which groups Mollusca and Entoprocta as sister taxa based on shared morphological characters (Wanninger 2009). This originally morphology-based hypothesis is not supported in ML analyses conducted herein.

H5: Lophophorata. ML analyses generally recovered Bryozoa sister to Entoprocta + Cycliophora and never recovered Lophophorata in ML analyses completed herein.

H6: **Aculifera-Conchifera** (Fig. 8e). Virtually all analyses strongly support the reciprocal monophyly of the two major lineages of Mollusca: Aculifera (Aplacophora + Polyplacophora) and Conchifera (all other shelled molluscs). Aculifera also received maximal support in the one BI analysis that converged.

Evolutionary Implications

H1: Annelida + Nemertea

ML analysis of the complete dataset and datasets based on the most stringent OGs in terms of missing data and RCFV recover Nemertea sister to Annelida. However, this result is weakly supported in most of our analyses and has received virtually no support from other molecular studies (but see Struck and Fisse 2008; Laumer et al. 2015). Morphological evidence supporting a clade of Annelida + Nemertea is scant, although this relationship has been hypothesized before. Cavalier-Smith (1998) viewed Annelida (including Echiura and Pogonophora but excluding Sipuncula) and Nemertea as sister taxa in a clade he called "Vermizoa" (Fig. 8a). Although both phyla have a prominent circulatory system in larger animals, the developmental origins and

organization of this system are quite different (Turbeville 1983, Ruppert and Carle 1983). Given the decreasing support for this hypothesis as various putative sources of systematic error are reduced, we view a clade of Annelida and Nemertea improbable.

H2: Mollusca, Annelida, and Brachiopoda + Phoronida

Brachiopoda and Phoronida were recovered as reciprocally monophyletic sister taxa with strong support in virtually all analyses, consistent with previous studies (e.g., Halanych et al. 1995; Dunn et al. 2008; Paps et al. 2009a, 2009b; Hausdorf et al. 2010; Sperling et al. 2011). Although we sampled only four brachiopods, we sampled all major lineages and all of our results are in contrast to studies recovering Phoronida as a subclade of Brachiopoda (Cohen, 2000, 2013; Cohen and Weydmann, 2005; Santagata and Cohen, 2009).

Several of our analyses place Brachiopoda + Phoronida in a clade with Mollusca and Annelida to the exclusion of other phyla (Fig. 8b). Some analyses place Annelida sister to Brachiopoda + Phoronida, thus allying phyla with *bona fide* chaetae. Chaetae are present in most members of the annelid radiation and brachiopods. The similar morphology of annelid and brachiopod chaetae suggests that they evolved in the last common ancestor of Annelida and Brachiopoda and were secondarily lost in phoronids (Orrhage 1971, 1973; Gustus and Cloney 1972; Westheide and Russell 1992; Lüter and Bartolomaeus 1997; Schulze 2002). However, chaetae-like structures are present in juvenile octopods (Brocco et al., 1974) and possibly also a fossil gastropod(-like?) mollusc (Tomas and Vinther 2012). Gene expression studies examining chaetogenesis in brachiopods, annelids, and octopods may be important to help elucidate these relationships. Likewise, studies comparing the development of chaetae and aculiferan mollusc sclerites, which are similar to chaetae in some respects, would also be of great interest.

Results placing Mollusca sister to Brachiopoda + Phoronida ally the trochozoan phyla with

external biomineralized structures (presumably also lost in phoronids under this hypothesis). We emphasize, however, that this topology was never strongly supported despite being recovered in numerous analyses. A clade of Mollusca and Brachiopoda + Phoronida is interesting with respect to evolution of biomineralization but the phylogenetic significance of biomineralization in Lophotrochozoa is also unclear. In addition to molluscs and brachiopods, many annelids (e.g., Szabó et al. 2014), bryozoans (reviewed by Taylor et al. 2014), nemerteans (Rieger and Sterrer 1975b), and even some flatworms (Rieger and Sterrer 1975a,b) also secrete calcareous structures. Recent transcriptomic and proteomic studies comparing shell biomineralization in brachiopods and molluscs indicate that, while there are some conserved genes involved in the process in both phyla and the general principles operating are the same, the genetic machinery involved differs substantially (Jackson et al. 2015; Luo et al. 2015; Isowa et al. 2015).

H3: Platyzoa

Platyzoa (Cavalier-Smith, 1998) is a grouping of generally small animals that lack a coelom or other spacious body cavity (as is common in very small metazoans), but no uniting synapomorphy for the group has been hypothesized. Most platyzoans are direct developers (also common in very small metazoans) with the parasitic acanthocephalans and large-bodied, free-living polyclad flatworms being notable exceptions. Molecular support for platyzoan monophyly has generally been weak (Passamaneck and Halanych 2006; Dunn et al. 2008 [Myzostomida was nested within Platyzoa]; Hejnol et al. 2009; Witek et al. 2009) or lacking (Glenner et al. 2004; Todaro et al. 2006; Paps et al. 2009a, 2009b), but relatively few molecular studies have had adequate taxon sampling to address the issue. Notably, platyzoans tend to have long branches in molecular phylogenies and, as noted above, Dunn et al. (2008) discussed the possibility that Platyzoa could be an artefact of long-branch attraction.

Our ML analyses typically recover Platyzoa monophyletic but support is variable and often decreased as sources of systematic error were reduced (Fig. 8c). Moreover, the converged BI analysis of Slope_106 supports platyzoan paraphyly with a clade including Gnathifera and Chaetognatha sister to a clade (pp = 1.0) including all other lophotrochozoans. Interestingly, Gnathifera and Chaetognatha form a clade in this analysis, but support (pp = 0.69) is too weak to draw much attention to this result. Otherwise, monophyly of sampled Gnathifera has been supported by numerous morphological (e.g., Kristensen and Funch 2000; Sørensen 2003; Funch et al. 2005) and some molecular studies (Zrzavý 2003; Witek et al. 2009).

H4: Tetraneuralia

Previous molecular studies and our own analyses herein generally recover Cycliophora sister to Entoprocta with strong support (Passamaneck and Halanych, 2006; Baguñà et al. 2008; Hejnol et al. 2009; Paps et al. 2009b; Fuchs et al. 2010; Mallatt et al. 2012), but placement of this clade has been ambiguous. Comparative morphological studies of the creeping trochophore larvae of entoprocts and larval and adult molluscs (particularly chitons and solenogaster aplacophorans) have prompted the Tetraneuralia hypothesis (Wanninger et al. 2009). In particular, there are similarities in the nervous system (both have tetraneury and similar flask cells in the apical region of the trochophore; Wanninger et al. 2007, 2009; Haszprunar and Wanninger, 2008). Earlier work also suggested this relationship under the names Lacunifera and Sinusoida (Bartolomaeus 1993, Ax 1999) owing to similarities in the musculature, cuticle, sinusal circulatory system, and foot.

Despite morphological characters suggesting a close relationship of entoprocts and molluscs, virtually no molecular studies including our own (e.g. Fig. 8d) have supported this

relationship. Instead, molecular phylogenetic studies have generally supported Polyzoa including Bryozoa, Entoprocta, and Cycliophora (Hausdorf et al. 2007, 2010; Helmkampf et al. 2008; Struck and Fisse 2008; Witek et al. 2008; Hejnol et al. 2009; Nesnidal et al. 2010; but see Nesnidal et al. 2013, below). Likewise, 0 out of 638 potentially informative OGs supported Entoprocta or Entoprocta + Cycliophora sister to Mollusca, respectively. Interestingly, Tetraneuralia with Entoprocta + Cycliophora sister to Mollusca was also recoveredin one BI analysis by Kocot et al. (2011), but with weak support.

Examination of amino acid composition in Entoprocta and Cycliophora (Supplementary Table 5) revealed that *Symbion* and entoprocts other than *Pedicellina* (as well as the bryozoan *Bugula*) were also among the most compositionally heterogeneous taxa (RCFV = 0.0005-0.0008; Supplementary Table 1). Amino acid compositional bias in these taxa, which has been reported previously (Nesnidal et al. 2010, 2013), could make their placement particularly sensitive to model choice (Jermiin et al. 2004, Delsuc et al. 2005, Rodríguez-Ezpeleta et al. 2007). Future studies addressing placement of these taxa may benefit from recoding of amino acids (Susko and Roger 2007) or employing models less sensitive to compositional heterogeneity such as CAT–BP (Blanquart and Lartillot 2008) as advocated by Nesnidal et al. (2010). However, the only current implementation of this model is the non-parallelized nhphylobayes (Blanquart and Lartillot 2008), which is impractical for datasets of this size.

H5: Lophophorata

Prior to molecular work, Brachiopoda, Phoronida, and Bryozoa were thought to form a clade, Lophophorata, on the basis of the shared presence of a horseshoe-shaped feeding tentacular apparatus that is invaded by the mesocoelom (lophophore; Hyman 1959, Halanych 1996, Lüter 1997, Ax 2001). However, most molecular studies to date have not recovered Lophophorata (e.g.,

Mackey et al. 1996; Cohen et al. 1998; Zrzavý et al. 1998; Cohen 2000; Giribet et al. 2000; Peterson and Eernisse 2001; Mallatt and Winchell 2002; Anderson et al. 2004; Cohen and Weydmann 2005; Passamaneck and Halanych 2006; Dunn et al. 2008; Helmkampf et al. 2008; Hausdorf et al. 2010; Nesnidal et al. 2010; Sperling et al. 2011).

Nesnidal et al. 2013 found strong support for Lophophorata with Phoronida sister to Bryozoa in both ML and BI analyses. Examination of compositional heterogeneity showed that the Nesnidal et al (2013) dataset had significantly less compositional heterogeneity than that of their previous study (Nesnidal et al. 2010). The authors concluded that Polyzoa was an artifact due to compositional heterogeneity in data from Entoprocta, Cycliophora, and Bryozoa. Laumer et al. (2015) recovered Polyzoa and Brachiopoda+Phoronida in most ML analyses although support was generally weak. However, BI analysis of a trimmed dataset by Laumer et al. (2015) that also excluded the two most unstable taxa (the entoproct *Barentsia* and cycliophoran *Symbion)* recovered Bryozoa sister to Phoronida and this clade sister to Brachiopoda, all with maximal support. In a BI analysis of the untrimmed matrix including *Barentsia* and *Symbion*, the aforementioned relationships were the same except that *Barentsia* (Entoprocta) was recovered sister to Bryozoa and *Symbion* (Cycliophora) was recovered sister to Trochozoa. This interesting result prompts an additional hypothesis that Polyzoa could be monophyletic and sister to Phoronida within Lophophorata, but this receives no support in our analyses and of course placement of Cycliophora in Laumer et al. (2015) is at odds with this idea. We also observed significant compositional heterogeneity in our polyzoan taxa and, consistent with most previous molecular studies, none of our analyses supported Lophophorata. Conflict between our results and those of Nesnidal et al. (2013) suggest that Lophophorata may need to be reevaluated in future studies with improved taxon sampling (especially for Bryozoa) and models that deal well with compositional heterogeneity, if such analyses can be made computationally achievable.

H6: Aculifera and Conchifera

Recent studies examining deep molluscan phylogeny (Kocot et al. 2011; Smith et al. 2011; Vinther et al. 2012; Osca et al. 2014) have supported a deep split dividing Mollusca into two clades: Aculifera (including chitons and aplacophorans) and Conchifera (all other mollusc taxa). However, strong support for Aculifera in phylogenomic studies has been met with skepticism by some (Salvini-Plawen and Steiner 2014, Schrödl and Stöger 2014), in part because previous studies based on datasets dominated by nuclear ribosomal genes and mitochondrial genes (Giribet et al. 2006; Wilson et al. 2010, Stöger et al. 2013) typically recovered chitons and monoplacophorans in a clade that has been termed Serialia. Every analysis we conducted herein recovered Aculifera, usually with maximal support. In the context of heterogenous signal from different data partitions, we note that 70 single-OG trees (out of 635 potentially informative trees) recovered Aculifera monophyletic (Supplementary Fig. 70) whereas only 4 (out of 76 potentially informative trees) recovered Serialia.

Results of Kocot et al. (2011) and Smith et al. (2011) differed in relationships among Gastropoda, Bivalvia, and Scaphopoda and the fact that Kocot et al. (2011) did not sample Monoplacophora. Whereas most analyses in Kocot et al. (2011) recovered Gastropoda + Bivalvia with strong support and some analyses in Smith et al. (2011) recovered Gastropoda + Scaphopoda with strong support, in our present analyses, relationships among these taxa were highly variable and rarely with strong support. Interestingly, the traditional, morphology-based Diasoma hypothesis, which unites scaphopods and bivalves (molluscs with a “through body”), was rarely recovered here and only received strong support in analyses of the two worst sextets based on saturation and the second best sextet based on LB (Diasoma was significantly rejected by

phylogenomic data by Kocot et al. 2011 and an earlier analysis of 18S and 28S rDNA by Passamaneck et al. 2004). The position of Monoplacophora (represented by *Laevipilina*) was also variable among analyses. The most commonly recovered topology placed Monoplacophora sister to the rest of Conchifera, consistent with traditional morphological views (reviewed by Haszprunar et al. 2008; Kocot 2013; Schrödl and Stöger 2014).

Other hypotheses

In particular, three other hypotheses of animal relationships were not supported by any of our analyses: Eutrochozoa, Neotrochozoa, and Kryptrochozoa. Eutrochozoa (*sensu* Peterson and Eernisse 2001; Mollusca, Annelida, and Nemertea) has been hypothesized based on the presence of lateral coelomic sacs that develop through schizocoely with the mesoderm formed directly from the primary mesoblasts (reviewed by Nielsen, 2011). However, none of our ML analyses supported Eutrochozoa. Likewise, we failed to find support for Neotrochozoa (Mollusca + Annelida), which unites the phyla that have a canonical trochophore larva. Support for this topology in Kocot et al. (2011) may have been due to relatively limited taxon sampling outside of Mollusca and Annelida. Additionally, the vast majority of our results are inconsistent with Kryptrochozoa, a grouping of Nemertea, Brachiopoda, and Phoronida, taxa with a "hidden" (modified) trochophore. This result has been recovered in previous phylogenomic studies with more limited taxon and gene sampling for Lophotrochozoa (e.g., Dunn et al. 2008; Hejnol et al. 2009; Hausdorf et al. 2010, Nesnidal et al. 2010). In the few analyses where we do recover this topology, it is always weakly supported.

Implications for Phylogenomics

Here, we conducted a rigorous set of analyses in order to identify and reduce putative sources of systematic error in our dataset. Removal of OGs with high branch-length heterogeneity (high values for LB) had the greatest impact on tree topology and support, consistent with observations by Struck et al. (2014). Support for relationships within Trochozoa increased as OGs with high LB were excluded, even though the overall number of OGs analyzed decreased. At the same time, support for Platyzoa, a grouping hypothesized to be the result of long-branch attraction (Dunn et al. 2008, Struck et al. 2014), decreased. These observations indicate that some of the more dubious results of the analysis of the complete dataset (unconventional relationships within Trochozoa and strong support for Platyzoa) may be artifacts caused by branch-length heterogeneity. Of interest, most sampled trochozoans have comparable, moderate branch lengths, but relationships within Trochozoa are sensitive to exclusion of OGs with high LB scores. Thus, excluding OGs with high LB scores may be sensible in phylogenomic studies even if the sampled taxa have comparable branch lengths and long-branch attraction is not suspected. In addition, our PCA revealed that measures of overall evolutionary rate might not be a predictor of which OGs are susceptible to long-branch attraction.

