Molecular Ecology Resources

WILEY

MOLECULAR ECOLOGY RESOURCES

RESOURCE ARTICLE OPEN ACCESS

Reliable Inference of Phylogenomic Relationship via Assembly-Based Strategy Accommodating Raw Reads and Proteins

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

[1]Key Laboratory of Evolution and Marine Biodiversity (Ministry of Education), Institute of Evolution and Marine Biodiversity, Ocean University of China, Qingdao, China | [2]Laboratory for Marine Biology and Biotechnology, Qingdao Marine Science and Technology Center, Laoshan Laboratory, Qingdao, China | [3]X-STAR, Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Yokosuka, 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, Alabama, USA

Correspondence: Yunlong Li (ylify@connect.ust.hk) | Jin Sun (jin_sun@ouc.edu.cn)

Received: 15 August 2025 | **Revised:** 20 January 2026 | **Accepted:** 17 February 2026

Keywords: deep sea | evolution | phylogenomics | phylogeny | pipeline | reads

ABSTRACT

Phylogenomics is a transformative approach in systematics, conservation biology, and biomedical research, enabling the inference of evolutionary relationships by leveraging hundreds to thousands of genes from genomic or transcriptomic data. However, acquiring high-quality genomes and transcriptomes necessitates samples with intact DNA and RNA, substantial sequencing investments, and extensive bioinformatic processing, such as genome/transcriptome assembly and annotation. This challenge is particularly pronounced for rare or difficult-to-collect species, such as those inhabiting the deep sea, where often only fragmented DNA reads are available due to environmental degradation or suboptimal preservation conditions. To address these limitations, we developed VEHoP (Versatile, Easy-to-use Homology-based Phylogenomic pipeline), a tool designed to infer protein-coding regions from diverse inputs, including raw reads (short and long), draft genomes, transcriptomes, and annotated genomes. VEHoP automates the generation of orthologous sequence alignments, concatenated matrices, and phylogenetic trees, streamlining phylogenomic analyses for researchers across disciplines. The tool expands taxonomic sampling by accommodating a wide range of input data types and simplifies phylogenomic workflows, making them accessible to researchers with varying levels of bioinformatic expertise. We validated VEHoP using datasets from oysters, catfish, and insects, demonstrating its ability to produce robust phylogenetic trees with strong bootstrap support, outperforming assembly-free methods. Additionally, we applied VEHoP to reconstruct the phylogeny of the enigmatic deep-sea gastropod order Neomphalida, resolving a well-supported phylogenetic backbone for this poorly understood group. VEHoP is freely available on GitHub (https://github.com/ylify/VEHoP) and easily installable via Bioconda or the configured container image via Docker, Singularity and Apptainer.

1 | 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, sequences of 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 markers were used, such as the

Yunlong Li and Xu Liu have equal contributions.

mitochondrial cytochrome *c* oxidase subunit I (COI), NADH dehydrogenase subunit 4 (NAD4), nuclear ribosomal RNA genes, or a combination of them (Hao et al. 2015; Ibáñez et al. 2019). The improvement of sequencing techniques 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). 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 to 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 phylogenetic questions. For one, some species inhabit inaccessible environments, such as the deep sea and polar regions, or are extremely rare, with only one or a few old, suboptimally preserved specimens available in natural history museums. Also, in certain groups, some species may be easier to sample than others, leading to biased sampling. In these cases, researchers would have to perform phylogenetic reconstruction using a dataset lacking some taxa. If those taxa happen to represent an important node, the tree topology may be changed based on such an imbalanced 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 a few gene markers (Chang et al. 2011). As next-generation sequencing (NGS) technology advanced, more and more genomes and transcriptomes have been sequenced and released to the public, at an elevated rate year after year (Turnbull et al. 2023). Nevertheless, many of these datasets were initially sequenced for organelle genome assembly, genome survey, genome annotation, gene expression level analysis, and so on. These are all potential data 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 practice. In many groups, only a few well-annotated genomes are available, while the rest are transcriptomes and raw Illumina DNA reads at best. To obtain a genome dataset for phylogenomic analyses from these, multiple steps of bioinformatics analyses must be performed, 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 reflect 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. Here, the 'matrix' refers to a phylogenomically critical supermatrix, constructed by integrating the filtered, high-quality orthologous gene alignments obtained from prior steps. This process 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).

Several existing tools for phylogenomic analysis are capable of utilising raw sequencing reads to construct phylogenetic trees. Among them, Read2Tree (Dylus et al. 2023) stands out as a software that enables direct tree inference from reads. However, it has notable limitations. The reference OMA ('Orthologous MAtrix') database specified in Read2Tree lacks customization options, which restricts users' ability to adapt it to specific research needs. Additionally, while Read2Tree is a useful tool for phylogenetic reconstruction, its current workflow for phylogenetic inference requires processing species one by one. It also involves numerous manual curation steps, which gives the workflow a certain degree of complexity in practice. aTRAM 2.0 (Allen et al. 2018) is a flexible locus assembler for NGS data, leveraging iterative BLAST searches and de novo assembly to target genes from short reads, yet only reads are supported as input, and it has been shown to be less efficient than another newer software called ALiBaSeq (Knyshov et al. 2021). ALiBaSeq is an alignment-based tool for extracting phylogenetic markers from low-coverage whole genome sequencing (WGS) and fragmented assemblies, using BLAST/HMMER to stitch homologous regions and trim non-coding sequences. A further alignment-based phylogenetic tool is Patchwork (Thalén et al. 2023), which can take both reads and contigs as the input data, but is sensitive to frameshifts and sequencing errors due to its reliance on intact reading frames for non-coding region trimming. Secapr (Ribeiro et al. 2021) is a bioinformatics pipeline designed for processing sequence capture data in phylogenetics, which can handle raw Illumina read data, and through steps like de novo assembly and reference-based assembly, convert it into multiple sequence alignments for downstream phylogenetic and phylogeographic analyses. It supports both read and assembly inputs, making it versatile for low-coverage WGS and target enrichment datasets. Yet it is narrowly tailored to exon-based targets, struggling with intron-spanning loci or protein-based references. MIKE (Wang et al. 2024) is a MinHash-based and *k*-mer phylogenetic algorithm developed for large-scale next-generation sequencing data, but it suffers from unstable topologies in deep or complex phylogenies, as it relies solely on distance-based methods without incorporating explicit evolutionary models. GeneMiner (Xie et al. 2024) is a toolkit developed for phylogenetic marker mining, which extracts markers from transcriptomic, genomic, or other next-generation sequencing (NGS) or third-generation sequencing (TGS) data. It could be used for multiple gene phylogeny, yet it is still inefficient for phylogenomic analysis due to vague instructions and low numbers of single-copy orthologs extracted.

To address the issues raised above, we here developed a new pipeline which we name 'VEHoP' (Versatile, Easy-to-use Homology-based Phylogenomic pipeline). The VEHoP workflow allows different types of datasets as input, including raw reads, 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 intermediate steps (transcriptomic and genomic assembly, trimming, and so on) and quality control in matrix assembly. For example, this includes occupancy and alignment threshold: occupancy refers to the occupancy threshold, which is the minimum proportion of input species that must have valid sequences for a given orthologous gene to be retained in the concatenated supermatrix; the default occupancy threshold is 2/3, meaning a gene is kept only if it has predicted proteins in at least 2/3 of the species, and the default alignment threshold is 100 amino acids (AAs). Alternative analyses can be specified if needed, such as more trees in CAT+GTR-based PhyloBayes and coalescence-based ASTRAL. 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, we tested it in three benchmarking groups with well-annotated genomes. 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. The other two groups of catfish and insects were also selected to verify how well VEHoP works across the different organism groups. To further test the applicability of VEHoP in resolving phylogenetic issues, we also used it to analyse 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 gastropods remained highly contentious. Our results lend support to the VEHoP as a user-friendly, efficient, and accurate workflow.

