Received: 9 June 2025 | Accepted: 22 December 2025

DOI: 10.1002/aps3.70059

Applications in Plant Sciences

SOFTWARE NOTE

HybSuite: An integrated pipeline for hybrid capture phylogenomics from reads to trees

Yu-Xuan Liu | **Zi-Jia Lu** | **Wyckliffe Omondi Omollo** | **Meng-Meng Wang** | **Li-Guo Zhang** | **Tao Xiong** | **Xue-Qin Wang** | **Yi-Ying Wang** | **Miao Sun**

National Key Laboratory for Germplasm Innovation and Utilization of Horticultural Crops, Huazhong Agricultural University, Wuhan 430070, China

Correspondence

Miao Sun, National Key Laboratory for Germplasm Innovation and Utilization of Horticultural Crops, Huazhong Agricultural University, Wuhan 430070, China.
Email: miaosun@mail.hzau.edu.cn

This article is part of the special issue "Branching out: Resolving plant evolution through phylogenetic networks."

Funding information

National Natural Science Foundation of China, Grant/Award Number: 32270222; Priority Research Programme of the National Key Laboratory for Germplasm Innovation & Utilization of Horticultural Crops, Grant/Award Number: Horti-PY-2023-003

Abstract

Premise: Hybrid capture sequencing (Hyb-Seq) is a widely used approach in phylogenomics, providing efficient access to targeted genomic regions. However, deriving high-quality phylogenetic trees from raw sequencing reads requires extensive bioinformatics processing, which increases complexity, the risk of errors, and challenges in file management, especially for users unfamiliar with bioinformatics workflows.

Methods and Results: We developed HybSuite, a streamlined Bash-based bioinformatics pipeline built upon mainstream tools such as HybPiper 2, designed to simplify the Hyb-Seq phylogenomic analysis from raw reads to species trees. Compared to existing tools (e.g., HybPiper 2, CAPTUS), it offers a modular yet integrated workflow covering all key steps from downloading from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA), adapter removal, data assembly, and paralog handling to species tree inference and extensive in-depth analysis. We validated HybSuite by reconstructing a robust phylogeny for the Elaeagnaceae family, using the Angiosperms353 probe set and a dataset of 100 single-copy nuclear loci from *Arabidopsis*.

Conclusions: HybSuite provides a flexible and user-friendly pipeline for Hyb-Seq phylogenomic analyses, and its high accuracy and efficiency were demonstrated through benchmarking with two empirical datasets. HybSuite is freely available at https://github.com/Yuxuanliu-HZAU/HybSuite. The pipeline is compatible with both the Linux and MacOS platforms.

KEYWORDS

bioinformatics pipeline, Elaeagnaceae, Hyb-Seq, paralog handling, phylogenomics

Hybrid capture (Hyb-Seq) is a sequencing strategy that integrates target enrichment with genome skimming (Weitemier et al., 2014). Compared to other next-generation sequencing (NGS) techniques, such as whole-genome sequencing (WGS), transcriptome sequencing (RNA-Seq), and restriction site–associated DNA sequencing (RAD-seq), it offers several key advantages: (1) it is a cost-effective and time-efficient method to obtain single or low-copy nuclear genes from the targeted loci and organelle genomes used for phylogenomics (Dodsworth et al., 2019); (2) numerous Hyb-Seq probe sets have been developed for different branches of the tree of life, such as Orchidaceae (Eserman et al., 2021) and Asteraceae (Mandel et al., 2014; Moore-Pollard et al., 2024), as well as universal probe sets (e.g., Angiosperms353, Johnson et al., 2019; GoFlag, Breinholt et al., 2021), enabling phylogenomic studies with expanded gene/taxa sampling at multiple taxonomic levels, from populations (Rincón-Barrado et al., 2024), to genera, families, orders, and even the entire angiosperm clade (Hendriks et al., 2021; Zuntini et al., 2021, 2024; Xu et al., 2024); and (3) the relatively low demand for high-quality DNA in Hyb-Seq enables the retrieval of genomic data from silica-dried and herbarium specimens, including those as old as 204 years (Brewer et al., 2019), thereby expanding sampling opportunities and providing a

powerful approach for reconstructing large-scale phylogenies with broader taxonomic coverage (Baker et al., 2022; Zuntini et al., 2024).

Due to these benefits, Hyb-Seq has become one of the most promising and mainstream approaches in phylogenomics (Dodsworth et al., 2019; Guo et al., 2022). However, despite its advantages and wide application, Hyb-Seq approaches require extensive bioinformatics analysis, including raw read trimming, data assembly, loci filtering, paralog identification and removal, multiple sequence alignment, alignment trimming, and phylogenetic tree inference. Although several published bioinformatics tools have been developed to streamline these processes, critical limitations persist. These available bioinformatics tools include HybPiper 2 (Johnson et al., 2016), CAPTUS (Ortiz et al., 2023), SECAPR (Andermann et al., 2018), and PHYLUCE (Faircloth, 2015), which facilitate sequence assembly and locus recovery through standardized workflows. Moreover, hybpiper-nf and paragone-nf (Jackson et al., 2023) containerized and extended the features of HybPiper while integrating four algorithms (Yang and Smith, 2014) to infer orthology groups directly from the paralog sequences outputted by HybPiper. Additionally, PhyloPyPruner (https://gitlab.com/fethalen/phylopypruner/) automates the process of direct orthology inference from alignments and gene trees via five algorithms (Kocot et al., 2013; Yang and Smith, 2014). Although these tools offer valuable functionality, they cannot complete a whole workflow from reads to trees independently or in supervised automation. HybPhyloMaker (Fér and Schmickl, 2018) is one of the few pipelines that provides a complete workflow, enabling comprehensive analyses from reads to trees and parallelization for efficiency. However, it requires sequential execution of ~10 separate scripts, making file management and parameter tracking challenging. Conversely, UnFATE (Ametrano et al., 2025) runs the complete Hyb-Seq phylogenomics workflow in a single command but lacks modular stagewise execution, increasing the risk of errors and limiting parameter flexibility. Taken together, while the currently available tools provide useful functions, their multi-step or inflexible nature can hinder reproducibility and increase the potential for errors, particularly for researchers without extensive computational experience.

Because of these limitations, a unified and streamlined framework is needed—one that integrates all essential steps of Hyb-Seq phylogenomics, allows modular stagewise execution, and ensures rigorous tracking of parameters and intermediate output files to guarantee reproducibility and usability. Thus, here we present HybSuite, a comprehensive Bash-based bioinformatics pipeline. HybSuite integrates and automates a suite of established bioinformatics tools into a unified workflow, enabling users to generate robust phylogenies directly from raw sequencing data through either single-command execution or modular stagewise operations for enhanced flexibility. The HybSuite workflow requires only a sample list, an input directory, and a target sequence file as inputs (details are discussed in the "Input files and parameters" section, below). Once the inputs are provided, users only need to specify the output directory and personalize the command line options; the desired analyses are then launched, such as different paralog-handling approaches (discussed in the section "Workflow description"), loci and sample filtering, and tree inference based on either concatenation- or coalescent-based methods. The outputs include both the final phylogenetic results and intermediate files (including, but not limited to, a paralog distribution heatmap and loci recovery, individual and concatenated sequence alignments for concatenation-based tree inference, locus-specific gene trees for coalescent-based tree inference, and other diagnostic outputs based on the user's choice). Importantly, HybSuite systematically manages the intermediate results and generates a detailed log file that records all executed commands, output file paths, and the runtime of each step, ensuring reproducibility, facilitating error tracking, and allowing users to review and control all the outputs. Additionally, HybSuite provides a dedicated subcommand to systematically retrieve key output files (e.g., the typical final species tree files, as well as the paralog heatmap exclusive to HybSuite), allowing users to quickly access essential results without manual searching.

To validate the functionality and reliability of HybSuite, we assembled two distinct nuclear genomic datasets from the oleaster family (Elaeagnaceae, Rosales, rosids) to reconstruct the phylogeny. The first dataset used the Angiosperms353 probe set (hereafter Angiosperms353; Johnson et al., 2019), and the second dataset used 100 single-copy nuclear loci from *Arabidopsis* designed for the rosid clade (hereafter Arabidopsis100; Folk et al., 2021). The results obtained from the two example datasets were highly consistent, demonstrating the accuracy and effectiveness of the pipeline. By automating the bioinformatics process and offering customizable parameters for various wrapped tools in the pipeline, HybSuite greatly simplifies phylogenomic analysis, making it accessible and reproducible to researchers at various bioinformatics levels.