Nesnidal et al. (2013) presented evidence that amino acid compositional heterogeneity, particularly in Entoprocta and Bryozoa, has misled previous phylogenomic investigations of lophotrochozoan relationships. Our exclusion of OGs with high compositional heterogeneity (high values of RCFV) still resulted in weakly supported trees inconsistent with the findings of Nesnidal et al. (2013). Although some OGs exhibit more compositional heterogeneity among taxa than others, compositional heterogeneity appears to be a more taxon-specific problem. This is of course problematic when we seek to place these taxa in a phylogenetic framework. Sampling phylogenetically and ecologically diverse representatives of such taxa and selecting only the least

compositionally heterogeneous exemplars could be one way to deal with this issue in future studies. The strategy of taxon rather than OG exclusion had already been successfully used in analyses of the placement of platyzoan taxa with respect to long-branched taxa (Struck et al. 2014). Although our taxon sampling spans much of the diversity of Entoprocta, improving taxon sampling and conducting analyses on taxa with less heterogeneous amino acid composition could be a game-changer for reliably placing Entoprocta and Bryozoa. Conducting analyses with models better suited for compositionally heterogeneous sequences could also help.

Significant amounts of missing data have been shown to be problematic in phylogenetic reconstruction (Roure et al. 2013). However, sensitivity analyses conducted herein excluding OGs with large amounts of missing data recovered the same general branching order among phyla as observed in the analysis of the complete dataset. Bootstrap support values tended to decrease as the number of OGs decreased, even though the percentage of missing data also decreased. In short, decreasing the proportion of missing data in a well-covered dataset like ours seems to have little influence but may be important for taxa with particularly poor coverage. Likewise, saturation can be problematic in phylogenomics (Philippe et al. 2011) but reduction of saturation in the present dataset also failed to yield strongly supported trees. However, the fact that the only BI analysis that converged was the analysis on the dataset in which saturation was reduced may suggest that reducing saturation is more important than it would seem.

CONCLUSIONS

In this study, we greatly expanded the amount of transcriptome data available for most major lineages of Lophotrochozoa, assembled a new taxon-specific HaMStR core ortholog set, and conducted a rigorous set of 67 phylogenomic analyses examining the evolutionary history of this group controlling for five factors known to cause systematic bias in phylogenetics. Although

branching pattern and support for some key nodes varied among analyses, we identified a reduced number of hypotheses of lophotrochozoan relationships that warrant further consideration. Best-supported ML analyses were recovered when branch-length heterogeneity (LB) was reduced. Specifically, when LB was reduced, we recover a sister taxon relationship between Annelida and Brachiopoda+Phoronida with Mollusca sister to this clade and Nemertea sister to the remainder of Trochozoa. Despite running for nearly six months, most BI analyses failed to converge. Interestingly, only the BI analysis in which saturation was reduced converged according to ASDSF. This analysis supported the paraphyly of Platyzoa, consistent with recent studies.

Unfortunately, pinpointing the one source of systematic error that seems to have the greatest impact on phylogenetic reconstruction for this area of the animal tree of life is difficult. However, branch-length heterogeneity was clearly problematic for many platyzoansand excluding genes with high LB scores yielded the overall best supported trees according to bootstrap support. Thus, exploring effects of reducing branch-length heterogeneity may be a good first step when trying to identify sources of systematic error in phylogenomic analyses. According to our principal component analysis, branch length heterogeneity and saturation are partially confounded. Interestingly, reducing saturation had the greatest effect on placement of Nemertea within Trochozoa. Compositional heterogeneity appears to also be an important issue to consider, at least in the context of this dataset. Specifically, compositional heterogeneity appeared to be a problem for entoprocts and bryozoans in this dataset (as seen before), and may have been an issue in other taxa as well.

Perhaps the two most important take-home messages from our sensitivity analyses are that lineage-specific issues such as branch-length heterogeneity (in platyzoans) and compositional heterogeneity (in entoprocts and bryozoans) likely need to be simultaneously addressed in order to resolve such difficult phylogenetic questions. Additionally, overall evolutionary rate is not

strongly correlated with branch-length heterogeneity as often assumed, but with amino acid compositional heterogeneity. Taken together, our results show that the five factors (branch-length heterogeneity/saturation, amino acid compositional heterogeneity/overall evolutionary rate, and precent missing data) examined can have important influence of topological reconstruction and should be routinely considered in phylogenomic studies. The approach employed here can be generally applied to any phylogenomic dataset to help identify and reduce sources of systematic error.

SUPPLEMENTARY MATERIAL

Supplementary material including all supplementary tables and figures, transcriptome assemblies, the "Lophotrochozoa-Kocot" core OG set, and all input/output files related to phylogenetic analyses can be accessed from the Dryad data repository (doi:10.5061/dryad.30k4v). Raw Illumina sequence data are available for download from NCBI SRA under Study numbers SRP059156 and SRP048758. Taxon-specific Experiment numbers are listed in Supplementary Table 1. Bioinformatic scripts used in this project can be downloaded from GitHub at https://github.com/kmkocot/.

ACKNOWLEDGMENTS

This work was funded by NSF DEB-1036537, IOS-0843473, DEB-1051106, and OCE-1155188 to K.M.H., NSF DEB-1210518 and DBI-1306538 to K.M.K, DFG STR-683/8-1 to T.H.S., DFG (Li 998/9-1/Feldbausch Foundation and the IUFF, both University of Mainz) to BL, and NSF grants IOS-1457162 and 0744649 to LLM. Antarctic collections were supported by NSF grants OPP9910164, OPP0338218, and ANT1043745 to K.M.H. Norwegian aplacophoran collection was funded by Norwegian Taxonomy Initiative project number 70184222 and The

Research Council of Norway project number 210460 to C.T. We thank Patrick Krug, Richard Heard, Jamie Baldwin, Dan Speiser, the crew and scientists of the BioSkag II cruise aboard the *R/V Håkon Mosby*, the crew of the *R/V Hans Brattström*, University of Bergen and Institute of Marine Research, Norway, the crew and scientists of the Icy Inverts cruises aboard the R/V L*awrence M. Gould* and *Nathaniel B. Palmer*, Friday Harbor Labs, and Heron Island Research Station for supporting specimen collection. We thank Nina Mikkelsen for identifying *Prochaetoderma californicum*. We thank Franzi Franke, Amanda Shaver, and Anja Anjuschka for help with cDNA library preparation and we thank Tatiana P. Moroz for library preparation and sequencing for *Sagitta*. We thank Nathan Whelan, Scott Santos, Jason Bond, and Christopher Laumer for helpful discussions and advice related to data analysis and interpretation. Animal images were downloaded without modification from PhyloPic.org under a creative commons license (http://creativecommons.org/licenses/by/3.0/). This is Auburn University Marine Biology Program contribution #146, Molette Lab contribution #X and NHM Evolutionary Genomics Lab contribution #2.

REFERENCES

Altenburger A., Wanninger A. 2010. Neuromuscular development in Novocrania anomala: evidence for the presence of serotonin and a spiralian-like apical organ in lecithotrophic brachiopod larvae. Evol. Dev. 12:16–24.

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.

Anderson F.E., Córdoba A.J., Thollesson M. 2004. Bilaterian phylogeny based on analyses of a region of the sodium–potassium ATPase beta-subunit gene. J. Mol. Evol. 58:252–268.

Ax P. 1999. Das System der Metazoa III. Stuttgart: Gustav Fischer Verlag.

Baguñà J., Martinez P., Paps J., Riutort M. 2008. Back in time: a new systematic proposal for the Bilateria. Phil. Trans. R. Soc. B. 363:1481-1491.

Bartolomaeus T. 1993. Die Leibeshöhlenverhältnisse und Verwandtschaftsbeziehungen der Spiralia. Verhandlungen Dtsch. Zool. Ges. 86:42.

Bergsten J. 2005. A review of long□branch attraction. Cladistics. 21:163-193.

Blanquart S., Lartillot N. 2008. A site-and time-heterogeneous model of amino acid replacement. Mol. Biol. Evol. 25:842–858.

Bleidorn C., Podsiadlowski L., Zhong M., Eeckhaut I., Hartmann S., Halanych K.M., Tiedemann R. 2009. On the phylogenetic position of Myzostomida: can 77 genes get it wrong? BMC Evol. Biol. 9:150.

Brocco S.L., O'Clair R.M., Cloney R.A. 1974. Cephalopod integument: The ultrastructure of Kölliker's organs and their relationship to setae. Cell Tissue Res. 151:293-308.

Brown C.T., Howe A., Zhang Q., Pyrkosz A.B., Brom T.H. 2012. A reference-free algorithm for computational normalization of shotgun sequencing data. arXiv:1203.4802.

Cavalier-Smith T. 1998. A revised six-kingdom system of life. Biol. Rev. 73:203-266.

Chou H.-H., Holmes M.H. 2001. DNA sequence quality trimming and vector removal. Bioinformatics. 17:1093-1104.

Cohen B.L., Gawthrop A., Cavalier-Smith T. 1998. Molecular phylogeny of brachiopods and phoronids based on nuclear-encoded small subunit ribosomal RNA gene sequences. Philos. Trans. R. Soc. Lond. Biol. 353:2039-2061.

Cohen B.L., Weydmann A. 2005. Molecular evidence that phoronids are a subtaxon of brachiopods (Brachiopoda: Phoronata) and that genetic divergence of metazoan phyla began long before the early Cambrian. Org. Divers. Evol. 5:253-273.

Cohen B.L. 2000. Monophyly of brachiopods and phoronids: reconciliation of molecular

evidence with Linnaean classification (the subphylum Phoroniformea nov.). Proc. R. Soc. B. 267:225-231.

Cohen B.L. 2013. Rerooting the rDNA gene tree reveals phoronids to be "brachiopods without shells"; dangers of wide taxon samples in metazoan phylogenetics (Phoronida; Brachiopoda). Zool. J. Linnean Soc. 167: 82-92.

Delsuc F., Brinkmann H., Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. 6:361–375.

Delsuc F., Brinkmann H., Chourrout D., Philippe H. 2006. Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature. 439:965–968.

Dordel, J., Fisse, F., Purschke, G., and Struck, T.H. 2010. Phylogenetic position of Sipuncula derived from multi-gene and phylogenomic data and its implication for the evolution of segmentation. *J. of Zool. Syst. and Evol. Res.* 48(3): 197-207.

Dunn C.W., Hejnol A., Matus D.Q., Pang K., Browne W.E., Smith S.A., Seaver E., Rouse G.W., Obst M., Edgecombe G.D., Sørensen M.V., Haddock S.H.D., Schmidt-Rhaesa A., Okusu A., Kristensen R.M., Wheeler W.C., Martindale M.Q., Giribet G. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature. 452:745–749.

Dunn C.W., Howison M., Zapata F. 2013. Agalma: an automated phylogenomics workflow. BMC Bioinformatics. 14:330.

Dunn C.W., Giribet G., Edgecombe G.D., Hejnol A. 2014. Animal Phylogeny and Its Evolutionary Implications*. Annu. Rev. Ecol. Evol. Syst. 45:371–395.

Ebersberger I., Strauss S., Von Haeseler A., 2009. HaMStR: Profile hidden Markov model based search for orthologs in ESTs. BMC Evol. Biol. 9:157.

Eddy S.R. 2011. Accelerated profile HMM searches. PLoS Comput. Biol. 7:e1002195.

Edgecombe G.D., Giribet G., Dunn C.W., Hejnol A., Kristensen R.M., Neves R.C., Rouse G.W., Worsaae K., Sørensen M.V. 2011. Higher-level metazoan relationships: recent progress and remaining questions. Org. Divers. Evol. 11:151.

Eernisse et al., 1992. Annelida and Arthropoda are not sister taxa: a phylogenetic analysis of spiralian metazoan morphology. Syst. Biol. 41:305-330

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Biol. 27:401–410.

Forment J., Gilabert F., Robles A., Conejero V., Nuez F., Blanca J. 2008. EST2uni: an open, parallel tool for automated EST analysis and database creation, with a data mining web interface and microarray expression data integration. BMC Bioinformatics. 9:5.

Fuchs J., Iseto T., Hirose M., Sundberg P., Obst M. 2010. The first internal molecular phylogeny of the animal phylum Entoprocta (Kamptozoa). Mol. Phylogenet. Evol. 56:370-379.

Funch P., Kristensen R.M., 1995. Cycliophora is a new phylum with affinities to Entoprocta and Ectoprocta. Nature 378:711-714.

Funch P., Sørensen M.V., Obst M. 2005. On the phylogenetic position of Rotifera – Have we come any further? Hydrobiologia 546:11-28.

Gadagkar, S.R., Kumar, S. 2005. Maximum likelihood outperforms maximum parsimony even when evolutionary rates are heterotachous. Mol. Biol. Evol. 22, 2139-2141.

Giribet G., 2008. Assembling the lophotrochozoan (=spiralian) tree of life. Phil. Trans. R. Soc. B. 363, 1513–1522.

Giribet G. 2014. On Aculifera: a review of hypotheses in tribute to Christoffer Schander. J. Nat. Hist. 48:2739–2749.

Giribet G. 2015. New animal phylogeny: future challenges for animal phylogeny in the age of

phylogenomics. Organisms Diversity & Evolution. doi:10.1007/s13127-015-0235-4.

Giribet G., Distel D.L., Polz M., Sterrer W., Wheeler W.C. 2000. Triploblastic relationships with emphasis on the Acoelomates and the position of Gnathostomulida, Cycliophora, Plathelminthes, and Chaetognatha: A combined approach of 18S rDNA sequences and morphology. Syst. Biol. 49:539-562.

Giribet G., Okusu A., Lindgren A.R., Huff S.W., Schrödl M., Nishiguchi M.K. 2006. Evidence for a clade composed of molluscs with serially repeated structures: Monoplacophorans are related to chitons. Proc. Natl. Acad. Sci. 103(20): 7723-7728.

Glenner H., Hansen A.J., Sørensen M.V., Ronquist F., Huelsenbeck J.P., Willerslev E. 2004. Bayesian inference of the metazoan phylogeny: A combined molecular and morphological approach. Current Biol. 14:1644-1649.

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., Palma F. di, 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. Nature Biotechnol. 29:644-652.

Gustus R.M., Cloney R.A. 1972. Ultrastructural similarities between setae of brachiopods and polychaetes. Acta Zool. 53:229-233.

Halanych K.M. 1996. Convergence in the feeding apparatuses of lophophorates and pterobranch hemichordates revealed by 18S rDNA: an interpretation. Biol. Bull. 190:1-5.

Halanych K.M. 2004. The new view of animal phylogeny. Annu. Rev. Ecol. Syst. 35:229-256.

Halanych K.M., Kocot K.M. 2014. Repurposed Transcriptomic Data Facilitate Discovery of Innate Immunity Toll-Like Receptor (TLR) Genes Across Lophotrochozoa. Biol. Bull. 227:201–209.

Halanych K.M., Bacheller J.D., Aguinaldo A.M.., Liva S.M., Hillis D.M., Lake J.A. 1995. Evidence from 18S ribosomal DNA that the lophophorates are protostome animals. Science 267:1641-1641.

Haszprunar G., Wanninger A. 2008. On the fine structure of the creeping larva of Loxosomella murmanica: additional evidence for a clade of Kamptozoa (Entoprocta) and Mollusca. Acta Zool. 89:137–148.

Hatschek B. 1878. Studien über Entwicklungsgeschichte der Anneliden: Ein Beitrag zur Morphologie der Bilaterien. A. Hölder.

Hausdorf B., Helmkampf M., Meyer A., Witek A., Herlyn H., Bruchhaus I., Hankeln T., Struck T.H., Lieb B., 2007. Spiralian phylogenomics supports the resurrection of Bryozoa comprising Ectoprocta and Entoprocta. Mol. Biol. Evol. 24:2723.

Hausdorf B., Helmkampf M., Nesnidal M.P., Bruchhaus I., 2010. Phylogenetic relationships within the lophophorate lineages (Ectoprocta, Brachiopoda and Phoronida). Mol. Phylogenet. Evol. 55:1121-1127.