2 | Description of VEHoP

2.1 | Input Files and Parameters

The VEHoP pipeline accepts raw reads, draft genomes, transcriptome sequencing data, well-annotated genomes, or any combination of these data types. Raw reads can be NGS or TGS, which could be configured in input with the tab-delimited text (prefix in output; supporting type: NGS, HiFi, ONT, or RNA; read path; read path). It also allows the use of SRA accession numbers instead of a local path, which is compiled to download data from NCBI automatically. The raw data will go through a simple, coarse assembly using a *de novo* assembler, such as MEGAHIT (Li et al. 2015) for genomic data (i.e., NGS), Trinity (Haas et al. 2013) for transcriptome data (i.e., RNA), hifiasm (Cheng et al. 2021) for HiFi and Shasta (Shafin et al. 2020) for nanopore reads (i.e., ONT), after quality control and trimming procedures. Other 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 genomic assemblies. All these assembling procedures can be customised in a VEHoP *.config* file, instead of sophisticated manual operations one by one. In addition, a database for homologue extraction is required when users input genomic reads, transcriptomic reads, or their respective assemblies. This database is critical for accurate homologue identification in subsequent steps, and it can be constructed by integrating protein files from close relatives with well-annotated genomes. By default, VEHoP uses 40 threads (–t 40) throughout, including de novo assembly, homologue-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).

2.2 | Checkpoints

Several checkpoints and the reservation of intermediate results are deployed in VEHoP, including the inference of amino acids from genomic or transcriptomic fragments, the homology among inputs, and their trimming results. It enables users to rerun the procedure in case of interruption and skip the duplicated steps for the adjusted parameters (e.g., occupancy and trimming threshold). The key factor is to run in the same directory as before. For example, it will skip the running of OrthoFinder if the sample inputs are provided.

2.3 | Workflow

The pipeline was coded in Python. All dependencies can be easily installed via Anaconda (Figure 1) and implemented as follows: except HmmCleaner v0.243280 (Di Franco et al. 2019) which the user can install optionally by following the instructions provided in the GitHub repository. Additionally, Docker, Singularity and Apptainer image support has been newly added to facilitate containerized deployment and ensure environment consistency across different systems.

The innovativeness of VEHoP is in the accommodation of raw reads in phylogenomic research, including the automatic data-fetching and assembly, as well as the prediction of coding regions from the newly assembled genomic or transcriptomic contigs. The ortholog inference procedure used in VEHoP has been shown to work well in both metazoan (Kocot et al. 2017, 2019; Sun et al. 2020, 2021; Li, Steenwyk, et al. 2021; Liu et al. 2023; Qi et al. 2024) and bacterial (Li et al. 2023) datasets. The workflow consists of the following steps (Figure 1), which can be implemented using a single command. The parameters of software for each step can be customised by editing the configuration file (*.config*).

(1) Draft assembly from reads based on the content of a configured text, with SRA download (if applicable) and *de novo* assembly, including Trinity v2.5.1 for RNA-seq (NGS in paired-end or single-end mode), Megahit v1.2.9 for metagenome from NGS raw reads, hifiasm v0.21.0-r686 for HiFi reads, and Shasta v0.14.0 for nanopore reads; (2) miniprot v0.13-r248 (Li 2023) is used to map protein sequences from the reference database (provided by users as input) to the coarsely assembled genomic or transcriptomic data to predict gene models; (3) TransDecoder v5.7.1 (Douglas 2018)

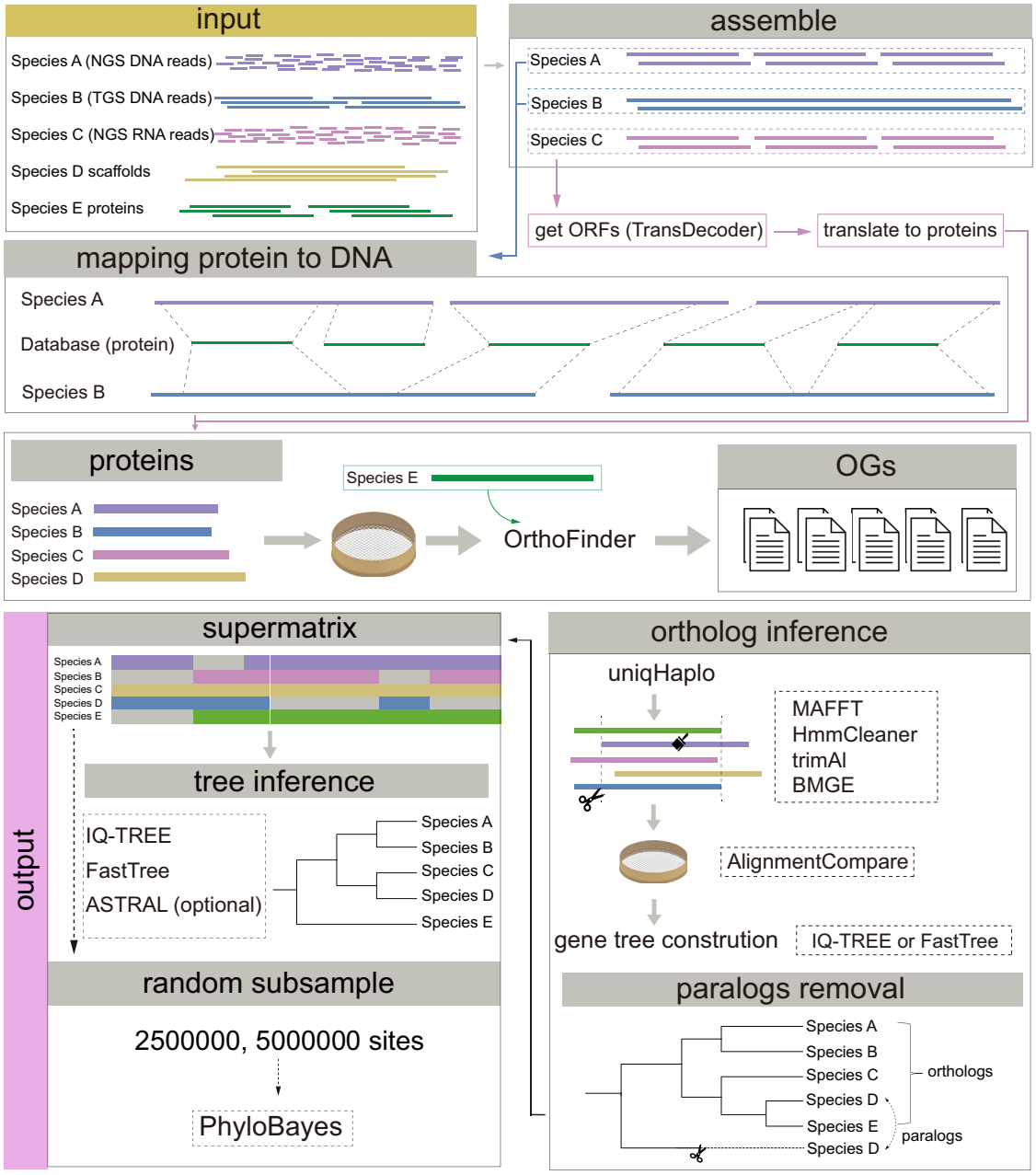

FIGURE 1 | The workflow of the VEHoP pipeline, including supported input data, assembling, homologue extraction, ortholog inference and phylogenetic analyses.