METHODS AND RESULTS

Input files and parameters

The HybSuite pipeline requires three structured inputs to initiate analysis: (1) a taxon-annotated sample list, (2) a designated input directory, and (3) a target sequence file. The sample list should document taxon names and their corresponding sequence sources, which can include one or a combination of any of the following three supported types: public raw reads from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA), user-provided raw reads, and pre-assembled sequences for target loci (the format specifications are available at https://yuxuanliu-hzau.github.io/HybSuite.docs/docs/tutorial/#1-the-sample-list-file). Accordingly, the input directory must contain either user-provided raw reads, pre-assembled sequences, or both, depending on the sequence types specified for each taxon in the

sample list. Notably, pre-assembled sequences must be obtained prior to HybSuite analysis, either through third-party software like CAPTUS (Ortiz et al., 2023), SECAPR (Andermann et al., 2018), and PHYLUCE (Faircloth, 2015) or downloaded from external databases such as the Plant and Fungal Trees of Life Project (PAFTOL) (Baker et al., 2022). HybSuite offers a unique functionality of integrating these sequences into the working dataset for orthology inference and phylogenetic analyses, significantly enhancing sample inclusion capacity. Unlike locally stored data, raw reads from the NCBI SRA are automatically downloaded via the HybSuite pipeline using accession numbers provided in the sample list and thus do not need to be provided in the input directory. All sequences in the target sequence file should be formatted according to HybPiper 2 specifications (see https://github.com/mossmatters/HybPiper/wiki#12-target-file). Additionally, to perform orthology inference and phylogenetic tree inference, users must designate at least one outgroup taxon in the sample list.

Once the inputs are prepared, users can configure their desired parameters for both the extended functions of HybSuite and the settings for the other integrated software/packages. For instance, setting the minimum sample coverage for each locus in HybSuite can reduce missing data, a practice that has been commonly used in previous studies (Herrando-Moraira, 2018; Fu et al., 2022), as loci represented in only a few samples are often uninformative and can bias downstream analyses (Hosner et al., 2015). Users can also select from among seven paralog-handling methods (discussed under "Comparison of the seven paralog-handling methods," below) and choose their preferred tools for species tree inference. For concatenated phylogenetic analysis, RAxML (Stamatakis, 2014), RAxML-NG (Kozlov et al., 2019), or IQ-TREE 2 (Minh et al., 2020) can be used, while for coalescent analysis, ASTRAL-IV (Zhang et al., 2025), wASTRAL (Zhang and Mirarab, 2022), or ASTRAL-Pro3 (multi-copy-genes aware; Zhang et al., 2025) are available.

To ensure efficient pipeline execution, HybSuite incorporates an automatic detection mechanism that optimizes the number of threads by default. Users may also manually adjust this setting to optimize the computational efficiency based on available resources. Furthermore, HybSuite leverages Bash's built-in job control together with a first-in, first-out (FIFO) scheduling system to implement parallel execution, which currently runs only on a single multi-core machine and does not support cluster or grid environments. Users can specify the number of subprocesses to run concurrently, allowing flexible balancing of computational load and runtime efficiency. The main pipeline steps that can be parallelized include public data downloading, raw reads trimming, data assembly, multiple sequence alignment, alignment trimming, and gene tree inference.

Workflow description

The HybSuite pipeline integrates Bash, Python, and R scripts for statistical analysis and data visualization, but users only need to execute a few simple Bash commands. The entire HybSuite workflow consists of four stages (Figure 1), with subcommands allowing either stagewise execution or running the full pipeline via a single command. Detailed step-by-step descriptions, together with the scripts and software used, are provided in Appendix S1, and a summary of the core software dependencies required prior to running HybSuite is provided in Appendix S2.

Stage 1: NGS dataset construction

This stage involves automatically downloading raw read files from NCBI and combining them with user-provided raw data (if available), followed by adapter removal to generate a clean dataset. Specifically, if the list of taxon names and their corresponding SRA accession numbers (e.g., SRR/ERR prefixes) are provided in the sample list file, the *prefetch* command from the NCBI SRA Toolkit (Edwards, 2022) is used to automatically download the raw read files in SRA format based on the provided SRA accession numbers. The fasterq-dump command-line tool from the NCBI SRA Toolkit is then applied to convert the SRA-formatted files into FASTQ format, followed by compression into FASTQ.GZ via pigz (https://zlib.net/pigz/) to minimize storage requirements. When raw data from both the user and public repository are detected, the pipeline will automatically merge them into a designated directory prior to the quality control steps. The combined raw reads then undergo adapter removal and quality control using Trimmomatic (Bolger et al., 2014), with customizable parameters to ensure the dataset is cleaned and ready for downstream analysis.

Stage 2: Data assembly and paralog retrieval

Using the cleaned reads generated in Stage 1 as input, data assembly of each sample is initially performed via the *hybpiper assemble* command in HybPiper 2 (Johnson et al., 2016) with customizable settings. Specifically, cleaned reads are first mapped to target loci using (1) BWA (Li and Durbin, 2010) as the default nucleotide mapper or (2) BLASTX (Altschul et al., 1990) as the default protein mapper, with DIAMOND (Buchfink et al., 2014) as an alternative protein-mapping option. Prior to this step, HybSuite automatically detects the sequence type (nucleotide/protein) from the target file, requiring users only to optionally specify preferred tools (without needing to specify the sequence type). The mapped reads are then de novo–assembled into contigs using SPAdes (Bankevich et al., 2012), followed by exon–intron boundary identification and contig-to-exon alignment with Exonerate (Slater and Birney, 2005). Next, all recovered paralog sequences including putative paralogs and single-copy genes are retrieved using the *hybpiper paralog_retriever* command in HybPiper 2. To address sequence misassembly, we

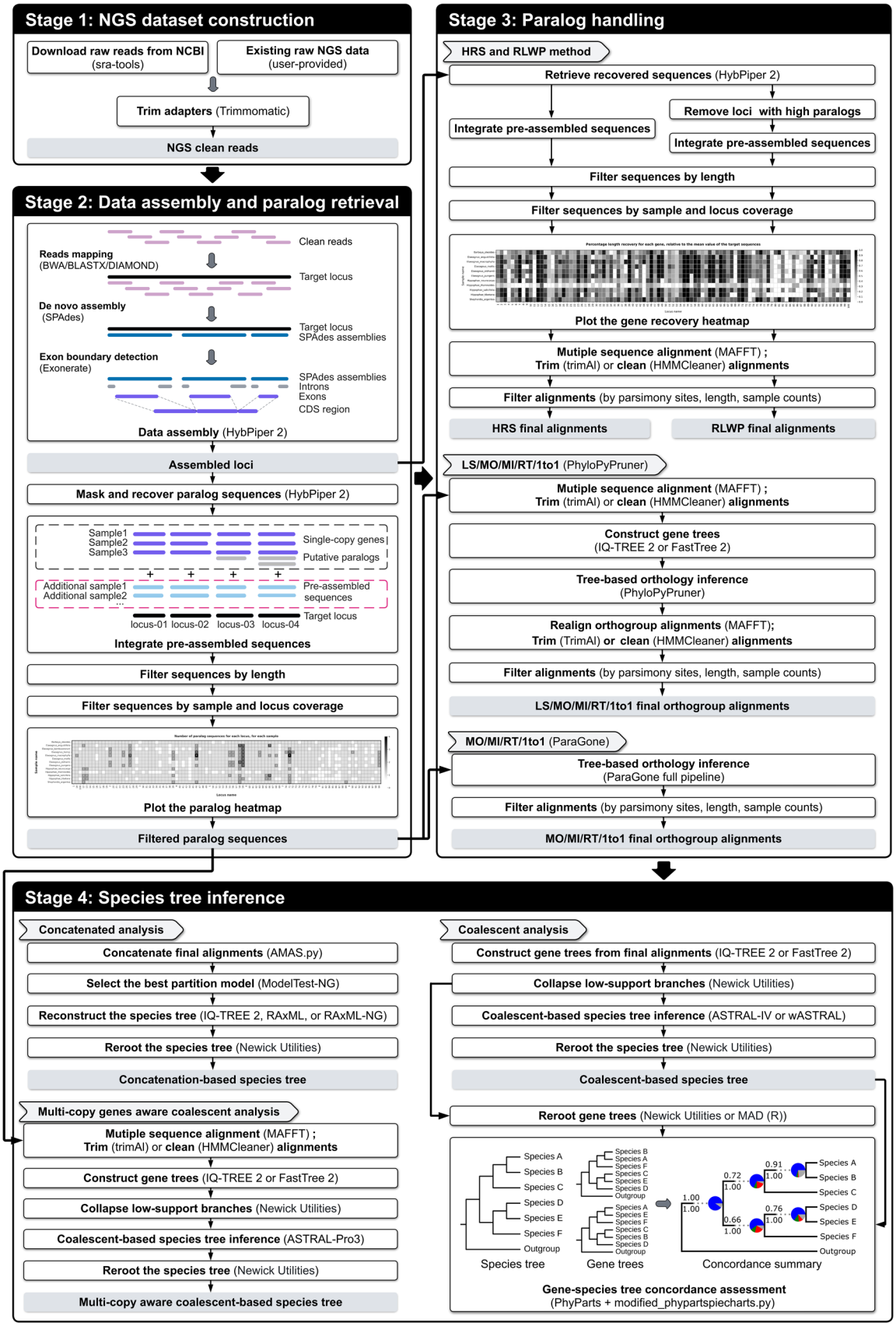

FIGURE 1 Flowchart of the HybSuite pipeline.

