VEHoP: A Versatile, Easy-to-use, and Homology-based Phylogenomic pipeline accommodating diverse sequences

Yunlong Li[1,2*], Xu Liu[1,2*], Chong Chen[3], Jian-Wen Qiu[4], Kevin Kocot[5], Jin Sun[1,2#]

[1] Key Laboratory of Evolution & Marine Biodiversity (Ministry of Education) and Institute of Evolution & Marine Biodiversity, Ocean University of China, Qingdao 266003, China

[2] Laboratory for Marine Biology and Biotechnology, Qingdao Marine Science and Technology Center, Laoshan Laboratory, Qingdao 266237, China

[3] X-STAR, Japan Agency for Marine-Earth Science and Technology (JAMSTEC), 2-15 Natsushima-cho, Yokosuka, Kanagawa 237-0061, Japan

[4] Department of Biology, Hong Kong Baptist University, Hong Kong, China

[5] Department of Biological Sciences and Alabama Museum of Natural History, University of Alabama, Tuscaloosa, AL 35487, USA

[*] Equal contribution

[#] Corresponding author: Jin Sun, jin_sun@ouc.edu.cn

Abstract

Phylogenomics has become a prominent method in systematics, conservation biology, and biomedicine, as it can leverage hundreds to thousands of genes derived from genomic or transcriptomic data to infer evolutionary relationships. However, obtaining high-quality genomes and transcriptomes requires samples preserved with high-quality DNA and RNA and demands considerable sequencing costs and lofty bioinformatic efforts (e.g., genome/transcriptome assembly and annotation). Notably, only fragmented DNA reads are accessible in some rare species due to the difficulty in sample collection and preservation, such as those inhabiting the deep sea. To address this issue, we here introduce the VEHoP (Versatile, Easy-to-use Homology-based Phylogenomic) pipeline, designed to infer protein-coding regions from DNA assemblies and generate alignments of orthologous sequences, concatenated matrices, and phylogenetic trees. This pipeline aims to 1) expand taxonomic sampling by accommodating a wide range of input files, including draft genomes, transcriptomes, and well-annotated genomes, and 2) simplify the process of conducting phylogenomic analyses and thus make it more accessible to researchers from diverse backgrounds. We first evaluated the performance of VEHoP using datasets of Ostreida, yielding robust phylogenetic trees with strong bootstrap support. We then applied VEHoP to reconstruct the phylogenetic relationship in the enigmatic deep-sea gastropod order Neomphalida, obtaining a robust phylogenetic backbone for this group. The VEHoP is freely available on GitHub (https://github.com/ylify/VEHoP), whose dependencies can be easily

39 installed using Bioconda.

40

41 **Keywords**: phylogenomics, NGS, phylogeny, evolution, pipeline, Ostreida, Neomphalida

42

Background

Phylogenetics is now the most fundamental method in evolutionary biology research to understand and illuminate the relationships between organisms. Multiple types of data can be used to infer phylogenetic relationships, including phenotypic and genotypic characteristics. Among them, biological molecules (i.e., nucleic acids and amino acids) are widely used for reconstructing phylogenetic trees. At the early stages of molecular phylogeny, one or a few gene markers were used, such as the mitochondrial cytochrome *c* oxidase subunit I (COI), NADH dehydrogenase subunit 4 (NAD4), nuclear ribosomal RNA genes, or the combination of them (Hao et al. 2015; Ibáñez et al. 2019). With the improvement of sequencing techniques, this was followed by mitogenome-based reconstructions (Donath et al. 2019; Irisarri et al. 2020; Ghiselli et al. 2021). However, these gene trees sometimes failed to reveal the true relationships among taxa due to introgression, different gene evolutional rates between groups, and long-branch attraction (Doolittle and Logsdon Jr 1998; Huynen and Bork 1998; Doolittle 1999; Degnan and Rosenberg 2006). This called for more sophisticated methods for phylogenetics that can address all such issues. Recently, with the development of next-generation sequencers, phylogenetics based on genome-level data (i.e., phylogenomics) has become a focus in many fields (Dunn et al. 2008; Young and Gillung 2020).

It has been shown that taxon sampling is key in reducing errors in phylogenetic inferences (Powell and Battistuzzi 2022). Despite this, in most cases, it is unrealistic to gather sufficient data on all target species to answer the phylogenetic questions. For one, some species inhabit inaccessible environments, such as the deep sea and polar regions, or maybe extremely rare that only one or few specimens are available as long-preserved samples in natural history museums. Also, species distribution in certain groups can be skewed which leads to biased sampling. In these cases, researchers would have to perform a phylogenetic reconstruction using a dataset lacking some species. If those species happen to represent an important node, the tree topology may change based on such an imbalanced taxon-sampling dataset. In addition, most extinct fossil species cannot be sequenced, thus rendering it impossible for molecular phylogenies to include all taxa on the tree of life across evolutionary history (Marshall 2017).

There is no doubt that genome-based phylogeny contains much more information than single or few gene makers (Chang et al. 2011). As next-generation sequencing (NGS) technology advanced, more and more sequenced genomes and transcriptomes have been released to the public at an elevated rate year after year (Turnbull et al. 2023). Nevertheless, many of these datasets were sequenced initially for organelle genome assembly, genome survey, genome annotation, gene expression level analysis, and so on. These are all potential sources for phylogenetics, yet they often remain buried deep in the public database.

The best datasets for phylogenomic analysis are whole genome data from different species (Cheon et al. 2020; Fleming et al. 2023). Yet, the situation is often complicated in practical use. In many groups, only a few well-annotated genomes are available while the rest are transcriptomes and raw Illumina DNA reads. To obtain a genome dataset for phylogenomic analyses from these, multiple steps of bioinformatics analyses must be performed (Liu et al. 2023), which always include quality control of the raw data, draft genome assembly and annotation (Simão et al. 2015). Apart from these, ortholog inference must be performed to identify sequences whose evolutionary history reflects that of the species, which may be the most important step for reliable phylogenomic reconstructions (Yang and Smith 2014; Mongiardino Koch 2021; Lozano-Fernandez 2022). Finally, matrix assembly must be performed, which involves further steps such as alignment, trimming of ambiguously aligned positions, concatenation, and tree reconstruction. The whole workflow is time-consuming and can be confusing for those researchers not from a bioinformatics background (Dylus et al. 2023).

Some tools for phylogenomic analysis can use raw sequencing reads to generate phylogenetic trees, such as Read2Tree (Dylus et al. 2023). However the reference OMA ("Orthologous MAtrix") database designated in Read2Tree is not fully customized, and the current procedure for the phylogenetic reconstruction is sophisticated with many manual curation steps. GeneMiner (Xie et al. 2024) is a toolkit developed for phylogenetic marker mining, which extracts markers from transcriptomic, genomic, or other NGS data. It could be used for multiple gene phylogeny, yet it is still inefficient in phylogenomic analysis due to vague instructions and low numbers of single-copy orthologs extracted.

To address these problems, we here developed a new pipeline which we name 'VEHoP' (Versatile, Easy-to-use Homology-based Phylogenomic). The VEHoP workflow allows different types of datasets as input, including genomic DNA assemblies, transcriptomes, well-annotated genomes, or any combinations thereof. After providing these files as the input, users only need to provide a prefix for the run, a path to the database (required if DNA assemblies or transcriptomes are provided), and the optional adjustment of quality control in matrix assembly (e.g., occupancy and alignment threshold, 2/3 and 100 AAs by default, respectively). Alternative analyses can be specified if needed, such as PhyloBayes, ASTRAL, set up occupancy. The output files include single-gene alignments, single-gene trees, a concatenated supermatrix, and results of phylogenetic analyses using the supertree and supermatrix-based approaches.

