OMA standalone: orthology inference among public and custom genomes and transcriptomes

Adrian M. Altenhoff,[1,2,10] Jeremy Levy,[3,4,10] Magdalena Zarowiecki,[5] Bartłomiej Tomiczek,[4,6] Alex Warwick Vesztrocy,[1,4] Daniel A. Dalquen,[2] Steven Müller,[4] Maximilian J. Telford,[4] Natasha M. Glover,[1,7,8] David Dylus,[1,7,8] and Christophe Dessimoz[1,4,7,8,9]

[1]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland; [2]Department of Computer Science, ETH Zurich, 8092 Zurich, Switzerland; [3]Centre for Mathematics and Physics in the Life Sciences and Experimental Biology (CoMPLEX), University College London, London WC1E 6BT, United Kingdom; [4]Centre for Life's Origins and Evolution, Department of Genetics, Evolution & Environment, University College London, London WC1E 6BT, United Kingdom; [5]Genomics England, Queen Mary University of London, London EC1M 6BQ, United Kingdom; [6]Intercollegiate Faculty of Biotechnology, University of Gdansk and Medical University of Gdansk, 80-307 Gdansk, Poland; [7]Department of Computational Biology, University of Lausanne, 1015 Lausanne, Switzerland; [8]Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland; [9]Department of Computer Science, University College London, London WC1E 6BT, United Kingdom

Genomes and transcriptomes are now typically sequenced by individual laboratories but analyzing them often remains challenging. One essential step in many analyses lies in identifying orthologs—corresponding genes across multiple species—but this is far from trivial. The Orthologous MAtrix (OMA) database is a leading resource for identifying orthologs among publicly available, complete genomes. Here, we describe the OMA pipeline available as a standalone program for Linux and Mac. When run on a cluster, it has native support for the LSF, SGE, PBS Pro, and Slurm job schedulers and can scale up to thousands of parallel processes. Another key feature of OMA standalone is that users can combine their own data with existing public data by exporting genomes and precomputed alignments from the OMA database, which currently contains over 2100 complete genomes. We compare OMA standalone to other methods in the context of phylogenetic tree inference, by inferring a phylogeny of Lophotrochozoa, a challenging clade within the protostomes. We also discuss other potential applications of OMA standalone, including identifying gene families having undergone duplications/losses in specific clades, and identifying potential drug targets in nonmodel organisms. OMA standalone is available under the permissive open source Mozilla Public License Version 2.0.

[Supplemental material is available for this article.]

The sequencing revolution is yielding a flood of genomes and transcriptomes, with thousands already sequenced and many more underway (Pagani et al. 2012). A powerful way of characterizing newly sequenced genes is to compare them with evolutionarily related genes—in particular, with orthologs in other species (Dessimoz et al. 2012; Sonnhammer et al. 2014; Forslund et al. 2018). In this way, experimental knowledge from model organisms can be propagated to nonmodel organisms. Elucidation of orthology and paralogy relationships is also essential to reconstruct species trees, to better understand the mechanics of gene/genome evolution, to study adaptation, or to pinpoint the emergence of new gene functions (Gabaldón and Koonin 2013).

The importance of determining orthology has led to the development of many inference methods and associated databases (for review, see Altenhoff and Dessimoz 2012). Some of the best established orthology resources include eggNOG (Huerta-Cepas et al. 2016b), Ensembl Compara (Zerbino et al. 2018), InParanoid (Sonnhammer and Östlund 2015), MBGD (Uchiyama et al. 2012), OrthoDB (Zdobnov et al. 2017), OrthoMCL (Chen et al. 2006), PANTHER (Mi et al. 2017), PhylomeDB (Huerta-Cepas et al. 2014), and OMA (Altenhoff et al. 2018).

Key distinctive features of OMA are the high specificity of its inference pipeline (Altenhoff and Dessimoz 2009; Boeckmann et al. 2011; Linard et al. 2011; Afrasiabi et al. 2013), the feature-rich web and programmatic interfaces, large size and taxonomic breadth of its precomputed data (currently 2167 genomes), its regular update schedule of two releases per year, and its sustained development over the last 13 yr. The algorithms underlying the OMA pipeline have been described and validated in multiple publications (Dessimoz et al. 2005, 2006; Roth et al. 2008; Altenhoff et al. 2013; Train et al. 2017). The quality of OMA is corroborated by a recent community benchmarking study, which highlighted the high specificity of orthologs predicted by the OMA pipeline (Altenhoff et al. 2016).

With genome and transcriptome sequencing rapidly becoming a commodity, there is an increasing need to analyze custom user data. Here, we present OMA standalone, an open-access software implementation of the OMA pipeline for Linux and Mac (http://omabrowser.org/standalone). We first outline some of the

[10]Joint first authors.
Corresponding author: Christophe.Dessimoz@unil.ch
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.243212.118. Freely available online through the *Genome Research* Open Access option.

29:1152–1163 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/19; www.genome.org

key features of OMA standalone. In the second part, we demonstrate the usefulness of OMA standalone in the context of species tree inference, by comparing its performance with state-of-the-art alternatives on the challenging Lophotrochozoa phylogeny.

Results

We first highlight the defining features of OMA standalone, then turn to the phylogeny of Lophotrochozoa, which we infer from orthologs inferred by OMA in comparison with alternative methods.

OMA standalone software

OMA standalone takes as input the coding sequences of genomes or transcriptomes, in FASTA format. The recommended input type is amino acid sequences, but OMA also supports nucleotide sequences. With amino acid sequences, users can combine their own data with publicly available genomes from the OMA database, including precomputed all-against-all sequence comparisons (the first and computationally most intensive step), using the export function on the OMA website (http://omabrowser.org/export).

OMA standalone produces several types of output (also summarized in Fig. 1):

1. Pairwise orthologs and their subtypes (one-to-one, one-to-many, many-to-one, many-to-many orthology). These orthologs are useful when comparing pairs of species or to identify orthologs to specific genes of interest.
2. OMA groups. These are sets of genes for which all pairs are inferred to be orthologous. These groups are inferred as cliques (fully connected subgraphs) of pairwise orthologs. These groups are not necessarily one-to-one orthologs, but, being inferred without assuming a species tree, they are particularly useful to identify marker genes for phylogenetic reconstruction.
3. Hierarchical orthologous groups (HOGs). These groups are defined for every internal node of the (rooted) species tree; each HOG contains the genes that are inferred to have descended from a common ancestral gene among the species attached to that internal node. Consider, for instance, gene *ADH1*, which duplicated within the primates (Carrigan et al. 2012): At the level of the last primate common ancestor, all genes that have descended from the ancestral *ADH1* belong to the same HOG. However, at the level of the common ancestor of all the great apes, because *ADH1* had at this point already duplicated into *ADH1A*, *ADH1B*, and *ADH1C*, these ancestral genes define three HOGs. A brief video tutorial on HOGs is available at https://youtu.be/5p5x5gxzhZA. The HOGs are stored in the standard OrthoXML format (Schmitt et al. 2011).
4. Gene Ontology annotations. OMA standalone annotates the input sequences with Gene Ontology annotations by propagating high-quality annotations across orthologs (Altenhoff et al. 2015). The annotations are provided in the standard GO Annotation File Format 2.1 (http://geneontology.org/docs/go-annotation-file-gaf-format-2.1).
5. Phylogenetic profiling. Orthology is also used to build phylogenetic profiling—patterns of presence and absence of genes across species (Pellegrini et al. 1999). We provide two forms of output: a binary matrix with species as rows and OMA groups as columns, indicating patterns of presence or absence of genes in each group; a count matrix with species as columns and HOGs as rows, indicating the number of genes in each deepest-level HOG (i.e., HOG defined at the broadest taxonomic level possible).
6. Species tree. Unless supplied with a (fully or partially resolved) reference species tree, OMA standalone computes a tree from the inferred OMA groups using the built-in distance tree procedure *MinSquareTree* in the programming environment *Darwin* (Gonnet et al. 2000). Note that, as with most tree inference methods, the rooting of the tree tends to be unreliable, so we encourage users to review and reroot the tree based on other information, if available.

OMA standalone supports parallel computation of the all-against-all sequence comparison phase. This phase, which computes Smith–Waterman (1981) alignments followed by pairwise maximum likelihood distance estimation for all significant pairs (Roth et al. 2008), is by far the most time-consuming step of the algorithm. To fully exploit parallelism, alignments are performed using single instruction multiple data (SIMD) instructions

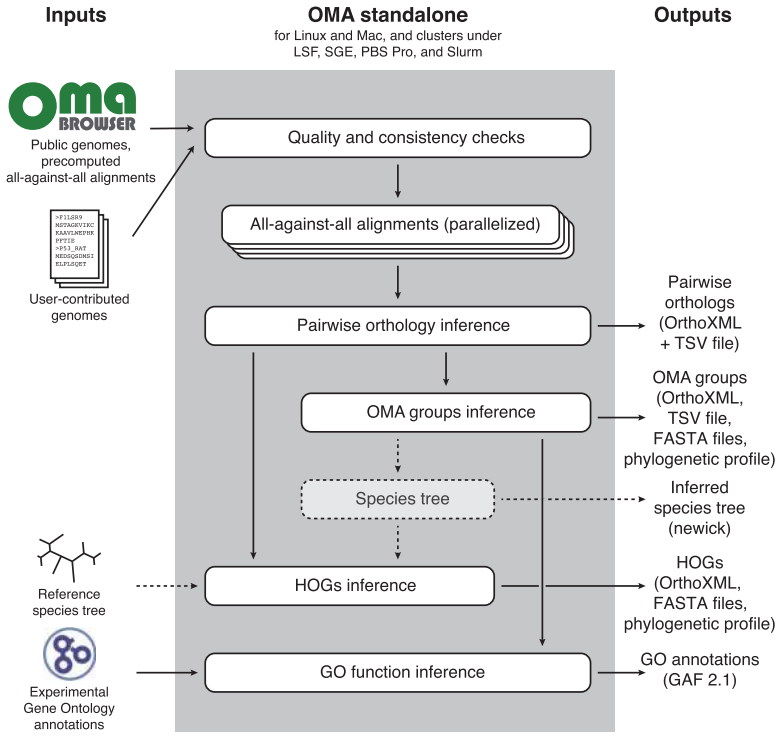

Figure 1. Conceptual overview of the OMA standalone software. Dotted arrows indicate alternative steps (reference species tree either specified as input or inferred from the data). The species tree inference step infers a distance tree but can be bypassed by supplying a reference tree.