Hejnol A., Obst M., Stamatakis A., Ott M., Rouse G.W., Edgecombe G.D., Martinez P., Baguñà J., Bailly X., Jondelius U., Wiens M., Müller W.E.G., Seaver E., Wheeler W.C., Martindale M.Q., Giribet G., Dunn C.W. 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. Proc. R. Soc. B. 276:4261–4270.

Helmkampf M., Bruchhaus I., Hausdorf B. 2008. Phylogenomic analyses of lophophorates (brachiopods, phoronids and bryozoans) confirm the Lophotrochozoa concept. Proc. R. Soc. B. 275:1927.

Henry, J. Q., Hejnol, A., Perry, K. J., & Martindale, M. Q. 2007. Homology of ciliary bands in spiralian trochophores. Int. Comp. Biol. 47(6):865-871.

Huang X., Madan A., 1999. CAP3: A DNA sequence assembly program. Genome Res. 9:868-877.

Hyman L.H., 1959. The Invertebrates: Smaller coelomate groups. McGraw-Hill.

Isowa Y., Sarashina, I., Oshima, K., Kito, K., Hattori, M., Endo, K. 2015. Proteome analysis of shell matrix proteins in the brachiopod *Laqueus rubellus*. Proteome Sci. 13(1):21.

Jackson, D.J., McDougall, C., Woodcroft, B., Moase, P., Rose, R.A., Kube, M., Reinhardt, R., Rokhsar, D.S. Montagnani, C., Joubert, C.,Piquemal, D., and Degnan, B.M. 2010. Parallel evolution of nacre building gene sets in molluscs. Mol. Biol. Evol. 27(3): 591-608.

Jackson D. J., Mann, K., Häussermann, V., Schilhabel, M. B., Lüter, C., Griesshaber, E., Schmahl, W., Wörheide, G. 2015. The *Magellania venosa* biomineralizing proteome: A window into brachiopod shell evolution. Genome Biol. Evol. 7(5):1349-1362.

James, M.A., Ansell, A.D., Collins, M. J., Curry, G. B., Peck, L.S. and Rhodes, M.C. 1992. Biology of living brachiopods. Adv. Mar. Biol. 28: 175-387.

Jeffroy O., Brinkmann H., Delsuc F., Philippe H. 2006. Phylogenomics: the beginning of incongruence? TRENDS Genet. 22:225–231.

Jermiin L.S., Ho S.Y., Ababneh F., Robinson J., Larkum A.W. 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. Syst. Biol. 53:638–643.

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.

Kocot K.M., Cannon J.T., Halanych K., 2010. Elucidating Animal Phylogeny., In: R. DeSalle, Schierwater B., editors. *Key Transitions in Animal Evolution*. Science Publishers, pp. 15-33.

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.

Kocot, K.M. Recent Advances and Unanswered Questions in Deep Molluscan Phylogenetics. 2013. Am. Malacological Bull. 31(1): 195-208.

Kocot, K.M. 2013. *A combined approach toward resolving the phylogeny of Mollusca.* Doctoral dissertation. Auburn University, Auburn, AL.

Kocot, K.M., Halanych, K.M., Krug, P.J. 2013. Phylogenomics supports Panpulmonata: Opisthobranch paraphyly and key evolutionary steps in a major radiation of gastropod molluscs. Mol. Phylogenet. Evol. 69(3):764-771.

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. Bioinformatics 2013:429.

Kocot, K.M. 2016. On 20 years of Lophotrochozoa. Org. Divers. Evol. 16(2): 329-343.

Kohn, A.B., Moroz T.P., Barnes J.P., Netherton M., Moroz L.L. 2013. Single-cell semiconductor sequencing. Methods Mol. Biol. 1048:247-284.

Kolaczkowski, B., Thornton, J.W. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431, 980-984.

Kristensen R.M., Funch P. 2000. Micrognathozoa: A new class with complicated jaws like those of Rotifera and Gnathostomulida. J. Morphol. 246:1-49.

Kück P. 2009. ALICUT: a Perl script which cuts ALISCORE identified RSS. Department of Bioinformatics, Zoologisches Forschungsmuseum A. Koenig (ZFMK), Bonn, Germany, version 2.

Kück, P., Meusemann, K. 2010. FASconCAT: convenient handling of data matrices. Mol. Phylogenet. Evol. 56(3):1115-1118.

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.

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(Suppl 1): S4.

Lartillot N., Rodrigue N., Stubbs D., Richer J. 2013. PhyloBayes MPI. Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst. Biol.:syt022.

Laumer C.E., Hejnol A., Giribet G. 2015. Nuclear genomic signals of the "microturbellarian"roots of platyhelminth evolutionary innovation. eLife. 4:e05503.

Laumer, C.E., Bekkouche, N., Kerbl, A., Goetz, F., Neves, R.C., Sørensen, M.V., Kristensen, R.M., Hejnol, A., Dunn, C.W., Giribet, G., Worsaae, K. 2015. Spiralian phylogeny informs the evolution of microscopic lineages. *Curr. Biol. 25*(15), 2000-2006.

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.

Li L., Stoeckert C.J., Roos D.S. 2003. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Res. 13:2178–2189.

Luo H., Arndt, W., Zhang, Y., Shi, G., Alekseyev, M. A., Tang, J., Hughes, A. L., Friedman, R. 2012. Phylogenetic analysis of genome rearrangements among five mammalian orders. Mol. Phylogenet. Evol. 65(3):871-882.

Lüter C., Bartolomaeus T., 1997. The phylogenetic position of Brachiopoda—a comparison of morphological and molecular data. Zoologica Scripta. 26:245-253.

Mackey L.Y., Winnepenninckx B., De Wachter R., Backeljau T., Emschermann P., Garey J.R., 1996. 18S rRNA suggests that Entoprocta are protostomes, unrelated to Ectoprocta.

J. Mol. Evol. 42:552-559.

Mallatt J., Craig C.W., Yoder M.J., 2012. Nearly complete rRNA genes from 371 Animalia: Updated structure-based alignment and detailed phylogenetic analysis. Mol. Phylogenet. Evol. 64(3):603-617.

Mallatt J., Winchell C.J., 2002. Testing the new animal phylogeny: First use of combined large-subunit and small-subunit rRNA gene sequences to classify the protostomes. Mol. Biol. Evol. 19:289-301.

Matus D.Q., Copley R.R., Dunn C.W., Hejnol A., Eccleston H., Halanych K.M., Martindale M.Q., Telford M.J. 2006. Broad taxon and gene sampling indicate that chaetognaths are protostomes. Curr. Biol. CB. 16:R575–576.

Meyer, E., G.V. Aglyamova, S. Wang, J. Buchanan-Carter, D. Abrego, J.K. Colbourne, B.L. Willis, and M.V. Matz. 2009. Sequencing and *de novo* analysis of a coral larval transcriptome using 454 GSFlx. BMC Genomics 10(1):219.

Minelli, Alessandro. 2009. *Perspectives in animal phylogeny and evolution*. Oxford: Oxford University Press.

Misof, B., Misof, K. 2009. A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: A more objective means of data exclusion. Syst. Biol. 58:21-34.

Moroz L.L., Kocot K.M., Citarella M.R., Dosung S., Norekian T.P., Povolotskaya I.S., Grigorenko A.P., Dailey C., Berezikov E., Buckley K.M. 2014. The ctenophore genome and the evolutionary origins of neural systems. Nature. 510:109–114.

Nesnidal M.P., Helmkampf M., Bruchhaus I., Hausdorf B. 2010. Compositional heterogeneity and phylogenomic inference of metazoan relationships. Mol. Biol. Evol. 27:2095–2104.

Nesnidal, M.P. Helmkampf, M., Meyer, A., Witek, A. Bruchhaus, I., Ebersberger, I., Hankeln, T., Lieb, B., Struck, T.H., and Hausdorf, B. 2013. New phylogenomic data support the monophyly of Lophophorata and an Ectoproct-Phoronid clade and indicate that Polyzoa and Kryptrochozoa are caused by systematic bias. BMC Evol. Biol. 13:253.

Nielsen C. 2011. *Animal Evolution: Interrelationships of the Living Phyla*. Oxford University Press.

Orrhage L. 1971. Light and electron microscope studies of some annelid setae. Acta Zool. 52:157-169.

Orrhage L. 1973. Light and electron microscope studies of some brachiopod and pogonophoran setae. Zoomorphology. 74:253-270.

Osca D., Irisarri I., Todt C., Grande C., Zardoya R. 2014. The complete mitochondrial genome of *Scutopus ventrolineatus* (Mollusca: Chaetodermomorpha) supports the Aculifera hypothesis. BMC Evol. Biol. 14:197.

Paps J., Bagunà J., Riutort M. 2009a. Bilaterian phylogeny: A broad sampling of 13 nuclear genes provides a new Lophotrochozoa phylogeny and supports a paraphyletic basal Acoelomorpha. Mol. Biol. Evol. 26:2397-2406.

Paps J., Baguñà J., Riutort M. 2009b. Lophotrochozoa internal phylogeny: new insights from an up-to-date analysis of nuclear ribosomal genes. Proc. R. Soc. B. 276:1245-1254.

Passamaneck Y. J., Halanych K.M. 2006. Lophotrochozoan phylogeny assessed with LSU and SSU data: evidence of lophophorate polyphyly. Mol. Phylogenet. Evol. 40:20-28.

Passamaneck Y. J., Schander C., Halanych K. M. 2004. Investigation of Molluscan Phylogeny Using Large-Subunit and Small-Subunit Nuclear rRNA Sequences. Mol. Phylogenet.

Evol. 32: 25-38.

Pérez-Huerta, A., Dauphin, Y., Cusack, M. 2013. Biogenic calcite granules—Are brachiopods different? *Micron* 44: 395-403.

Peterson K.J., Eernisse D.J. 2001. Animal phylogeny and the ancestry of bilaterians: inferences from morphology and 18S rDNA gene sequences. Evol. Dev. 3:170-205.

Philippe H., Snell E.A., Bapteste E., Lopez P., Holland P.W.H., Casane D. 2004. Phylogenomics of eukaryotes: Impact of missing data on large alignments. Mol. Biol. Evol. 21:1740 –1752.

Philippe H., Lartillot N., Brinkmann H. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol. Biol. Evol. 22:1246–1253.

Philippe, H., Brinkmann, H., Lavrov, D.V., Littlewood, T.J., Manuel, M., Wörheide, G., Baurain, D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Viology 9(3): e1000602.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rice, P., Longden, I., Bleasby, A. 2000. EMBOSS: the European molecular biology open software suite. TIG 16(6):276-277.

Rieger R. M., Sterrer, W. 1975a. New spicular skeletons in Turbellaria, and the occurrence of spicules in marine meiofauna. Part I. *Zeitschrift fur Zoologische Systematik und Evolutionsforschung.* 13:207-248.

Rieger R. M., Sterrer, W. 1975b. New spicular skeletons in Turbellaria, and the occurrence of spicules in marine meiofauna. Part II. *Zeitschrift fur Zoologische Systematik und Evolutionsforschung.* 13:249-278.

Rodríguez-Ezpeleta N., Brinkmann H., Roure B., Lartillot N., Lang B.F., Philippe H. 2007.

Detecting and overcoming systematic errors in genome-scale phylogenies. Syst. Biol. 56:389–399.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., Huelsenbeck, J.P. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol., 61(3), 539-542.

Roule L. 1891. Considerations sur I'embranchement des Trochozoaires. Annales des sciences naturelles, Paris series 7 11: 121-178.

Roure, B., Philippe, H. 2011. Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference. BMC Evol. Biol. 11, 17.

Roure B., Baurain D., Philippe H. 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic datasets. Mol. Biol. and Evol. 30(1):197-214.

Rouse G.W. 1999. Trochophore concepts: ciliary bands and the evolution of larvae in spiralian Metazoa. Biol. J. Linnean Soc. 66:411-464.

Ruppert E.E., Carle K.J. 1983. Morphology of metazoan circulatory systems. Zoomorphology. 103:193-208.

Ryan J.F., Pang K., Schnitzler C.E., Nguyen A.-D., Moreland R.T., Simmons D.K., Koch B.J., Francis W.R., Havlak P., Smith S.A. 2013. The genome of the ctenophore Mnemiopsis leidyi and its implications for cell type evolution. Science. 342:1242592.

Salichos L., Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature. 497:327–331.

Salvini-Plawen L., Steiner G. 2014. The Testaria concept (Polyplacophora+ Conchifera) updated. J. Nat. Hist. 48:2751–2772.

Sanderson M.J., McMahon M.M., Steel M. 2011. Terraces in Phylogenetic Tree Space. Science. 333:448–450.

Santagata S., Cohen B. 2009. Phoronid phylogenetics (Brachiopoda; Phoronata): evidence from morphological cladistics, small and large subunit rDNA sequences, and mitochondrial cox1. Zool. J. Linnean Soc. 157:34-50.

Schrödl M., Stöger I. 2014. A review on deep molluscan phylogeny: old markers, integrative approaches, persistent problems. J. Nat. Hist. 48:2773–2804.

Schulze A., 2002. Ultrastructure of opisthosomal chaetae in Vestimentifera (Pogonophora, Obturata) and implications for phylogeny. Acta Zoologica. 82:127-135.

Sharma, P. P., Kaluziak, S. T., Pérez-Porro, A. R., González, V. L., Hormiga, G., Wheeler, W. C., & Giribet, G. 2014. Phylogenomic interrogation of Arachnida reveals systemic conflicts in phylogenetic signal. Mol. Biol. Evol. 31(11): 2963-2984

Simakov O., Marletaz F., Cho S.-J., Edsinger-Gonzales E., Havlak P., Hellsten U., Kuo D.-H., Larsson T., Lv J., Arendt D., Savage R., Osoegawa K., de Jong P., Grimwood J., Chapman J.A., Shapiro H., Aerts A., Otillar R.P., Terry A.Y., Boore J.L., Grigoriev I.V., Lindberg D.R., Seaver E.C., Weisblat D.A., Putnam N.H., Rokhsar D.S. 2013. Insights into bilaterian evolution from three spiralian genomes. Nature. 493:526–531.

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.

Sørensen M.V. 2003. Further structures in the jaw apparatus of Limnognathia maerski (Micrognathozoa), with notes on the phylogeny of the Gnathifera. J. Morphol. 255:131–145.

Sperling E.A., Pisani D., Peterson K.J. 2011. Molecular paleobiological insights into the origin of the Brachiopoda. Evol. Dev. 13:290-303.

Stamatakis A. 2014. RAxML Version 8: A tool for phylogenetic analysis and post-analysis of

large phylogenies. Bioinformatics 30(9):1312-1313.

Stöger I., Sigwart J.D., Kano Y., Knebelsberger T., Marshall B.A., Schwabe E., Schrödl M. 2013. The Continuing Debate on Deep Molluscan Phylogeny: Evidence for Serialia (Mollusca, Monoplacophora. BioMed Res. Int. 2013:407072.

Struck, T. H. 2013. The impact of paralogy on phylogenomic studies–a case study on annelid relationships. PLoS One. 10.1371/journal.pone.0062892.

Struck T. H. 2014. TreSpEx – detection of misleading signal in phylogenetic reconstructions based on tree information. Evol. Bioinformatics, 10: 51.

Struck T.H., Fisse F. 2008. Phylogenetic Position of Nemertea derived from phylogenomic data. Mol. Biol. and Evol. 25:728-736.

Struck T.H., Schult N., Kusen T., Hickman E., Bleidorn C., McHugh D., Halanych K.M. 2007. Annelid phylogeny and the status of Sipuncula and Echiura. BMC Evol. Biol. 7:57.

Struck T.H., Paul C., Hill N., Hartmann S., Hosel C., Kube M., Lieb B., Meyer A., Tiedemann R., Purschke G., Bleidorn C. 2011. Phylogenomic analyses unravel annelid evolution. Nature. 471:95–98.