To assess and benchmark the performance of the VEHoP pipeline, we tested it in two case

studies. Ostreida (the 'oyster' order) is a well-studied group of animals in phylum Mollusca with 10 high-quality and well-annotated genomes plus a range of transcriptome datasets, making it an ideal clade for benchmarking the performance and reliability of VEHoP. To further test the applicability of VEHoP in resolving phylogenetic issues, we also used it to analyze a dataset of the gastropod order Neomphalida which is a deep-sea clade of typically small-sized animals. Previously phylogenetic analyses did not fully resolve the internal relationships within this order, due to the lack of high-quality genomes and transcriptomes required by traditional phylogenomic pipelines, and thus the evolutionary relationships among the neomphalidan taxa remained highly contentious. Our results on both case studies lend support to the VEHoP as a user-friendly, efficient, and accurate workflow.

Description of VEHoP

Input files and parameters

The VEHoP pipeline accepts genomic sequencing data, transcriptome sequencing data, and well-annotated genomes, or any combination of these data types. All inputs should be in *.fasta* format, but with different suffixes: *.pep.fasta* for proteomes from quality datasets, *.transcript.fasta* for transcriptomes, and *.genomic.fasta* for DNA assemblies. The raw data only need to go through a simple, coarse assembly using a *de novo* assembler, such as MEGAHIT (Li et al. 2015) for genomic data and Trinity (Haas et al. 2013) for transcriptome data after quality control and trimming procedures. In addition, the user also needs to prepare a database for homolog extraction, if draft genomes or transcriptomes are provided. The reference database could be a concatenation of protein files suggested from close relatives with well-annotated genomes. By default, VEHoP uses 40 threads (-t 40) throughout, including homolog-inference using miniprot, OrthoFinder processing, matrix assembly and tree construction. During the matrix assembly, VEHoP keeps the quality single-gene alignments with the threshold of alignment length (-l 100) and taxonomy occupancy (2/3, users could adjust manually via setting the minimum samples, -m #s).

Workflow

The pipeline was written in Python. All dependencies can be easily installed via Anaconda (Fig. 1) and implemented as follows, except HmmCleaner (Di Franco et al. 2019) which the user can install optionally by the instructions provided in the GitHub repository.

The workflow consists of the following steps (Fig. 1), which can be implemented using a single command:

1) Miniprot (Li 2023) is used to map protein sequences from the reference database to the coarsely assembled genomic or transcriptomic data to predict gene models; 2) TransDecoder (Douglas 2018) is used to extract proteins based on the predicted gene models; 3), cd-hit (Fu et al. 2012) is performed to remove redundant sequences with the threshold of 85% similarity;

4) the filtered protein sequences are submitted to OrthoFinder (Emms and Kelly 2019) to identify orthogroups (OGs), with the occupancy assigned by the user (default 2/3, and only orthologs matching the standard will be kept); 5) redundant sequences are removed with uniqHaplo while the remaining sequences are aligned with MAFFT (Katoh and Standley 2013) with default settings; 6) the misaligned regions are removed with HmmCleaner and the aligned files are trimmed with BMGE (Criscuolo and Gribaldo 2010) and trimAL (Capella-Gutiérrez et al. 2009); 7) AlignmentCompare (https://github.com/DamienWaits/Alignment_Compare) is then used to remove sequences shorter than 20 amino acids (AAs), followed by a second occupancy check to make sure all sequences overlap, which is necessary for single-gene tree reconstructions; 8) IQ-TREE or FastTree (default being FastTree) is used to build trees for each filtered OGs. 9) PhyloPyPruner is used to remove paralogs in the filtered alignments; 10) The generated supermatrix is used to reconstruct phylogenetic trees, using IQ-TREE (Minh et al. 2020), FastTree (Price et al. 2010), and PhyloBayes (Lartillot et al. 2013); 11) A random subsample of the initial matrix to 2,500,000 and 5,000,000 sites can also performed for the reconstruction of phylogenetic relationships using IQ-TREE and PhyloBayes. Apart from concatenation-based phylogeny, the pipeline provides a coalescent phylogenetic approach (default: off) implemented via ASTRAL (Mirarab et al. 2014).

Output files

The output files of the workflow include an initial data matrix in .fasta format, an IQ-TREE tree file, and a FastTree tree file. Apart from the above-mentioned default outputs, the results of ASTRAL and PhyloBayes can also be found in the final output directory if related settings are specified in the commands. If users want to attempt more phylogenetic analyses, they can perform additional custom analyses using the initial data matrix.

Results

Case study 1: Benchmarking VEHoP with the oyster dataset

To benchmark the usability and efficiency of the workflow, we collected data from representatives of Ostreida as an example. The datasets include 10 species from Ostreida including *Pinctada fucata*, *Crassostrea hongkongensis*, *C. angulata*, *C. ariakensis*, *C. nippona*, *Ostrea edulis*, *O. denselamellosa*, *C. virginica*, *C. gigas*, *Saccostrea glomerata*, and two species from the closely related order Pectinida (as the outgroup), *Pecten maximus* and *Mizuhopecten yessoensis*. The data included well-annotated genomes, draft genomes from NGS reads, and *de novo* transcriptomes from RNA-seq. The sources of these data are included in the Supplementary Table S1.

We tested our workflow with different datasets, including dataset 1: well-annotated genomes;

dataset 2: DNA reads, preliminarily assembled with MEGAHIT; dataset 3: transcriptome reads, assembled with Trinity; and dataset 4: a dataset combination including all three types of abovementioned data. For each dataset, the occupancy was set to 2/3, and phylogenetic analyses were performed with two efficient algorithms IQ-TREE (MFP) and FastTree, based on maximum likelihood estimation. The analyses resulted in the same branching order when using both datasets 1 and 2 (Fig. 2a and b). All bootstrap values reached 100 in these two trees, except for two nodes in dataset 2, with a bootstrap value of 68 within the genus *Crassostrea*. However, the position of *C. nippona* was different when using dataset 3, though the bootstrap of all nodes reached 100 (Fig. 2c). Furthermore, the same phylogenetic methods were performed on the matrix of 2973 orthologs generated from genome-wide proteins, genome sequences, and transcriptomes, which showed that most of the terminals were clustered by species, except that *C. gigas* was mixed with its most closely related species *C. angulata* (Fig 2e).

Benchmark of two strategies in VEHoP when processing transcriptomes: miniprot and TransDecoder

Previous works typically predict the coding regions of transcripts via TransDecoder and further reduce the redundancy via CD-HIT. Generally, this takes about 20 minutes for a single transcriptome dataset, which could be a time-consuming step if dozens or hundreds of datasets are to be included in the tree reconstruction. In VEHoP, we introduce miniprot as an alternative method via the alignment strategy. In our benchmarks (12 transcriptional profiles), it took only 25 minutes to call all coding regions, compared to 169 minutes when using TransDecoder (Supplementary Table S2). Meanwhile, we recovered a consistent branching order as that of the genome dataset on the genus level with robust support for all nodes (Fig. 2, evaluated in IQ-TREE -m MFP).