developed a new Python script, *filter_seqs_by_length.py*, which allows users to optionally remove sequences from FASTA files in the directory containing all recovered paralog sequences with low base pair length, low ratio relative to the mean length, or low ratio relative to the maximum length per locus in the target file (threshold values can be set by users). Subsequently, another newly developed Python script, *filter_seqs_by_sample_and_locus_coverage.py*, is applied to remove samples with low locus coverage and loci with low sample coverage, based on user-defined thresholds, thereby limiting missing data for the downstream analyses. The information regarding all removed sequences, samples, and loci is reported in the output directory. Additionally, the paralog heatmaps for both the original paralogs and final filtered paralogs are also generated by our newly developed Python script *plot_paralog_heatmap.py*. Compared to HybPiper 2's paralog heatmap plotting command *hybpiper paralog_retriever*, this script accepts a folder of FASTA files as input—broadening its applicability beyond HybPiper 2 outputs—and offers enhanced visualization options, such as customizable colors and some other optional displays, for publication-ready figures.

Notably, chimeric contigs—sequences mistakenly formed by stitching together reads from different gene variants (including paralogs)—may introduce misleading phylogenetic signals and are likely to be generated in pipelines implementing de novo assembly such as PHYLUCE, HybPiper, and SECAPR (Nauheimer et al., 2021; Joyce et al., 2025). By default, HybSuite integrates HybPiper 2's *--chimeric_stitched_contig_check* option (via the *hybpiper assemble* command) to identify and flag putative chimeric contigs during data assembly. When retrieving recovered paralog sequences, users may optionally exclude these flagged contigs from downstream analyses (default: included). Nevertheless, users are advised to inspect outputs carefully, as additional signs of chimeric assemblies can be detected through alignment cleaning, unusually long branches in gene trees, or read remapping (Joyce et al., 2025).

Stage 3: Paralog handling

This stage offers seven paralog-handling methods, from which users may choose one or more; the first of these methods, Hybpiper-Retrieved Sequences (hereafter HRS), is set as the default. The HRS method retrieves coding regions (default), introns, or supercontigs from assembled contigs using the *hybpiper retrieve_sequences* command in HybPiper 2, following the assembly phase in Stage 2, where HybPiper 2 is applied to detect contigs assembled by SPAdes that match each target locus. For any target locus with a single contig detected in a sample, the corresponding sequence is directly retrieved. For any locus with multiple contigs, the contig with the maximum coverage depth and greatest percent identity to the reference is selected, following the methods detailed in HybPiper 2 (see https://github.com/mossmatters/HybPiper/wiki/Paralogs#scenario-1-long-paralogs).

The second method, Removing Loci With high Paralog risk (hereafter RLWP), follows the strategy proposed by Villaverde et al. (2018). It removes loci containing paralogs in more than a user-defined number of samples (default: one). Based on the paralog sequences generated in Stage 2, if a locus shows putative paralogs in only one sample it is likely due to allelic variation, but if there are putative paralogs in multiple samples that locus is considered high-risk for duplication and paralogy (Villaverde et al., 2018). Thus, by excluding such high-risk loci, this method may help reduce phylogenetic noise and improve orthology confidence, particularly for taxa characterized by polyploidy or frequent gene duplications.

The remaining five methods are all tree-based orthology inference algorithms: (1) largest subtree (LS) as implemented by PhyloTreePruner (Kocot et al., 2013), which searches the gene tree from the root to tips and selects the subtree with the most non-duplicated taxa as an orthogroup, choosing the first such subtree if there are multiple equally large subtrees; (2) maximum inclusion (MI; as introduced by Yang and Smith, 2014), which searches the gene tree from the root to tips and iteratively extracts the largest subtrees without taxon duplication as orthogroups, stopping when no subtree meets the minimum taxon threshold; (3) monophyletic outgroups (MO; Yang and Smith, 2014), which selects gene trees with monophyletic outgroups from all the input trees, roots the selected trees, and iteratively cuts off the branches with fewer taxa at bifurcations where duplicated taxa are split between the two sides, stopping when no taxon duplication remains and no subtree meets the minimum taxon threshold to infer orthogroups; (4) rooted ingroups (RT; Yang and Smith, 2014), which identifies, extracts, and roots the subtree containing the largest number of ingroup taxa from the input gene tree, then iteratively cuts off the branches of the subtree in a similar way to MO to infer orthogroups; and (5) one-to-one orthologs (1to1; Yang and Smith, 2014), which retains only subtrees without duplicated taxa as orthogroups.

If users select one or more of these five methods, HybSuite will apply PhyloPyPruner 1.2.4 by default to execute the chosen algorithms. Before running PhyloPyPruner, all filtered paralog sequences generated in Stage 2 are aligned using MAFFT (Katoh and Standley, 2013). The resulting alignments are then trimmed with the block-filtering software trimAl (Capella-Gutierrez et al., 2009) by default or, alternatively, cleaned by the segment-filtering software HMMCleaner (Di Franco et al., 2019). By default, MAFFT uses the *auto* algorithm, trimAl uses the *automated1* option, and HMMCleaner runs with its default cost parameters (*-0.15, -0.08, 0.15, 0.45*). Some important parameters can be personalized, such as the running algorithm in MAFFT, the running mode of trimAl, and the cost parameters of HMMCleaner. Next, gene trees are inferred using IQ-TREE 2 by default, with the *-m MFP* option for automatic model detection and ultrafast bootstrap, or FastTree 2 (Price et al., 2010) under the general time-reversible (GTR) model with a gamma rate heterogeneity approximation (*-gamma*

option) for the balance between model realism and computational efficiency, using user-defined bootstrap replicates (default: 1000). The resulting alignments and gene trees are then used as input for PhyloPyPruner. Notably, PhyloPyPruner's implementation of the RT algorithm retains putative 1:1 orthologs during the initial step; this is a key modification from the original RT algorithm described by Yang and Smith (2014), leading to the systematic inclusion of outgroup sequences. As an alternative option, ParaGone (Jackson et al., 2023) is also integrated into our pipeline. ParaGone uses the filtered paralog files generated during Stage 2 as input and allows users to execute the MO, MI, RT, and/or 1to1 algorithms based on their preference, with more detailed intermediate outputs than possible in PhyloPyPruner.

The final outputs of Stage 3 are alignments that are prepared for both concatenated and coalescent analyses in the next stage. To generate these alignments, the outputs from all seven methods are further processed. Specifically, sequences produced by the HRS and RLWP methods initially undergo a two-step filtering procedure: first using *filter_seqs_by_length.py* to remove potentially misassembled sequences based on length-related criteria (as described in Stage 2), and then using *filter_seqs_by_sample_and_locus_coverage.py* to remove samples with low locus coverage and loci with low sample coverage, thereby minimizing missing data. Threshold values for both of these filtering steps can be defined by the user. Filtered sequences are then aligned using MAFFT and subsequently processed using trimAl (default) or HMMCleaner. Alternatively, orthogroups produced by the five algorithms in PhyloPyPruner (LS, MI, MO, RT, and 1to1) are realigned using MAFFT and processed using trimAl (default) or HMMCleaner. In contrast, orthogroups generated using ParaGone (with alternative MI, MO, RT, and 1to1 algorithms) are directly retrieved from the corresponding output directories, as these alignments have already been realigned using MAFFT and trimmed using trimAl in the ParaGone pipeline.