(Szalkowski et al. 2008) on multiple cores. OMA standalone natively supports common cluster schedulers—LSF, SGE, PBS, and Slurm—and has been successfully run with several thousand jobs in parallel. Figure 2 shows typical runtimes and memory usage for data sets of various sizes.

Application: the phylogenetic relationships within Lophotrochozoa

Resolving the relationships of ancient lineages is a major challenge for molecular phylogenetics. Although some aspects of the phylogeny of the major animal clades are well-resolved, the relative positions of the deeper lying clades are often disputed. The construction of large phylogenomic supermatrices has been the method of choice for resolving the deepest nodes in the tree of life (Dunn et al. 2008; Hejnol et al. 2009; Fernández et al. 2014; Egger et al. 2015).

Fundamental to the analyses of phylogenetic relationships is the use of sequences that have descended from a single common gene in their last common ancestor, that is, orthologous sequences. Ensuring that we correctly infer orthologs is therefore vital if we are to reconstruct difficult to resolve phylogenies. The limitations of automated orthology and paralogy prediction methods with regard to phylogenetic analysis have previously been highlighted (Philippe et al. 2011b); simplistic orthology inference methods may miss orthologs (Dalquen and Dessimoz 2013) or erroneously identify paralogous pairs of genes as orthologs as a result of differential gene losses (Dessimoz et al. 2006).

One notoriously difficult to resolve phylogeny is that of Lophotrochozoa (Kocot 2016), a clade of animals positioned sister to Ecdysozoa, within the protostomes, and which, for instance, includes segmented worms and molluscs. Lophotrochozoa contains about 10 different phyla, each of which is clearly monophyletic, but the relationships among these phyla are far from clear, with many different topologies having been supported by different analyses. The inference is that the phyla are likely to have emerged in an ancient and rapid radiation resulting in weak phylogenetic signal for interphylum relationships. These circumstances make the solving of this problem particularly difficult and mean that the use of accurately identified orthologs is particularly relevant.

We used OMA standalone to identify orthologous marker genes among the proteomes of 19 lophotrochozoans and, as outgroups, four deuterostomes, four ecdysozoans, and three nonbilaterians, totaling 894,528 input sequences (see Methods). As a basis of comparison, we also repeated the analysis using orthology inference pipelines, on the same data set, based on OrthoMCL (Li et al. 2003), BUSCO (Simão et al. 2015), HaMStR (Ebersberger et al. 2009), and OrthoFinder (Emms and Kelly 2015). Like OMA, these methods do not require prior specification of a species tree, are available as standalone programs, and have all been used in phylogenetic analyses previously. Species trees were then constructed using these orthologs with both maximum likelihood and Bayesian tree reconstruction packages, IQ-TREE (Nguyen et al. 2015) and PhyloBayes (Lartillot et al. 2013), on the resultant supermatrices. In terms of computational cost, OMA is by far the most costly of the orthology methods tested, due to its reliance on full Smith–Waterman (1981) alignments and evolutionary distance in the all-against-all phase (~85 k CPU hours). By comparison, OrthoMCL and OrthoFinder, which rely on BLAST for all-against-all comparisons, are much faster (~2 k CPU hours). Finally, BUSCO (11 CPU hours) and HaMStR (230 CPU hours) are the fastest, owing to their reliance on predefined hidden Markov models of the orthologous markers.

We first consider the amount of orthology information recovered by the various methods. OMA inferred 2162 orthologous groups containing 15 or more species (Fig. 3A). By comparison, the HaMStR pipeline inferred 1241 orthologous groups, the OrthoMCL pipeline inferred 484 orthologous groups, BUSCO

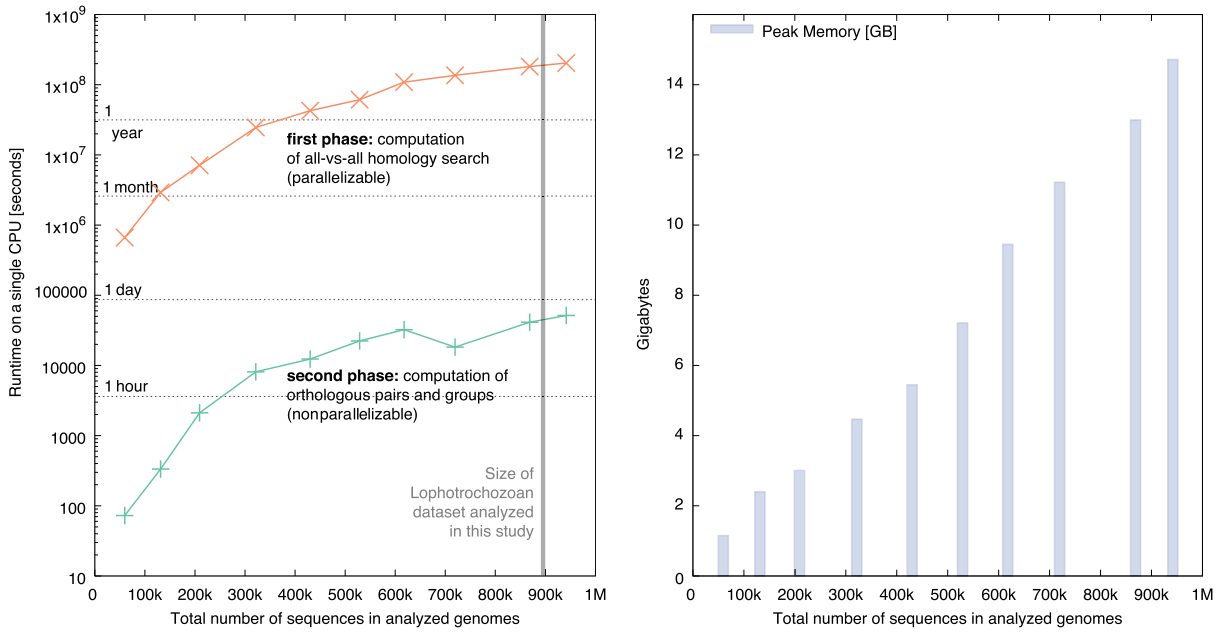

Figure 2. Resource measurements for various data sets of increasing sizes as total number of protein sequences. The data sets have been sampled from the public OMA Browser to maintain a constant composition of 20% fungi, 10% archaea, 10% plants, 20% metazoan, and 40% bacteria genomes. (*Left*) Runtime of the all-against-all phase (orange) on a single CPU, and the inference of the orthologous pairs and various groups (green). (*Right*) Peak memory usage of OMA standalone in gigabytes (GB).

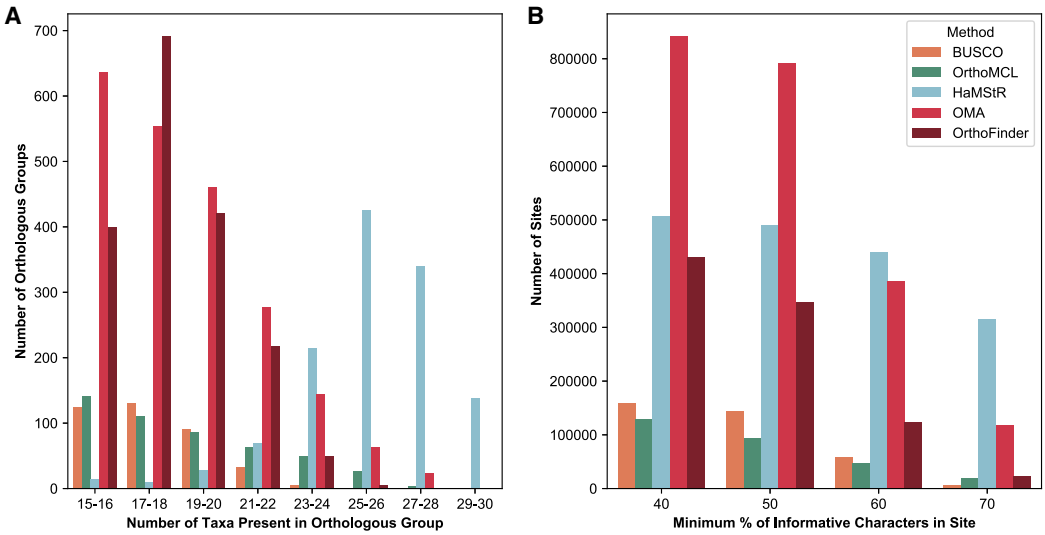

Figure 3. Comparison of amount of orthologous data inferred by the different pipelines. (*A*) OMA and OrthoFinder infer more orthologous groups than other methods, whereas the groups inferred by HaMStR are considerably larger on average than for the other methods. (*B*) The resulting supermatrix has most sites for OMA, whether the minimum site occupancy threshold is 40% or 50%, and most sites for HaMStR at the 60% cutoff (used for phylogenomic reconstruction) and 70% cutoff.

inferred 384 orthologous groups, and OrthoFinder yielded 1784 groups. Although OMA identifies more orthologous genes than other methods, it infers fewer larger groups than HaMStR and produces a less dense data matrix (Supplemental Fig. S1). This difference in group size distribution is likely to be the result of different trade-offs in terms of precision (proportion of predicted orthologs that are correct) and recall (proportion of true orthologs that are correctly predicted). These trade-offs have been observed in multiple benchmarking studies (e.g., Boeckmann et al. 2011; Altenhoff et al. 2016). Indeed, the OMA algorithm is known for having higher precision but lower recall than most other methods (Altenhoff et al. 2016).