Benchmark of VEHoP and the other two published methods

To better understand how much data is sufficient to reconstruct reliable phylogenetic relationships, we subsampled the *C. hongkongensis* data into 2X, 4X, 6X, and 8X of its genome size. Based on these datasets, we performed phylogenetic analyses using IQ-TREE (MFP) and FastTree. The results showed that the pipeline worked well with all the datasets: the branching order of the trees were identical, and all node supports were 100% (Fig. 2d and Supplementary Fig.1). Reduced datasets for every species (1 Gb, 2 Gb, 4 Gb, 6 Gb, and 8 Gb) were also made and phylogenetic analyses conducted (see Supplementary Table S3 for details). The results showed that branch order became unstable for the 1 Gb and 2 Gb datasets, resulting in paraphyly within *Crassostrea*. For datasets larger than 2 Gb, the VEHoP was able to recover phylogenetic relationships at least at the genus level (Supplementary Fig.2). The total run time was also recorded for these different datasets to test the performance. For the

reduced datasets, it took 4.24, 10.22, 18.60, 25.46, and 27.32 hours to obtain the two tree files, one generated by FastTree and another one generated by IQ-TREE, respectively. As for the full-size mixed dataset, it took VEHoP 54.38 hours to obtain the results (Supplementary Table S4).

Read2Tree (Dylus et al. 2023) was also performed on the reduced datasets and full-size genomic datasets. Marker genes of the only three mollusc species available on the OMA database, including the oyster *C. gigas*, the octopus *Octopus bimaculoides*, and the true limpet *Lottia gigantea*, were downloaded from the OMA Orthology database as mapping references. For the 1G dataset, Read2Tree took 7.79 hours to get a *.nwk* format tree file, with *Pecten maximus* incorrectly grouped with two reference species from OMA database (Supplementary Fig. 3). As for the 2G dataset, 19.56 hours were used to generate the tree, yet the position of *C. nippona* was inconsistent with the genome-based tree, though the bootstrap of this node was 100%. In the 4G dataset, a total of 18.5 hours was used, resulting in the same branching order as that in the 2G dataset. For 6G, 8G, and full-size datasets, 21.75, 27.55, and 43.83 hours were used for each dataset, respectively, and they all shared the same branching order as that of the 2G dataset. The total run time for each dataset can also be found in Supplementary Table S4.

MIKE (Wang et al. 2024) is a MinHash-based and *k*-mer phylogenetic algorithm developed for large-scale next-generation sequencing data. MIKE was also performed to benchmark the performance of the VEHoP pipeline with different sizes of datasets, in addition to Read2Tree. In 1G, 2G, 4G, 8G, and full-size datasets, *Saccostrea glomerata* nested with *Crassostrea* or within *C. virginica*, causing paraphyly. Only in the 6G dataset, the topology was well-resolved and consistent with the current understanding of oyster phylogeny at the genus level (Li et al. 2021) (Supplementary Fig. 4). The run time of MIKE for different sizes of datasets can been found in Supplementary Table S5.

Case study 2: Neomphalidan snails

The molecular phylogeny of deep-sea endemic neomphalidan gastropods has long been contentious, partially due to insufficient sampling, small body size and tissue quantity, and lacking many sequences. Here, we applied VEHoP on the original Illumina sequencing dataset (see Supplementary Table S6 for details) used to assemble the mitochondrial genomes from a previous study (see Zhang et al., 2024), which generated a matrix consisting of 1899 orthologs with an occupancy of 2/3. In addition, to improve taxon sampling, we newly sequenced a specimen of *Neomphalus fretterae* (collected from Tempus Fugit vent field, Galápagos Rift, 0°46.1954’N / 85°54.6869’W, 2561 m deep, R/V *Falkor (too)* cruise FKt231024, remotely operated vehicle (ROV) *SuBastian* dive #609, 2023/Nov/02) following the same

methods as Zhang et al. (2024). Four species of Cocculinida (*Cocculina enigmadonta, C. tenuitesta, C. japonica, C. subcompressa*), the sister-order of Neomphalida were used as outgroups, as well as the more distantly related vetigastropod snails *Tristichotrochus unicus* and *Steromphala cineraria*. The data of *C. enigmadonta*, *C. tenuitesta*, *Lamellomphalus manusensis*, *Lirapex politus*, *Symmetriapelta wreni*, *Melanodrymia laurelin*, *Melanodrymia telperion*, Neomphalidae gen *et* sp. Hatoma *sensu* Zhong et al., 2022, *Nodopelta heminoda*, and *Symmetriapelta becki* were gathered from previous studies, which were used to assemble mitochondrial genomes for phylogeny (Zhong et al. 2022; Zhang et al. 2024).

We first attempted to reconstruct the molecular phylogeny of Neomphalida using mitochondrial genomes with multiple models in IQ-TREE, including MFP, C20, C40, and C60 based on the matrices from Zhang et al. (2024). This revealed two distinct tree branching orders with nearly equal support from different sequencing matrices (see Supplementary Fig. 5), confirming the same situation encountered also in a previous study (Zhang et al., 2024). We then conducted multiple phylogenetic analyses through VEHoP based on the assemblies of the abovementioned datasets, including IQ-TREE with the MFP model, Site-specific frequency models (including C20, C40, and C60), and FastTree. All these analyses resulted in the same tree branching order with maximum support in each node, except for one node in Peltospiridae which had the bootstrap value of 85 in the C20 model (Fig. 3). Apart from VEHoP, Read2Tree and MIKE were also performed on the same dataset of Neomphalida. However, these two methods were unable to resolve a consistent topology, even at the order level (Supplementary Fig. 6).

Discussion

We present VEHoP, a new pipeline for phylogenomic analyses with the flexibility of using draft genomes (the ultra-fast and memory-efficient NGS assembler MEGAHIT is feasible if NGS reads), transcriptomes, well-annotated genomes, or a combination of these data. This workflow allows users to reconstruct phylogenetic trees with one single command, significantly lowering the technical hurdle for researchers to carry out phylogenomic inferences. VEHoP is able to reconstruct congruent and robust relationships among taxa using fragmented draft genomes that were rapidly assembled from NGS reads, with results comparable with trees generated from well-annotated genome datasets.

Currently, most available phylogenomic pipelines are based on protein datasets (Kocot et al. 2011; Sun et al. 2021), which require cumbersome steps and are time-consuming to prepare. To obtain high-quality protein files, high-quality DNA sequencing data is inevitably needed. Furthermore, it is necessary to conduct genome assembly to get a contig- or scaffold-level draft genome, followed by gene model prediction. This workflow usually takes several days just for one single species even with ample computing resources.

There is a vast amount of data in public databases, including unannotated genomes and raw NGS reads (genome skimming projects previously used in organelle assemblies or genome surveys), which have been underutilized in phylogenomic studies. Understandably, these data vary in quality and coverage, and thus it has been challenging to use them for phylogenetic analysis. With VEHoP, however, researchers can extract homologs from these genomic data at ease, with the potential to greatly enhance taxon sampling and produce a more robust and consistent tree topology in phylogenetic analyses. As an example, we generated pie charts for major lophotrochozoan animal phyla to show the potential of these 'buried' data in phylogenetics based on NCBI data (Fig. 4 and Supplementary Table S7, data up to May 2024). Among Mollusca, for example, there are only 286 species with genome assemblies (only a small fraction of these is annotated) while an additional 896 species have transcriptomic data. These two data types are mostly commonly used source data for phylogenomic analysis. With VEHoP, we can further include 325 species which lack both genome and transcriptome data but with DNA genomic data, greatly expanding the taxon coverage.