Once these processed alignments are generated, AMAS (Borowiec, 2016) is initiated via the *summary* function to compute and export statistics. Based on these statistics, HybSuite removes alignments that (1) lack parsimony-informative sites, (2) have sequence lengths shorter than a user-defined threshold (default: 4 bp, to ensure tree inference is possible), and (3) contain fewer taxa than a user-defined threshold (default: 5, to avoid unstable tree inference), thereby retaining the final alignments prepared for the next stage. Finally, the *summary* function in AMAS is also used to provide statistics for these final alignments, allowing users to inspect the datasets prepared for species tree inference.

Stage 4: Species tree inference

In this stage, users can construct phylogenetic trees by either a concatenation-based or coalescent-based approach, depending on the final alignments generated in Stage 3. For the concatenation-based approach, all final alignments generated via each paralog-handling method in Stage 3 are first concatenated into a supermatrix using the *concat* function from AMAS; this is exported as a partition file for downstream analyses. Next, HybSuite allows users to execute IQ-TREE 2, RAxML, or RAxML-NG under the best-fit model to infer the species tree. For IQ-TREE 2 analyses, HybSuite employs IQ-TREE 2's built-in model selection function (*-m MFP*) by default to select the best model for each gene based on the partition file generated in the previous concatenation step, but it also allows ModelTest-NG (Darriba et al., 2019) to be used as an alternative option. For RAxML-NG analyses, HybSuite employs the parsing function (*--parse* in RAxML-NG) by default to select the best model or, alternatively, uses ModelTest-NG. RAxML does not include built-in model auto-selection features, so for these analyses HybSuite automatically executes ModelTest-NG (based on the corrected Akaike information criterion score) to determine the best-fit model prior to species tree inference. Once the best model is determined, IQ-TREE 2 is executed by default using the Shimodaira–Hasegawa-like approximate likelihood ratio test (SH-aLRT) (Guindon et al., 2010) and ultrafast bootstrapping (UFBoot) (Hoang et al., 2017), with user-defined replicates (default: 1000) to estimate branch support. Similarly, RAxML is executed with 1000 rapid bootstrap replicates by default using the *-f a* option, followed by a best maximum likelihood (ML) tree search, and RAxML-NG is executed with 1000 rapid bootstrap replicates by default using the *--bs-trees* parameter. Notably, HybSuite allows users to customize various parameters regarding species tree inference in these three programs, such as specifying constraint trees and setting the number of bootstrap replicates. These features enhance flexibility and enable users to obtain robust results tailored to their specific datasets and phylogeny reconstruction needs. Finally, all trees generated are rerooted using the *nw_reroot* function in Newick Utilities (Junier and Zdobnov, 2010) to produce the final output tree files, with user-specified outgroup(s) and the *-s* option, which ensures that inner node labels are correctly interpreted as branch support values (Czech et al., 2017).

For the coalescent-based approach, individual gene trees are first inferred from orthogroup alignments generated in Stage 3, using IQ-TREE 2 (default) or FastTree 2 (alternative), following the same parameter settings described in Stage 3. Next, branches of these unrooted gene trees with support less than a threshold value (default: 0, user-definable) are collapsed using the *nw_ed* function in Newick Utilities. The resulting trees are then combined and analyzed using ASTRAL-IV or wASTRAL with the user-definable number of initial rounds of placements (*-r* in ASTRAL-IV or wASTRAL, default: 4) and number of subsampling rounds per exploration step (*-s* in ASTRAL-IV or wASTRAL, default: 4) for species tree inference. Notably, ASTRAL-IV lacks the bootstrapping feature, thus we also integrate ASTRAL-III (Zhang et al., 2018) to execute bootstrapping on the ASTRAL-IV output tree, following the

ASTRAL-IV manual (https://github.com/chaoszhang/ASTER/blob/master/tutorial/astral4.md). Finally, the species tree inferred by ASTRAL-IV or wASTRAL is rerooted via the *nw_reroot* function with the *-s* option in Newick Utilities.

In the HybSuite pipeline, after the coalescent-based species tree is generated, the user can choose whether to conduct the gene–species tree conflict assessment (by specifying *-run_phyparts* as TRUE, default: FALSE). Then, the unrooted gene trees generated in the coalescent analysis are rerooted using either the *nw_reroot* function in Newick Utilities with the *-s* option or the minimal ancestor deviation (MAD) method via an R script (Tria et al., 2017), depending on whether outgroup taxa are present in the gene trees. Next, PhyParts is employed to summarize the number of conflicting and concordant bipartitions in the coalescent-based species tree (Smith et al., 2015); the results are visualized via a Python script developed in HybSuite (*modified_phypartspiecharts.py*, available at: https://yuxuanliu-hzau.github.io/HybSuite.docs/docs/extension_tools/#6-modified_phypartspiechartspy) that enhances visualization (e.g., supporting adjusting text size, changing node label display mode, rotating internal nodes) compared to the publicly available *phypartspiecharts.py* script (https://github.com/mossmatters/phyloscripts/tree/master/phypartspiecharts).

HybSuite also supports multi-copy-genes–aware coalescent-based species tree inference using ASTRAL-Pro3 (Zhang et al., 2025), with the same settings as ASTRAL-IV and wASTRAL. If users enabled the ASTRAL-Pro3 option, HybSuite takes the filtered paralog files generated in Stage 2, aligns them with MAFFT, processes the alignments using trimAl (default) or HMMCleaner, infers gene trees using IQ-TREE 2 (default) or FastTree 2 following the same parameter settings as described in Stage 3, and then collapses branches with bootstrap support values lower than the user-defined threshold value. The resulting unrooted gene trees (which may include multi-copy genes) are then combined as input for ASTRAL-Pro3. Following the same workflows mentioned above for single-copy species tree inference, the inferred species tree is rerooted using the *nw_reroot* function from Newick Utilities with the *-s* option specified.

Output

The output files of the HybSuite workflow vary according to the analysis chosen by the user to run at each stage. In Stage 1, the outputs include the downloaded raw NGS data in FASTQ.GZ format, as well as cleaned sequence data with adapters removed. In Stage 2, the outputs consist of assembled sequence data for all samples, original and filtered paralog sequence files for each locus, related statistical files, and visualization documents (e.g., paralog heatmaps generated by *plot_paralog_heatmap.py*). In Stage 3, the outputs consist of orthogroup alignments in FASTA format inferred using user-chosen paralog-handling methods and alignment statistics. These alignments are ready for downstream species tree inference. In Stage 4, the outputs include rerooted species tree files (either concatenation-based or coalescent-based, depending on the user's choice), as well as gene-tree-conflict-related visualization documents from the coalescent-based analysis if the PhyParts option was enabled. Importantly, all intermediate and final output files from all stages are organized within a single main directory specified by the option *-output_dir*, which must be set when running HybSuite. Within this directory, files and subfolders are systematically named to facilitate easy access to both the intermediate and final results. Detailed information on the output directory structure and file contents can be found on the HybSuite Wiki page (https://yuxuanliu-hzau.github.io/HybSuite.docs/docs/output_files/).

Example datasets

Elaeagnaceae is a small family consisting of three genera (*Elaeagnus* L., *Hippophae* L., and *Shepherdia* Nutt.) and 103 species of evergreen and deciduous shrubs/trees, mainly distributed in the Northern Hemisphere (Wu et al., 2007; Sun and Lin, 2010; Jia and Bartish, 2018; Gu et al., 2024). To test the performance of the HybSuite pipeline described above, we assembled two NGS datasets of Elaeagnaceae from public repositories and one new sequence generated in this study. Both datasets comprise 10 species, covering all three accepted genera of the family, and one outgroup, *Barbeya oleoides* Schweinf., based on the Angiosperm Phylogeny Group IV (Angiosperm Phylogeny Group et al., 2016). The first dataset, Angiosperms353, was prepared using the Angiosperms353 target file (Johnson et al., 2019), and the second dataset, Arabidopsis100, was assembled using the *Arabidopsis thaliana* 100 proteins target file (Folk et al., 2021). Raw data for all samples from these two datasets were obtained from GenBank (https://www.ncbi.nlm.nih.gov/) and PAFTOL (https://treeoflife.kew.org/), except for *Elaeagnus oldhamii* Maxim., which was sequenced in this study (PRJCA047010, see the Data Availability Statement). Detailed sampling information is provided in Appendix S3.