Struck T.H., Wey-Fabrizius A.R., Golombek A., Hering L., Weigert A., Bleidorn C., Klebow S., Iakovenko N., Hausdorf B., Petersen M., Kück P., Herlyn H., Hankeln T. 2014. Platyzoan Paraphyly Based on Phylogenomic Data Supports a Noncoelomate Ancestry of Spiralia. Mol. Biol. Evol. 31:1833–1849.

Susko, E., Roger, A.J., 2007. On reduced amino acid alphabets for phylogenetic inference. Mol. Biol. Evol. 24(9):2139-2150.

Szabó R., Calder A.C., Ferrier D.E. 2014. Biomineralisation during operculum regeneration in the polychaete *Spirobranchus lamarcki*. Mar. Biol. 161:2621–2629.

Taylor P.D., Lombardi C., Cocito S. 2014. Biomineralization in bryozoans: present, past and

future. Biol. Rev. 90(4):1118-1150.

Temereva, E.N. 2012. Ventral nerve cord in *Phoronopsis harmeri larvae*. J. Exp. Zool. 318B(1): 26-34.

Thomas R.D.K., Vinther J. 2012. Implications of the occurrence of paired anterior chaetae in the Late Early Cambrian mollusc Pelagiella from the Kinziers Formation of Pennsylvania for Relationships among taxa and early evolution of the Mollusca. Geol. Soc. Am. Abstr. Programs. 44:326.

Todaro M.A., Telford M.J., Lockyer A.E., Littlewood D.T.J. 2006. Interrelationships of the Gastrotricha and their place among the Metazoa inferred from 18S rRNA genes. Zoologica Scripta. 35:251-259.

Torruella, G., de Mendoza, A., Grau-Bové, X., Antó, M., Chaplin, M.A., del Campo, J., Eme, L., Pérez-Cordón, G.,Whipps, C. M., Nichols, K.M., Paley, R., Roger, A.J., Sitjà-Bobadilla, A., Donachie, S., Ruiz-Trillo, I. (2015). Phylogenomics Reveals Convergent Evolution of Lifestyles in Close Relatives of Animals and Fungi. Curr. Biol., *25*(18), 2404-2410.

Turbeville J.M. 1983. An ultrastructural analysis of coelomogenesis in the hoplonemertine *Prosorhochmus* americanus and the polychaete *Magelona* sp. J. Morphol. 187:51-60.

Vinther J., Sperling E.A., Briggs D.E., Peterson K.J. 2012. A molecular palaeobiological hypothesis for the origin of aplacophoran molluscs and their derivation from chiton-like ancestors. Proc. Roy. Soc. B. 279:1259–1268.

Wanninger A. 2009. Shaping the things to come: Ontogeny of lophotrochozoan neuromuscular systems and the Tetraneuralia concept. Biol. Bull. 216:293-306.

Wanninger A., Fuchs J., Haszprunar G. 2007. Anatomy of the serotonergic nervous system of an

entoproct creeping type larva and its phylogenetic implications. Invertebr. Biol. 126:268–278.

Westheide W., Russell C.W. 1992. Ultrastructure of chrysopetalid paleal chaetae (Annelida, Polychaeta). Acta Zoologica. 73:197-202.

Whelan N.V., Kocot K.M., Santos S.R., Halanych K.M. 2014. Nemertean Toxin Genes Revealed through Transcriptome Sequencing. Genome Biol. Evol. 6:3314-3325.

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. 112:5773-5778.

Whelan, N. V., Halanych, K.M. Accepted with minor revisions. Who let the CAT out of the bag? Accurately dealing with substitution heterogeneity in phylogenomics. Syst. Biol.

Winnepenninckx B., Backeljau T., Mackey L.Y., Brooks J.M., De Wachter R., Kumar S., Garey J.R. 1995. 18S rRNA data indicate that Aschelminthes are polyphyletic in origin and consist of at least three distinct clades. Mol. Biol. Evol. 12:1132-1137.

Wiens J.J. 2006. Missing data and the design of phylogenetic analyses. J. Biomed. Inform. 39:34–42.

Wiens J.J., Moen D.S. 2008. Missing data and the accuracy of Bayesian phylogenetics. J Syst Evol. 46:307–314.

Witek, A., Herlyn, H., Meyer, A., Boell, L., Bucher, G., Hankeln, T. 2008. EST based phylogenomics of Syndermata questions monophyly of Eurotatoria. BMC Evol. Biol. 8(1): 345.

Witek A., Herlyn H., Ebersberger I., Mark Welch D.B., Hankeln T. 2009. Support for the monophyletic origin of Gnathifera from phylogenomics. Mol. Phylogenet. Evol.

53:1037-1041.

Wourms J.P. 1976. Structure, composition, and unicellular origin of nemertean stylets. Am. Zool. 16:213–213.

Yang Y., Smith S.A. 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.

Zhang G., Fang X., Guo X., Li L., Luo R., Xu F., Yang P., Zhang L., Wang X., Qi H., Xiong Z., Que H., Xie Y., Holland P.W.H., Paps J., Zhu Y., Wu F., Chen Y., Wang J., Peng C., Meng J., Yang L., Liu J., Wen B., Zhang N., Huang Z., Zhu Q., Feng Y., Mount A., Hedgecock D., Xu Z., Liu Y., Domazet-Lošo T., Du Y., Sun X., Zhang S., Liu B., Cheng P., Jiang X., Li J., Fan D., Wang W., Fu W., Wang T., Wang B., Zhang J., Peng Z., Li Y., Li N., Wang J., Chen M., He Y., Tan F., Song X., Zheng Q., Huang R., Yang H., Du X., Chen L., Yang M., Gaffney P.M., Wang S., Luo L., She Z., Ming Y., Huang W., Zhang S., Huang B., Zhang Y., Qu T., Ni P., Miao G., Wang J., Wang Q., Steinberg C.E.W., Wang H., Li N., Qian L., Zhang G., Li Y., Yang H., Liu X., Wang J., Yin Y., Wang J. 2012. The oyster genome reveals stress adaptation and complexity of shell formation. Nature. 490:49–54.

Zhong, B., Deusch, O., Goremykin, V. V., Penny, D., Biggs, P. J., Atherton, R. A., Nikiforova, S. V., Lockhart, P. J. 2011a. Systematic error in seed plant phylogenomics. Genome Biol. Evol., *3*, 1340-1348.

Zhong, M., Hansen, B., Nesnidal, M., Golombek, A., Halanych, K. M., Struck, T. H. 2011b. Detecting the symplesiomorphy trap: a multigene phylogenetic analysis of terebelliform annelids. BMC Evol. Biol. 11(1):369.

Zrzavý J., 2001. The interrelationships of metazoan parasites: a review of phylum-and higher-

level hypotheses from recent morphological and molecular phylogenetic analyses. Folia Parasitologica. 48:81-103.

Zrzavý J., 2003. Gastrotricha and metazoan phylogeny. Zoologica Scripta. 32:61-81.

Zrzavý J., Mihulka S., Kepka P., Bezděk A., Tietz D. 1998. Phylogeny of the Metazoa Based on Morphological and 18S Ribosomal DNA Evidence. Cladistics. 14:249–285.

TABLE CAPTIONS

Table 1. Details of data matrices.

Table 2. Summary of results. Bootstrap support values for selected hypotheses are presented. Values >90 are shaded with dark gray. Values >70 are shaded with light gray. This table reflects strict definitions of clade membership as described herein. For example, recovery of Bryozoa within an otherwise monophyletic Trochozoa would result in an "X" in the Trochozoa column given the definition of Trochozoa used herein. N/A = not applicable given the taxon sampling.

Supplementary Table 1. Taxon sampling.

Supplementary Table 2. Specimen collection data.

Supplementary Table 3. Data matrix characteristics. Length = length of alignment in amino acids. Taxa = number of taxa sampled. LB = branch-length heterogeneity score. PD = patristic distance. RCFV = relative composition frequency variability. Slope = slope of patristic distance versus uncorrected p-distance. R^2 = Pearson correlation coefficient of patristic distance versus uncorrected p-distance. Avg. BS = average bootstrap support in single-OG tree.

Supplementary Table 4. Values for principal component analysis.

Supplementary Table 5. Amino acid properties of complete dataset.

Supplementary Table 6. Correlation of number of positions to bootstrap support. In contrast to the other measurements, a positive correlation for saturation and bootstrap support or positions means that as saturation decreases bootstrap support or the number of positions increases.

FIGURE CAPTIONS

Figure 1. Maximum likelihood phylogeny of Lophotrochozoa based on complete dataset of 638 OGs. The dataset was partitioned by gene and the PROTGAMMALGF model was used for each partition. Bootstrap support values are listed at each node. Taxa from which new data were collected are shown in bold. Bars represent proportion of genes sampled per taxon. The bar for *Lottia* corresponds to all 636 genes sampled. Below: Graphical representation of complete data matrix. OGs are ordered along the X-axis from left to right based on number of taxa sampled (most completely sampled OGs on left). Taxa are ordered along the Y-axis from top to bottom from most genes sampled to fewest genes sampled. Black squares represent a sampled gene fragment and white squares represent a missing gene fragment.

Figure 2. Results of principal component analysis. PC1 is plotted along the X-axis and PC2 is plotted along the Y-axis. Arrows indicate the direction and magnitude of the eigenvectors for each of the five factors examined and the dots indicate individual OGs. The values on the lower x axis and left y axis show the coordinates for the first two principal components for the individual OG's,

while the upper and right ones show the coordinates for the eigenvectors of the variables.

Figure 3. Maximum likelihood phylogeny of Lophotrochozoa based on most stringent 1/6 OGs for LB. Bootstrap support values are listed at each node.

Figure 4. Maximum likelihood phylogeny of Lophotrochozoa based on most stringent 1/6 OGs for PD. Bootstrap support values are listed at each node.

Figure 5. Maximum likelihood phylogeny of Lophotrochozoa based on most stringent 1/6 OGs for missing data. Bootstrap support values are listed at each node.

Figure 6. Maximum likelihood phylogeny of Lophotrochozoa based on most stringent 1/6 OGs for RCFV. Bootstrap support values are listed at each node.

Figure 7. Maximum likelihood phylogeny of Lophotrochozoa based on most stringent 1/6 OGs for saturation. Bootstrap support values are listed at each node.

Figure 8. Hypotheses for relationships within Lophotrochozoa. a) Annelida sister to Nemertea. b) Clade of Annelida, Mollusca, and Brachiopoda + Phoronida. c) Platyzoa. d) Tetraneuralia. e) Mollusca with Aculifera and Conchifera. Filled colored rectangles indicate that the relationship was recovered in the corresponding ML analysis with at least 70% bootstrap support. Images from phylopic.org.

Supplementary Figure 1. Density distribution for each of the five factors examined across all 638

OGs. a) LB. Higher values indicate more branch-length heterogeneity. b) PD. Higher values indicate a greater average patristic distance among taxa. c) Missing data. Higher values indicate a greater percentage of missing data. d) RCFV. Higher values indicate more compositional heterogeneity. e) Saturation. Higher values indicate less saturation.

Supplementary Figure 2. Density distribution for LB, missing data, and RCFV across all 74 taxa. a) LB. Higher values indicate more branch-length heterogeneity. b) Missing data. Higher values indicate a greater percentage of missing data. c) RCFV. Higher values indicate more compositional heterogeneity.

Supplementary Figure 3. Phylogeny of Lophotrochozoa based on most stringent 532 OGs according to LB. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 4. Phylogeny of Lophotrochozoa based on most stringent 425 OGs according to LB. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 5. Phylogeny of Lophotrochozoa based on most stringent 319 OGs according to LB. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 6. Phylogeny of Lophotrochozoa based on most stringent 213 OGs according to LB. Maximum likelihood topology shown with bootstrap support values listed at each

node.

Supplementary Figure 7. Phylogeny of Trochozoa based on most stringent 106 OGs according to LB with no outgroups. Tree is arbitrarily rooted with Mollusca. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 8. Phylogeny of Lophotrochozoa based on sixth-most stringent (worst) sextile of OGs according to LB. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 9. Phylogeny of Lophotrochozoa based on fifth-most stringent sextile of OGs according to LB. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 10. Phylogeny of Lophotrochozoa based on fourth-most stringent sextile of OGs according to LB. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 11. Phylogeny of Lophotrochozoa based on third-most stringent sextile of OGs according to LB. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 12. Phylogeny of Lophotrochozoa based on second-most stringent sextile of OGs according to LB. Maximum likelihood topology shown with bootstrap support values listed at

each node.

Supplementary Figure 13. Phylogeny of Lophotrochozoa based on taxa with LB score < 15.90. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 14. Phylogeny of Lophotrochozoa based on taxa with LB score < 15.90 plus *Bugula*. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 15. Phylogeny of Lophotrochozoa based on taxa with LB score < 15.90 plus *Symbion*. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 16. Phylogeny of Lophotrochozoa based on taxa with LB score < 15.90 plus *Bugula* and *Symbion*. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 17. Phylogeny of Lophotrochozoa based on taxa with LB score < 39.21. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 18. Phylogeny of Lophotrochozoa based on most stringent 532 OGs according to PD. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 19. Phylogeny of Lophotrochozoa based on most stringent 425 OGs according to PD. Maximum likelihood topology shown with bootstrap support values listed at each

node.

Supplementary Figure 20. Phylogeny of Lophotrochozoa based on most stringent 319 OGs according to PD. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 21. Phylogeny of Lophotrochozoa based on most stringent 213 OGs according to PD. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 22. Phylogeny of Lophotrochozoa based on sixth-most stringent (worst) sextile of OGs according to PD. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 23. Phylogeny of Lophotrochozoa based on fifth-most stringent sextile of OGs according to PD. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 24. Phylogeny of Lophotrochozoa based on fourth-most stringent sextile of OGs according to PD. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 25. Phylogeny of Lophotrochozoa based on third-most stringent sextile of OGs according to PD. Maximum likelihood topology shown with bootstrap support values listed at

each node.

Supplementary Figure 26. Phylogeny of Lophotrochozoa based on second-most stringent sextile of OGs according to PD. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 27. Phylogeny of Trochozoa based on most stringent 106 OGs according to PD with no outgroups. Tree is arbitrarily rooted with Mollusca. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 28. Phylogeny of Lophotrochozoa based on most stringent 532 OGs according to percent missing data. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 29. Phylogeny of Lophotrochozoa based on most stringent 425 OGs according to percent missing data. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 30. Phylogeny of Lophotrochozoa based on most stringent 319 OGs according to percent missing data. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 31. Phylogeny of Lophotrochozoa based on most stringent 213 OGs according to percent missing data. Maximum likelihood topology shown with bootstrap support

values listed at each node.

Supplementary Figure 32. Phylogeny of Lophotrochozoa based on sixth-most stringent (worst) sextile of OGs according to missing data. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 33. Phylogeny of Lophotrochozoa based on fifth-most stringent sextile of OGs according to missing data. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 34. Phylogeny of Lophotrochozoa based on fourth-most stringent sextile of OGs according to missing data. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 35. Phylogeny of Lophotrochozoa based on third-most stringent sextile of OGs according to missing data. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 36. Phylogeny of Lophotrochozoa based on second-most stringent sextile of OGs according to missing data. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 37. Phylogeny of Trochozoa based on most stringent 106 OGs according to missing data with no outgroups. Tree is arbitrarily rooted with Mollusca. Maximum likelihood

topology shown with bootstrap support values listed at each node.

Supplementary Figure 38. Phylogeny of Lophotrochozoa based on taxa with less than 80% missing data. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 39. Phylogeny of Lophotrochozoa based on taxa with less than 37.8% missing data. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 40. Heat map with hierarchical clustering to visualize shared missing data among taxa. Lack of correlation between the observed pattern of shared missing data and any recovered tree topology indicates that shared missing data is not influencing our results.