In our benchmarking study using various data types from oysters (case study 1), VEHoP showed a high speed and accuracy in inferring phylogeny. The branching order inferred based on unannotated genomic data was the same as that based on well-annotated genomes, though not all node support reached 100%. For the RNA data, we attempted two strategies: 1) extracting homologs directly from assembled transcripts with miniprot; 2) predicting proteins with TransDecoder. Those two strategies resulted in the same branching order, and each node reached 100% support. By comparison, a strategy based on miniprot was more time-efficient than the TransDecoder-based prediction. In detail, the miniprot-based method took a total of 5.38 hours to complete (within which the prediction process took 0.41 hours), while the TransDecoder-based method took 7 hours to obtain the tree file (of which 2.81 hours were taken just to obtain the coding potential regions). However, the branching order from this analysis differed from those based on well-annotated genomes. This discrepancy was probably due to the presence of isoforms in the transcriptomes, which made it difficult to distinguish homologs from paralogs, leading to the different branching orders in the transcriptome-based trees (Cheon et al. 2020). Thus, genomic data is still recommended when available. Nonetheless, the miniprot-based strategy in transcripts could be a time-efficient way in tree construction and still highly robust at the genus level.

We also tested Read2Tree (Dylus et al. 2023) with the same datasets and made a comparison with VEHoP. Read2Tree only accepts marker genes from the OMA database, where only three mollusc species are currently available. We used marker genes of these abovementioned

348 species as a reference to reconstruct phylogenetic trees with Read2Tree. Both Read2Tree and
349 VEHoP were not able to reveal the same branching order as that of the high-quality genome
350 dataset. The position of *Crassostrea nippona* was unstable. However, VEHoP successfully
351 recovered the same branching order as the same as the high-quality dataset, while Read2Tree
352 retained the branching order with low-coverage datasets. As for run time comparison, VEHoP
353 performed much quicker with dataset less than 4G. After 4G, Read2Tree took less time than
354 VEHoP, since it reconstructed trees directly from raw sequencing reads, and VEHoP needed
355 to assemble the reads first before proceeding with phylogenetic reconstruction. Apart from
356 Read2Tree, MIKE was also tested with the same datasets mentioned above. Though the total
357 run time of MIKE was much less than both Read2Tree and VEHoP, the branching orders
358 generated by MIKE were unstable in most datasets. *Saccostrea glomerata* grouped within
359 *Crassostrea* in most cases (Supplementary Fig. 3), with the sole exception of the 6G dataset,
360 where *S. glomerata* grouped with *Ostrea*. Besides, none of the branch order were the same as
361 that of the whole genome dataset. Compared with Read2Tree and MIKE, VEHoP accepts all
362 three types of input data, including proteins from well-annotated genomes, transcriptomes
363 and DNA genomic data, rather than raw Illumina reads, which highly improved the taxon
364 sampling in the phylogenetic analysis.
365
366 The topology shown on Fig. 3 obtained by VEHoP is identical to 'topology 1' in a former
367 study using mitochondrial genomes (Zhang et al. 2024), which lends further support to the
368 hypothesis of multiple habitat transitions from non-chemosynthetic deep sea to various
369 chemosynthetic habitats, i.e., hot vent, sunken wood, or even inactive vent, over the
370 evolutionary history of Neomphalida (Chen et al. 2024). These results indicate that
371 phylogenomic analyses using VEHoP are more robust than phylogenetic analyses using
372 mitochondrial genomes and the other two published software (i.e., MIKE and Read2Tree).
373
374 We acknowledge that VEHoP currently has several limitations: 1) In some uncommon cases
375 (not shown in this work), HmmCleaner.pl or BMGE appeared to get 'stuck' on a single OG,
376 taking up to thousands of minutes on a single OG. 2) The data size imbalance of raw reads
377 may result in unstable topology through VEHoP, such as data from organisms with extremely
378 low read coverage (< 2X). This might also lead to the expurgation of some taxa, if the strict
379 occupancy criteria (e.g. >80%) is applied. Therefore, adjustment of occupancy and length
380 thresholds are recommended when processing low-coverage sequenced samples. 3) So far,
381 VEHoP is only compiled for use in the Linux system. We are improving the pipeline to make
382 it more widely accessible (e.g., on Windows system).
383
384 With VEHoP, users can define a highly customizable dataset for reference, and it can be a
385 concatenation of high-quality genomes of related species, not limited by an online orthology

database, which might result in much more homologs for ortholog inference. The ortholog inference procedure used in VEHoP has been shown to work well in metazoan (Kocot et al. 2019; Sun et al. 2020; Sun et al. 2021) and bacterial (Li et al. 2023) datasets. With VEHoP, every ortholog that passes the filtering steps is kept, and the user can determine which ones to eliminate based on other criteria if desired, after the process has been completed. In the output folder, the orthologs, concatenated matrix, as well as related partition file will be available for further deep-phylogeny analyses if necessary. Overall, VEHoP shows many advantages, including fast, accurate, and user-friendly. Importantly, VEHoP makes it possible to utilize and combine genomic DNA and transcriptome data widely available in SRAs. We foresee that a wide application of VEHoP would alleviate the problem of low taxon sampling in the phylogenetic analysis of many groups of organisms.

Author Contributions

JS and YL conceived the project. YL coded the pipeline. CC collected the samples. YL and XL carried out the phylogenetic analyses (i.e., draft genome assembly, benchmarks, reanalysis of public data) and manuscript preparation. All authors contributed to the revision of the manuscript.

Acknowledgements

This research was financially supported by the Science and Technology Innovation Project of Laoshan Laboratory (LSKJ202203104), Natural Science Foundation of Shandong Province (ZR2023IQ014), Fundamental Research Funds for the Central Universities (202172002 and 202241002), and the Young Taishan Scholars Program of Shandong Province (tsqn202103036). The *Neomphalus fretterae* specimen used herein was collected during R/V *Falkor (too)* cruise FKt231024 (Project Zombie: Bringing dead vents to life – Ultra fine-scale seafloor mapping") funded by the Schmidt Ocean Institute. We thank the captain and crew of R/V *Falkor (too)* as well as the ROV *SuBastian* team for their immense support of our science. John W. Jamieson (Memorial University of Newfoundland), the chief scientist of cruise FKt231024, is gratefully acknowledged for his diligent execution of the research cruise.

Data Availability

The raw reads from the newly sequenced Neomphalida are deposited in NCBI BioProject (accession number: PRJNA1129887). All the raw inputs (draft genomes, transcripts, and proteins) used, and matrixes generated in this work are available at GitHub (https://github.com/ylify/VEHoP/). For further enquiries on how to use the VEHoP pipeline, please feel free to contact the corresponding author.

Code Availability

The package of VEHoP is available at https://github.com/ylify/VEHoP/.

425

References

Ahmed M, Roberts NG, Adediran F, Smythe AB, Kocot KM, Holovachov O (2022) Phylogenomic Analysis of the Phylum Nematoda: Conflicts and Congruences With Morphology, 18S rRNA, and Mitogenomes. Frontiers in Ecology and Evolution 9, 769565.