Input parameters

For both Angiosperms353 and Arabidopsis100, we employed the *full_pipeline* subcommand in HybSuite to execute the entire workflow, from raw reads assembly directly to phylogenetic tree reconstruction. The parameters *-seqs_min_sample_coverage* and *-seqs_min_length* were set to 0.1 and 100, respectively, to remove low-quality sequences as a conservative strategy. Additionally, all seven paralog-handling methods were applied by setting the *-OI* parameter to 1234567 to generate all orthology group datasets using PhyloPyPruner, with the goal of thoroughly validating the effectiveness of the HybSuite workflow and ensuring the robustness of the phylogenetic results derived from the example datasets. Subsequently, the *-mafft_algorithm* option in HybSuite was set to linsi, enabling MAFFT to

perform alignments using the L-INS-i algorithm for higher accuracy, and the *-sp_tree* option in HybSuite was set to 14 to run IQ-TREE 2 for concatenation-based analysis and ASTRAL-IV for coalescent-based analysis. When executing concatenation-based analysis, we used the default settings of HybSuite. For the coalescent-based analysis, we set *-collapse_threshold* to 50 to collapse branches of gene trees with bootstrap support values lower than 50 and *-run_phyparts* to TRUE to enable the execution of PhyParts and *modified_phypartspiecharts.py*, in order to quantify and visualize the conflict and concordance between gene trees and the species tree.

Gene recovery and paralog distribution

The gene recovery and paralog distribution of the two example datasets are shown in Appendix S4, and the data assembly statistics are summarized in Appendix S5. Both Angiosperms353 and Arabidopsis100 showed successful locus recovery, with Angiosperms353 exhibiting higher recovery numbers (137–351 loci per sample, mean = 322) compared to Arabidopsis100 (55–96 loci per sample, mean = 82). Paralog identification revealed that 118 loci had paralogy warnings in Angiosperms353 and 19 in Arabidopsis100.

Comparison of the seven paralog-handling methods

Statistical analyses (Appendix S6) and evaluation of the final alignments (Figure 2) for seven paralog-handling methods illustrated that MI produced the highest number of final alignments, while 1to1 retained the fewest in both datasets. The other five methods (HRS, RLWP, LS, MO, RT) yielded intermediate numbers, with MO and RT performing identically across datasets. The tree-based orthology inference methods (LS, MI, MO, RT, 1to1) generally produced more compact alignments with lower gap percentages compared to HRS and RLWP (Figure 2A, C; Appendix S6). Despite these differences, all seven methods displayed similar distributions of parsimony informative sites (Figure 2B, D; Appendix S6), indicating that phylogenetic signal was largely preserved regardless of the paralog-handling approach.

Phylogenetic results

Based on the parameter configuration of the HybSuite pipeline, seven concatenation-based and seven coalescent-based trees were constructed for both Angiosperms353 and Arabidopsis100, each derived from distinct orthology

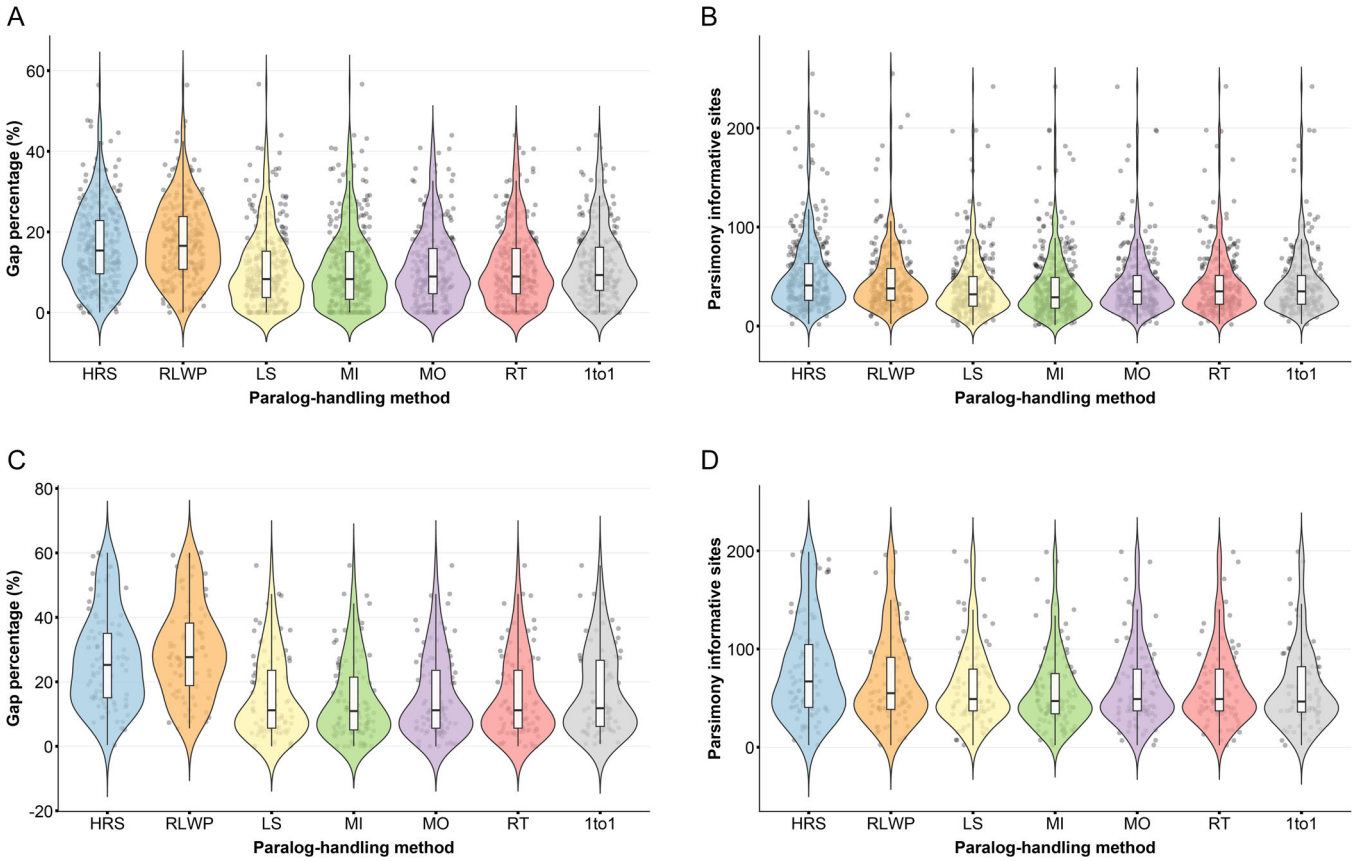

FIGURE 2 The orthogroup alignment summary for seven paralog-handling methods. (A, B) The gap percentage distribution (A) and the number of parsimony-informative sites distribution (B) of Angiosperms353 final alignments used for species tree estimation. (C, D) The gap percentage distribution (C) and the number of parsimony-informative sites distribution (D) for Arabidopsis100.

groups. To systematically assess the pipeline's performance in reconstructing robust phylogenies, the phylogenetic results were comprehensively evaluated across three key dimensions: topological variation (Angiosperms353 vs. Arabidopsis100), tree inference approach (concatenation-based method vs. coalescent-based method), and the seven paralog-handling methods.

Phylogenetic analyses yielded strong (UFBoot/SH-aLRT > 95 in IQ-TREE analyses; local posterior probability [LPP] > 0.95 in ASTRAL-IV analyses) and consistent topology support for the monophyly of the three genera in Elaeagnaceae and their intergeneric relationships across Angiosperms353 and Arabidopsis100, using both concatenation-based and coalescent-based methods (Figures 3–5). However, some topological variations were observed within the *Elaeagnus* and *Hippophae* clades when comparing different methods with the same dataset or the same method tested on different datasets.