and embedded Python are used to extract quality proteins based on the predicted gene models (no stop codon in the sequences except for the last one and length above threshold); (4) cd-hit v4.8.1 (Fu et al. 2012) is performed to remove redundant sequences with the threshold of 85% similarity; (5) the filtered protein sequences are submitted to OrthoFinder v3.0.1b1 (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); (6) redundant sequences are removed with uniqHaplo while the remaining sequences are aligned with MAFFT v7.525 (Katoh and Standley 2013) with default settings; (7) the misaligned regions are removed with HmmCleaner and the aligned files are further trimmed with BMGE v1.12 (Criscuolo and Gribaldo 2010) and trimAL v1.2rev57 (Capella-Gutiérrez et al. 2009); (8) AlignmentCompare (https://github.com/DamienWaits/Alignment_Compare) is then used to remove sequences shorter than 20 AAs, followed by a second occupancy check to make sure all sequences overlap, which is necessary for single-gene tree reconstructions; (9) IQ-TREE v2.2.0.3 (Minh et al. 2020) or FastTree v2.1.11 (Price et al. 2010) (default being FastTree) is used to build trees for each filtered OGs. (10) PhyloPyPruner v1.2.6 (https://pypi.org/project/phylopypruner) is used to remove paralogs in the filtered alignments; (11) The generated supermatrix is used to reconstruct phylogenetic trees, using IQ-TREE, and FastTree; (12) 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 v1.8c (Lartillot et al. 2013). Apart from concatenation-based phylogeny, the pipeline provides a coalescent phylogenetic approach (default: off) implemented via ASTRAL v1.22.4.6, part of aster v1.16 (Mirarab et al. 2014).

2.4 | 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.

3 | Results

3.1 | Benchmark Test 1: Oyster Dataset

To benchmark the usability and efficiency of the workflow, we collected data from representatives of the order Ostreida (true oysters) 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*, and *Saccostrea glomerata*; plus 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 Table S1.

We tested our workflow with different datasets, including the following. Dataset 1: well-annotated genomes, whose output was labelled as 'reference topology' in Figure 2; Dataset 2: NGS raw reads; Dataset 3: transcriptome reads, assembled with Trinity; and Dataset 4: a dataset combination including all three aforementioned types of 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 between the reference topology from well-annotated genomes and that from NGS raw reads (Figure 2a). All bootstrap values reached 100 in these two trees, except for two nodes in the NGS reads dataset, with a bootstrap value of 68 within the genus *Crassostrea*. However, the position of *C. nippona* was different from the reference topology when using Dataset 3 (transcriptomes), though the bootstrap of all nodes reached 100 (Figure 2a). 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* (Figure S1). To simulate a scenario where high-quality genomes are available for most species but *C. hongkongensis* lacks pre-existing data, we subsampled its sequencing data to 1×, 2×, 4×, 6×, and 8× genome coverage, systematically validating the minimum data volume required to reconstruct reliable phylogenetic trees under such incomplete sampling conditions. This approach mirrors real-world challenges in phylogenomic studies of understudied taxa, where targeted sequencing of missing species often demands empirical evaluation of data sufficiency. 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 was identical to the reference topology, and all node supports were 100% (Figure S2). Reduced datasets for every species (1, 2, 4, 6, and 8 Gb of next-generation sequencing (NGS) reads) were also made and phylogenetic analyses conducted (see Table S2 for details). The results showed that branching order became unstable for the 1 and 2 Gb datasets, resulting in paraphyly within *Crassostrea*. For datasets larger than 2 Gb, the VEHoP was able to recover phylogenetic relationships from well-annotated genomes, at least at the genus level (Figure S1). The total run time was also recorded for these different datasets to test the performance using a Dell PowerEdge R7525 server equipped with two AMD EPYC 7702 CPUs (limited to 40 threads) and 512 GB of RAM. For the reduced datasets, it took 4.24, 10.22, 18.60, 25.46, and 27.32 h 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 h to obtain the results, showing clusters from the same species (except the NGS data from *C. angulata*) and a consistent branching order with the reference tree (Table S3).

Read2Tree 0.1.5 (Dylus et al. 2023), as a representative NGS-based assembly-free algorithm, was chosen to compare with VEHoP. It was 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 h to get a *.nwk* format tree file, with *Pecten maximus* incorrectly grouped with two reference species from the OMA database (Figure S3). As for the 2G dataset, 19.56 h 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 h 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 h 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 Table S3.

To further compare VEHoP and Read2Tree objectively, VEHoP was performed with the same OMA database as Read2Tree based on the same NGS dataset. The results showed that the topologies generated by VEHoP based on the OMA database were the same as those of the custom database. Additionally, the missing gene rate, gap rate, and average sequence length of the supermatrices generated by the two tools were statistically analysed. Read2Tree produced a supermatrix comprising 2055 OGs, featuring an average missing gene rate of 23.88%, a gap rate of 48.16%, and an average sequence length of 259 AAs. In contrast, VEHoP, when using the same OMA database, generated a supermatrix with 4506 OGs, demonstrating a lower average missing gene rate (18.27%), gap rate (20.18%), and a slightly shorter average sequence length of 254 AAs (detailed in Table S4). Notably, the average Gene Concordance Factor (GCF) further highlighted VEHoP's superiority: Read2Tree's matrix showed a GCF of 46.76%, whereas VEHoP achieved a much higher GCF of 63.36% (73.47% in the reference topology based on well-annotated genomes, Table S4), indicating stronger topological consistency with the reference trees.