A priori, the effect of the number of orthologous groups and completeness on tree inference is not obvious. The effect of missing data in even large supermatrices has been shown to have a detrimental effect on the quality of trees inferred from them (Roure et al. 2013). Other studies have shown that more complete supermatrices do not necessarily yield better results (Fernández et al. 2016). The latter study found that when using a stringent minimal site completeness cut-off, resulting in fewer sites, phylogenetic inference was in disagreement with established classifications of taxa.

If we consider both the number of sites above a minimum occupancy rate threshold (i.e., minimum proportion of informative characters in each site), OMA standalone yields the largest data matrices (i.e., the most alignment columns) with at least 40% or 50% occupancy, while HaMStR yields the largest data matrices for 60% and 70% (Fig. 3B).

Using the aligned sets of orthologs identified in the previous step, we reconstructed species trees using Maximum Likelihood (IQ-TREE [Nguyen et al. 2015], a model selected with ModelFinder [Kalyaanamoorthy et al. 2017]), and Bayesian analysis (PhyloBayes, CAT + GTR + G4 [Lartillot et al. 2013]) on supermatrices that had been filtered to include only alignment columns with at least 60% site occupancy. In the rest of our analyses, we chose to infer trees from matrices with a minimum occupancy rate of 60%, for pragmatic reasons: With higher thresholds, some methods recover too few sites (e.g., BUSCO yields 7135 positions only if we require at least 70% occupancy). With a lower cutoff, the increase in data matrix size renders Bayesian tree inference analyses prohibitively costly.

With OMA, both the Bayesian tree (using PhyloBayes; Fig. 4) and the ML tree (using IQ-TREE; Supplemental Fig. S2) had high branch support values. The Bayesian tree had branch posterior probabilities of 1 across the tree apart from the Lophotrochozoa clade, with a posterior probability of 0.82. The ML tree had bootstrap support of 100 for all but eight of 27 branches. Deuterostomes were recovered with full bootstrap support, while Lophotrochozoa, with the exception of Rotifera, were recovered with bootstrap support of 92.

The OMA tree inferred using the ML inference method found that the Rotifera (*Adineta ricciae, Brachionus plicatilis*) are grouped with the Nematoda (*Caenorhabditis elegans*, *Pristionchus pacificus*), as part of the ecdysozoans. This is in disagreement with the current consensus (Giribet and Edgecombe 2017). In contrast, the tree constructed using Bayesian inference found the Rotifera to be sister to the rest of the lophotrochozoans, in agreement with recent studies (Philippe et al. 2011a; Egger et al. 2015). The discrepancy in the ML tree is likely due to the long branched Rotifera being attracted to the long branched Nematoda—a problem to which PhyloBayes under the CAT model has been previously shown to be more robust (Lartillot et al. 2013).

Both the ML and Bayesian trees found the rest of the lophotrochozoans to consist of two monophyletic groups. The first group comprises the Gastrotricha (*Mesodasys laticaudatus*) and the Platyhelminthes (flatworms). This relationship is consistent with recent studies (Dunn et al. 2008; Edgecombe et al. 2011; Struck et al. 2014; Laumer et al. 2015). Because of their seemingly simple morphology, with characteristics such as having no body cavity, no respiratory organs, and having only a single opening for both the intake of nutrients and excretion of waste, they were originally thought to be among the most basally branching Bilateria, until molecular studies on 18S rDNA sequence data was carried out, placing them within the protostomes (Baguñà and Riutort 2004).

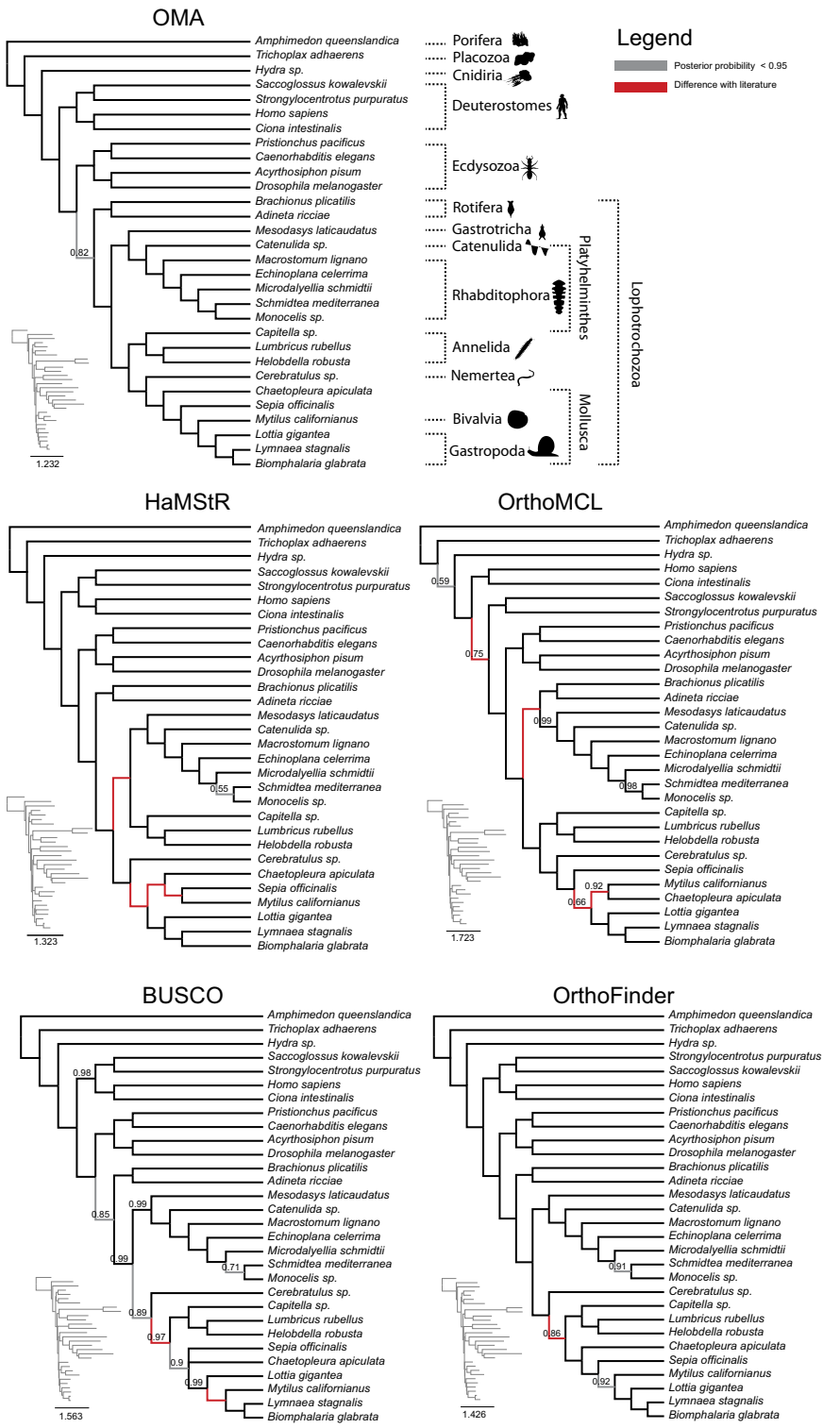

Figure 4. Comparison of trees obtained using PhyloBayes with the CAT-GTR-G4 model from the different orthology methods. OMA tree is in congruence with published results (see main text). Branches that are at odds with the literature are in red; otherwise they are displayed in gray (posterior probability < 0.95) or else in black. Only posterior probabilities below one are displayed. Please note that the PhyloBayes tree computed from HaMStR data did not converge after 900,000 CPU hours and thus should be interpreted with caution.

Authors now divide the Platyhelminthes into the Catenulida, with currently no known synapomorphies (i.e., no shared distinctive character), and the Rhabditophora, which has uniting characteristics such as the presence of lamellated rhabdites, a common structure of the epidermis (Egger et al. 2015; Laumer et al. 2015). Our ML and Bayesian trees corroborated this and found the Catenulida (*Catenulida* sp.) to be sister to Rhabditophora (*Macrostomum lignano, Echinoplana celerrima, Microdalyellia schmidtii, Monocelis* sp., *Schmidtea mediterranea*).

Within the Rhabditophora, the most basal branches of the OMA-inferred trees are those of the Macrostomorpha (*Macrostomum lignano*), followed by the Polycladida (*Echinoplana celerrima*), also in agreement with recent studies (Egger et al. 2015; Laumer et al. 2015). We also inferred the Rhabdocoela (*Microdalyellia schmidti*) to be the most basally branching, followed by the Proseriata (*Monocelis sp.*) and Acentrosomata (*Schmidtea mediterranea*). This too is in agreement with recently published phylogenies (Egger et al. 2015; Laumer et al. 2015).

The second monophyletic group found within the rest of Lophotrochozoa contains the Annelida (*Lumbricus rubellus*, *Helobdella robusta*, *Capitella* sp.), segmented worms, the Mollusca (*Biomphalaria glabrata*, *Lymnaea stagnalis*, *Lottia gigantea*, *Mytilus californianus*, *Sepia officinalis*, *Chaetopleura apiculata*), the largest marine phylum, and Nemertea (*Cerebratulus* sp.), also known as ribbon worms or proboscis worms, to form the Trochozoa (Dunn et al. 2014). However, there is disagreement on the positioning of these clades within the group (Dunn et al. 2008; Struck and Fisse 2008; Struck et al. 2014; Laumer et al. 2015). Both tree reconstruction methods find the Gastropoda (*Lottia gigantea*, *Lymnaea stagnalis*, *Biomphalaria glabrata*) to be sister to the Bivalvia (*Mytilus californianus*). Both methods also found the Annelida to be sister to (Mollusca + Nemertea), with high support (posterior probability of 1 and bootstrap of 96).