Chang C-W, Lyu P-C, Arita M (2011) Reconstructing phylogeny from metabolic substrate-product relationships. BMC Bioinformatics 12:S27.

Chen C, Li Y, Sun J, Beaulieu SE, Mullineaux LS (2024) Two new melanodrymiid snails from the East Pacific Rise indicate the potential role of inactive vents as evolutionary stepping-stones. Systematics and Biodiversity 22:2294014.

Cheon S, Zhang J, Park C (2020) Is phylotranscriptomics as reliable as phylogenomics? Molecular Biology and Evolution 37:3672-3683.

Criscuolo A, Gribaldo S (2010) BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Ecology and Evolution 10:210.

Degnan JH, Rosenberg NA (2006) Discordance of species trees with their most likely gene trees. PLOS Genetics 2:e68.

Di Franco A, Poujol R, Baurain D, Philippe H (2019) Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. BMC Evol Biol 19:21.

Donath A, Jühling F, Al-Arab M, Bernhart SH, Reinhardt F, Stadler PF, Middendorf M, Bernt M (2019) Improved annotation of protein-coding genes boundaries in metazoan mitochondrial genomes. Nucleic Acids Research 47:10543-10552.

Doolittle WF (1999) Phylogenetic Classification and the Universal Tree. Science 284:2124-2128.

Doolittle WF, Logsdon Jr JM (1998) Archaeal genomics: Do archaea have a mixed heritage? Current Biology 8:R209-R211.

Douglas (2018) TransDecoder/TransDecoder. GitHub. Available from: https://github.com/TransDecoder/TransDecoder (accessed March 23, 2020).

Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, Sørensen MV, Haddock SH, Schmidt-Rhaesa A, Okusu A, Kristensen RM, Wheeler WC, Martindale MQ, Giribet G (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452:745-749.

Dylus D, Altenhoff A, Majidian S, Sedlazeck FJ, Dessimoz C (2024) Inference of phylogenetic trees directly from raw sequencing reads using Read2Tree. Nature Biotechnology 42:139–147

Emms DM, Kelly S (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology 20:238.

Fleming JF, Valero-Gracia A, Struck TH (2023) Identifying and addressing methodological incongruence in phylogenomics: A review. Evolutionary Applications 16:1087-1104.

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.

Ghiselli F, Gomes-Dos-Santos A, Adema CM, Lopes-Lima M, Sharbrough J, Boore JL (2021) Molluscan mitochondrial genomes break the rules. Philosophical Transactions of the Royal Society B 376:20200159.

Hao Y, Kajihara H, Chernyshev AV, Okazaki RK, Sun SC (2015) DNA Taxonomy of Paranemertes (Nemertea: Hoplonemertea) with spirally fluted stylets. Zoology 32:571-578.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, MacManes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, LeDuc RD, Friedman N, Regev A (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8:1494-1512.Huynen MA, Bork P (1998) Measuring genome evolution. Proceedings of the National Academy of Sciences of the United States of America 95:5849-5856.

Ibáñez CM, Eernisse DJ, Méndez MA, Valladares M, Sellanes J, Sirenko BI, Pardo-Gandarillas MC (2019) Phylogeny, divergence times and species delimitation of Tonicia (Polyplacophora: Chitonidae) from the eastern Pacific Ocean. Zoological Journal of the Linnean Society 186:915-933.

Irisarri I, Uribe JE, Eernisse DJ, Zardoya R (2020) A mitogenomic phylogeny of chitons (Mollusca: Polyplacophora). BMC Ecology and Evolution 20:22.
Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution 30:772-780.
Kocot KM, Cannon JT, Todt C, Citarella MR, Kohn AB, Meyer A, Santos SR, Schander C, Moroz LL, Lieb B, Halanych KM (2011) Phylogenomics reveals deep molluscan relationships. Nature 477:452-456.
Kocot KM, Todt C, Mikkelsen NT, Halanych KM (2019) Phylogenomics of Aplacophora (Mollusca, Aculifera) and a solenogaster without a foot. Proceedings of the Royal Society B: Biological Sciences 286:20190115.
Lartillot N, Rodrigue N, Stubbs D, Richer J (2013) PhyloBayes MPI: Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Systematic Biology 62:611-615.
Lee Michael SY, Palci A (2015) Morphological phylogenetics in the genomic age. Current Biology 25:R922-R929.
Li C, Kou Q, Zhang Z, Hu L, Huang W, Cui Z, Liu Y, Ma P, Wang H (2021) Reconstruction of the evolutionary biogeography reveal the origins and diversification of oysters (Bivalvia: Ostreidae). Mol Phylogen Evol 164:107268.
Li D, Liu C-M, Luo R, Sadakane K, Lam T-W (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31:1674-1676.
Li H (2023) Protein-to-genome alignment with miniprot. Bioinformatics 39:btad014.
Li Y, He X, Lin Y, Li YX, Kamenev GM, Li J, Qiu JW, Sun J (2023) Reduced chemosymbiont genome in the methane seep thyasirid and the cooperated metabolisms in the holobiont under anaerobic sediment. Molecular Ecology Resources 23:1853-1867.
Liu X, Sigwart JD, Sun J (2023) Phylogenomic analyses shed light on the relationships of chiton superfamilies and shell-eye evolution. Marine Life Science & Technology 5:525-537.
Lozano-Fernandez J (2022) A practical guide to design and assess a phylogenomic study. Genome Biology and Evolution 14:evac129.
Marshall CR (2017) Five palaeobiological laws needed to understand the evolution of the living biota. Nature Ecology & Evolution 1:0165.
Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R (2020) IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Molecular Biology and Evolution 37:1530-1534.
Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T (2014) ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30:i541-i548.
Mongiardino Koch N (2021) Phylogenomic subsampling and the search for phylogenetically reliable loci. Molecular Biology and Evolution 38:4025-4038.
Nei M, Kumar S (2000). Molecular evolution and phylogenetics, Oxford University Press, USA.
Powell CLE, Battistuzzi FU (2022). Testing Phylogenetic Stability with Variable Taxon Sampling. Environmental Microbial Evolution: Methods and Protocols. H. Luo. New York, NY, Springer US**:** 167-188.
Price MN, Dehal PS, Arkin AP (2010) FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLoS One 5:e9490.
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.
Sun J, Chen C, Miyamoto N, Li R, Sigwart JD, Xu T, Sun Y, Wong WC, Ip JCH, Zhang W, Lan Y, Bissessur D, Watsuji TO, Watanabe HK, Takaki Y, Ikeo K, Fujii N, Yoshitake K, Qiu JW, Takai K, Qian PY (2020) The Scaly-foot Snail genome and implications for the origins of biomineralised armour. Nature Communications 11:1657.
Sun J, Li R, Chen C, Sigwart JD, Kocot KM (2021) Benchmarking Oxford Nanopore read assemblers for high-quality molluscan genomes. Proceedings of the Royal Society B: Biological Sciences 376:20200160.
Turnbull R, Steenwyk J, Mutch S, Scholten P, Salazar V, Birch J, Verbruggen H (2023). OrthoFlow: phylogenomic analysis and diagnostics with one command. https://doi.org/10.21203/rs.3.rs-3699210/v1

Wang F, Wang Y, Zeng X, Zhang S, Yu J, Li D, Zhang X (2024) MIKE: an ultrafast, assembly-, and alignment-free approach for phylogenetic tree construction. Bioinformatics 40:btae154.