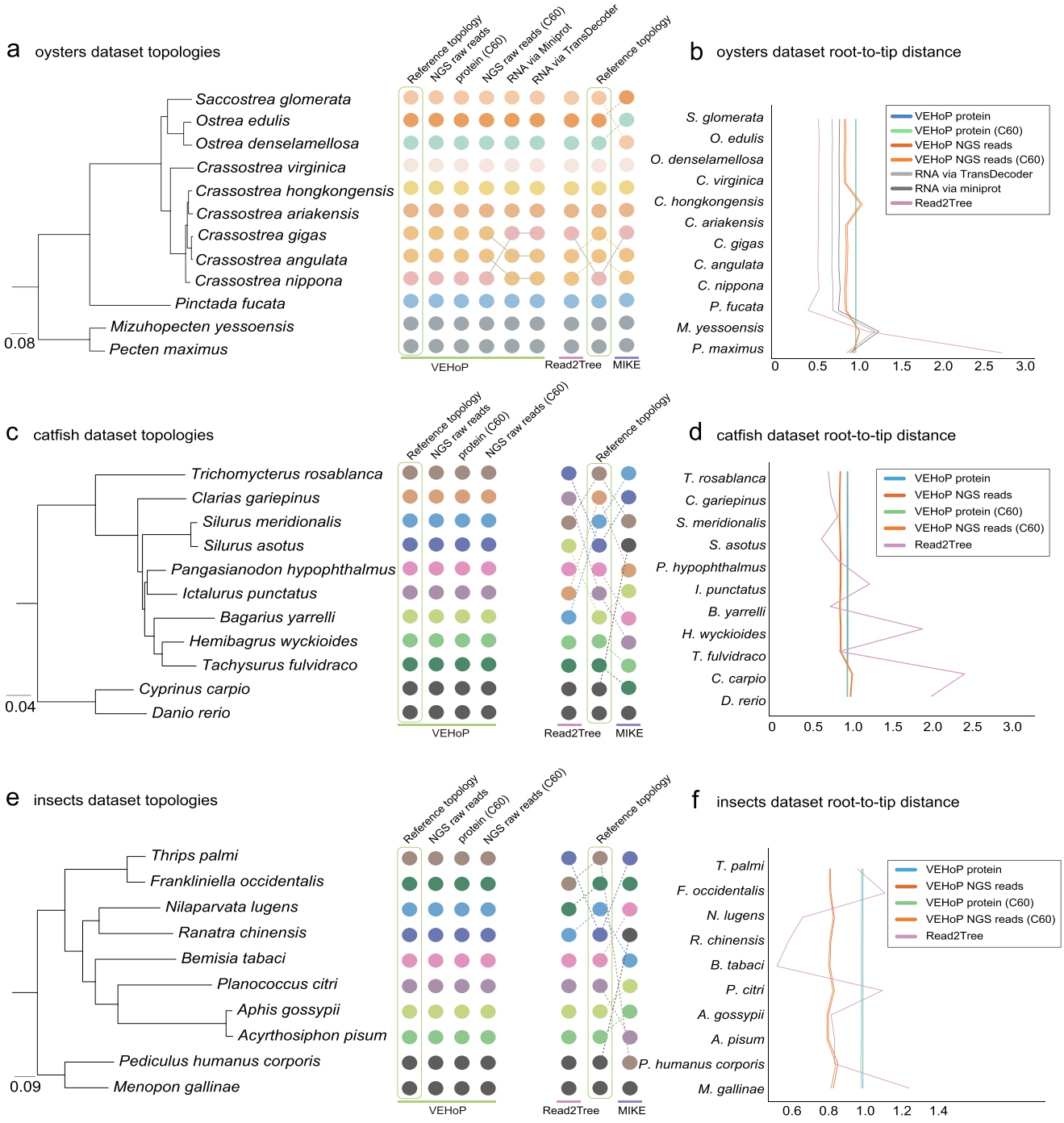

FIGURE 2 | Results of phylogenomic analyses with different datasets, including true oysters (Ostreida), catfish (Siluriformes) and insects (Condylognatha). All depicted topologies in the comparison subfigures (a, c, e) take the phylogenies inferred from well-annotated genomes as the reference topologies. The root-to-tip distances in subfigures (b, d, f) are standardised to the reference topology, with 1.0 representing the reference topology. (a) True oyster dataset topology comparison between different methods and the reference topologies. (b) True oyster dataset root-to-tip distance analysis. (c) Catfish dataset topology comparison between different methods and the reference topologies. (d) Catfish dataset root-ot-tip distance analysis. (e) Insect dataset topology comparison between different methods and the reference topologies. (f) Insect dataset root-to-tip distance analysis. In all topology comparison subfigures (a, c, e): Each coloured dot denotes an individual species (consistent colour coding for the same species across subfigures), and dashed lines connect species to their positions in the topologies inferred by different methods. These dashed lines indicate that the species' placement in the tree generated by the corresponding dataset/method is inconsistent with the reference topology.

MIKE was chosen for comparisons with VEHoP as a representative of a typical k-mer based phylogenetic tool. It was 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* was nested with *Crassostrea* or clustered with *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, Kou, et al. 2021; Figure S4). The run time of MIKE for different sizes of datasets can be found in Table S3. Notably, a comparative summary among MIKE, Read2Tree, and VEHoP reveals key trade-offs: MIKE runs significantly faster than both counterparts but only produces cladograms (without branch lengths or other quantitative phylogenetic metrics) and fails to generate consistent topologies across most dataset sizes. In contrast, while Read2Tree and VEHoP require more runtime (especially for large datasets), they produce complete phylogenetic trees with robust topological support. VEHoP further outperforms Read2Tree in terms of congruence with reference topologies, lower missing gene rates, and higher gene concordance factors (Table S4).

Additionally, ASTER (Zhang et al. 2025) and ROADIES (Gupta et al. 2025) were included in the comparative analyses with VEHoP. We used datasets of varying sizes: 1G, 2G, 4G, 6G, 8G and full-size. For ASTER, polytomy was consistently observed across datasets. It occurred in the *C. virginica* clade in the 1G dataset and was present in the *C. nippona* clade in the 2G, 6G, 8G and full-size datasets. Polytomies further appeared in both the *S. glomerata* and *C. nippona* clades in the 4G dataset. For ROADIES, none of the datasets successfully recovered the reference topology. Notable misplacements were observed across multiple taxa. Both *S. glomerata* and *P. maximus* deviated significantly from their expected phylogenetic positions.

The root-to-tip distances for each species were calculated using various tree files to assess tree quality. The distances of each tip in reference topology (well-annotated genomes) were employed to normalise the corresponding distances in other trees (Table S5). The findings showed similar root-to-tip distances to reference topology in the results of VEHoP, whereas the trees produced by Read2Tree displayed significant variance compared to the other results (Figure 2b). The root-to-tip distances in MIKE are not applicable for quantification.

3.2 | Benchmark Test 2: Catfish Dataset

The fish datasets include 9 catfishes from the order Siluriformes: *Bagarius yarrelli*, *Clarias gariepinus*, *Hemibagrus wyckioides*, *Ictalurus punctatus*, *Pangasianodon hypophthalmus*, *Silurus asotus*, *Silurus meridionalis*, *Tachysurus fulvidraco*, and *Trichomycterus rosablanca*. The common carp *Cyprinus carpio* and the zebrafish *Danio rerio* were used as outgroups. The datasets include well-annotated genomes and NGS raw reads. The source of this data can be found in Table S1.

Two datasets were used in this benchmark test, including Dataset 1: well-annotated genomes and Dataset 2: NGS raw reads. VEHoP, Read2Tree, MIKE, ASTER and ROADIES were applied to Dataset 2, whose topologies can be found in Figure 2c. Furthermore, an additional IQ-TREE (MFP) procedure was also performed on the matrix generated by Read2Tree. IQ-TREE (MFP) and IQ-TREE (C60) were performed based on the matrix generated by VEHoP in both datasets. For comparison, the topology generated by IQ-TREE (MFP) based on well-annotated genomes, in this case, Dataset 1, was chosen to be the reference topology. VEHoP showed great consistency and stability in NGS reads (Dataset 2), while Read2Tree, MIKE, ASTER and ROADIES all resulted in rather different topologies compared to the reference topology (Figure 2c). The topology generated by Read2Tree showed that *S. asotus* came to the basal position of the ingroup instead of *T. rosablanca*. The genus *Silurus* was recovered as paraphyletic. In MIKE, the outgroup species *Cyprinus carpio* nested within catfishes. The genus *Silurus* became basal instead of *T. rosablanca*. *Clarias gariepinus* was nested deep inside the ingroup catfish species, while in the reference topology, it was positioned at the location of the secondary basal node. Both ASTER and ROADIES misassigned the phylogenetic position of *C. gariepinus*. ASTER additionally produced unresolved polytomies across the analysed datasets (Figure S9). Root-to-tip distance was also calculated and normalised for each tree file generated (Table S5); all results generated by VEHoP demonstrated a high level of consistency (Figure 2d). Furthermore, we performed VEHoP based on the OMA database and topologies were the same as that as the custom database. Read2Tree generated a supermatrix consists of 200 OGs, with an average missing gene rate of 40.86%, gap rate of 67.35%, average sequence length of 543 and an average gCF of 30.21%, while VEHoP generated a matrix consists of 5770 OGs, with an average missing gene rate of 21.28%, gap rate of 25.48%, average sequence length of 254 and an average gCF of 46.73% (57.35% in the reference topology based on well-annotated genomes, Table S4). All original tree topologies can be found in Figure S5.