In contrast, on this lophotrochozoan data set, trees obtained from other orthology pipelines had more unresolved nodes and/or more discrepancies with the literature (Fig. 4; Supplemental Table S1; Dunn et al. 2008; Kocot et al. 2011; Egger et al. 2015; Laumer et al. 2015; Telford et al. 2015; Kocot et al. 2017).

The BUSCO Bayesian tree had slightly less support throughout than

the OMA tree, although it only had one branch with support of less than pp = 0.80. The relationship between the Proseriata, Rhabdocoela, and the Acentrosomata agrees with the OMA Bayesian tree, as does the relationship between the Gastrotricha and the Platyhelminthes. However, the BUSCO tree indicates Gastropoda to be paraphyletic with high support (pp = 0.99), with *Lottia gigantea* more basally branching to the Bivalvia and the rest of the Gastropoda. This is in contrast to both the OMA tree and other studies (Dunn et al. 2008; Struck et al. 2014). The BUSCO tree found the Nemertea as sister to (Annelida + Mollusca), with a support value of pp = 0.89. This is in disagreement with the current consensus and the OMA tree (Dunn et al. 2008; Struck et al. 2014; Laumer et al. 2015).

The HaMStR tree had high support throughout but differed markedly from the OMA tree. The HaMStR method placed *Sepia officinalis, Mytilus californianus*, and *Chaetopleura apiculata* in a clade together, sister to the Gastropoda. This is in disagreement with Kocot et al. (2011) and the OMA trees, which place the Polyplacophora (*Chaetopleura apiculata*) to be the most basally branching, followed by the Cephalopoda (*Sepia officinalis*), with the Bivalvia sister to the Gastropoda. The Bayesian tree also fails to recover Trochozoa, placing the Annelida with the (Platyhelminthes + Gastrotricha), as opposed to full support found in the OMA tree. One caveat with the Bayesian HaMStR tree is that the tree reported is unconverged (even after 22,230 iterations); thus, we cannot rule out that some of these differences might ultimately disappear. However, the ML tree also shows substantial disagreement with the OMA tree and the literature (Supplemental Table S1).

The OrthoMCL trees had the most issues, with the lowest support values. Deuterostomes, comprising a well-established relationship between the chordates and the Ambulacraria (Philippe et al. 2011a), are paraphyletic in the Phylobayes tree, which places chordates (*Ciona intestinalis, Homo sapiens*) more basally branching than the Ambulacraria (*Strongylocentrotus purpuratus, Saccoglossus kowalevskii*), with the latter sister to the Protostomes with pp = 0.75. Rotifera were incorrectly placed as sister to (Gastrotricha + Platyhelminthes) with full support. This is in disagreement with both the OMA tree and recent studies. The tree was able to correctly infer the (Mollusca + Nemertea) relationship with full support. Within the Mollusca, in contrast to the OMA tree, the Bayesian tree inferred *Sepia officinalis* to be the most basally branching, with *Chaetopleura apiculata* and *Mytilus californianus* forming a clade sister to the rest of the Mollusca. However, this has low support with pp = 0.66 for the Bayesian tree.

The OrthoFinder Bayesian tree was less supported than the OMA tree, with three values below pp = 1. The Nemertea were found to be sister to (Annelida + Mollusca), in contrast to the OMA Bayesian tree. The ML tree was also weakly supported, with nine branches with less than full support and six below bs = 80. The Rotifera were found to be sister to Platyhelminthes, as part of a clade with the Gastrotricha. This is in disagreement with recent analysis, which places them as sister to the rest of the lophotrochozoans. The phylogeny of the Mollusca differed from the OMA tree, with *Chaetopleura apiculata* and *Sepia officinalis* inferred as sister to one another, with bs = 44, which were in turn sister to (Bivalvia + Gastropoda).

The different data matrices used to build phylogenies span an almost 10-fold difference in terms of informative sites. To better understand the potential impact of these differences, we sought to compare the quality of trees obtained from matrices subsampled to similar sizes, still on the lophotrochozoan data set. From each of the sets of orthologous groups produced by each method, a number of orthologous groups were selected at random, without replacement, but nevertheless ensuring that every species was represented at least once. For this analysis, which required the reconstruction of many species trees, we used IQ-TREE under the WAG + I model—which we found to be a reasonable trade-off between speed and accuracy. To gauge the accuracy of the resulting trees, we compared them with a partially resolved reference tree derived from the literature (see Supplemental Table S1). We observed that the lower accuracy of trees reconstructed from BUSCO and OrthoMCL is not solely due to the lower number of orthologous groups they infer: The resulting trees were less accurate even when we considered the same number of groups for all methods (Fig. 5). More generally, the analysis shows the merit of including more orthologous groups, as for most methods this leads to an increase in tree accuracy.

Discussion

OMA standalone enables researchers to infer high-quality orthologs among genomes or transcriptomes, on public and in-house data. It runs on a wide range of hardware, from a single computer to large clusters with thousands of parallel processes.

A key application of OMA standalone lies in the identification of genome-wide orthologous marker sequences to infer difficult species phylogenies. On the lophotrochozoan data set, compared with other approaches, OMA yielded more orthologous information for phylogenetic species tree inference and resulted in better resolved trees, which are also more consistent with the existing literature. BUSCO finds orthologs by comparing sequence data to a predefined set of genes present in at least 90% of the species in a given data set. This relies on preexisting knowledge of orthology relationships in a set of reference species, in this case, the species present in the Metazoa data set. Therefore, the number of orthologous groups is limited to 843. Similarly, HaMStR relies on predefined core orthologs, which in our case were obtained from a

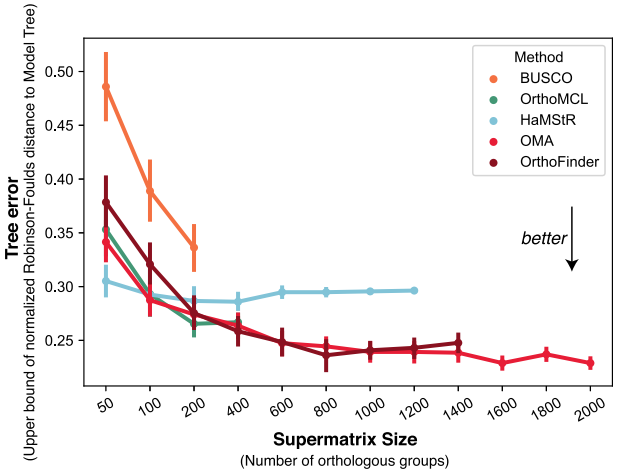

Figure 5. Accuracy of trees reconstructed with varying number of orthologous groups, on the lophotrochozoan data set, using IQ-TREE with a WAG + I model. Each point is obtained by averaging over results obtained from 50 random group subsets of varying size, drawn without replacement. Even if all methods are downsampled to have the same number of groups, trees obtained from OMA are consistently among the most accurate ones (measured in terms of the Robinson-Foulds distance to a partially resolved reference tree) (see Methods). Error bars depict one standard error on each side.

high-profile previous study of the Annelida phylogeny (Weigert et al. 2014b). One advantage of such predefined sets is that phylogenetically uninformative or misleading genes might have already been excluded. The downside is that the number of orthologous groups is limited to 1253 orthologous groups. OMA is advantageous in this regard because it infers orthologs across potentially all protein sequences. This may in part explain why OMA standalone has been adopted in other phylogenomic studies, such as for centipedes (Fernández et al. 2014), arachnids (Sharma et al. 2014; Fernández and Giribet 2015), assassin flies (Dikow et al. 2017), scorpions (Sharma et al. 2015), spiders (Garrison et al. 2016), flatworms (Egger et al. 2015; Laumer et al. 2015), tapeworms (Tsai et al. 2013), Spiralian (Marlétaz et al. 2019), or Archaea (Williams et al. 2017).

Our comparison also has methodological implications for phylogenomic studies. These studies are typically greatly concerned about the impact of the evolutionary model on tree inference (e.g., Song et al. 2012), as well as that of taxon sampling (e.g., Dunn et al. 2008), but the impact of orthology inference methods has not nearly been as commonly investigated. Our comparison of orthology methods on the lophotrochozoan data set highlights the considerable impact orthology inference can have on phylogenetic tree inference. Thus, a more systematic investigation of the impact of orthology inference on phylogenetic tree inference may be required to resolve the most vexing phylogenetic questions, such as that of the ctenophore placement (Pisani et al. 2015, 2016; Whelan et al. 2015, 2017; Halanych et al. 2016; Feuda et al. 2017).

One drawback of the current OMA algorithm is its high computational cost compared to the other methods. It would be possible to replace the costly Smith–Waterman alignments by fast heuristics such as DIAMOND (Buchfink et al. 2015) or MMSeq2 (Steinegger and Söding 2017). However, currently about half of the time spent in the all-against-all phase is to compute pairwise evolutionary distances, which would still be needed—thus fast heuristics would only provide a 2× speed-up to the OMA pipeline at best. Instead, we see potential in avoiding the computation of some pairs altogether by exploiting the transitivity property of homology (Wittwer et al. 2014).

The comparative analysis has some limitations. First, the taxon sampling is far from optimal, with several clades, such as the rotifers, suffering from long branches. Since we started this study, more lophotrochozoan genomes have become available; their inclusion would likely improve the resolution of the trees. Second, while running and comparing five orthology inference methods on a data set of nearly 900,000 sequences already represents a major undertaking, other orthology methods would be interesting as well —in particular, tree-based approaches that require no prior species tree knowledge (Yang and Smith 2014; Huerta-Cepas et al. 2016a).