Xie P, Guo Y, Teng Y, Zhou W, Yu Y (2024) GeneMiner: A tool for extracting phylogenetic markers from next-generation sequencing data. Molecular Ecology Resources:e13924.

Yang Y, Smith SA (2014) Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. Molecular Biology and Evolution 31:3081-3092.

Young AD, Gillung JP (2020) Phylogenomics — principles, opportunities and pitfalls of big-data phylogenetics. Syst Entomol 45:225-247.

Zhang L, Gu X, Chen C, He X, Qi Y, Sun J (2024) Mitogenome-based phylogeny of the gastropod order Neomphalida points to multiple habitat shifts and a Pacific origin. Frontiers in Marine Science 10:1341869.

Zhong Z, Lan Y, Chen C, Zhou Y, Linse K, Li R, Sun J (2022) New mitogenomes in deep-water endemic Cocculinida and Neomphalida shed light on lineage-specific gene orders in major gastropod clades. Frontiers in Ecology and Evolution 10:973485.

Figure Legends

Fig. 1. The workflow of the VEHoP pipeline. a) supported input data; b) homolog extraction; c) ortholog inference; d) phylogenetic analyses.

Fig. 2. Results of phylogenomic analyses with different Ostreida (oysters) datasets. a) annotated genomes; b) DNA reads assembled with MEGAHIT; c) assembled transcriptomes, with proteins generated by miniprot and TransDecoder; d) single species data (*Crassostrea hongkongensis*) size test, including 1X, 2X, 4X, 8X reads; e) the combination of all three types of input data. Red star represents genome data, blue spot represents DNA genomic data, and orange block stands for RNA data.

Fig. 3. Results of phylogenomic analysis using VEHoP on short Illumina sequencing data from Neomphalida. Nodes with blue dots indicate maximal support in all analyses using different methods. *Neomphalus fretterae* was newly sequenced in this study.

Fig. 4. Available phylogenomic resources for major phyla in the major animal clade Lophotrochozoa enumerated in terms of the number of taxa with published genomes (red), RNA-seq datasets (orange), and DNA genomic assemblies (blue). Sizes of the circles are proportional to the number of species in each phylum.

Supplementary Fig. 1. Ostreida phylogeny by VEHoP of subsampled *Crassostrea hongkongensis* data size test.

Supplementary Fig. 2. Ostreida phylogeny by VEHoP of full-size dataset and reduced datasets, including 1G, 2G, 4G, 6G and 8G.

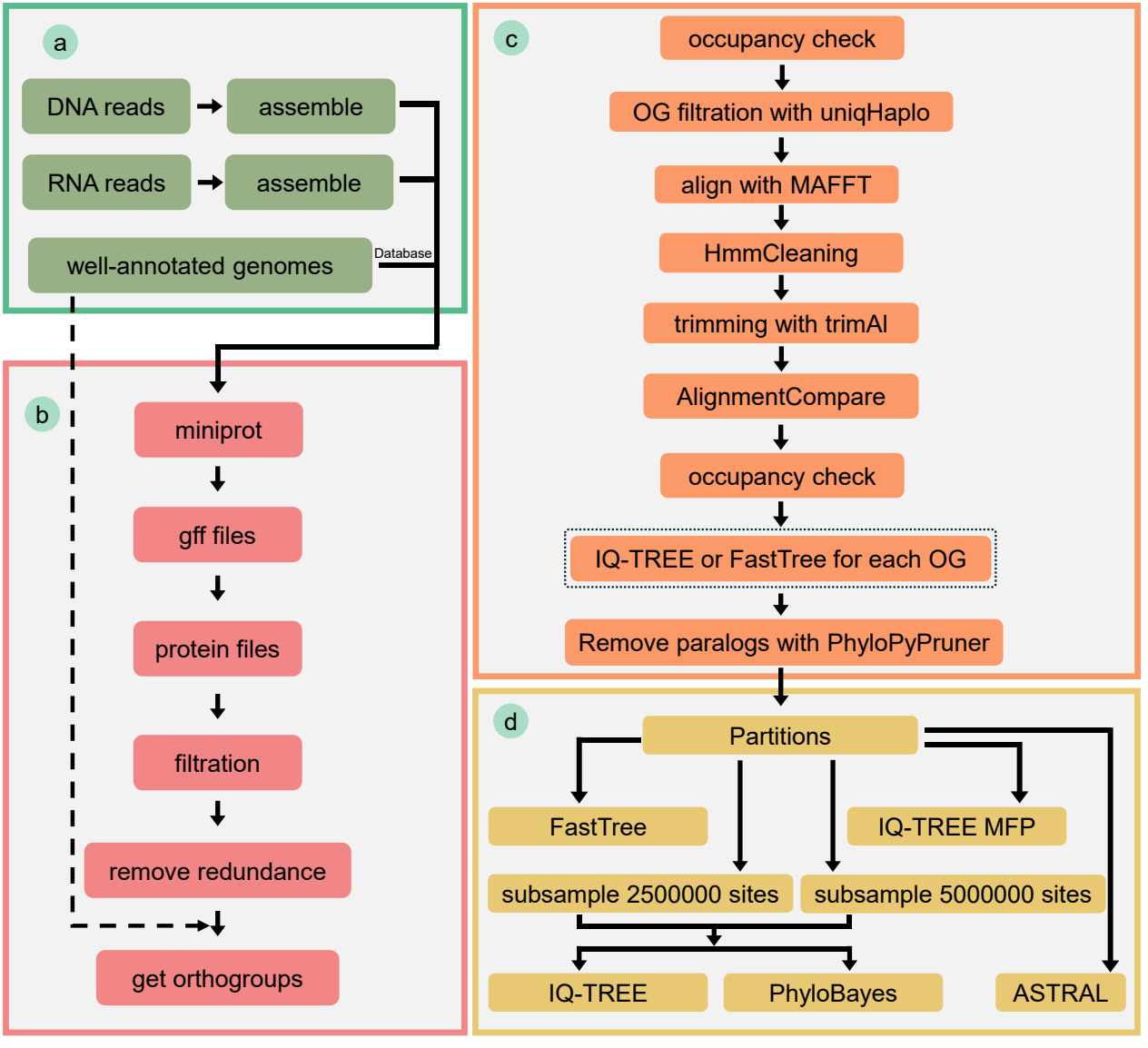
Supplementary Fig. 3. Ostreida phylogeny by ReadTree of full-size dataset and reduced datasets, including 1G, 2G, 4G, 6G and 8G.