3.3 | Benchmark Test 3: Insect Dataset

The insect datasets include 8 species from the superorder Condylognatha, which comprises two orders: Thysanoptera (thrips) and Hemiptera (true bugs). These species are *Acyrthosiphon pisum*, *Aphis gossypiii*, *Bemisia tabaci*, *Frankliniella occidentalis*, *Nilaparvata lugens*, *Planococcus citri*, *Ranatra chinensis*, and *Thrips palmi*. Two taxa from order Psocodea were selected as outgroups: *Pediculus humanus corporis* and *Menopon gallinae*. The datasets used in this study also include well-annotated genomes and NGS raw reads, whose data source can also be found in Table S1.

The dataset composition in this benchmark test was the same as that in benchmark test 2: including Dataset 1: well-annotated genomes; and Dataset 2: NGS raw reads. The methodology used was the same as for the catfish case study. The results generated by VEHoP shared the same branching order as the reference topology (i.e., inferred from well-annotated genomes), who successfully recovered the monophyletic relationship of Thysanoptera and Hemiptera. But in Read2Tree, MIKE, ASTER and ROADIES, they all resulted in inconsistent branching orders compared to reference topology (Figure 2e) In Read2Tree, Hemiptera was successfully recovered as monophyletic, yet *N. lugens* was assigned to Thysanoptera incorrectly. In terms of MIKE, *M. gallinae* nested within Hemiptera as an outgroup species. Moreover, both Hemiptera and Thysanoptera were not recovered as monophyletic groups. The inconsistency can also be found in the root-to-tip distance results in Figure 2f. Compared with the reference topology, both ASTER and ROADIES failed to recover the clade comprising *T. palmi* and *F. occidentalis* as the root of the ingroups. Additionally, ROADIES exhibited unresolved polytomies across the entire phylogenetic tree (Figure S9).

In contrast, VEHoP, when leveraging the OMA database, yielded a phylogenetic topology identical to that derived from a custom database. Read2Tree generated a supermatrix comprising 200 orthologous groups (OGs), characterised by an average missing gene rate of 77.30%, a gap rate of 93.79% and an average sequence length of 758. Conversely, VEHoP produced a substantially larger matrix of 2733 OGs, demonstrating notably lower average missing gene (17.41%), gap (21.70%) rates, highlighting its superior performance in data completeness and alignment quality though with a relatively shorter average sequence length of 261. Furthermore, the gCF value was calculated for the dataset, yielding a value of 57.08%, slightly lower than 62.16% based on well-annotated genomes. Notably, attempts to compute the GCF for the matrix generated by Read2Tree were unsuccessful (Table S4), probably due to topological discordance with the reference topology. All original tree topologies can be found in Figure S5.

3.4 | Case Study: 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 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 ingroups, 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 Figure S6), 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 with the tree of MFP as the guide tree), 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 (Figure 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 (Figure S7).

4 | Discussion

We present VEHoP, a new pipeline for phylogenomic analyses with the flexibility of using genomic assemblies, well-annotated genomes, NGS raw reads, RNA-seq raw reads, 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 can 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 complicated 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 underutilised in phylogenomic studies. Understandably, these data vary in quality and coverage, and thus it has been challenging to use them together in phylogenetic analyses. With VEHoP, however, researchers can now extract homologues 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 phylogenomics based on NCBI data (Figure 4 and 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 are annotated) while an additional 896 species have transcriptomic data. These two data types are the most commonly used source data for phylogenomic analysis. With VEHoP, we can further include 325 species which lack both genome and transcriptome data but with genomic data, greatly expanding the taxon coverage.

In our benchmarking study using various data types from true oysters (benchmark test 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 homologues directly from assembled transcripts with miniprot; and (2) predicting proteins with TransDecoder. These two strategies resulted in the same branching order, and each node reached 100% support. However, the branching

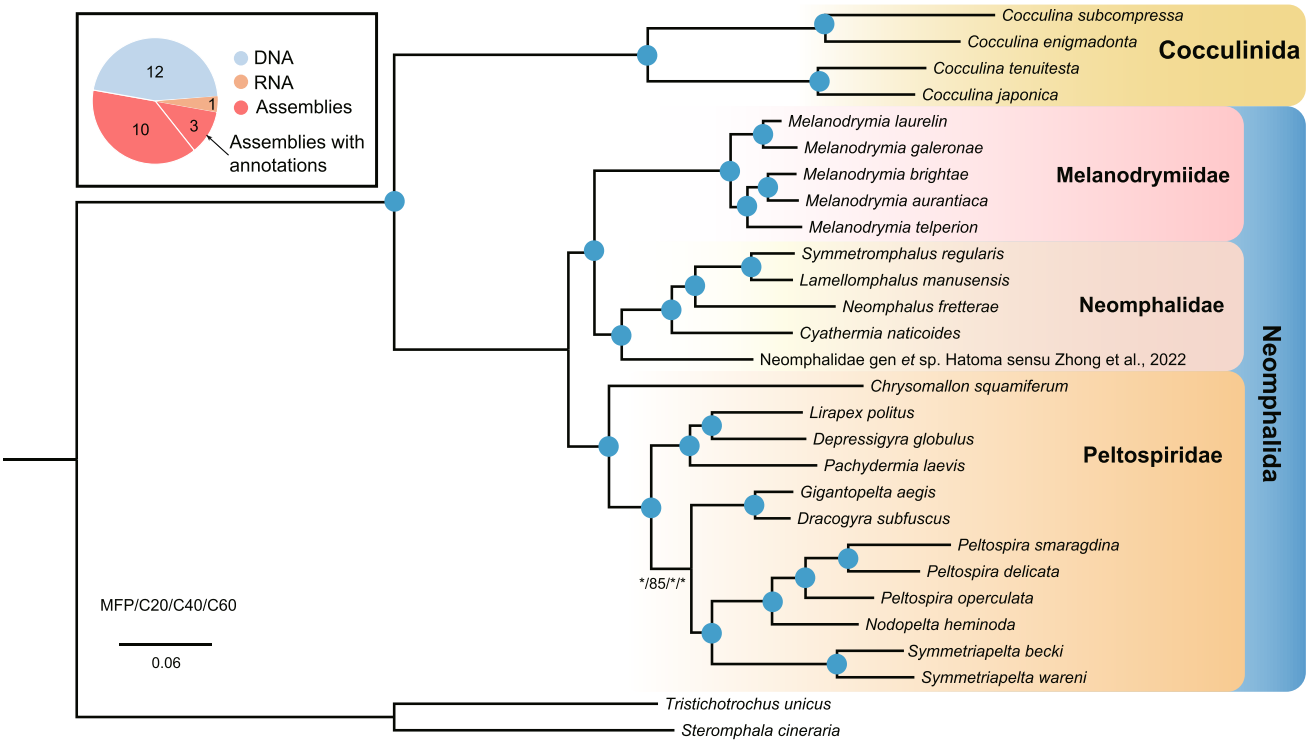

FIGURE 3 | Results of phylogenomic analysis using VEHoP on short Illumina sequencing data from Neomphalida (with IQ-TREE). ModelFinder Plus (MFP) was first used to select the optimal evolutionary model, and site-specific frequency models (C20, C40, C60) were also applied. The scale bar (0.06) indicates the number of amino acid substitutions per site. Nodes with blue dots indicate maximal support in all analyses. For nodes without blue dots, their support degrees are denoted in the format of "MFP/C20/C40/C60" (where "*" stands for 100% support in the corresponding model). *Neomphalus fretterae* was newly sequenced in this study.

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 homologues 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 more accurate compared with the TransDecoder strategy in tree construction and still highly robust at the genus level, since the transcripts were obtained by blasting with close relatives, which in some cases could reduce the impact of contamination.