Beyond species tree inference, OMA can also be used to pinpoint the emergence of gene families in evolution, an approach that is sometimes referred to as phylostratigraphy (Domazet-Lošo et al. 2007). Conventional approaches work by considering all the genes annotated in a species of reference and performing BLAST searches against increasingly distant sets of taxa. The point at which no homolog can be found is inferred to immediately precede the emergence of the gene. However, such an approach does not differentiate between orthologs and paralogs and thus has a limited resolution in terms of subfamilies. Alternatively, it is possible to extract more fine-grained information from reconciled gene trees—i.e., gene trees with internal nodes labeled as speciation or duplication nodes (Vilella et al. 2008; e.g., Huerta-Cepas et al. 2014)—but this is computationally demanding and there is a lack of tools to perform such analyses on custom data.

By inferring high-quality hierarchical orthologous groups, OMA standalone provides a way to map gene emergence, gene duplication, and gene loss onto species phylogenies. For instance, OMA standalone has been used to contrast gene families that have expanded and contracted in the common ancestors of echolocating and nonecholocating bats. The emergence of echolocation coincides with a decrease in chemosensory genes, while secondary loss of echolocation coincides with an increase in chemosensory genes (Tsagkogeorga et al. 2017). The hierarchical orthologous groups inferred by OMA standalone can be further analyzed using the *iHam* visualization tool and the *pyHam* Python library (Train et al. 2018).

Orthology is also key to integrating biological knowledge among model and nonmodel species. Particularly when dealing with deep timescales, it can be challenging to identify genes with or without orthologous counterparts. By reconstructing fine-grained orthology between mice and protostomes, OMA standalone could identify new drug targets for neglected tropical diseases (Tsai et al. 2013). With such diseases, which disproportionately affect poorer people, it can be challenging to develop new medicines. To accelerate drug development in such cases, drug repurposing has been suggested whereby an already existing and approved medicine, or a well-researched lead, is used to combat neglected tropical diseases (Ekins et al. 2011). As a first-pass bioinformatic identification of drug targets in four newly sequenced tapeworm genomes, OMA standalone was used to identify orthologs of known human drug targets (Tsai et al. 2013): Human genes targeted by drugs were retrieved from various databases, and their orthologs in tapeworms were inferred using OMA standalone. To identify targets likely to be essential across animals, orthologs present in both mice and nematodes were also identified: If both mice and nematode orthologs had knock-out phenotypes, we inferred that the orthologous group was essential across animals. Together with other indicators, such as gene expression data, we were able to rank every gene in these largely unexplored genomes for their suitability as a drug target and associate lead compounds to them. As drugs could exhibit off-target effects on paralogs, the analysis focused on orthologs, which tend to be functionally more conserved (e.g., Altenhoff et al. 2012). The importance of investigating orthologs was illustrated by the drug Praziquantel, which is efficient against adult tapeworms but not against the more dangerous larval form (Nogi et al. 2009). Praziquantel targets one particular voltage-gated calcium channel subunit. Using OMA standalone, we could identify the precise subunit ortholog in tapeworms and show that it is not expressed in the larval form, thereby providing a plausible explanation for the drug's low efficacy.

To conclude, orthology inference is a key step in integrating biological knowledge across multiple species. OMA standalone is a versatile orthology inference software with a proven track record. Contrary to some of the orthology methods considered in this study, it was designed from the onset with species tree inference in mind, though it has since been applied for a broad range of other applications. The OMA standalone software implementation has been continuously improved and maintained over the past 5 yr, undergoing two major and 25 minor releases—in the course of which a considerable number of bugs were identified and fixed (https://omabrowser.org/standalone/release_notes.txt). We intend to keep developing and maintaining it. For support enquiries or bug reporting, we encourage users to use the biostars.org forum using the keyword "oma" (https://www.biostars.org).

Table 1. List of input parameters of OMA standalone

Parameter	Meaning	Default
InputDataType	Type of input sequences. This can be set either to 'AA' for amino acid sequences or 'DNA' for nucleotide sequences	AA
OutputFolder	Folder to which the output is written. At each run, the content of this folder will be overwritten. Don't store any important files in it. The OutputFolder must not contain any spaces.	Output
ReuseCachedResults	If you want to recompute everything from scratch every time the script is run, set this to false.	True
AlignBatchSize	In the all-against-all phase, each genome pair is split in smaller chunks of AlignBatchSize protein comparisons. The larger this number, the longer each unit runs, and the fewer files get produced. This allows to adjust the frequency of milestone steps (e.g., in case of computer crash) or to process few but large genomes with many CPUs efficiently.	1,000,000
MinScore	Alignments that have a score lower than MinScore will not be considered. The scores are in Gonnet PAM matrices units.	181
LengthTol	Length tolerance ratio. If the length of the effective alignment is less than LengthTol × min[length(s1), length(s2)], then the alignment is not considered.	0.61
StablePairTol	During the stable pair formation, if a pair has a distance provable higher than another pair (i.e., StablePairTol standard deviations away), then it is discarded.	1.81
VerifiedPairTol	Tolerance parameter for the detection of differential gene losses using a third genome. The larger the tolerance, the more liberal the algorithm assigns orthologous relations. A detailed description is provided in Dessimoz et al. (2006).	1.53
MinSeqLen	Any sequence that is less than MinSeqLen amino acids long in regular genomes is not considered.	50
UseOnlyOneSplicingVariant	Enables/disables the filtering on a single representative splicing variant. If enabled, OMA selects the variant that has the most homologous matches with all other genomes. Orthology inference is then only based on this variant. If disabled, alternative splicing variants will usually be inferred as paralogs.	True
StableIdsForGroups	Enables/disables the generation of stable identifiers for OMA groups (and Hierarchical Groups if the top-down algorithm is selected). The identifier consists of a prefix to determine the type of the group ('OMA' or 'HOG') and a subsequence of the amino acid sequence uniquely present in this group. The computation of these IDs might require a substantial amount of time. The IDs are stored in the OrthoXML files only (Schmitt et al. 2011).	False
GuessIdType	Enable/disable guessing of the ID types while generating the OrthoXML file (Schmitt et al. 2011). In this context, we refer to ID type guessing as the task to guessing whether an ID should be stored in the geneId, protId, or transcriptId tag. If the flag is set to false, the whole FASTA header is used and stored as is in the protId tag.	False
DoHierarchicalGroups	Enables/disables and selects the algorithm to compute the Hierarchical Orthologous Groups (HOGs). Valid parameters are false, 'top-down,' and 'bottom-up.' The top-down approach was the only algorithm until OMA standalone 2.0. The bottom-up approach is as of now still an experimental feature but will become the default choice in the future.	'Top-down'
MaxTimePerLevel	Define maximum amount of time (in sec) spent by the program for breaking every connected component of the orthology graph at its weakest link on a given taxonomic level. If set to a negative value, no time limit is enforced. Once the time limit is reached, OMA will treat the remaining connected component at the lower level (groups will not span over the deeper node).	1200
SpeciesTree	The hierarchical groups require a (partially) resolved species phylogeny. With the parameter SpeciesTree, the user can specify a phylogeny in Newick-format, or, by setting the variable to "estimate," compute a species tree based on the OMA Groups and use this one.	Estimate
ReachabilityCutoff	The cutoff of "average reachability within two steps" defines up to what point a cluster is split into subclusters. Details on this parameter are explained in Altenhoff et al. (2013). This parameter applies only to the top-down HOG inference approach. See parameter DoHierarchicalGroups for additional information.	0.65
MinEdgeCompletenessFraction	The cutoff in GETHOGs bottom-up algorithm to make an edge trusted in the orthology graph among HOGs. This parameter applies only to the bottom-up approach. See parameter DoHierarchicalGroups for additional information.	0.80
DoGroupFunctionPrediction	Compute Gene Ontology function predictions based on the OMA Groups assignments. The predictions are then stored in a GAF file. Computing these predictions can take a substantial amount of time. Note: Predictions are based on transferring existing annotations from genomes. Only genomes exported through the OMA Browser export interface (https://omabrowser.org/export) have usable functional input annotations.	True
GroupFunctionCutoff	Parameter to specify the fraction of genes in a group that need to be annotated with a GO term in order to propagate the annotation to the unannotated group members. The parameter ensures that predictions are not propagated too liberally. Note that it requires at least the specified fraction of genomes to be exported from the OMA Browser.	0.5
CladeDefinition	Path to tab-separated file that provides a mapping from the species names to the clade/group to which annotations should at most be propagated. If set to default, the algorithm infers a species tree and propagates GO annotations to user genomes only within some predefined clades. These predefined clades are 'Amphibia,' 'Archaea,' 'Arthropoda,' 'Bacteria,' 'Clupeocephala,' 'Dictyostelium,' 'Fungi,' 'Mammalia,' 'Nematoda,' 'Sauria,' and 'Viridiplantae.' If the parameter is set to false or none, no limitations on the clades are used to propagate the function annotations.	Default

Methods

OMA standalone

The list of all parameters of OMA standalone, their meaning, and default values is provided in Table 1.

Large-scale species phylogenetic reconstruction: Lophotrochozoa

Transcriptome assembly and peptide prediction

We used transcriptomes from seven Lophotrochozoa species published in Egger et al. (2015): *Mesodasys laticaudatus* (Gastrotricha), *Catenulida* sp., *Macrostomum ligano*, *Echinoplana celerrima*, *Microdalyellia schmidtii*, *Monocelis* sp. (Platyhelminthes), and *Cerebratulus* sp. (Nemertea). In addition, 12 sets of genomic and transcriptomic protein predictions from *Saccoglossus kowalevskii*, *Brachionus plicatilis*, *Adineta ricciae*, *Schmidtea mediterranea*, *Lumbricus rubellus*, *Chaetopleura apiculata*, *Sepia officinalis*, *Mytilus californianus*, *Biomphalaria glabrata*, *Lymnaea stagnalis*, *Hydra magnipapillata*, and *Amphimedon queenslandica* were downloaded from the NCBI RefSeq repository (ftp://ftp.ncbi.nlm.nih.gov/refseq/).