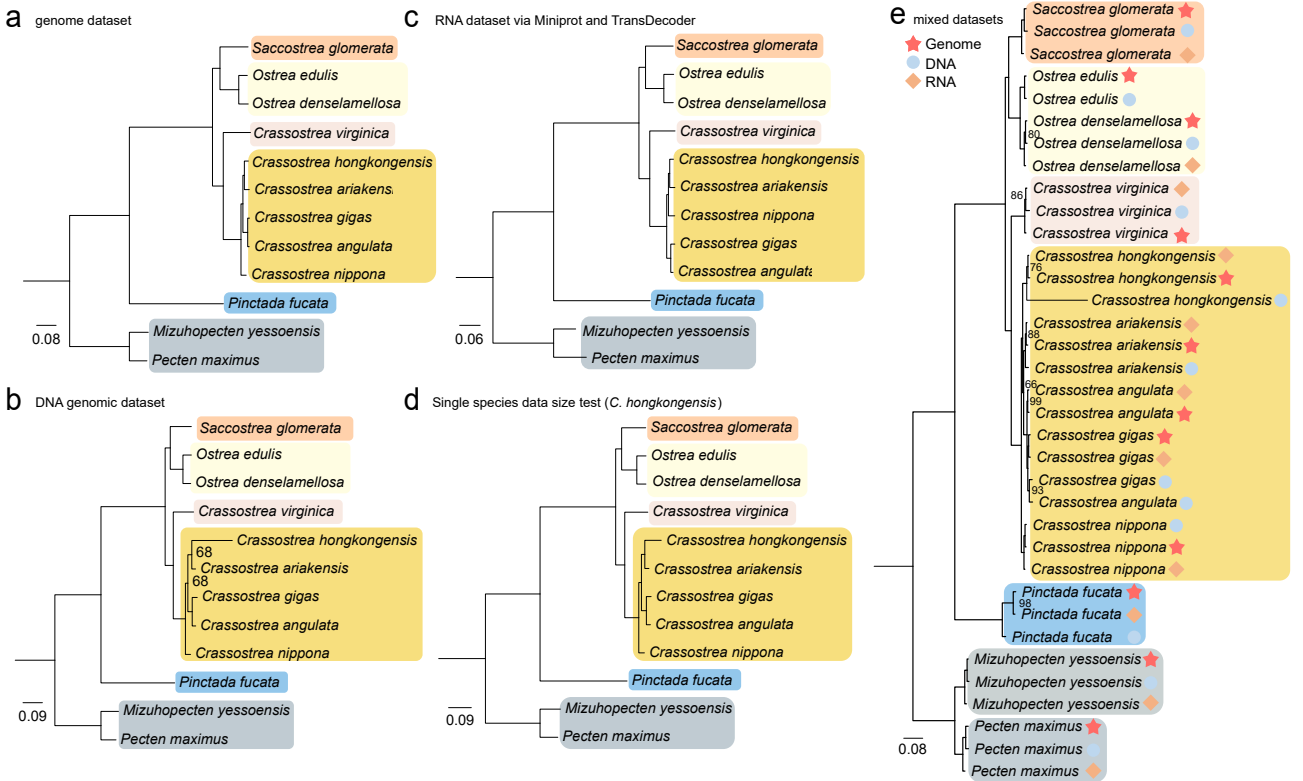
Supplementary Fig. 4. Ostreida phylogeny by MIKE of full-size dataset and reduced datasets, including 1G, 2G, 4G, 6G and 8G.

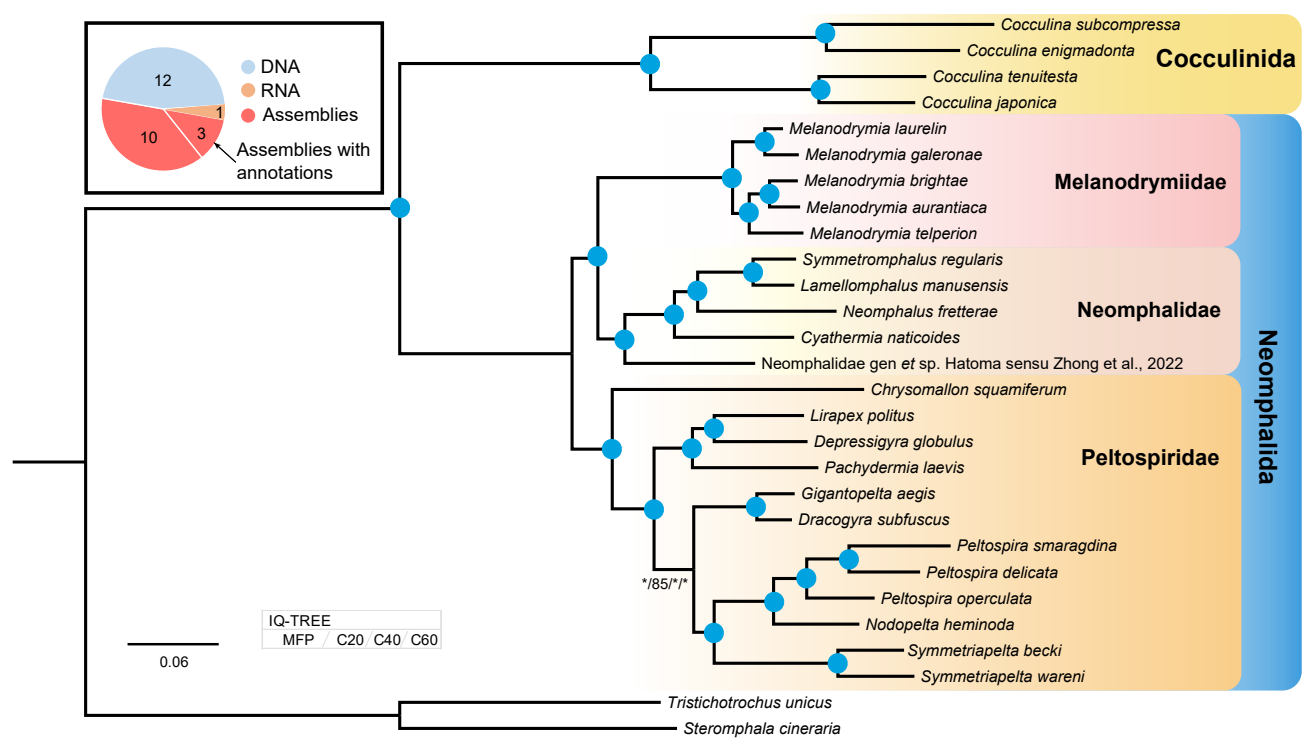
Supplementary Fig. 5. Mitochondrial genome-based phylogeny of Neomphalida.