We also tested Read2Tree 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 species as a reference to reconstruct phylogenetic trees with Read2Tree. Both Read2Tree and VEHoP could not reveal the same branching order as that of the high-quality genome dataset. The position of *Crassostrea nippona* was unstable. However, VEHoP successfully recovered the same branching order as the same as reference topology inferred from well-annotated genomes, while Read2Tree failed to recover the branching order with low-coverage datasets. As for run time comparison, VEHoP performed much quicker with datasets less than 4G. After 4G, Read2Tree took less time than VEHoP, since it reconstructed trees directly from raw sequencing reads, and VEHoP needed to assemble the reads first before proceeding with phylogenetic reconstruction. Apart from Read2Tree, MIKE was also tested with the same datasets mentioned above. Though the total run time of MIKE was much less than both Read2Tree and VEHoP, the branching orders generated by MIKE were unstable in most datasets. *Saccostrea glomerata* was grouped within *Crassostrea* in most cases (Figure S3), with the sole exception of the 6G dataset, where *S. glomerata* grouped with *Ostrea*. Besides, none of the branching orders were the same as the reference topology. Missing gene rate, gap rate, average sequence length and gCF of matrix generated by VEHoP and Read2Tree were compared. VEHoP showed higher quality with lower missing gene rate and gap rate, and higher gCF value. Compared with Read2Tree and MIKE, VEHoP accepts all three types of input data, including proteins from well-annotated genomes, transcriptomes, and DNA genomic data, as well as raw Illumina reads, which highly improved the taxon sampling in the phylogenetic analysis.

To test the universality of VEHoP across diverse taxa, we further employed catfish and insect datasets for testing, comparing the results with those of Read2Tree and MIKE. In these two benchmark tests, the tree generated based on well-annotated genomes was chosen as the reference topology. VEHoP successfully reproduced the same branching orders as that of the reference topology in both cases. In the catfish datasets, all results generated by VEHoP exhibited a high degree of consistency with 100% support for all nodes. In contrast, Read2Tree resulted in a paraphyletic genus, and MIKE misclassified

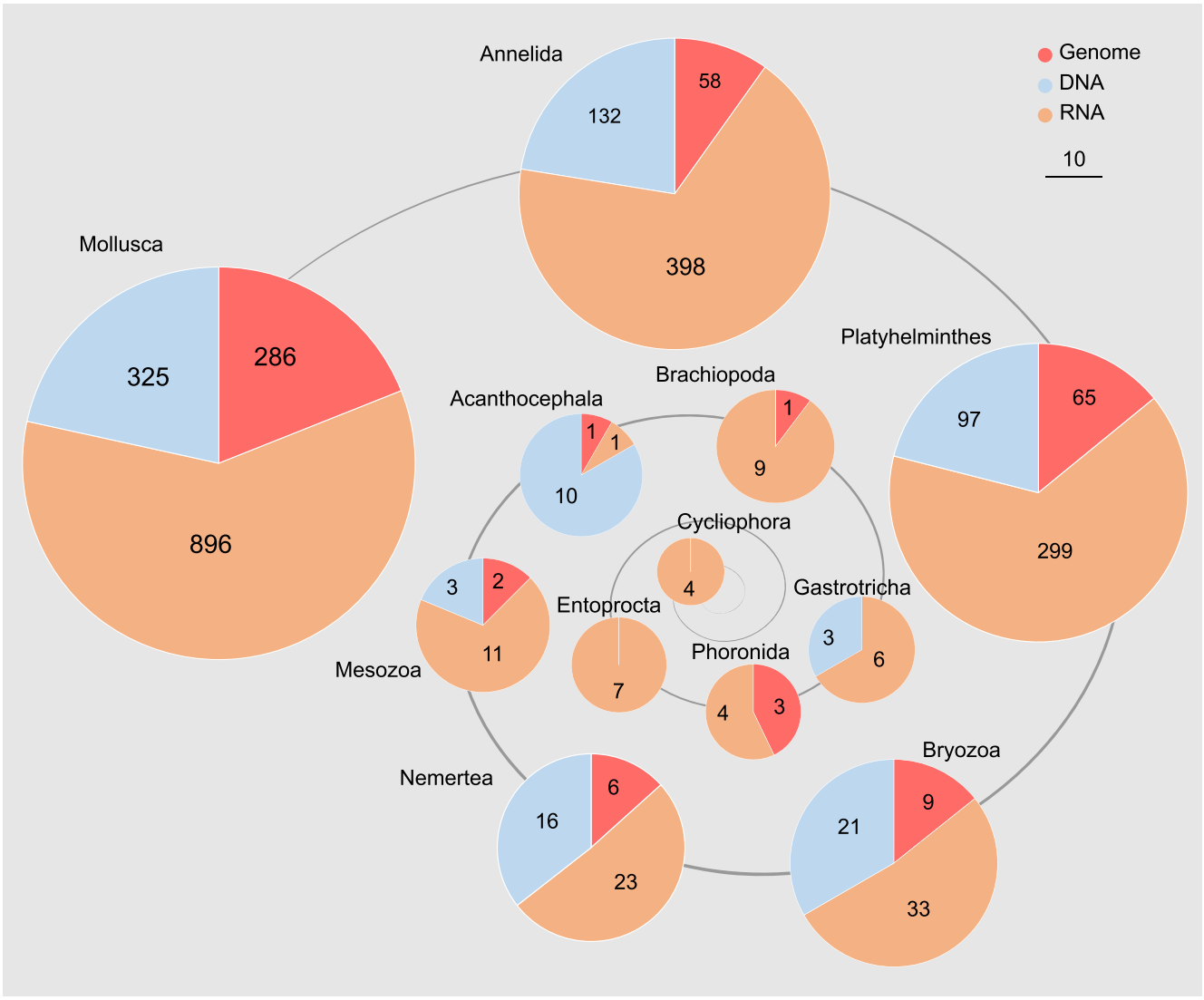

FIGURE 4 | Available phylogenomic resources for major phyla in the major animal clade Lophotrochozoa enumerated in terms of the number of taxa with published genomes (red), RNA-seq datasets (orange), and DNA genomic assemblies (blue). The scale bar (labelled "10") indicates the radius of the circles. Sizes of the circles are proportional to the number of species in each phylum.

outgroup species within the ingroup catfishes. Regarding the insect datasets, VEHoP effectively recovered both Hemiptera and Thysanoptera as monophyletic groups with high bootstrap support values. But in Read2Tree, Thysanoptera was found to be paraphyletic. The difference might be ascribed to the lower missing gene and gap rate, along with higher gCF value, indicating the better matrix quality of VEHoP. In MIKE, none of the orders were recovered as monophyletic. These findings indicate that assembly-free phylogenomic methods still have certain limitations. The inconsistency was also shown in the root-to-tip distance analysis.

We then applied VEHoP to resolve the evolutionary history of the deep-sea gastropod order Neomphalida, which mostly lacks high-quality genome assemblies (unlike the three benchmarking tests). The topology shown in Figure 3 obtained by VEHoP is identical to 'topology 1' in a former study using mitochondrial genomes (Zhang et al. 2024), which lends further support to the hypothesis of multiple habitat transitions from non-chemosynthetic deep sea to various chemosynthetic habitats, that is, hot vent, sunken wood, and even inactive vent, over the evolutionary history of Neomphalida (Chen et al. 2024). These results indicate that phylogenomic analyses using VEHoP are more robust than phylogenetic analyses using mitochondrial genomes and the other two published software (i.e., MIKE and Read2Tree).