Supplementary Figure 41. Phylogeny of Lophotrochozoa based on most stringent 532 OGs according to RCFV. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 42. Phylogeny of Lophotrochozoa based on most stringent 423 OGs according to RCFV. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 43. Phylogeny of Lophotrochozoa based on most stringent 319 OGs according to RCFV. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 44. Phylogeny of Lophotrochozoa based on most stringent 213 OGs according to RCFV. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 45. Phylogeny of Lophotrochozoa based on sixth-most stringent (worst) sextile of OGs according to RCFV. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 46. Phylogeny of Lophotrochozoa based on fifth-most stringent sextile of OGs according to RCFV. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 47. Phylogeny of Lophotrochozoa based on fourth-most stringent sextile of OGs according to RCFV. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 48. Phylogeny of Lophotrochozoa based on third-most stringent sextile of OGs according to RCFV. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 49. Phylogeny of Lophotrochozoa based on second-most stringent sextile of OGs according to RCFV. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 50. Phylogeny of Trochozoa based on most stringent 106 OGs according to RCFV with no outgroups. Tree is arbitrarily rooted with Mollusca. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 51. Phylogeny of Lophotrochozoa based on taxa with RCFV < 0.00107. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 52. Phylogeny of Lophotrochozoa based on taxa with RCFV < 0.00063. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 53. Phylogeny of Lophotrochozoa based on most stringent 532 OGs according to slope. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 54. Phylogeny of Lophotrochozoa based on most stringent 425 OGs according to slope. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 55. Phylogeny of Lophotrochozoa based on most stringent 319 OGs according to slope. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 56. Phylogeny of Lophotrochozoa based on most stringent 214 OGs

according to slope. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 57. Phylogeny of Lophotrochozoa based on sixth-most stringent (worst) sextile of OGs according to slope. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 58. Phylogeny of Lophotrochozoa based on fifth-most stringent sextile of OGs according to slope. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 59. Phylogeny of Lophotrochozoa based on fourth-most stringent sextile of OGs according to slope. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 60. Phylogeny of Lophotrochozoa based on third-most stringent sextile of OGs according to slope. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 61. Phylogeny of Lophotrochozoa based on second-most stringent sextile of OGs according to slope. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 62. Phylogeny of Trochozoa based on most stringent 106 OGs according to

slope with no outgroups. Tree is arbitrarily rooted with Mollusca. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 63. Phylogeny of Lophotrochozoa based on 296 OGs ranked among the most stringent 532 OGs for all factors examined. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 64. Phylogeny of Lophotrochozoa based on 135 OGs ranked among the most stringent 425 OGs for all factors examined. Maximum likelihood topology shown with bootstrap support values listed at each node.

Supplementary Figure 65. Preliminary Bayesian inference phylogeny of Lophotrochozoa based on most stringent 106 OGs according to LB. Posterior probabilities listed at each node.

Supplementary Figure 66. Preliminary Bayesian inference phylogeny of Lophotrochozoa based on most stringent 107 OGs according to RCFV. Posterior probabilities listed at each node.

Supplementary Figure 67. Preliminary Bayesian inference phylogeny of Lophotrochozoa based on most stringent 106 OGs according to percent missing data. Posterior probabilities listed at each node.

Supplementary Figure 68. Preliminary Bayesian inference phylogeny of Lophotrochozoa based on most stringent 106 OGs according to PD. Posterior probabilities listed at each node.

Supplementary Figure 69. Preliminary Bayesian inference phylogeny of Lophotrochozoa based on most stringent 106 OGs according to slope. Posterior probabilities listed at each node.

Supplementary Figure 70. Maximum likelihood phylogeny of Lophotrochozoa based on complete dataset of 638 OGs. Values at nodes are number of single-gene trees supporting each node / number of single-gene trees with sampling potentially informative for that node.

Data file(s):

assemblies

Preliminary_BI_results

ML_results

Thank you for submitting your data package to Dryad for journal review. Please read the following information carefully, so you will know what to expect during the rest of the data archiving process.

YOUR DRYAD DOI

Your data package has been assigned a unique identifier, called a DOI. This DOI is provisional for now, but may be included in the article manuscript. It will be fully registered with the DOI system when your submission has been approved by Dryad curation staff.

doi:10.5061/dryad.30k4v

REVIEWER ACCESS TO YOUR DRYAD DATA

Journal editors and anonymous peer reviewers may view the submission for review purposes using the following url:

http://datadryad.org/review?doi=doi:10.5061/dryad.30k4v

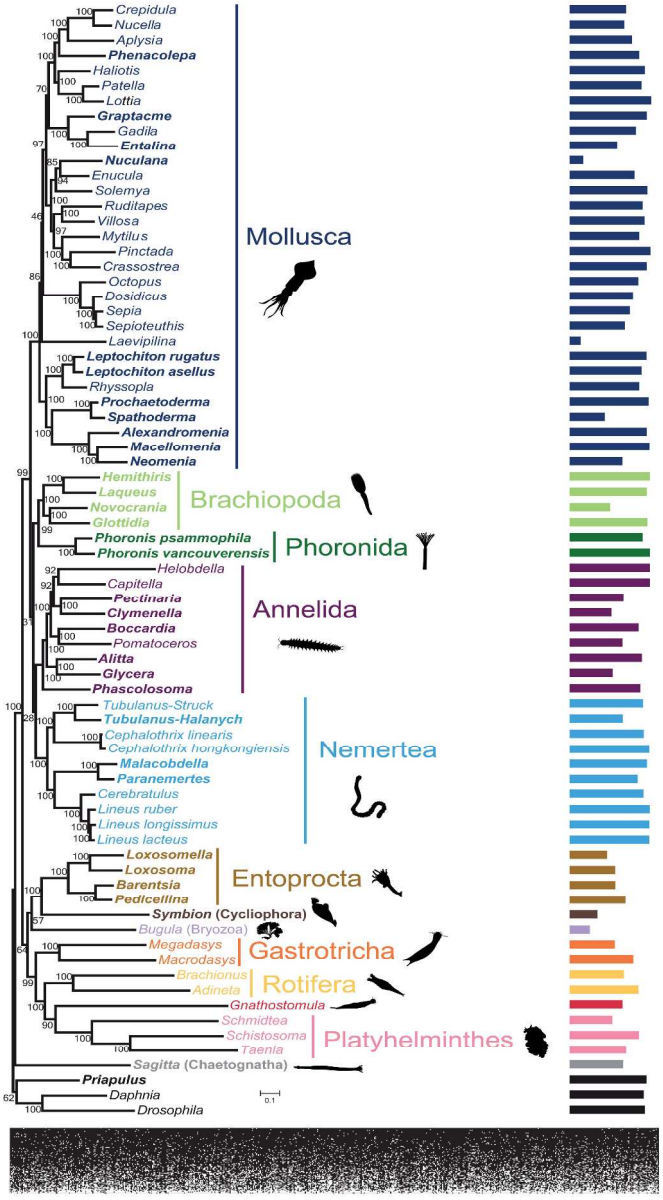
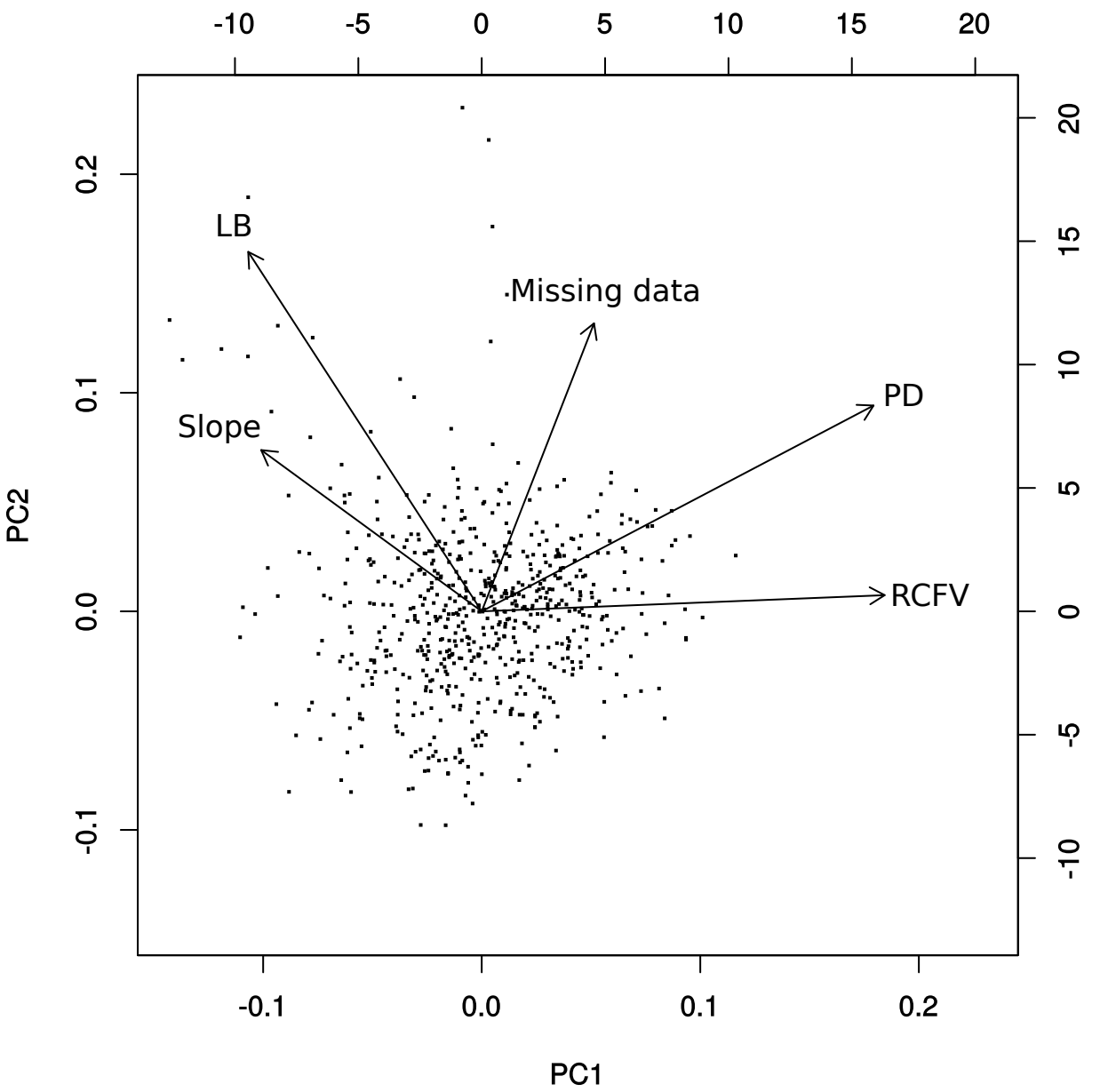
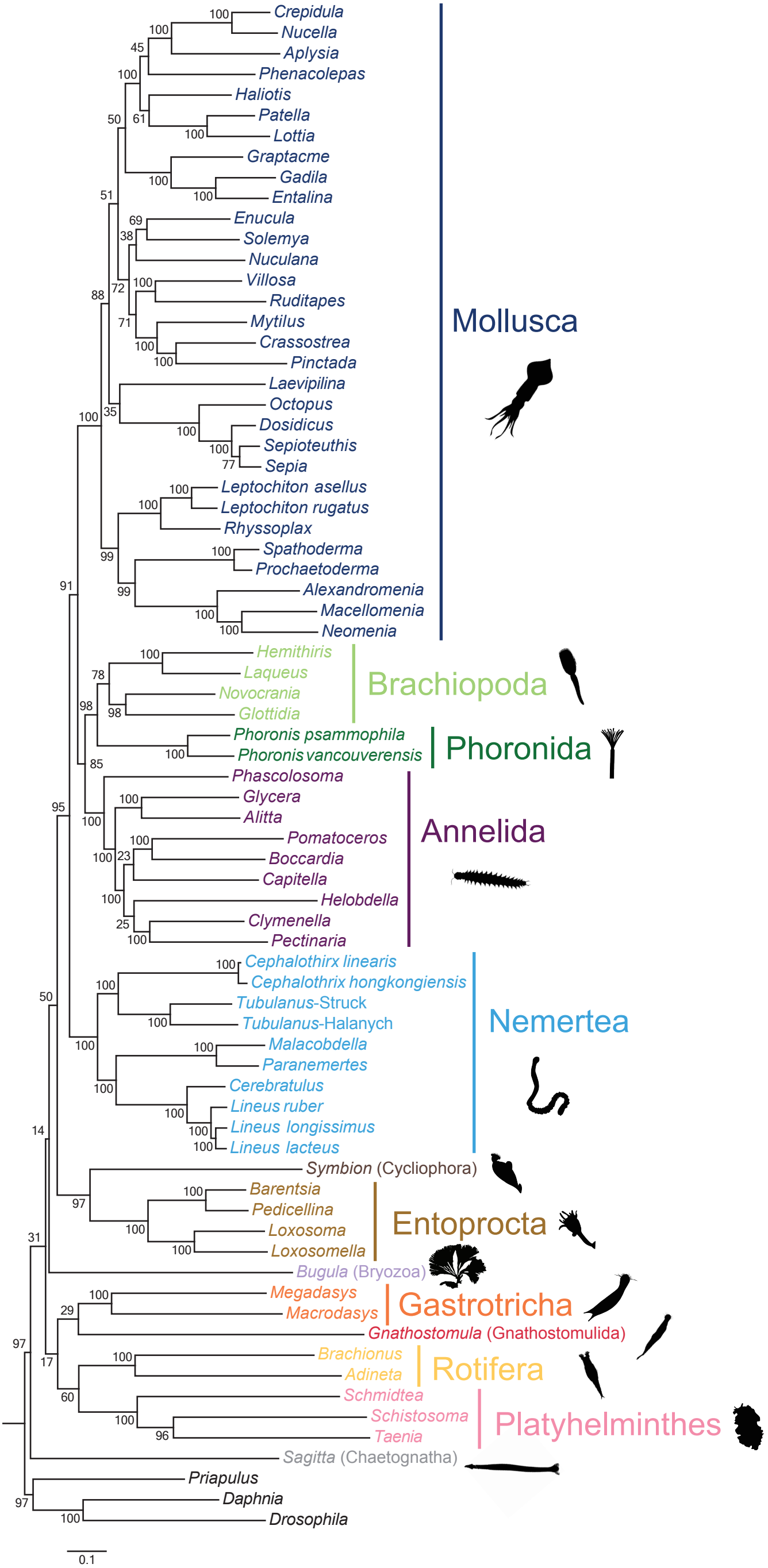
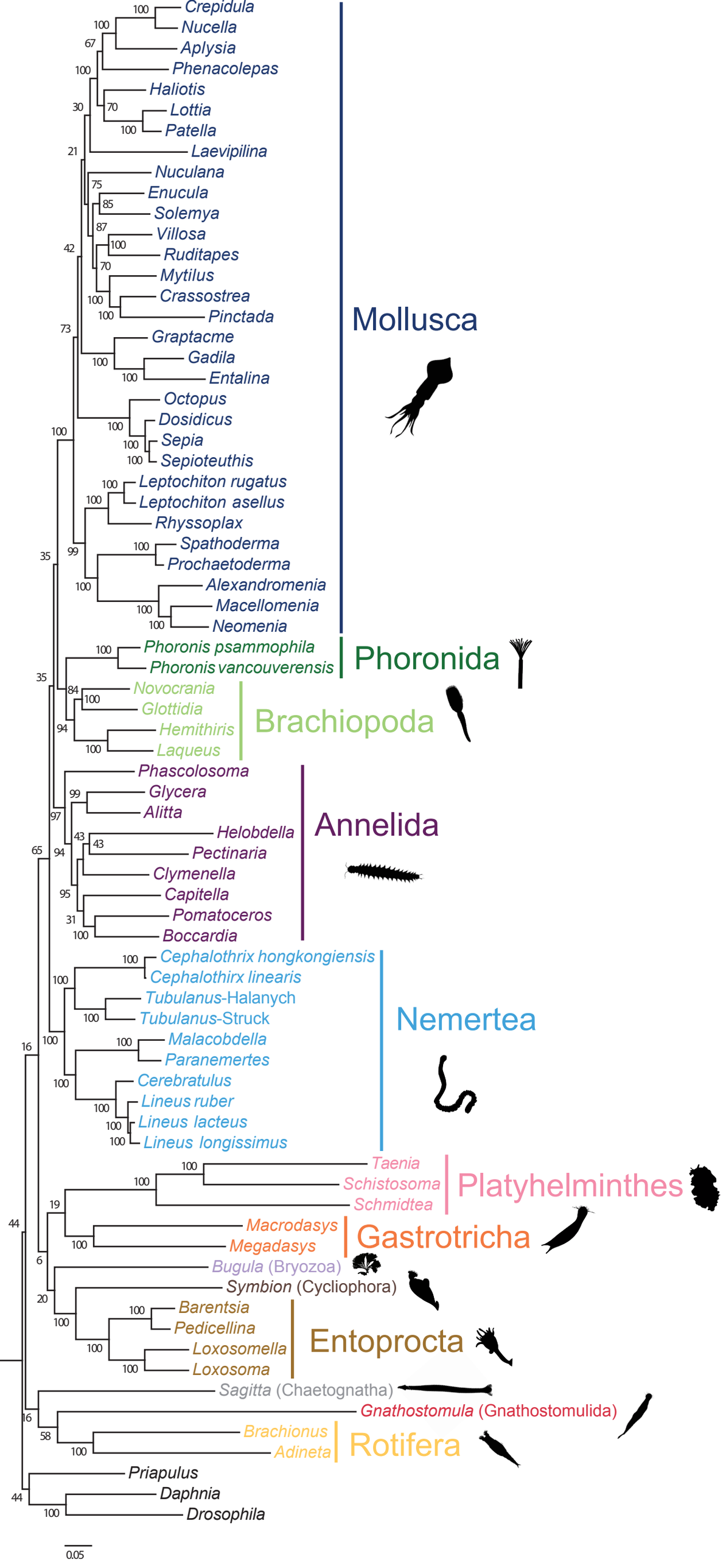