Specifically, in concatenation-based analyses, Angiosperms353 yielded the same Elaeagnaceae phylogeny under the HRS, RLWP, and 1to1 methods, and a different but consistent topology under the LS, MI, MO, and RT methods (Figure 3: Angiosperms353). The only small topological difference was observed in a sister group within *Elaeagnus*: the HRS, RLWP, and 1to1 methods strongly support *E. pungens* + (*E. oldhamii* + *E. macrophylla*), whereas the LS, MI, MO, and RT methods support *E. oldhamii* + (*E. pungens* + *E. macrophylla*), although with significantly lower support (32.6/57, 67/75, 27.4/61, and 30.4/57, respectively; Figure 3C–F). Arabidopsis100 yielded the same Elaeagnaceae topology with strong support under all seven paralog-handling methods without exceptions (Figure 3: Arabidopsis100). However, compared to the Angiosperms353 topology, it differs in two placements: (1) within *Elaeagnus*, *E. pungens* is sister to *E. oldhamii* instead of *E. macrophylla*;

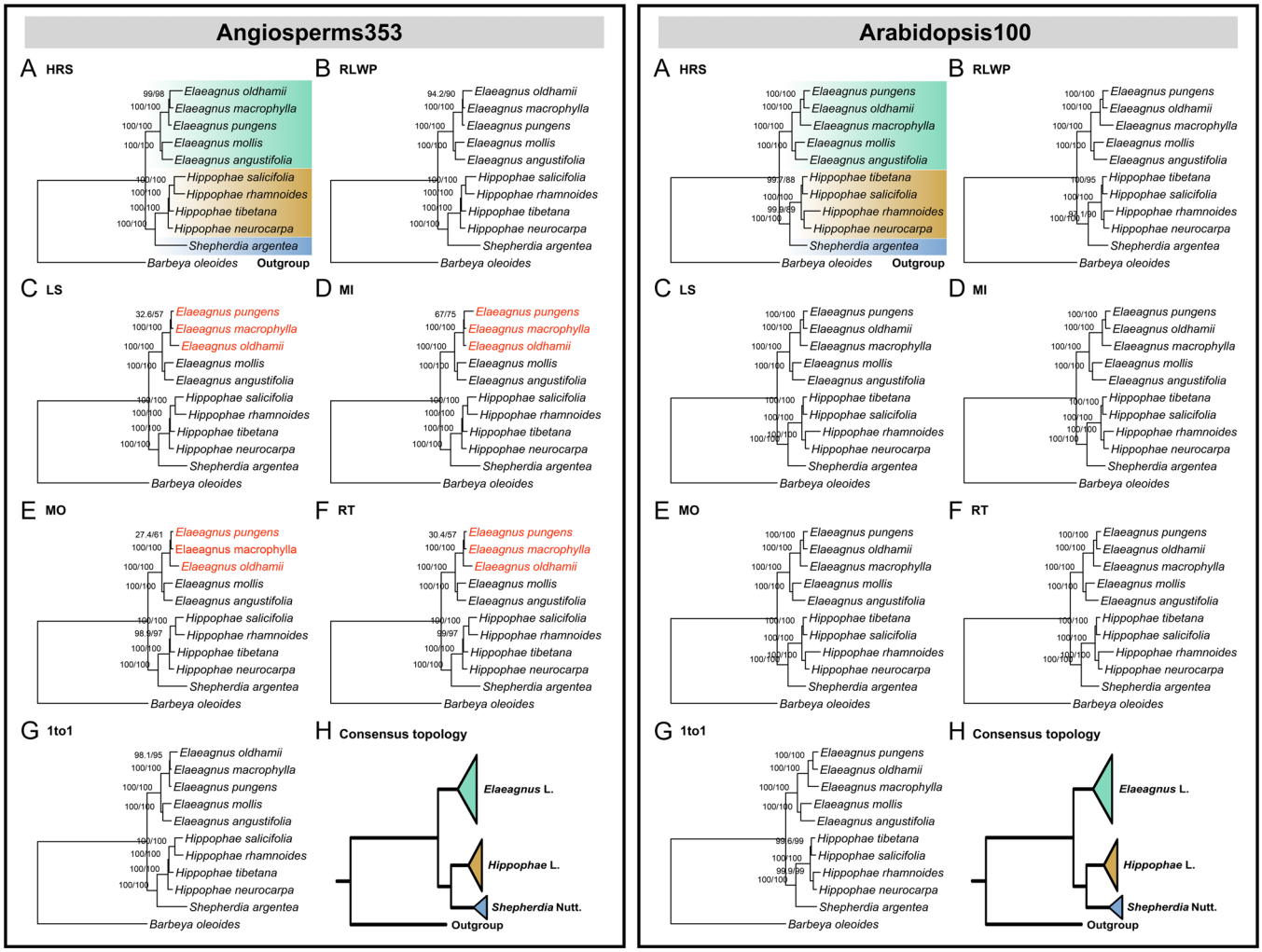

FIGURE 3 Concatenation-based phylogenetic results from Angiosperms353 (left) and Arabidopsis100 (right) using IQ-TREE 2, based on ortholog datasets inferred by seven paralog-handling methods in HybSuite: (A) Hybpiper-Retrieved Sequences (HRS), (B) Removing Loci With high Paralog risk (RLWP), (C) largest subtree (LS), (D) maximum inclusion (MI), (E) monophyletic outgroups (MO), (F) rooted ingroups (RT), (G) filtered one-to-one orthologs (1to1). (H) The consensus topology derived from all methods. Branch support values (UFBoot/SH-aLRT) are shown above the branches. Taxon names are highlighted in red when their phylogenetic positions conflict with the HRS results, while those consistent with HRS placements are in black. Each genus clade is distinguished by background colors: *Elaeagnus* in green, *Hippophae* in brown, and *Shepherdia* in blue.