Quality assessment of sequencing reads was carried out with FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Subsequent to this, it was determined, using PRINSEQ lite (Schmieder and Edwards 2011), that the first 12 nucleotides should be trimmed off the 100-bp reads. The assembly of the trimmed paired reads was done using Trinity v20130225 (Haas et al. 2013), with the flag '--min_kmer_cov 2', with default parameters. Open reading frames (ORFs) were predicted using TransDecoder (Haas et al. 2013). All ORFs greater than 100 amino acids were retained. Redundant sequences with higher than 97% identity at the amino acid level were removed by clustering with CD-HIT (Fu et al. 2012).

In order to detect the presence of cross contaminations between the various libraries run on the same flow cell, we used the CroCo package (Simion et al. 2018). This identified any assembled transcripts with fewer than four read matches, which were subsequently discarded. Furthermore, this also discarded all transcripts in which the number of reads, from the intended species matching the transcript, was not at least five times greater than the number of matches to the transcript, from reads from any of the other potentially contaminating species.

Additionally, 11 precomputed proteomes for *Homo sapiens*, *Strongylocentrotus purpuratus*, *Ciona intestinalis*, *Trichoplax adhaerens*, *Pristionchus pacificus*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Acyrthosiphon pisum*, *Capitella* sp., *Helobdella robusta*, and *Lottia gigantea* were downloaded from the OMA database website. The combined set of 30 nonredundant protein sets contained 19 lophotrochozoans, four deuterostomes, four ecdysozoans, and proteomes from three nonbilaterian animals.

Orthology inference

For the HaMStR analysis, putative orthologs were determined for each species using HaMStR v13.1 (Ebersberger et al. 2009) using the Lophotrochozoa core ortholog reference data set (Weigert et al. 2014a,b) as required by the HaMStR tool, with default parameters. HaMStR was run with the "-representative" option to pick at most one sequence per species, with all other parameters as default.

Orthologous groups were inferred by running BUSCO v1.22 (Simão et al. 2015) on the Metazoa data set found at https://busco.ezlab.org/v1/. We created orthologous groups made up of the protein sequences which BUSCO deemed to have had complete matches with their own highly conserved genes. At most, one species containing multiple sequences was allowed per group. There was only a single occurrence of a group containing more than one species with multiple sequences. In this case, we retained only the longest sequence.

The set of 30 proteomes were first filtered to remove low-quality protein sequences using the OrthoMCL script "orthomclFilterFasta.pl" (Chen et al. 2006). The "orthomclFilterFasta.pl" script filters away poor-quality sequences based on their length and percent stop codons. Default parameters were used, which retains only sequences with a minimum length of 20 characters, and fewer than 10% stop codons. This step resulted in the exclusion of 29 out of 894,528 input sequences (0.0032%). An all-versus-all NCBI BLAST v2.7.1 was then used with default parameters, in order to find the similarity score between sequences. Matches with an *E*-value $<10^{-6}$ were retained. Orthologs, in-paralogs, and co-orthologs were then identified using the OrthoMCL script "OrthomclPairs.pl" (Chen et al. 2006) before clustering using MCL. An MCL inflation parameter of 2.2 was used in order to identify clusters. Each group was required to have at most one species containing multiple sequences. When more than one sequence from a single species was present, the longest sequence was selected to remain in the group, with the others removed.

Putative orthologs were inferred using OrthoFinder v2.2.7, used in conjunction with BLAST, with default parameters. When applying the same criteria as for OrthoMCL for generating single copy orthologs (i.e., at most sequence per species), no orthologous groups were recovered. As a workaround, within each orthologous group, we removed all sequences from species that appeared multiple times.

Table 2. Best fit model found by ModelFinder

Method	IQ-TREE model
OMA	LG + F + G4
HaMStR	LG + F + G4
BUSCO	LG + F + G4
OrthoMCL	LG + F + G4
OrthoFinder	LG + F + G4

Table 3. Convergence of the PhyloBayes runs

Method	Num Cycles	MaxDiff	MeanDiff
OMA	7080	0.297691	0.00522266
HaMStR	*22,230*	*1*	*0.0175439*[a]
BUSCO	47,281	0.0957351	0.00435825
OrthoMCL	1543	0.104071	0.0548237
OrthoFinder	2260	0.117054	0.00368342

In italics: MaxDiff > 0.3, thus not converged.
[a]Unchanged for at least 15,000 cycles.

Phylogenetic inference

Each orthologous group that contained a minimum of 15 protein sequences, of the 30 total, representing unique species were aligned using MUSCLE (Edgar 2004), using default parameters. All spurious sequences, and poorly aligned regions of the multiple sequence alignments, were then removed using trimAl (Capella-Gutiérrez et al. 2009), using the -automated1 flag. Supermatrices were then constructed by concatenating all of the remaining alignments, with missing sequences treated as gaps. The final alignment was subsequently reduced to only contain sites in which more than 60% were occupied by amino acids.

Species trees were constructed using IQ-TREE, with 1000 ultrafast bootstrap replicates (Hoang et al. 2018). Model selection was determined by ModelFinder, with a gamma rate of heterogeneity, which found the best fitting model for each supermatrix (Table 2). We also computed IQ-TREE trees using the C20 mixture model (site-specific frequency model) (Wang et al. 2018), but the support values were low across all methods (Supplemental Fig. S3), and thus we decided against using them further in our analyses. In addition to the maximum likelihood trees, we constructed Bayesian trees using PhyloBayes MPI v1.5a, using the CAT + GTR + G4 model. Convergence information is provided in Table 3.

Group subsampling analyses

As the number of orthologous groups can depend on the parameters for each of the inference methods, we subsampled the data so that the supermatrices were of equivalent size. This allows us to assess the quality of each of the groups. For each orthology method, from the set of predicted orthologous groups with at least 50% of the species, a number of groups were selected at random, without repeats, but ensuring that every species was represented in at least one group. The groups were concatenated in order to construct supermatrices using the same process mentioned previously, when constructing full species trees. Species trees were then constructed using IQ-TREE, with a WAG + I model of evolution. WAG + I was chosen because, after preliminary tests on a selection of trees, it was found to give good trees in a relatively short amount of time. This process was repeated 50 times for each orthology inference method. The number of orthologous groups were 50, 100, 200, and every further 200 up to 2000. When the number of orthologous groups to select exceeded the total number of orthologous groups a method inferred (i.e., over 400 groups for BUSCO and over 600 groups for OrthoMCL, etc.), no further supermatrices could be constructed.

The Robinson-Foulds distances between the model tree (Fig. 6) and each of the species trees were computed. In order to account for polytomies, the upper bound for the Robinson-Foulds distance was calculated. This is achieved by counting each missing split as a contribution to the Robinson-Foulds score, assuming that each missing split resulted in a conflicting topology. The distance was normalized by dividing by the maximum possible Robinson-Foulds score ($2\cdot(n-3)$, where n is the number of taxa).

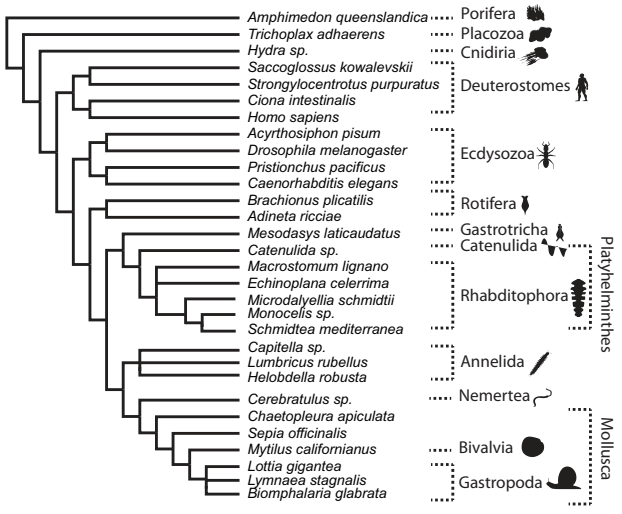

Figure 6. Model tree based on the literature (see Methods).

Software availability

To facilitate reproducibility, we are providing custom Python script as Supplemental Code. Intermediate and output data of the Lophotrochozoan phylogenomic analysis are provided as Supplemental Data.

Acknowledgments

Computations were performed on the University College London Computer Science cluster and at the Vital-IT Center for high-performance computing at the SIB Swiss Institute of Bioinformatics. J.L. is funded by EPSRC Centre for Doctoral Training studentship at UCL CoMPLEX (EP/F500351/1). M.J.T. acknowledges support by a Biotechnology and Biological Sciences Research Council grant (BB/H006966/1) and the European Research Council (ERC-2012-AdG 322790). C.D. acknowledges support by Swiss National Science Foundation grant 150654, UK BBSRC grant BB/M015009/1, and the Swiss State Secretariat for Education, Research and Innovation (SERI).