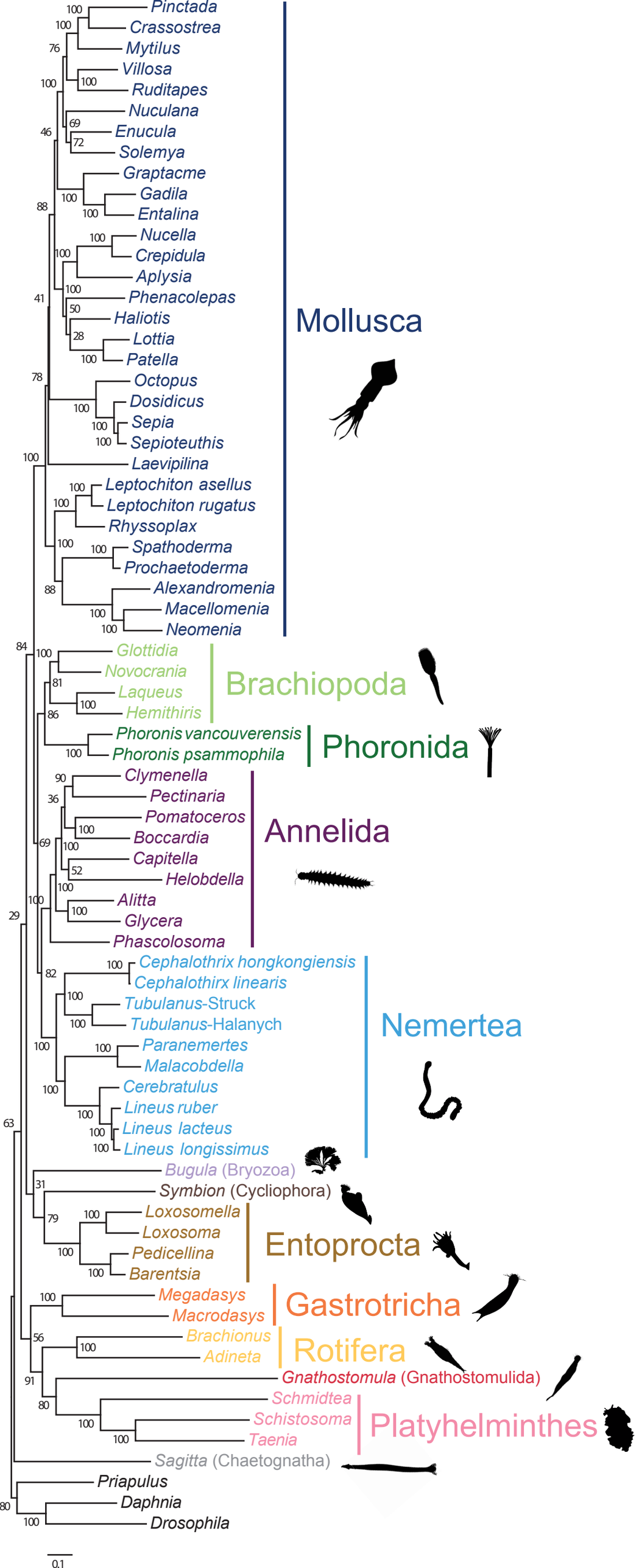
Figure 1. Maximum likelihood phylogeny of Lophotrochozoa based on complete dataset of 638 OGs. The dataset was partitioned by gene and the PROTGAMMALGF model was used for each partition. Bootstrap support values are listed at each node. Taxa from which new data were collected are shown in bold. Bars represent proportion of genes sampled per taxon. The bar for Lottia corresponds to all 636 genes sampled. Below: Graphical representation of complete data matrix. OGs are ordered along the X-axis from left to right based on number of taxa sampled (most completely sampled OGs on left). Taxa are ordered along the Y-axis from top to bottom from most genes sampled to fewest genes sampled. Black squares represent a sampled gene fragment and white squares represent a missing gene fragment.

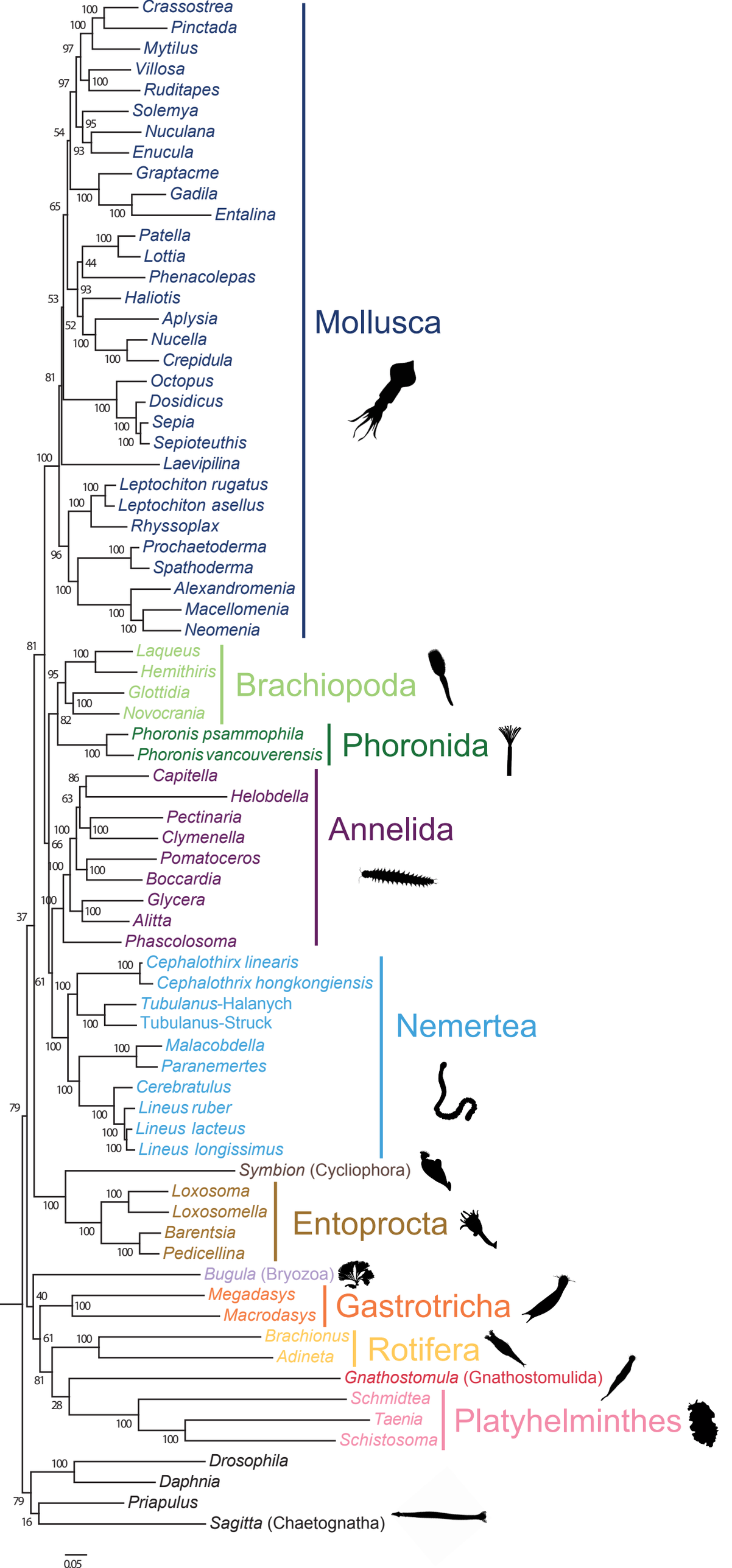
183x329mm (300 x 300 DPI)

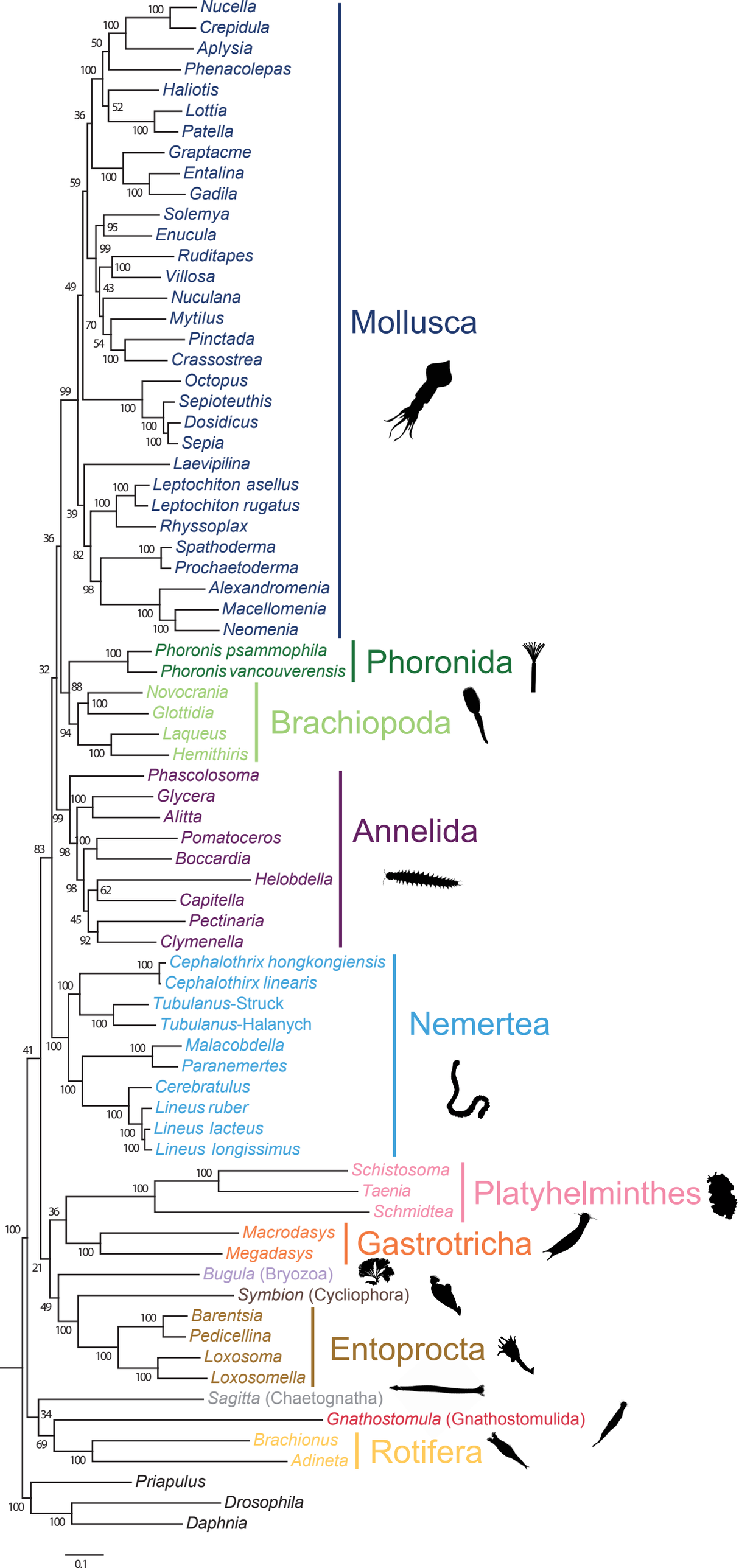

-10
-5
0
5
10
15
20
0.2
0.1
0.0
-0.1
LB
Missing data
PD
Slope
RCFV
PC2
PC1
-0.1
0.0
0.1
0.2
20
15
10
5
0
-5
-10