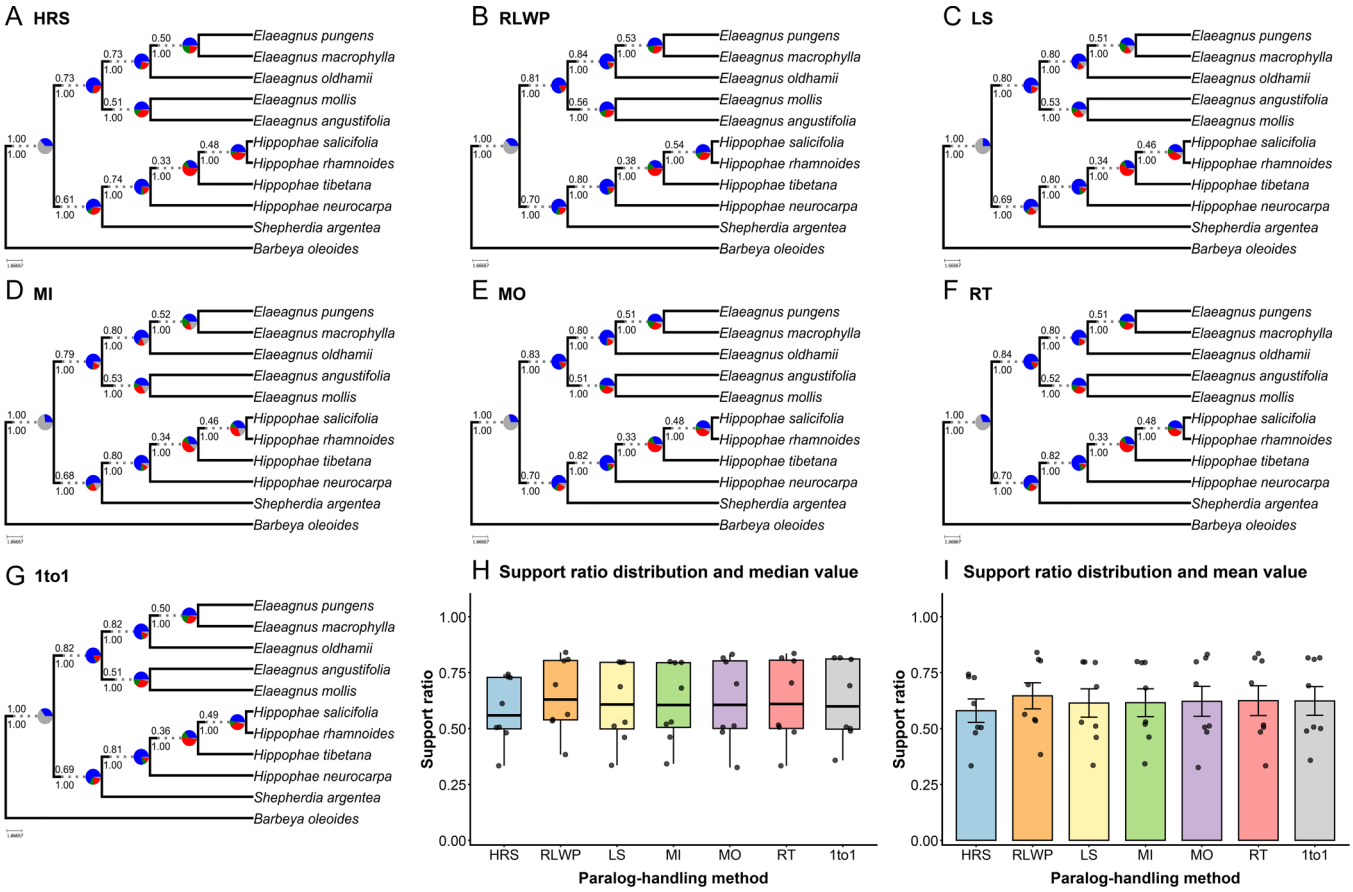

FIGURE 4 Coalescent-based phylogenetic results from Angiosperms353, using seven paralog-handling methods in HybSuite: (A) Hybpiper-Retrieved Sequences (HRS), (B) Removing Loci With high Paralog risk (RLWP), (C) largest subtree (LS), (D) maximum inclusion (MI), (E) monophyletic outgroups (MO), (F) rooted ingroups (RT), (G) filtered one-to-one orthologs (1to1). Node pie charts show gene tree congruence (blue), conflict (green/red), and uninformative data (gray). The numbers along both sides of the branch indicate congruent gene counts (above) and support values (below). (H, I) Support ratio (blue/[blue+red]) distributions across seven paralog-handling methods, showing median values (box centers) in a box plot (H) and mean values (bar tops) in a bar plot (I), where blue points represent node-level support ratios in all resulting phylogenetic trees.

(2) within *Hippophae*, *H. tibetana* is sister to *H. salicifolia* rather than *H. rhamnoides*) (Figure 3).

In coalescent-based analyses, Angiosperms353 yielded the same Elaeagnaceae phylogeny under all seven paralog-handling methods, as well as the majority topology generated by the concatenation-based method mentioned earlier (Figures 3A, B, G and 4A–G), with the one difference described above (i.e., *E. oldhamii* + [*E. pungens* + *E. macrophylla*], in agreement with the topology supported by LS, MI, MO, and RT; Figure 3C–F). Likewise, in coalescent-based analyses, Arabidopsis100 also yielded the same Elaeagnaceae phylogeny under all seven paralog-handling methods (Figure 5), in agreement with the species tree inferred from Angiosperms353 (Figure 4), except for a small group within *Elaeagnus* (i.e., *E. pungens* + [*E. oldhamii* + *E. macrophylla*]), but differed from the one inferred from the concatenation-based method (Figure 3: Arabidopsis100; Figure 5). Additionally, for coalescent-based analyses, we also calculated the median and mean gene tree support ratios (i.e., the proportion of gene trees supporting the species tree topology) for all internal nodes across trees generated by all seven paralog-handling methods (Figures 4H, I and 5H, I). For both Angiosperms353 and Arabidopsis100, the median and mean support ratios from the HRS method were generally lower than those from most of the other six paralog-handling methods, except the 1to1 method in the Arabidopsis100 dataset (Figure 5H, I). This suggests that excluding loci with high paralogs or applying tree-based orthology inference may, in many cases, alleviate conflicts between gene trees and the species tree. Nevertheless, overly stringent paralog removal, as implemented in the 1to1 algorithm in the Arabidopsis100 dataset, may also eliminate gene trees that support the species tree, thereby increasing conflicts between gene trees and the species tree.

Overall, the HybSuite pipeline demonstrates robust paralog-handling capabilities and a consistent generic phylogeny of Elaeagnaceae across both example datasets and tree inference approaches under almost all seven paralog-handling methods (Figures 3–5). However, future studies with expanded taxon sampling and/or lineage-specific probe sets are needed to further reconcile the discordance among the orthology groups as well as gene trees, given the

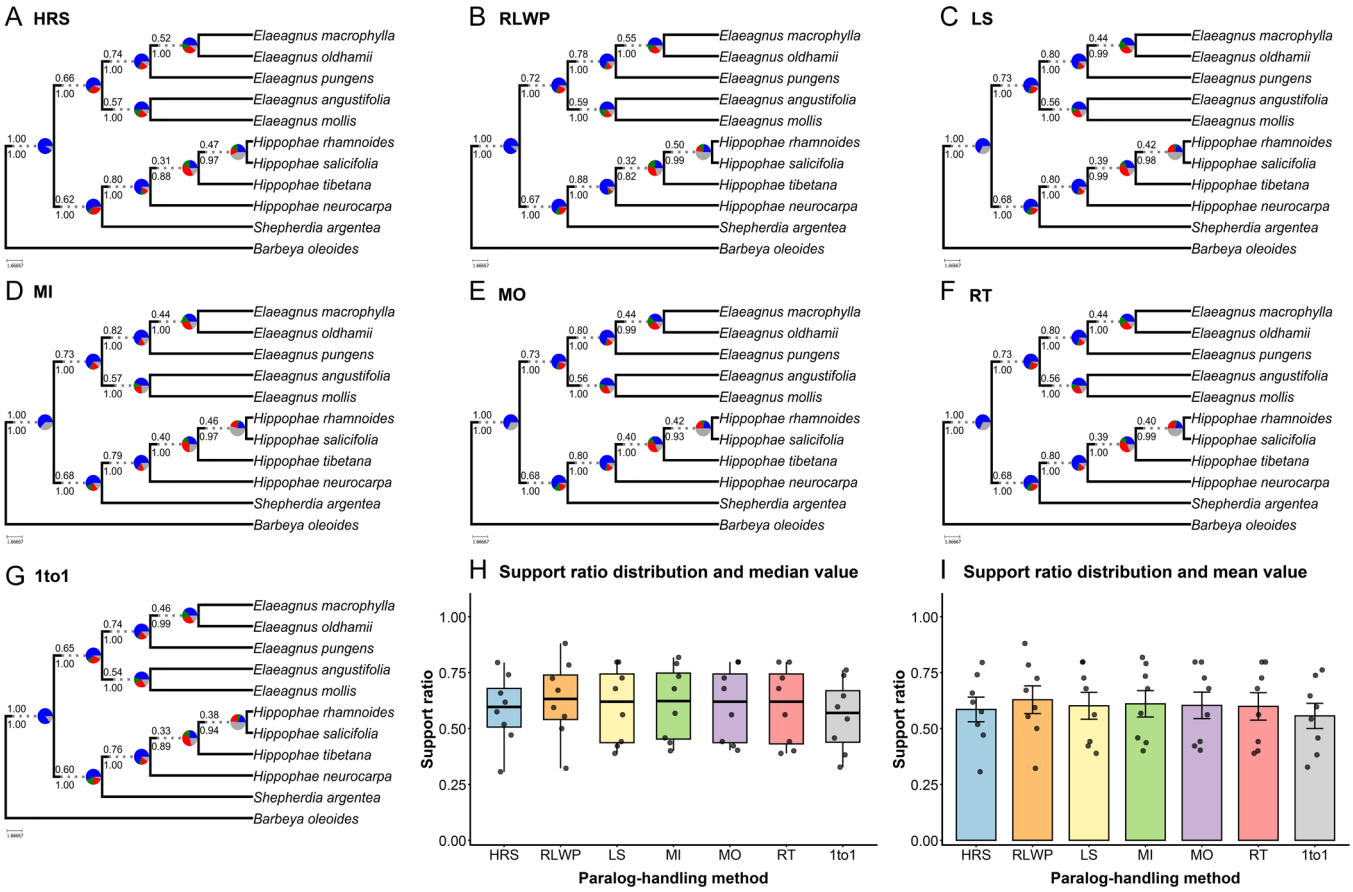

FIGURE 5 Coalescent-based phylogenetic analysis results from Arabidopsis100, using seven paralog-handling methods in HybSuite: (A) Hybpiper-Retrieved Sequences (HRS), (B) Removing Loci With high Paralog risk (RLWP), (C) largest subtree (LS), (D) maximum inclusion (MI), (E) monophyletic outgroups (MO), (F) rooted ingroups (RT), (G) filtered one-to-one orthologs (1to1). Node pie charts show gene tree congruence (blue), conflict (green/red), and uninformative data (gray). The numbers along both sides of the branch indicate congruent gene counts (above) and support values (below). (H, I) Support ratio (blue/[blue+red]) distributions across seven paralog-handling methods, showing median values (box centers) in a box plot (H) and mean values (bar tops) in a bar plot (I), where blue points represent node-level support ratios in all resulting phylogenetic trees.

topology variations observed among species in both *Elaeagnus* and *Hippophae* in our example datasets.