References

Afrasiabi C, Samad B, Dineen D, Meacham C, Sjölander K. 2013. The PhyloFacts FAT-CAT web server: ortholog identification and function prediction using fast approximate tree classification. *Nucleic Acids Res* **41:** W242–W248. doi:10.1093/nar/gkt399

Altenhoff AM, Dessimoz C. 2009. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol* **5:** e1000262. doi:10.1371/journal.pcbi.1000262

Altenhoff AM, Dessimoz C. 2012. Inferring orthology and paralogy. *Methods Mol Biol* 855**:** 259–279. doi:10.1007/978-1-61779-582-4_9

Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C. 2012. Resolving the ortholog conjecture: Orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Comput Biol* **8:** e1002514. doi:10.1371/journal.pcbi.1002514

Altenhoff AM, Gil M, Gonnet GH, Dessimoz C. 2013. Inferring hierarchical orthologous groups from orthologous gene pairs. *PLoS One* **8:** e53786. doi:10.1371/journal.pone.0053786

Altenhoff AM, Škunca N, Glover N, Train C-M, Sueki A, Piližota I, Gori K, Tomiczek B, Müller S, Redestig H, et al. 2015. The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. *Nucleic Acids Res* **43:** D240–D249. doi:10.1093/nar/gku1158

Altenhoff AM, Boeckmann B, Capella-Gutierrez S, Dalquen DA, DeLuca T, Forslund K, Huerta-Cepas J, Linard B, Pereira C, Pryszcz LP, et al. 2016. Standardized benchmarking in the quest for orthologs. *Nat Methods* **13:** 425–430. doi:10.1038/nmeth.3830

Altenhoff AM, Glover NM, Train C-M, Kaleb K, Warwick Vesztrocy A, Dylus D, de Farias TM, Zile K, Stevenson C, Long J, et al. 2018. The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Res* **46:** D477–D485. doi:10.1093/nar/gkx1019

Baguñà J, Riutort M. 2004. Molecular phylogeny of the platyhelminthes. *Can J Zool* **82:** 168–193. doi:10.1139/z03-214

Boeckmann B, Robinson-Rechavi M, Xenarios I, Dessimoz C. 2011. Conceptual framework and pilot study to benchmark phylogenomic databases based on reference gene trees. *Brief Bioinform* **12:** 423–435. doi:10.1093/bib/bbr034

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12:** 59–60. doi:10.1038/nmeth.3176

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25:** 1972–1973. doi:10.1093/bioinformatics/btp348

Carrigan MA, Uryasev O, Davis RP, Zhai L, Hurley TD, Benner SA. 2012. The natural history of class I primate alcohol dehydrogenases includes gene duplication, gene loss, and gene conversion. *PLoS One* **7:** e41175. doi:10.1371/journal.pone.0041175

Chen F, Mackey AJ, Stoeckert CJ, Roos DS. 2006. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* **34:** D363–D368. doi:10.1093/nar/gkj123

Dalquen DA, Dessimoz C. 2013. Bidirectional best hits miss many orthologs in duplication-rich clades such as plants and animals. *Genome Biol Evol* **5:** 1800–1806. doi:10.1093/gbe/evt132

Dessimoz C, Cannarozzi G, Gil M, Margadant D, Roth A, Schneider A, Gonnet G. 2005. OMA, a comprehensive, automated project for the identification of orthologs from complete genome data: introduction and first achievements. In *Proceedings of the RECOMB 2005 Workshop on Comparative Genomics* (ed. McLysaght A, Huson DH), pp. 61–72. Springer-Verlag, Berlin.

Dessimoz C, Boeckmann B, Roth ACJ, Gonnet GH. 2006. Detecting non-orthology in the COGs database and other approaches grouping orthologs using genome-specific best hits. *Nucleic Acids Res* **34:** 3309–3316. doi:10.1093/nar/gkl433

Dessimoz C, Gabaldón T, Roos DS, Sonnhammer ELL, Herrero J, Quest for Orthologs Consortium. 2012. Toward community standards in the quest for orthologs. *Bioinformatics* **28:** 900–904. doi:10.1093/bioinformatics/bts050

Dikow RB, Frandsen PB, Turcatel M, Dikow T. 2017. Genomic and transcriptomic resources for assassin flies including the complete genome sequence of *Proctacanthus coquilletti* (Insecta: Diptera: Asilidae) and 16 representative transcriptomes. *PeerJ* **5:** e2951. doi:10.7717/peerj.2951

Domazet-Lošo T, Brajković J, Tautz D. 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet* **23:** 533–539. doi:10.1016/j.tig.2007.08.014

Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, et al. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* **452:** 745–749. doi:10.1038/nature06614

Dunn CW, Giribet G, Edgecombe GD, Hejnol A. 2014. Animal phylogeny and its evolutionary implications. *Annu Rev Ecol Evol Syst* **45:** 371–395. doi:10.1146/annurev-ecolsys-120213-091627

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

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32:** 1792–1797. doi:10.1093/nar/gkh340

Edgecombe GD, Giribet G, Dunn CW, Hejnol A, Kristensen RM, Neves RC, Rouse GW, Worsaae K, Sørensen MV. 2011. Higher-level metazoan relationships: recent progress and remaining questions. *Org Divers Evol* **11:** 151–172. doi:10.1007/s13127-011-0044-4

Egger B, Lapraz F, Tomiczek B, Müller S, Dessimoz C, Girstmair J, Škunca N, Rawlinson KA, Cameron CB, Beli E, et al. 2015. A transcriptomic-phylogenomic analysis of the evolutionary relationships of flatworms. *Curr Biol* **25:** 1347–1353. doi:10.1016/j.cub.2015.03.034

Ekins S, Williams AJ, Krasowski MD, Freundlich JS. 2011. In silico repositioning of approved drugs for rare and neglected diseases. *Drug Discov Today* **16:** 298–310. doi:10.1016/j.drudis.2011.02.016

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* **16:** 157. doi:10.1186/s13059-015-0721-2

Fernández R, Giribet G. 2015. Unnoticed in the tropics: phylogenomic resolution of the poorly known arachnid order Ricinulei (Arachnida). *R Soc Open Sci* **2:** 150065. doi:10.1098/rsos.150065

Fernández R, Laumer CE, Vahtera V, Libro S, Kaluziak S, Sharma PP, Pérez-Porro AR, Edgecombe GD, Giribet G. 2014. Evaluating topological conflict in centipede phylogeny using transcriptomic data sets. *Mol Biol Evol* **31:** 1500–1513. doi:10.1093/molbev/msu108

Fernández R, Edgecombe GD, Giribet G. 2016. Exploring phylogenetic relationships within Myriapoda and the effects of matrix composition and occupancy on phylogenomic reconstruction. *Syst Biol* **65:** 871–889. doi:10.1093/sysbio/syw041

Feuda R, Dohrmann M, Pett W, Philippe H, Rota-Stabelli O, Lartillot N, Wörheide G, Pisani D. 2017. Improved modeling of compositional heterogeneity supports sponges as sister to all other animals. *Curr Biol* **27:** 3864–3870.e4. doi:10.1016/j.cub.2017.11.008

Forslund K, Pereira C, Capella-Gutierrez S, Sousa da Silva A, Altenhoff A, Huerta-Cepas J, Muffato M, Patricio M, Vandepoele K, Ebersberger I, et al. 2018. Gearing up to handle the mosaic nature of life in the quest for orthologs. *Bioinformatics* **34:** 323–329. doi:10.1093/bioinformatics/btx542

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28:** 3150–3152. doi:10.1093/bioinformatics/bts565

Gabaldón T, Koonin EV. 2013. Functional and evolutionary implications of gene orthology. *Nat Rev Genet* **14:** 360–366. doi:10.1038/nrg3456

Garrison NL, Rodriguez J, Agnarsson I, Coddington JA, Griswold CE, Hamilton CA, Hedin M, Kocot KM, Ledford JM, Bond JE. 2016. Spider phylogenomics: untangling the spider tree of life. *PeerJ* **4:** e1719. doi:10.7717/peerj.1719

Giribet G, Edgecombe GD. 2017. Current understanding of ecdysozoa and its internal phylogenetic relationships. *Integr Comp Biol* **57:** 455–466. doi:10.1093/icb/icx072

Gonnet GH, Hallett MT, Korostensky C, Bernardin L. 2000. Darwin v. 2.0: an interpreted computer language for the biosciences. *Bioinformatics* **16:** 101–103. doi:10.1093/bioinformatics/16.2.101

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8:** 1494–1512. doi:10.1038/nprot.2013.084

Halanych KM, Whelan NV, Kocot KM, Kohn AB, Moroz LL. 2016. Miscues misplace sponges. *Proc Natl Acad Sci* **113:** E946–E947. doi:10.1073/pnas.1525332113

Hejnol A, Obst M, Stamatakis A, Ott M, Rouse GW, Edgecombe GD, Martinez P, Baguñà J, Bailly X, Jondelius U, et al. 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proc Biol Sci* **276:** 4261–4270. doi:10.1098/rspb.2009.0896

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol* **35:** 518–522. doi:10.1093/molbev/msx281

Huerta-Cepas J, Capella-Gutiérrez S, Pryszcz LP, Marcet-Houben M, Gabaldón T. 2014. PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res* **42:** D897–D902. doi:10.1093/nar/gkt1177

Huerta-Cepas J, Serra F, Bork P. 2016a. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol* **33:** 1635–1638. doi:10.1093/molbev/msw046

Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, et al. 2016b. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res* **44:** D286–D293. doi:10.1093/nar/gkv1248

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* **14:** 587–589. doi:10.1038/nmeth.4285

Kocot KM. 2016. On 20 years of Lophotrochozoa. *Org Divers Evol* **16:** 329–343. doi:10.1007/s13127-015-0261-3

Kocot KM, Cannon JT, Todt C, Citarella MR, Kohn AB, Meyer A, Santos SR, Schander C, Moroz LL, Lieb B, et al. 2011. Phylogenomics reveals deep molluscan relationships. *Nature* **477:** 452–456. doi:10.1038/nature10382

Kocot KM, Struck TH, Merkel J, Waits DS, Todt C, Brannock PM, Weese DA, Cannon JT, Moroz LL, Lieb B, et al. 2017. Phylogenomics of lophotrochozoa with consideration of systematic error. *Syst Biol* **66:** 256–282. doi:10.1093/sysbio/syw079