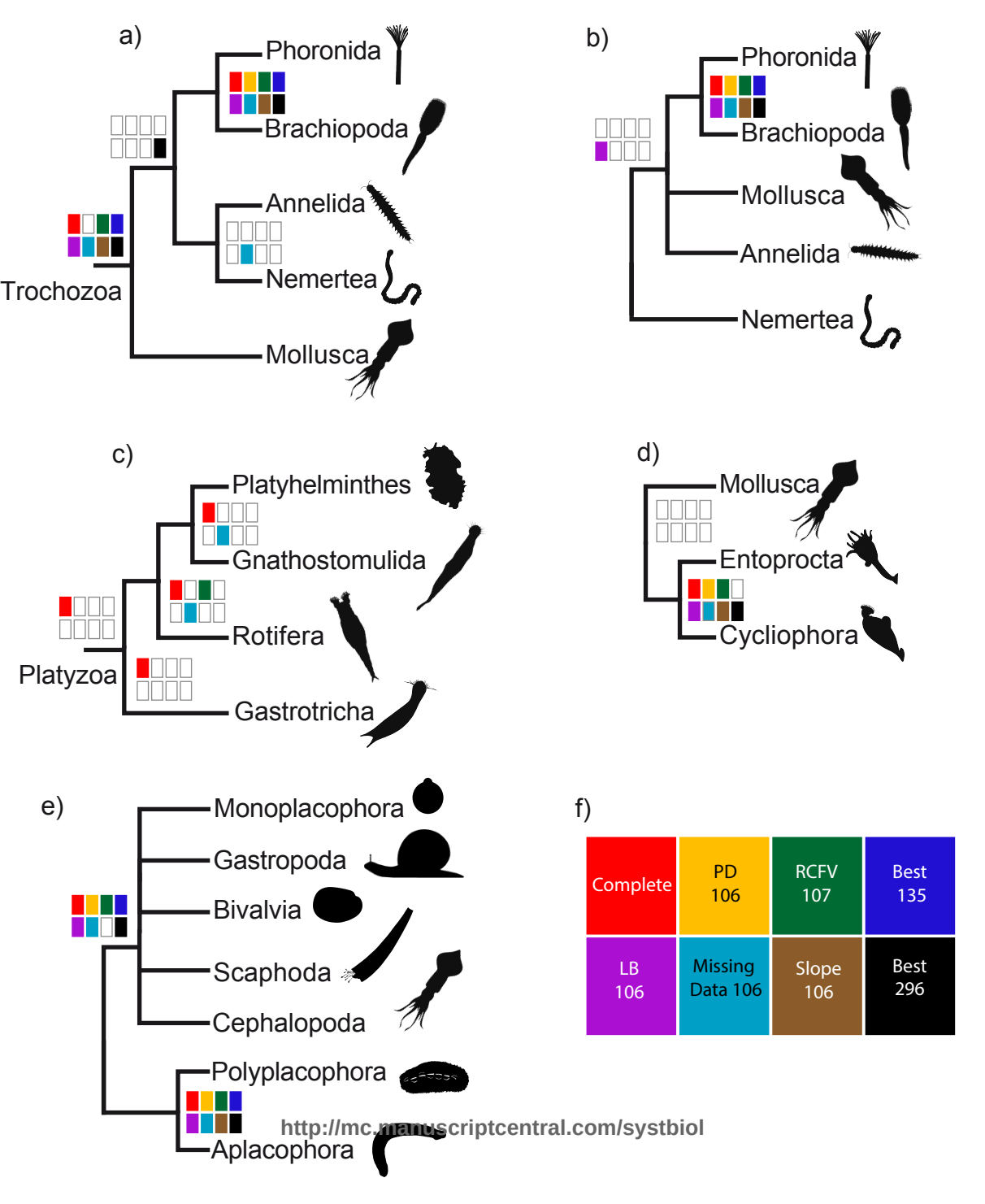

Crepidula
Nucella
Aplysia
Phenacolepas
Haliotis
Patella
Lottia
Graptacme
Gadila
Entalina
Enucula
Solemya
Nuculana
Villosa
Ruditapes
Mytilus
Crassostrea
Pinctada
Laevipilina
Octopus
Dosidicus
Sepioteuthis
Sepia
Leptochiton asellus
Leptochiton rugatus
Rhyssoplax
Spathoderma
Prochaetoderma
Alexandromenia
Macellomenia
Neomenia
Mollusca
Hemithiris
Laqueus
Novocrania
Glottidia
Brachiopoda
Phoronis psammophila
Phoronis vancouverensis
Phoronida
Phascolosoma
Glycera
Alitta
Pomatoceros
Boccardia
Capitella
Helobdella
Clymenella
Pectinaria
Annelida
Cephalothirx linearis
Cephalothrix hongkongiensis
Tubulanus-Struck
Tubulanus-Halanych
Malacobdella
Paranemertes
Cerebratulus
Lineus ruber
Lineus longissimus
Lineus lacteus
Nemertea
Symbion (Cycliophora)
Barentsia
Pedicellina
Loxosoma
Loxosomella
Entoprocta
Bugula (Bryozoa)
Megadasys
Macrodasys
Gastrotricha
Gnathostomula (Gnathostomulida)
Brachionus
Adineta
Rotifera
Schmidtea
Schistosoma
Taenia
Platyhelminthes
Sagitta (Chaetognatha)
Priapulus
Daphnia
Drosophila
0.1


Crepidula
Nucella
Aplysia
Phenacolepas
Haliotis
Lottia
Patella
Laevipilina
Nuculana
Enucula
Solemya
Villosa
Ruditapes
Mytilus
Crassostrea
Pinctada
Graptacme
Gadila
Entalina
Octopus
Dosidicus
Sepia
Sepioteuthis
Leptochiton rugatus
Leptochiton asellus
Rhyssoplax
Spathoderma
Prochaetoderma
Alexandromenia
Macellomenia
Neomenia
Mollusca
Phoronis psammophila
Phoronis vancouverensis
Phoronida
Novocrania
Glottidia
Hemithiris
Laqueus
Brachiopoda
Phascolosoma
Glycera
Alitta
Helobdella
Pectinaria
Clymenella
Capitella
Pomatoceros
Boccardia
Annelida
Cephalothrix hongkongiensis
Cephalothirx linearis
Tubulanus-Halanych
Tubulanus-Struck
Malacobdella
Paranemertes
Cerebratulus
Lineus ruber
Lineus lacteus
Lineus longissimus
Nemertea
Taenia
Schistosoma
Schmidtea
Platyhelminthes
Macrodasys
Megadasys
Gastrotricha
Bugula (Bryozoa)
Symbion (Cycliophora)
Barentsia
Pedicellina
Loxosomella
Loxosoma
Entoprocta
Sagitta (Chaetognatha)
Gnathostomula (Gnathostomulida)
Brachionus
Adineta
Rotifera
Priapulus
Daphnia
Drosophila
0.05


Pinctada
Crassostrea
Mytilus
Villosa
Ruditapes
Nuculana
Enucula
Solemya
Graptacme
Gadila
Entalina
Nucella
Crepidula
Aplysia
Phenacolepas
Haliotis
Lottia
Patella
Octopus
Dosidicus
Sepia
Sepioteuthis
Laevipilina
Leptochiton asellus
Leptochiton rugatus
Rhyssoplax
Spathoderma
Prochaetoderma
Alexandromenia
Macellomenia
Neomenia
Mollusca
Glottidia
Novocrania
Laqueus
Hemithiris
Brachiopoda
Phoronis vancouverensis
Phoronis psammophila
Phoronida
Clymenella
Pectinaria
Pomatoceros
Boccardia
Capitella
Helobdella
Alitta
Glycera
Phascolosoma
Annelida
Cephalothrix hongkongiensis
Cephalothirx linearis
Tubulanus-Struck
Tubulanus-Halanych
Paranemertes
Malacobdella
Cerebratulus
Lineus ruber
Lineus lacteus
Lineus longissimus
Nemertea
Bugula (Bryozoa)
Symbion (Cycliophora)
Loxosomella
Loxosoma
Pedicellina
Barentsia
Entoprocta
Megadasys
Macrodasys
Gastrotricha
Brachionus
Adineta
Rotifera
Gnathostomula (Gnathostomulida)
Schmidtea
Schistosoma
Taenia
Platyhelminthes
Sagitta (Chaetognatha)
Priapulus
Daphnia
Drosophila
0.1


Crassostrea
Pinctada
Mytilus
Villosa
Ruditapes
Solemya
Nuculana
Enucula
Graptacme
Gadila
Entalina
Patella
Lottia
Phenacolepas
Haliotis
Aplysia
Nucella
Crepidula
Octopus
Dosidicus
Sepia
Sepioteuthis
Laevipilina
Leptochiton rugatus
Leptochiton asellus
Rhyssoplax
Prochaetoderma
Spathoderma
Alexandromenia
Macellomenia
Neomenia
Mollusca
Laqueus
Hemithiris
Glottidia
Novocrania
Brachiopoda
Phoronis psammophila
Phoronis vancouverensis
Phoronida
Capitella
Helobdella
Pectinaria
Clymenella
Pomatoceros
Boccardia
Glycera
Alitta
Phascolosoma
Annelida
Cephalothirx linearis
Cephalothrix hongkongiensis
Tubulanus-Halanych
Tubulanus-Struck
Malacobdella
Paranemertes
Cerebratulus
Lineus ruber
Lineus lacteus
Lineus longissimus
Nemertea
Symbion (Cycliophora)
Loxosoma
Loxosomella
Barentsia
Pedicellina
Entoprocta
Bugula (Bryozoa)
Megadasys
Macrodasys
Gastrotricha
Brachionus
Adineta
Rotifera
Gnathostomula (Gnathostomulida)
Schmidtea
Taenia
Schistosoma
Platyhelminthes
Drosophila
Daphnia
Priapulus
Sagitta (Chaetognatha)
0.05


Nucella
Crepidula
Aplysia
Phenacolepas
Haliotis
Lottia
Patella
Graptacme
Entalina
Gadila
Solemya
Enucula
Ruditapes
Villosa
Nuculana
Mytilus
Pinctada
Crassostrea
Octopus
Sepioteuthis
Dosidicus
Sepia
Laevipilina
Leptochiton asellus
Leptochiton rugatus
Rhyssoplax
Spathoderma
Prochaetoderma
Alexandromenia
Macellomenia
Neomenia
Mollusca
Phoronis psammophila
Phoronis vancouverensis
Phoronida
Novocrania
Glottidia
Laqueus
Hemithiris
Brachiopoda
Phascolosoma
Glycera
Alitta
Pomatoceros
Boccardia
Helobdella
Capitella
Pectinaria
Clymenella
Annelida
Cephalothrix hongkongiensis
Cephalothirx linearis
Tubulanus-Struck
Tubulanus-Halanych
Malacobdella
Paranemertes
Cerebratulus
Lineus ruber
Lineus lacteus
Lineus longissimus
Nemertea
Schistosoma
Taenia
Schmidtea
Platyhelminthes
Macrodasys
Megadasys
Gastrotricha
Bugula (Bryozoa)
Symbion (Cycliophora)
Barentsia
Pedicellina
Loxosoma
Loxosomella
Entoprocta
Sagitta (Chaetognatha)
Gnathostomula (Gnathostomulida)
Brachionus
Adineta
Rotifera
Priapulus
Drosophila
Daphnia
0.1



Matrix name	Description	OGs	Taxa	Positions	Missing data	LB	PD	RCFV	Saturation
complete_dataset	All 638 OGs	638	74	121,980	28.88%	22.7501	0.8788	0.0300	0.0845
LB_106	Best 106 OGs based on LB	106	74	21,510	27.13%	14.0162	0.9985	0.0406	0.0870
LB_106_no_outgroup	Best 106 OGs based on LB, trochozoans only	103	56	21,510	23.16%	7.4364	0.8458	0.0324	0.2064
LB_213	Best 213 OGs based on LB	213	74	41,319	28.32%	16.3965	0.9862	0.0368	0.0843
LB_319	Best 319 OGs based on LB	319	74	61,228	28.40%	18.2766	0.9885	0.0355	0.0849
LB_425	Best 425 OGs based on LB	425	74	81,280	28.49%	19.9389	0.9782	0.0335	0.0799
LB_532	Best 532 OGs based on LB	532	74	100,917	28.78%	21.7678	0.9590	0.0319	0.0803
LB_2of6	Second-best sextile of OGs based on LB	107	74	19,809	29.62%	19.1943	0.9772	0.0426	0.0858
LB_3of6	Third-best sextile of OGs based on LB	106	74	19,909	28.55%	22.6152	0.9981	0.0402	0.0853
LB_4of6	Fourth-best sextile of OGs based on LB	106	74	20,052	28.78%	25.6390	0.9478	0.0411	0.0715
LB_5of6	Fifth-best sextile of OGs based on LB	107	74	19,637	29.98%	31.4273	0.8602	0.0381	0.0866
LB_6of6	Sixth-best (worst) sextile of OGs based on LB	106	74	21,063	29.35%	29.5406	0.7594	0.0342	0.0832
Missing_106	Best 106 OGs based on missing data	106	74	23,275	18.17%	20.6086	0.8237	0.0284	0.1149
Missing_106_no_outgroup	Best 106 OGs based on missing data; trochozoans only	106	56	23,275	14.82%	8.3556	0.6445	0.0222	0.2388
Missing_213	Best 213 OGs based on missing data	213	74	44,784	21.43%	21.9150	0.8545	0.0284	0.1038
Missing_319	Best 319 OGs based on missing data	319	74	64,886	23.65%	22.0602	0.8804	0.0280	0.0940
Missing_425	Best 425 OGs based on missing data	425	74	84,189	25.47%	22.3060	0.8994	0.0290	0.0894
Missing_532	Best 532 OGs based on missing data	532	74	102,446	27.06%	22.6512	0.9247	0.0296	0.0848
Missing_2of6	Second-best sextile of OGs based on missing data	107	74	21,509	24.95%	23.5781	0.8912	0.0372	0.0945
Missing_3of6	Third-best sextile of OGs based on missing data	106	74	20,102	28.59%	22.9817	0.9526	0.0391	0.0696
Missing_4of6	Fourth-best sextile of OGs based on missing data	106	74	19,303	31.61%	23.7070	0.9713	0.0423	0.0702
Missing_5of6	Fifth-best sextile of OGs based on missing data	107	74	18,257	34.37%	24.6476	1.0377	0.0474	0.0652
Missing_6of6	Sixth-best (worst) sextile of OGs based on missing data	106	74	19,534	38.43%	23.8757	0.9852	0.0487	0.0475
PD_106	Best 106 OGs based on PD	106	74	23,332	27.11%	24.5804	0.4871	0.0270	0.1171
PD_106_no_outgroup	Best 106 OGs based on PD; trochozoans only	106	56	23,332	23.64%	10.8190	0.3659	0.0220	0.2116
PD_213	Best 213 OGs based on PD	213	74	45,954	27.24%	24.3800	0.6053	0.0256	0.1108
PD_319	Best 319 OGs based on PD	319	74	66,727	27.69%	23.9602	0.6863	0.0264	0.0990
PD_425	Best 425 OGs based on PD	425	74	86,144	28.14%	23.6146	0.7656	0.0276	0.0904
PD_532	Best 532 OGs based on PD	532	74	104,650	28.64%	23.0825	0.8491	0.0293	0.0852
PD_2of6	Second-best sextile of OGs based on PD	107	74	20,435	28.04%	21.7065	0.9546	0.0383	0.0888
PD_3of6	Third-best sextile of OGs based on PD	106	74	20,466	27.93%	21.7538	0.9057	0.0371	0.0763
PD_4of6	Fourth-best sextile of OGs based on PD	106	74	20,933	28.09%	24.3545	0.8855	0.0370	0.0811
PD_5of6	Fifth-best sextile of OGs based on PD	107	74	20,873	30.29%	23.7067	0.9149	0.0405	0.0769
PD_6of6	Sixth-best (worst) sextile of OGs based on PD	106	74	18,919	30.76%	24.8555	0.9594	0.0438	0.0646
RCFV_107	Best 107 OGs based on RCFV	107	74	28,490	28.52%	26.5036	0.5941	0.0283	0.0925
RCFV_107_no_outgroup	Best 107 OGs based on RCFV; trochozoans only	107	56	28,490	25.11%	11.6235	0.4312	0.0242	0.1602
RCFV_213	Best 213 OGs based on RCFV	213	74	52,527	28.50%	24.2258	0.6839	0.0270	0.0936
RCFV_319	Best 319 OGs based on RCFV	319	74	74,298	28.68%	24.5305	0.7639	0.0277	0.0871
RCFV_423	Best 423 OGs based on RCFV	423	74	93,509	28.88%	23.4065	0.8302	0.0286	0.0856
RCFV_532	Best 532 OGs based on RCFV	532	74	110,167	28.87%	23.1981	0.8841	0.0292	0.0844
RCFV_2of6	Second-best sextile of OGs based on RCFV	106	74	24,037	28.46%	22.5279	0.7607	0.0345	0.0986
RCFV_3of6	Third-best sextile of OGs based on RCFV	106	74	21,771	29.13%	25.1913	0.8978	0.0399	0.0807
RCFV_4of6	Fourth-best sextile of OGs based on RCFV	104	74	19,211	29.67%	20.3277	1.0007	0.0444	0.0866
RCFV_5of6	Fifth-best sextile of OGs based on RCFV	109	74	16,658	28.80%	22.3875	1.0699	0.0475	0.0884
RCFV_6of6	Sixth-best (worst) sextile of OGs based on RCFV	106	74	11,813	28.96%	20.2015	1.1996	0.0542	0.0610
Slope_106	Best 106 OGs based on saturation	106	74	19,779	28.48%	22.6141	0.7213	0.0352	0.0931
Slope_106_no_outgroup	Best 106 OGs based on saturation; trochoozans only	106	56	19,779	24.42%	9.9431	0.5513	0.0273	0.2429
Slope_214	Best 214 OGs based on saturation	214	74	40,341	28.24%	22.2854	0.8004	0.0330	0.1013
Slope_319	Best 319 OGs based on saturation	319	74	60,267	28.32%	22.1557	0.8670	0.0322	0.0919
Slope_425	Best 425 OGs based on saturation	425	74	80,092	28.34%	22.1845	0.8968	0.0317	0.0882
Slope_532	Best 532 OGs based on saturation	532	74	100,396	28.67%	22.9460	0.9307	0.0313	0.0825
Slope_2of6	Second-best sextile of OGs based on saturation	108	74	20,562	28.00%	22.3989	0.8575	0.0405	0.1052
Slope_3of6	Third-best sextile of OGs based on saturation	105	74	19,926	28.48%	21.8976	0.9917	0.0398	0.0777
Slope_4of6	Fourth-best sextile of OGs based on saturation	106	74	19,825	28.41%	22.6180	0.9674	0.0403	0.0797
Slope_5of6	Fifth-best sextile of OGs based on saturation	107	74	20,304	29.94%	25.6276	1.0682	0.0408	0.0676
Slope_6of6	Sixth-best (worst) sextile of OGs based on saturation	106	74	21,584	29.87%	23.2223	0.9588	0.0371	0.0598
Best_296_all_cat	296 OGs in Best 425 category for all 5 metrics	135	74	32,257	24.42%	20.4052	0.7422	0.0293	0.1055
Best_135_all_cat	135 OGs in Best 532 category for all 5 metrics	296	74	61,963	26.76%	22.3077	0.7930	0.0302	0.1007
Missing_<_0.8	Taxa with missing data < 80.0%	638	71	121,980	26.35%	23.2125	0.9461	0.0299	0.0808
Missing_<_0.378	Taxa with missing data < 37.8%	638	49	121,980	14.82%	22.0291	0.8309	0.0232	0.1756
RCFV_<_0.00063	Taxa with RCFV < 0.00063	638	61	121,980	25.22%	18.6451	0.8238	0.0223	0.1043
RCFV_<_0.00107	Taxa with RCFV < 0.00107	638	72	121,980	28.34%	22.3680	0.9220	0.0283	0.0858
LB_<_39.21	Taxa with LB score < 39.21	638	68	121,980	28.03%	15.5361	0.8451	0.0260	0.0909
LB_<_15.90	Taxa with LB score < 15.90	638	61	121,980	26.37%	10.3864	0.7652	0.0239	0.1392
LB_<_15.90+Bugula	Taxa with LB score < 15.90 + Bugula	638	62	121,980	27.32%	11.2563	0.7750	0.0242	0.1033
LB_<_15.90+Symbion	Taxa with LB score < 15.90 + Symbion	638	62	121,980	27.05%	11.7446	0.7777	0.0244	0.1089
LB_<_15.90+Bugula+Symbion	Taxa with LB score < 15.90 + Bugula + Symbion	638	63	121,980	27.97%	12.3686	0.7870	0.0244	0.0801