We acknowledge that VEHoP currently has several limitations: (1) In some uncommon cases (not shown in this work), HmmCleaner.pl. or BMGE appeared to get 'stuck' on a single OG, taking up to thousands of minutes on a single OG. (2) The data size imbalance of raw reads may result in unstable topology through VEHoP, such as data from organisms with extremely low read coverage (< 2×). This might also lead to the expurgation of some taxa, if the strict occupancy criteria (e.g., > 80%) is applied. Therefore, adjustment of occupancy and length thresholds are recommended when processing low-coverage sequenced samples. (3) So far, VEHoP is only compiled for use in the Linux system. We are improving the pipeline to make it more widely accessible in the future (e.g., on Windows system).

With VEHoP, users can define a highly customizable dataset for reference, and it can be a mixture of high-quality genomes of related species, not limited by an online orthology database, which might result in many more homologues for ortholog inference. 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 being fast, accurate, user-friendly, and being highly customizable, including the reference database and parameters. Importantly, VEHoP makes it possible to utilise 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 taxonomy sampling in the phylogenetic analysis of many organismal groups.

Author Contributions

J.S. and Y.L. conceived the project. Y.L. coded the pipeline. C.C. collected the samples. Y.L. and X.L. carried out the phylogenetic analyses (i.e., draft genome assembly, benchmarks, reanalysis of public data) and original manuscript preparation. All authors contributed to the revision of the manuscript.

Acknowledgements

This research was financially supported by the National Key Research and Development Program of China (2024YFC2816100), Science and Technology Innovation Project of Laoshan Laboratory (LSKJ202203104), Natural Science Foundation of Shandong Province (ZR2023JQ014), Fundamental Research Funds for the Central Universities (202172002 and 202241002), and the Young Taishan Scholars Program of Shandong Province (tsqn202103036).

We sincerely thank signed reviewers, Sina Majidian and Alana Alexander for their constructive comments, which have significantly improved the quality of this manuscript. 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.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

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 matrices generated in this work are available at Figshare (https://doi.org/10.6084/m9.figshare.26370955 including oyster dataset and https://doi.org/10.6084/m9.figshare.28189616 for fish and insect datasets). For further enquiries on how to use the VEHoP pipeline, please feel free to contact the corresponding authors.

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

References

Allen, J. M., R. LaFrance, R. A. Folk, K. P. Johnson, and R. P. Guralnick. 2018. "aTRAM 2.0: An Improved, Flexible Locus Assembler for NGS Data." *Evolutionary Bioinformatics Online* 14: 4546.

Capella-Gutiérrez, S., J. M. Silla-Martínez, and T. Gabaldón. 2009. "trimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses." *Bioinformatics* 25: 1972–1973.

Chang, C.-W., P.-C. Lyu, and M. Arita. 2011. "Reconstructing Phylogeny From Metabolic Substrate-Product Relationships." *BMC Bioinformatics* 12: S27.

Chen, C., Y. Li, J. Sun, S. E. Beaulieu, and L. S. Mullineaux. 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.

Cheng, H., G. T. Concepcion, X. Feng, H. Zhang, and H. Li. 2021. "Haplotype-Resolved De Novo Assembly Using Phased Assembly Graphs With Hifiasm." *Nature Methods* 18: 170–175.

Cheon, S., J. Zhang, and C. Park. 2020. "Is Phylotranscriptomics as Reliable as Phylogenomics?" *Molecular Biology and Evolution* 37: 3672–3683.

Criscuolo, A., and S. Gribaldo. 2010. "BMGE (Block Mapping and Gathering With Entropy): A New Software for Selection of Phylogenetic Informative Regions From Multiple Sequence Alignments." *BMC Evolutionary Biology* 10: 210.

Di Franco, A., R. Poujol, D. Baurain, and H. Philippe. 2019. "Evaluating the Usefulness of Alignment Filtering Methods to Reduce the Impact of Errors on Evolutionary Inferences." *BMC Evolutionary Biology* 19: 21.

Donath, A., F. Jühling, M. Al-Arab, et al. 2019. "Improved Annotation of Protein-Coding Genes Boundaries in Metazoan Mitochondrial Genomes." *Nucleic Acids Research* 47: 10543–10552.

Doolittle, W. F. 1999. "Phylogenetic Classification and the Universal Tree." *Science* 284: 2124–2128.

Doolittle, W. F., and J. M. Logsdon Jr. 1998. "Archaeal Genomics: Do Archaea Have a Mixed Heritage?" *Current Biology* 8: R209–R211.

Douglas, P. 2018. "TransDecoder/TransDecoder." GitHub. Accessed March 23, 2020. https://github.com/TransDecoder/TransDecoder.

Dunn, C. W., A. Hejnol, D. Q. Matus, et al. 2008. "Broad Phylogenomic Sampling Improves Resolution of the Animal Tree of Life." *Nature* 452: 745–749.

Dylus, D., A. Altenhoff, S. Majidian, F. J. Sedlazeck, and C. Dessimoz. 2023. "Inference of Phylogenetic Trees Directly From Raw Sequencing Reads Using Read2Tree." *Nature Biotechnology* 42: 139–147.

Emms, D. M., and S. Kelly. 2019. "OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics." *Genome Biology* 20: 238.

Fleming, J. F., A. Valero-Gracia, and T. H. Struck. 2023. "Identifying and Addressing Methodological Incongruence in Phylogenomics: A Review." *Evolutionary Applications* 16: 1087–1104.

Fu, L., B. Niu, Z. Zhu, S. Wu, and W. Li. 2012. "CD-HIT: Accelerated for Clustering the Next-Generation Sequencing Data." *Bioinformatics* 28: 3150–3152.

Ghiselli, F., A. Gomes-Dos-Santos, C. M. Adema, M. Lopes-Lima, J. Sharbrough, and J. L. Boore. 2021. "Molluscan Mitochondrial Genomes Break the Rules." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 376: 20200159.

Gupta, A., S. Mirarab, and Y. Turakhia. 2025. "Accurate, Scalable, and Fully Automated Inference of Species Trees From Raw Genome Assemblies Using ROADIES." *Proceedings of the National Academy of Sciences of the United States of America* 122, no. 19: e2500553122.

Haas, B. J., A. Papanicolaou, M. Yassour, et al. 2013. "De Novo Transcript Sequence Reconstruction From RNA-Seq Using the Trinity Platform for Reference Generation and Analysis." *Nature Protocols* 8: 1494–1512.

Hao, Y., H. Kajihara, A. V. Chernyshev, R. K. Okazaki, and S. C. Sun. 2015. "DNA Taxonomy of Paranemertes (Nemertea: Hoplonemertea) With Spirally Fluted Stylets." *Zoological Science* 32: 571–578.

Huynen, M. A., and P. Bork. 1998. "Measuring Genome Evolution." *Proceedings of the National Academy of Sciences of the United States of America* 95: 5849–5856.

Ibáñez, C. M., D. J. Eernisse, M. A. Méndez, et al. 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., J. E. Uribe, D. J. Eernisse, and R. Zardoya. 2020. "A Mitogenomic Phylogeny of Chitons (Mollusca: Polyplacophora)." *BMC Evolutionary Biology* 20: 22.

Katoh, K., and D. M. Standley. 2013. "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability." *Molecular Biology and Evolution* 30: 772–780.

Knyshov, A., E. R. L. Gordon, and C. Weirauch. 2021. "New Alignment-Based Sequence Extraction Software (ALiBaSeq) and Its Utility for Deep Level Phylogenetics." *PeerJ* 9: e11019.

Kocot, K. M., J. T. Cannon, C. Todt, et al. 2011. "Phylogenomics Reveals Deep Molluscan Relationships." *Nature* 477: 452–456.