Lartillot N, Rodrigue N, Stubbs D, Richer J. 2013. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst Biol* **62:** 611–615. doi:10.1093/sysbio/syt022

Laumer CE, Hejnol A, Giribet G. 2015. Nuclear genomic signals of the 'microturbellarian' roots of platyhelminth evolutionary innovation. *eLife* **4:** e05503. doi:10.7554/eLife.05503

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13:** 2178–2189. doi:10.1101/gr.1224503

Linard B, Thompson JD, Poch O, Lecompte O. 2011. OrthoInspector: comprehensive orthology analysis and visual exploration. *BMC Bioinformatics* **12:** 11. doi:10.1186/1471-2105-12-11

Marlétaz F, Peijnenburg KTCA, Goto T, Satoh N, Rokhsar DS. 2019. A new spiralian phylogeny places the enigmatic arrow worms among gnathiferans. *Curr Biol* **29:** 312–318.e3. doi:10.1016/j.cub.2018.11.042

Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD. 2017. PANTHER version 11: expanded annotation data from gene ontology and reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res* **45:** D183–D189. doi:10.1093/nar/gkw1138

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32:** 268–274. doi:10.1093/molbev/msu300

Nogi T, Zhang D, Chan JD, Marchant JS. 2009. A novel biological activity of praziquantel requiring voltage-operated Ca^{2+} channel β subunits: subversion of flatworm regenerative polarity. *PLoS Negl Trop Dis* **3:** e464. doi:10.1371/journal.pntd.0000464

Pagani I, Liolios K, Jansson J, Chen I-MA, Smirnova T, Nosrat B, Markowitz VM, Kyrpides NC. 2012. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* **40:** D571–D579. doi:10.1093/nar/gkr1100

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci* **96:** 4285–4288. doi:10.1073/pnas.96.8.4285

Philippe H, Brinkmann H, Copley RR, Moroz LL, Nakano H, Poustka AJ, Wallberg A, Peterson KJ, Telford MJ. 2011a. Acoelomorph flatworms are deuterostomes related to *Xenoturbella*. *Nature* **470:** 255–258. doi:10.1038/nature09676

Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D. 2011b. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol* **9:** e1000602. doi:10.1371/journal.pbio.1000602

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

Pisani D, Pett W, Dohrmann M, Feuda R, Rota-Stabelli O, Philippe H, Lartillot N, Wörheide G. 2016. Reply to Halanych et al.: Ctenophore misplacement is corroborated by independent datasets. *Proc Natl Acad Sci* **113:** E948–E949. doi:10.1073/pnas.1525718113

Roth AC, Gonnet GH, Dessimoz C. 2008. Correction: Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics* **9:** 518. doi:10.1186/1471-2105-9-518

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

Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27:** 863–864. doi:10.1093/bioinformatics/btr026

Schmitt T, Messina DN, Schreiber F, Sonnhammer ELL. 2011. Letter to the editor: SeqXML and OrthoXML: standards for sequence and orthology information. *Brief Bioinform* **12:** 485–488. doi:10.1093/bib/bbr025

Sharma PP, Kaluziak ST, Pérez-Porro AR, González VL, Hormiga G, Wheeler WC, Giribet G. 2014. Phylogenomic interrogation of arachnida reveals systemic conflicts in phylogenetic signal. *Mol Biol Evol* **31:** 2963–2984. doi:10.1093/molbev/msu235

Sharma PP, Fernández R, Esposito LA, González-Santillán E, Monod L. 2015. Phylogenomic resolution of scorpions reveals multilevel discordance with morphological phylogenetic signal. *Proc Biol Sci* **282:** 20142953. doi:10.1098/rsbp.2014.2953

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31:** 3210–3212. doi:10.1093/bioinformatics/btv351

Simion P, Belkhir K, François C, Veyssier J, Rink JC, Manuel M, Philippe H, Telford MJ. 2018. A software tool 'CroCo' detects pervasive cross-species contamination in next generation sequencing data. *BMC Biol* **16:** 28. doi:10.1186/s12915-018-0486-7

Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. *J Mol Biol* **147:** 195–197. doi:10.1016/0022-2836(81)90087-5

Song S, Liu L, Edwards SV, Wu S. 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proc Natl Acad Sci* **109:** 14942–14947. doi:10.1073/pnas.1211733109

Sonnhammer ELL, Östlund G. 2015. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res* **43:** D234–D239. doi:10.1093/nar/gku1203

Sonnhammer ELL, Gabaldón T, da Silva AW S, Martin M, Robinson-Rechavi M, Boeckmann B, Thomas PD, Dessimoz C, Quest for Orthologs consortium. 2014. Big data and other challenges in the quest for orthologs. *Bioinformatics* **30:** 2993–2998. doi:10.1093/bioinformatics/btu492

Steinegger M, Söding J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35:** 1026–1028. doi:10.1038/nbt.3988

Struck TH, Fisse F. 2008. Phylogenetic position of Nemertea derived from phylogenomic data. *Mol Biol Evol* **25:** 728–736. doi:10.1093/molbev/msn019

Struck TH, Wey-Fabrizius AR, Golombek A, Hering L, Weigert A, Bleidorn C, Klebow S, Iakovenko N, Hausdorf B, Petersen M, et al. 2014. Platyzoan paraphyly based on phylogenomic data supports a noncoelomate ancestry of spiralia. *Mol Biol Evol* **31:** 1833–1849. doi:10.1093/molbev/msu143

Szalkowski A, Ledergerber C, Krähenbühl P, Dessimoz C. 2008. SWPS3–fast multi-threaded vectorized Smith-Waterman for IBM Cell/B.E. and x86/SSE2. *BMC Res Notes* **1:** 107. doi:10.1186/1756-0500-1-107

Telford MJ, Budd GE, Philippe H. 2015. Phylogenomic insights into animal evolution. *Curr Biol* **25:** R876–R887. doi:10.1016/j.cub.2015.07.060

Train C-M, Glover NM, Gonnet GH, Altenhoff AM, Dessimoz C. 2017. Orthologous Matrix (OMA) algorithm 2.0: more robust to asymmetric evolutionary rates and more scalable hierarchical orthologous group inference. *Bioinformatics* **33:** i75–i82. doi:10.1093/bioinformatics/btx229

Train C-M, Pignatelli M, Altenhoff A, Dessimoz C. 2018. iHam & pyHam: visualizing and processing hierarchical orthologous groups. *Bioinformatics*. doi:10.1093/bioinformatics/bty994

Tsagkogeorga G, Müller S, Dessimoz C, Rossiter SJ. 2017. Comparative genomics reveals contraction in olfactory receptor genes in bats. *Sci Rep* **7:** 259. doi:10.1038/s41598-017-00132-9

Tsai IJ, Zarowiecki M, Holroyd N, Garciarrubio A, Sanchez-Flores A, Brooks KL, Tracey A, Bobes RJ, Fragoso G, Sciutto E, et al. 2013. The genomes of four tapeworm species reveal adaptations to parasitism. *Nature* **496:** 57–63. doi:10.1038/nature12031

Uchiyama I, Mihara M, Nishide H, Chiba H. 2012. MBGD update 2013: the microbial genome database for exploring the diversity of microbial world. *Nucleic Acids Res* **41:** D631–D635. doi:10.1093/nar/gks1006

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2008. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* **19:** 327–335. doi:10.1101/gr.073585.107

Wang H-C, Minh BQ, Susko E, Roger AJ. 2018. Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. *Syst Biol* **67:** 216–235. doi:10.1093/sysbio/syx068

Weigert A, Helm C, Meyer M, Nickel B, Arendt D, Hausdorf B, Santos SR, Halanych KM, Purschke G, Bleidorn C, et al. 2014a. Data from: Illuminating the base of the annelid tree using transcriptomics. https://datadryad.org/resource/doi:10.5061/dryad.g2qp5.

Weigert A, Helm C, Meyer M, Nickel B, Arendt D, Hausdorf B, Santos SR, Halanych KM, Purschke G, Bleidorn C, et al. 2014b. Illuminating the base of the annelid tree using transcriptomics. *Mol Biol Evol* **31:** 1391–1401. doi:10.1093/molbev/msu080

Whelan NV, Kocot KM, Moroz LL, Halanych KM. 2015. Error, signal, and the placement of Ctenophora sister to all other animals. *Proc Natl Acad Sci* **112:** 5773–5778. doi:10.1073/pnas.1503453112

Whelan NV, Kocot KM, Moroz TP, Mukherjee K, Williams P, Paulay G, Moroz LL, Halanych KM. 2017. Ctenophore relationships and their placement as the sister group to all other animals. *Nat Ecol Evol* **1:** 1737–1746. doi:10.1038/s41559-017-0331-3

Williams TA, Szöllősi GJ, Spang A, Foster PG, Heaps SE, Boussau B, Ettema TJG, Embley TM. 2017. Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc Natl Acad Sci* **114:** E4602–E4611. doi:10.1073/pnas.1618463114

Wittwer LD, Piližota I, Altenhoff AM, Dessimoz C. 2014. Speeding up all-against-all protein comparisons while maintaining sensitivity by considering subsequence-level homology. *PeerJ* **2:** e607. doi:10.7717/peerj.607

Yang Y, Smith SA. 2014. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. *Mol Biol Evol* **31:** 3081–3092. doi:10.1093/molbev/msu245

Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simão FA, Ioannidis P, Seppey M, Loetscher A, Kriventseva EV. 2017. OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res* **45:** D744–D749. doi:10.1093/nar/gkw1119

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, et al. 2018. Ensembl 2018. *Nucleic Acids Res* **46:** D754–D761. doi:10.1093/nar/gkx1098

Received August 22, 2018; accepted in revised form May 24, 2019.