	Lophotrochozoa (+/- *Sagitta*)	Trochozoa	Brachiopoda + Phoronida	Annelida + Nemertea + Brachiopoda + Phoronida	Annelida + Nemertea	Annelida + Brachiopoda + Phoronida	Annelida + Brachiopoda + Phoronida + Mollusca	Kryptrochozoa	Mollusca + Brachiopoda + Phoronida	Mollusca + Nemertea	Eutrochozoa	Tetraneuralia (+/- Cyclilophora)	Lophophorata	Entoprocta + Cyclilophora	Platyzoa	Polyzoa	Platyzoa + Polyzoa	Platyzoa + Entoprocta + Cyclilophora	Trochozoa + Polyzoa	Trochozoa + Entoprocta + Cyclilophora	Trochozoa + Bryozoa	Mollusca	Aculifera	Conchifera	Gastropoda + Bivalvia	Gastropoda + Scaphopoda	Diasoma (Bivalvia + Scaphopoda)	Annelida	Nemertea	Brachiopoda	Phoronida
complete_dataset	100	99	99	31	28	X	X	X	X	X	X	X	X	100	99	57	64	X	X	X	X	100	100	86	X	70	X	100	100	100	100
LB_106 (=LB_1of6)	31	95	98	X	X	85	91	X	X	X	X	X	X	97	17	X	X	X	14	50	X	100	99	88	X	50	X	100	100	78	100
LB_106_no_outgroup	N/A	N/A	100	N/A	X	58	N/A	X	N/A	N/A	N/A	X	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	100	98	81	X	69	X	100	100	87	100
LB_213	86	80	100	X	X	76	97	X	X	X	X	X	X	100	90	38	X	X	72	X	X	100	85	71	X	42	X	100	100	99	100
LB_319	90	96	100	X	X	79	84	X	X	X	X	X	X	100	89	69	X	X	93	X	X	100	100	89	60	X	X	100	100	100	100
LB_425	100	100	100	X	X	65	70	X	X	X	X	X	X	100	98	63	X	X	59	X	X	100	100	84	69	X	X	100	100	100	100
LB_532	100	100	100	48	44	X	X	X	X	X	X	X	X	100	99	71	43	X	X	X	X	100	100	87	X	40	X	100	100	100	100
LB_2of6	49	X	99	X	X	X	53	X	48	X	X	X	X	69	98	X	X	X	58	X	51	100	54	X	X	X	80	100	100	98	100
LB_3of6	X	87	92	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	25	100	100	93	X	X	39	100	100	100	100
LB_4of6	100	66	66	47	61	X	X	X	X	X	X	X	X	100	65	X	63	X	X	X	X	100	100	100	X	X	57	99	100	100	100
LB_5of6	98	100	100	68	37	X	X	X	X	X	X	X	X	99	54	X	93	X	X	X	X	100	98	86	X	64	X	100	100	100	100
LB_6of6	X	65	76	X	X	X	38	X	40	X	X	X	X	99	48	X	X	36	X	X	X	100	100	55	X	62	X	100	100	98	100
PD_106 (=PD_1of6)	X	65	84	X	X	X	35	X	35	X	X	X	X	100	X	20	X	X	X	X	X	100	99	73	X	X	X	97	100	94	100
PD_106_no_outgroup	N/A	N/A	100	N/A	95	X	N/A	X	N/A	N/A	N/A	X	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	100	95	67	X	X	X	100	100	100	100
PD_213	93	88	92	68	66	X	X	X	X	X	X	X	X	100	56	27	X	X	22	X	X	100	93	66	X	33	X	100	100	99	100
PD_319	97	91	92	73	66	X	X	X	X	X	X	X	X	100	92	57	X	X	52	X	X	100	98	70	X	63	X	100	100	100	100
PD_425	99	100	100	87	75	X	X	X	X	X	X	X	X	99	99	67	X	X	41	X	X	100	100	92	X	76	X	100	100	100	100
PD_532	99	100	100	72	68	X	X	X	X	X	X	X	X	100	100	50	77	X	X	X	X	100	100	94	X	56	X	100	100	100	100
PD_2of6	79	96	97	X	X	X	84	X	69	X	X	X	X	93	90	X	65	48	X	X	X	100	89	X	55	X	X	100	100	98	100
PD_3of6	X	66	80	32	24	X	X	X	X	X	X	X	X	X	76	X	X	X	X	X	X	100	100	88	X	34	X	100	100	99	100
PD_4of6	83	90	99	34	26	X	X	X	X	X	X	X	X	99	94	X	X	57	X	X	48	100	100	91	X	78	X	100	100	98	100
PD_5of6	37	61	60	30	X	X	X	39	X	X	X	X	X	99	33	32	24	X	X	X	X	100	100	91	X	X	46	100	100	93	100
PD_6of6	69	71	100	X	X	X	X	X	X	66	X	X	X	98	96	X	X	53	X	X	X	100	88	82	X	X	X	85	95	100	100
Missing_106 (=Missing_1of6	63	84	86	69	82	X	X	X	X	X	X	X	X	79	56	31	X	X	X	X	X	100	100	78	X	X	46	100	100	81	100
Missing_106_no_outgroup	N/A	N/A	100	N/A	94	X	N/A	X	N/A	N/A	N/A	X	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	100	99	81	X	X	36	100	100	58	100
Missing_213	X	90	94	66	61	X	X	X	X	X	X	X	X	99	X	X	X	X	X	X	X	100	100	73	X	73	X	100	100	100	100
Missing_319	91	100	100	66	54	X	X	X	X	X	X	X	X	100	82	34	53	X	X	X	X	100	100	83	X	38	X	100	100	100	100
Missing_425	98	99	99	42	39	X	X	X	X	X	X	X	X	100	99	42	48	X	X	X	X	100	100	81	X	54	X	100	100	100	100
Missing_532	100	98	98	43	42	X	X	X	X	X	X	X	X	100	99	X	54	X	X	X	X	100	100	82	X	71	X	100	100	100	100
Missing_2of6	X	91	96	X	X	28	36	X	X	X	X	X	X	100	X	X	X	X	X	X	X	100	80	X	X	77	X	99	100	100	100
Missing_3of6	29	100	100	X	X	40	37	X	X	X	X	X	X	42	81	X	X	28	X	X	X	100	100	83	56	X	X	100	100	100	100
Missing_4of6	64	X	X	X	X	X	X	X	X	X	X	X	43	85	78	X	X	X	48	X	47	100	100	X	X	41	X	100	100	100	100
Missing_5of6	95	77	79	X	X	X	X	X	22	X	X	X	X	98	80	X	46	X	X	X	X	100	93	67	X	X	57	98	100	98	100
Missing_6of6	73	95	99	X	X	X	X	41	X	X	X	X	X	97	92	83	X	X	42	X	X	99	79	73	42	X	X	99	99	100	100
RCFV_107 (=RCFV_1of6)	79	81	82	66	61	X	X	X	X	X	X	X	X	100	61	X	X	X	X	37	X	100	96	81	X	X	X	54	100	100	95	100
RCFV_107_no_outgroup	N/A	N/A	100	N/A	72	X	N/A	X	N/A	N/A	N/A	X	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	100	92	80	X	X	43	100	100	100	100
RCFV_213	79	88	93	40	27	X	X	X	X	X	X	X	X	100	83	48	23	X	X	X	X	100	98	77	X	45	X	100	100	100	100
RCFV_319	95	99	99	69	42	X	X	X	X	X	X	X	X	100	100	61	53	X	X	X	X	100	100	97	X	59	X	100	100	100	100
RCFV_423	94	100	100	57	37	X	X	X	X	X	X	X	X	100	96	50	X	X	22	X	X	100	100	98	66	X	X	100	100	100	100
RCFV_532	100	98	98	55	42	X	X	X	X	X	X	X	X	100	99	60	X	X	32	X	X	100	100	84	64	X	X	100	100	100	100
RCFV_2of6	X	88	94	X	X	X	X	43	X	X	X	X	X	60	59	X	X	X	X	X	X	100	94	68	X	90	X	98	100	100	100
RCFV_3of6	80	100	99	42	34	X	X	X	X	X	X	X	X	94	97	49	54	X	X	X	X	100	100	99	X	51	X	100	100	88	100
RCFV_4of6	X	94	96	33	X	25	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	100	90	X	54	X	X	99	100	99	100
RCFV_5of6	91	81	83	X	X	X	37	X	54	X	X	X	X	99	100	X	X	35	X	X	53	100	95	X	X	X	43	98	100	100	100
RCFV_6of6	73	87	84	X	X	62	80	X	X	X	X	X	X	99	56	X	X	X	X	39	X	98	79	75	X	60	X	98	100	98	100
Slope_106 (=Slope_1of6)	X	83	88	X	X	X	32	X	36	X	X	X	X	100	X	49	X	X	X	X	X	99	82	X	X	36	X	99	100	94	100
Slope_106_no_outgroup	N/A	N/A	100	N/A	89	X	N/A	X	N/A	N/A	N/A	X	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	100	86	X	X	38	X	100	100	100	100
Slope_214	92	79	82	59	42	X	X	X	X	X	X	X	X	100	90	54	43	X	X	X	X	100	93	76	X	47	X	100	100	100	100
Slope_319	96	87	88	78	77	X	X	X	X	X	X	X	X	97	98	52	X	X	51	X	X	100	100	90	X	48	X	100	100	100	100
Slope_425	100	98	98	53	50	X	X	X	X	X	X	X	X	100	98	52	45	X	X	X	X	100	100	84	X	32	X	100	100	100	100
Slope_532	100	100	100	X	X	41	55	X	X	X	X	X	X	100	100	43	X	X	26	X	X	100	100	73	X	59	X	100	100	100	100
Slope_2of6	X	56	87	46	37	X	X	X	X	X	X	X	X	X	82	X	X	X	X	X	X	100	93	90	X	56	X	78	100	98	100
Slope_3of6	X	72	87	53	63	X	X	X	X	X	X	X	X	X	46	X	X	X	X	X	X	100	100	93	40	X	X	100	100	100	100
Slope_4of6	85	90	97	X	X	74	86	X	X	X	X	X	X	94	72	X	54	X	X	X	X	100	98	X	62	X	X	100	100	97	100
Slope_5of6	91	59	63	X	X	X	X	46	X	X	X	X	X	99	99	X	X	X	X	19	X	100	98	54	X	X	80	100	100	100	100
Slope_6of6	X	99	99	35	30	X	X	X	X	X	X	X	X	100	X	X	X	X	X	X	X	100	100	95	X	X	85	100	100	100	100
Best_135_all_cat	X	71	90	49	52	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	100	100	94	X	45	X	100	100	100	100
Best_296_all_cat	82	100	100	77	66	X	X	X	X	X	X	X	X	83	96	56	68	X	X	X	X	100	100	87	X	50	X	100	100	100	100
Missing_<_0.8	100	100	100	X	X	X	50	X	25	X	X	X	X	100	100	N/A	N/A	84	N/A	X	N/A	100	100	100	X	73	X	100	100	100	100
Missing_<_0.378	N/A	100	100	50	49	X	X	X	X	X	X	X	X	N/A	100	N/A	N/A	N/A	N/A	N/A	N/A	100	100	100	X	79	X	100	100	100	100
RCFV_>_0.00063	97	97	97	X	X	X	58	X	44	X	X	X	X	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	100	100	67	X	45	X	100	100	100	100
RCFV_>_0.00107	99	94	94	X	X	X	47	X	25	X	X	X	X	100	100	X	62	50	X	X	X	100	100	73	X	70	X	100	100	100	100
LB_>_39.21	100	98	98	50	49	X	X	X	X	X	X	X	X	100	N/A	X	N/A	N/A	X	X	X	100	100	75	X	74	X	100	100	100	100
LB_>_15.90	N/A	100	100	70	69	X	X	X	X	X	X	X	X	N/A	N/A	N/A	N/A	N/A	N/A	N/A	X	100	100	76	X	76	X	100	100	100	100
LB_>_15.90+BNER	N/A	99	99	67	67	X	X	X	X	X	X	X	X	N/A	N/A	N/A	N/A	N/A	N/A	N/A	X	100	100	74	X	74	X	100	100	100	100
LB_>_15.90+SAME	N/A	100	100	65	64	X	X	X	X	X	X	X	X	100	N/A	N/A	N/A	N/A	N/A	X	X	100	100	73	X	73	X	100	100	100	100
LB_>_15.90+BNER+SAME	N/A	100	100	61	61	X	X	X	X	X	X	X	X	100	N/A	100	N/A	N/A	N/A	X	X	100	100	76	X	75	X	100	100	100	100