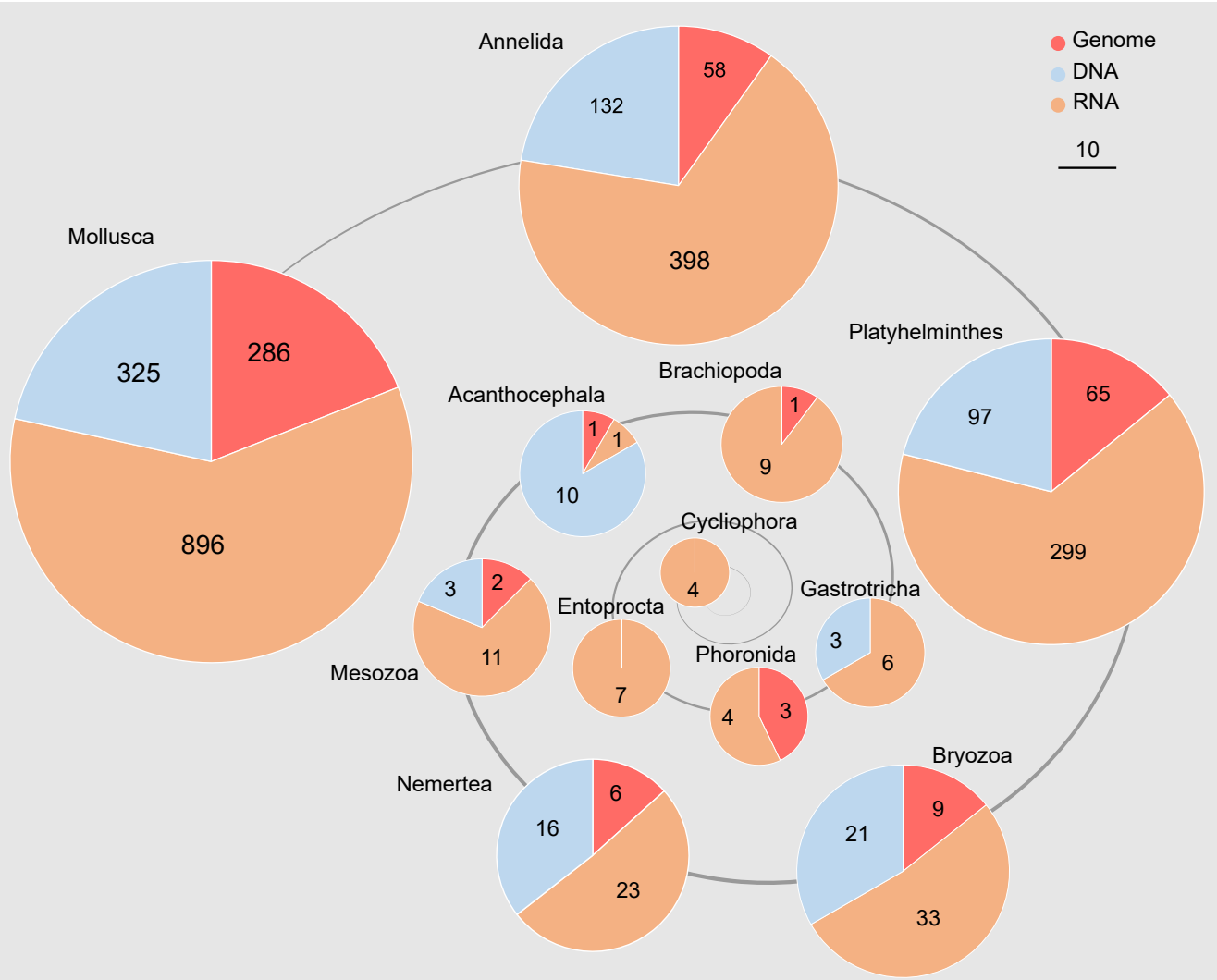
Supplementary Fig. 6. Neomphalida phylogeny based on NGS data, including VEHoP (multiple models), MIKE and Read2Tree.

Supplementary Fig. 7. Occupancy of matrix generated by VEHoP.


a
DNA reads
assemble
RNA reads
assemble
well-annotated genomes
Database
b
miniprot
gff files
protein files
filtration
remove redundance
get orthogroups
c
occupancy check
OG filtration with uniqHaplo
align with MAFFT
HmmCleaning
trimming with trimAl
AlignmentCompare
occupancy check
IQ-TREE or FastTree for each OG
Remove paralogs with PhyloPyPruner
d
Partitions
FastTree
IQ-TREE MFP
subsample 2500000 sites
subsample 5000000 sites
IQ-TREE
PhyloBayes
ASTRAL


a
genome dataset
Saccostrea glomerata
Ostrea edulis
Ostrea denselamellosa
Crassostrea virginica
Crassostrea hongkongensis
Crassostrea ariakens.
Crassostrea gigas
Crassostrea angulata
Crassostrea nippona
Pinctada fucata
Mizuhopecten yessoensis
Pecten maximus
0.08
b
DNA genomic dataset
Saccostrea glomerata
Ostrea edulis
Ostrea denselamellosa
Crassostrea virginica
Crassostrea hongkongensis
68
Crassostrea ariakensis
68
Crassostrea gigas
Crassostrea angulata
Crassostrea nippona
Pinctada fucata
Mizuhopecten yessoensis
Pecten maximus
0.09
c
RNA dataset via Miniprot and TransDecoder
Saccostrea glomerata
Ostrea edulis
Ostrea denselamellosa
Crassostrea virginica
Crassostrea hongkongensis
Crassostrea ariakensis
Crassostrea nippona
Crassostrea gigas
Crassostrea angulata
Pinctada fucata
Mizuhopecten yessoensis
Pecten maximus
0.06
d
Single species data size test (C. hongkongensis)
Saccostrea glomerata
Ostrea edulis
Ostrea denselamellosa
Crassostrea virginica
Crassostrea hongkongensis
Crassostrea ariakensis
Crassostrea gigas
Crassostrea angulata
Crassostrea nippona
Pinctada fucata
Mizuhopecten yessoensis
Pecten maximus
0.09
e
mixed datasets
Genome
DNA
RNA
Saccostrea glomerata
Saccostrea glomerata
Saccostrea glomerata
Ostrea edulis
Ostrea edulis
Ostrea denselamellosa
80
Ostrea denselamellosa
Ostrea denselamellosa
Crassostrea virginica
86
Crassostrea virginica
Crassostrea virginica
Crassostrea hongkongensis
76
Crassostrea hongkongensis
Crassostrea hongkongensis
Crassostrea ariakensis
88
Crassostrea ariakensis
Crassostrea ariakensis
Crassostrea angulata
66
99
Crassostrea angulata
Crassostrea gigas
Crassostrea gigas
Crassostrea gigas
93
Crassostrea angulata
Crassostrea nippona
Crassostrea nippona
Crassostrea nippona
Pinctada fucata
98
Pinctada fucata
Pinctada fucata
Mizuhopecten yessoensis
Mizuhopecten yessoensis
Mizuhopecten yessoensis
Pecten maximus
Pecten maximus
Pecten maximus
0.08


12
1
10
3
DNA
RNA
Assemblies
Assemblies with annotations
Cocculina subcompressa
Cocculina enigmadonta
Cocculina tenuitesta
Cocculina japonica
Cocculinida
Melanodrymia laurelin
Melanodrymia galeronae
Melanodrymia brightae
Melanodrymia aurantiaca
Melanodrymia telperion
Melanodrymiidae
Symmetromphalus regularis
Lamellomphalus manusensis
Neomphalus fretterae
Cyathermia naticoides
Neomphalidae gen et sp. Hatoma sensu Zhong et al., 2022
Neomphalidae
Chrysomallon squamiferum
Lirapex politus
Depressigyra globulus
Pachydermia laevis
Gigantopelta aegis
Dracogyra subfuscus
Peltospira smaragdina
Peltospira delicata
Peltospira operculata
Nodopelta heminoda
Symmetriapelta becki
Symmetriapelta wareni
Peltospiridae
Neomphalida
/85//*
IQ-TREE
MFP / C20 / C40 / C60
0.06
Tristichotrochus unicus
Steromphala cineraria


Genome
DNA
RNA
10
Annelida
58
132
398
Mollusca
286
325
896
Platyhelminthes
65
97
299
Brachiopoda
1
9
Acanthocephala
1
1
10
Cycliophora
4
Gastrotricha
3
6
Entoprocta
7
Phoronida
4
3
Mesozoa
3
2
11
Nemertea
6
16
23
Bryozoa
9
21
33