Kocot, K. M., T. H. Struck, J. Merkel, et al. 2017. "Phylogenomics of Lophotrochozoa With Consideration of Systematic Error." *Systematic Biology* 66: 256–282.

Kocot, K. M., C. Todt, N. T. Mikkelsen, and K. M. Halanych. 2019. "Phylogenomics of Aplacophora (Mollusca, Aculifera) and a Solenogaster Without a Foot." *Proceedings of the Royal Society B: Biological Sciences* 286: 20190115.

Lartillot, N., N. Rodrigue, D. Stubbs, and J. Richer. 2013. "PhyloBayes MPI: Phylogenetic Reconstruction With Infinite Mixtures of Profiles in a Parallel Environment." *Systematic Biology* 62: 611–615.

Li, C., Q. Kou, Z. Zhang, et al. 2021. "Reconstruction of the Evolutionary Biogeography Reveal the Origins and Diversification of Oysters (Bivalvia: Ostreidae)." *Molecular Phylogenetics and Evolution* 164: 107268.

Li, D., C.-M. Liu, R. Luo, K. Sadakane, and T.-W. Lam. 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: 14.

Li, Y., X. He, Y. Lin, et al. 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.

Li, Y., J. L. Steenwyk, Y. Chang, et al. 2021. "A Genome-Scale Phylogeny of the Kingdom Fungi." *Current Biology* 31: 1653–1665.

Liu, X., J. D. Sigwart, and J. Sun. 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, C. R. 2017. "Five Palaeobiological Laws Needed to Understand the Evolution of the Living Biota." *Nature Ecology & Evolution* 1: 0165.

Minh, B. Q., H. A. Schmidt, O. Chernomor, et al. 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., R. Reaz, M. S. Bayzid, T. Zimmermann, M. S. Swenson, and T. Warnow. 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.

Powell, C. L. E., and F. U. Battistuzzi. 2022. "Testing Phylogenetic Stability With Variable Taxon Sampling." In *Environmental Microbial Evolution: Methods and Protocols*, edited by H. Luo, 167–188. Springer.

Price, M. N., P. S. Dehal, and A. P. Arkin. 2010. "FastTree 2—Approximately Maximum-Likelihood Trees for Large Alignments." *PLoS One* 5: e9490.

Qi, Y., Z. Zhong, X. Liu, et al. 2024. "Phylogenomic Analyses Reveal a Single Deep-Water Colonisation in Patellogastropoda." *Molecular Phylogenetics and Evolution* 190: 107968.

Ribeiro, P., M. F. Torres Jiménez, T. Andermann, A. Antonelli, C. D. Bacon, and P. Matos-Maraví. 2021. "A Bioinformatic Platform to Integrate Target Capture and Whole Genome Sequences of Various Read Depths for Phylogenomics." *Molecular Ecology* 30: 6021–6035.

Shafin, K., T. Pesout, R. Lorig-Roach, et al. 2020. "Nanopore Sequencing and the Shasta Toolkit Enable Efficient De Novo Assembly of Eleven Human Genomes." *Nature Biotechnology* 38: 1044–1053.

Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov. 2015. "BUSCO: Assessing Genome Assembly and Annotation Completeness With Single-Copy Orthologs." *Bioinformatics* 31: 3210–3212.

Sun, J., C. Chen, N. Miyamoto, et al. 2020. "The Scaly-Foot Snail Genome and Implications for the Origins of Biomineralised Armour." *Nature Communications* 11: 1657.

Sun, J., R. Li, C. Chen, J. D. Sigwart, and K. M. Kocot. 2021. "Benchmarking Oxford Nanopore Read Assemblers for High-Quality Molluscan Genomes." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 376: 20200160.

Thalén, F., C. G. Köhne, and C. Bleidorn. 2023. "Patchwork: Alignment-Based Retrieval and Concatenation of Phylogenetic Markers From Genomic Data." *Genome Biology and Evolution* 15: 227.

Turnbull, R., J. Steenwyk, S. Mutch, et al. 2023. "OrthoFlow: Phylogenomic Analysis and Diagnostics With One Command."

Wang, F., Y. Wang, X. Zeng, et al. 2024. "MIKE: An Ultrafast, Assembly-, and Alignment-Free Approach for Phylogenetic Tree Construction." *Bioinformatics* 40: 154.

Xie, P., Y. Guo, Y. Teng, W. Zhou, and Y. Yu. 2024. "GeneMiner: A Tool for Extracting Phylogenetic Markers From Next-Generation Sequencing Data." *Molecular Ecology Resources* 24: e13924.

Yang, Y., and S. A. Smith. 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, A. D., and J. P. Gillung. 2020. "Phylogenomics—Principles, Opportunities and Pitfalls of Big-Data Phylogenetics." *Systematic Entomology* 45: 225–247.

Zhang, C., R. Nielsen, and S. Mirarab. 2025. "ASTER: A Package for Large-Scale Phylogenomic Reconstructions." *Molecular Biology and Evolution* 42, no. 8: 1–4.

Zhang, L., X. Gu, C. Chen, X. He, Y. Qi, and J. Sun. 2024. "Mitogenome-Based Phylogeny of the Gastropod Order Neomphalida Points to Multiple Habitat Shifts and a Pacific Origin." *Frontiers in Marine Science* 10: 869.

Zhong, Z., Y. Lan, C. Chen, et al. 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: 485.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Figure S1:** Ostreida phylogeny by VEHoP of full-size dataset and reduced datasets, including 1G, 2G, 4G, 6G and 8G. **Figure S2:** Ostreida phylogeny by VEHoP of subsampled *Crassostrea hongkongensis* data size test. **Figure S3:** Ostreida phylogeny by ReadTree of full-size dataset and reduced datasets, including 1G, 2G, 4G, 6G and 8G. The shading of taxa in the trees indicates species whose phylogenetic placement in the corresponding topology is inconsistent with the reference topology. **Figure S4:** Ostreida phylogeny by MIKE of full-size dataset and reduced datasets, including 1G, 2G, 4G, 6G and 8G. The shading of taxa in the trees indicates species whose phylogenetic placement in the corresponding topology is inconsistent with the reference topology. **Figure S5:** Tree topologies of catfish and insect datasets from different software. **Figure S6:** Mitochondrial genome-based phylogeny of Neomphalida. **Figure S7:** Neomphalida phylogeny based on NGS data, including VEHoP (multiple models), MIKE and Read2Tree. **Figure S8:** Occupancy of matrix generated by VEHoP based on different subsampled datasets. **Figure S9:** Benchmark results for ASTER and ROADIES based on Ostreida, catfish and insects datasets. Red boxes in the trees mark incorrect clades or polytomies, indicating inconsistencies between the phylogenetic placement of taxa and the reference topology. **Table S1:** Summarises sequencing data sources and NCBI SRA accessions of all taxa to support data reproducibility. **Table S2:** Presents genome or transcriptome assembly (e.g., N50, BUSCO) and protein annotation metrics of different datasets. **Table S3:** Compares runtime (hours) of VEHoP and Read2Tree across datasets of different sizes. **Table S4:** Shows missing gene rate, gap rate, and GCF of the all datasets. **Table S5:** Calculates root-to-tip distance of different datasets. **Table S6:** Presents genome assembly (e.g., N50, BUSCO) and protein annotation metrics of Neomphalida dataset. **Table S7:** Summarises the number of NCBI sequencing datasets of non-model invertebrates for reference of public data availability. **Table S8:** Calculates RF distance between trees from different pipelines and the reference topology for fish/insect datasets to compare topological similarity.