CONCLUSIONS

HybSuite is a Bash-based integrated pipeline designed to simplify Hyb-Seq phylogenomic analyses by consolidating bioinformatics workflows. Compared with existing Hyb-Seq analysis tools, HybSuite offers the following key advantages: (1) it provides a unified framework that streamlines the Hyb-Seq phylogenomics workflow from reads to trees into four stages—allowing users to execute each stage independently or run the entire pipeline at once (assuming proficiency in the workflow and parameter settings), rather than juggling multiple standalone scripts, thereby minimizing manual intervention and scripting errors; (2) it allows integration of pre-assembled target loci sequences into the working datasets for subsequent paralog handling and species tree inference, thereby enhancing the ability to incorporate data from diverse genomic resources (Wang et al., 2024); (3) it includes customizable filtering strategies for both loci and samples to optimize dataset quality; (4) it incorporates seven different paralog-handling methods to improve the accuracy of phylogenetic analyses; (5) it supports multiple tree inference approaches, including concatenation-based methods (IQ-TREE 2, RAxML, and RAxML-NG) and coalescent-based methods (ASTRAL-IV, wASTRAL, and ASTRAL-Pro3), accommodating varied evolutionary hypotheses and datasets; and (6) it also offers several additional features, such as automated SRA data downloading, integrated visualization of phylogenetic results, and built-in parallelization for processing multiple samples or trees. The feature comparison between HybSuite and other existing Hyb-Seq-based software is summarized in Appendix S7.

It should be noted that, while running the entire HybSuite pipeline with a single command provides convenience, executing the full pipeline without examining the intermediate results can be risky, as assembly errors or misconfigured parameters may go undetected and influence the final results. We therefore recommend that users first run

and inspect each stage independently to evaluate intermediate outputs and verify parameter settings. Once users are familiar with the pipeline's behavior and confident in its configuration, the one-command execution can then be used to maximize analytical efficiency.

Moreover, although primarily designed for Hyb-Seq reads, HybSuite can in principle accommodate other NGS data types, such as WGS, genome skimming, and RNA-Seq. This flexibility stems from its use of HybPiper 2 for sequence assembly, which has been shown to perform well on these types of data (Wang et al., 2024). However, we advise users to carefully evaluate the suitability of their input raw reads for downstream phylogenomic analyses based on factors such as gene recovery and putative paralog statistics generated after locus assembly (see "Stage 2" under the "Workflow description" section).

We evaluated HybSuite's reliability and efficiency using two Elaeagnaceae NGS datasets (Angiosperms353 and Arabidopsis100). Our example datasets generated well-supported phylogenetic relationships at the generic level consistent with previous studies (Sun et al., 2016; Gu et al., 2024), although a few topological differences were observed at the species level, particularly within *Elaeagnus*. This is not surprising given the current sampling; in this study we focused on the performance of HybSuite rather than resolving phylogenetic relationships within Elaeagnaceae. Notably, almost all of the six alternative paralog-handling methods outperformed the HRS method in providing stronger gene tree support for species tree topologies in coalescent-based analyses (Figures 4H, I and 5H, I). Moreover, the HRS method prioritizes sequences with maximum coverage depth and reference sequence identity, but it may inadvertently retain recently duplicated paralogs with high genomic copy number, allelic variants erroneously split during assembly (Nauheimer et al., 2021), or chimeric sequences. Without phylogenetic verification, these non-orthologous sequences can satisfy the selection criteria while introducing systematic errors in downstream analyses. We therefore recommend employing the other six methods prior to species tree inference, especially for coalescent-based analyses. However, given our limited sample size in the example datasets, users are advised to explore the optimal paralog-handling strategies with their own data. Additionally, we encourage researchers to evaluate different paralog-handling methods by examining the statistic table of orthogroup alignments generated in Stage 3 and the log files produced by PhyloPyPruner/ParaGone, while integrating topological support, morphological evidence, and taxonomic knowledge in their specific research context. Collectively, these findings validate HybSuite as a robust and easy-to-use phylogenomic tool while highlighting the importance of paralog handling in phylogenomic analysis.

HybSuite is freely available as a Linux/MacOS-compatible package in both Anaconda (https://anaconda.org/yuxuanliu/hybsuite) and GitHub (https://github.com/Yuxuanliu-HZAU/HybSuite) and will be maintained along with updates to integrated tools such as HybPiper and PhyloPyPruner. Researchers citing this pipeline should also cite the original publications for all the wrapped software and tools. By lowering technical barriers and standardizing best practices, HybSuite not only advances methodological reproducibility, but also empowers the broader adoption of Hyb-Seq in resolving complex phylogenetic relationships across the tree of life.

AUTHOR CONTRIBUTIONS

M.S. and Y.X.L. conceived and coordinated this project; Y.X.L. wrote the manuscript under the supervision of M.S.; Y.X.L. conducted the analyses and coding for this project; M.S. reviewed/revised the code and GitHub page; Z.J.L., M.M.W., L.G.Z., T.X., X.Q.W., and Y.Y.W. tested the HybSuite pipeline, provided feedback, and helped improve the workflow; Z.J.L. uploaded the raw data used in our example datasets to the China National Center for Bioinformation; Y.X.L., M.S., and W.O.O. revised the manuscript, together with the rest of the authors. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (grant no. 32270222 to M.S.) and the Priority Research Programme of the National Key Laboratory for Germplasm Innovation and Utilization of Horticultural Crops (Horti-PY-2023-003 to M.S.), as well as by startup funds from Huazhong Agricultural University. We thank Dr. Ryan Folk (Mississippi State University) for sharing the Arabidopsis100 reference sequences, and the members of the Sun Lab for their insightful comments, multiple rounds of tests of the HybSuite pipeline, and feedback on the manuscript.

DATA AVAILABILITY STATEMENT

The HybSuite pipeline, all extension tools, and the input files of our two example datasets are available through the HybSuite repository on GitHub (https://github.com/Yuxuanliu-HZAU/HybSuite). All raw read files used in the example datasets are available from the NCBI Sequence Read Archive under the SRA accessions, except our newly sequenced *Elaeagnus oldhamii*, which has been uploaded to the China National Center for Bioinformation (https://ngdc.cncb.ac.cn/gsub/) (BioProject ID PRJCA047010). All additional technical details of our example datasets are available on our GitHub Wiki page: https://yuxuanliu-hzau.github.io/HybSuite.docs/docs/example_dataset/.

ORCID

Wyckliffe Omondi Omollo https://orcid.org/0000-0001-9091-0742
Miao Sun https://orcid.org/0000-0001-5701-0478

REFERENCES

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215: 403–410.

Ametrano, C. G., J. Jensen, H. T. Lumbsch, and F. Grewe. 2025. UnFATE: A comprehensive probe set and bioinformatics pipeline for phylogeny

reconstruction and multilocus barcoding of filamentous Ascomycetes (Ascomycota, Pezizomycotina). *Systematic Biology* 74: 740–757.

Andermann, T., Á. Cano, A. Zizka, C. Bacon, and A. Antonelli. 2018. SECAPR—A bioinformatics pipeline for the rapid and user-friendly processing of targeted enriched Illumina sequences, from raw reads to alignments. *PeerJ* 6: e5175.

Angiosperm Phylogeny Group, M. W. Chase, M. J. M. Christenhusz, M. F. Fay, J. W. Byng, W. S. Judd, D. E. Soltis, et al. 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society* 181: 1–20.

Baker, W. J., P. Bailey, V. Barber, A. Barker, S. Bellot, D. Bishop, L. R. Botigué, et al. 2022. A comprehensive phylogenomic platform for exploring the Angiosperm Tree of Life. *Systematic Biology* 71: 301–319.

Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, et al. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* 19: 455–477.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114–2120.

Borowiec, M. L. 2016. AMAS: A fast tool for alignment manipulation and computing of summary statistics. *PeerJ* 4: e1660.

Breinholt, J. W., S. B. Carey, G. P. Tiley, E. C. Davis, L. Endara, S. F. McDaniel, L. G. Neves, et al. 2021. A target enrichment probe set for resolving the flagellate land plant tree of life. *Applications in Plant Sciences* 9: e11406.

Brewer, G. E., J. J. Clarkson, O. Maurin, A. R. Zuntini, V. Barber, S. Bellot, N. Biggs, et al. 2019. Factors affecting targeted sequencing of 353 nuclear genes from herbarium specimens spanning the diversity of angiosperms. *Frontiers in Plant Science* 10: e1102.

Buchfink, B., C. Xie, and D. H. Huson. 2014. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* 12: 59–60.

Capella-Gutierrez, S., J. M. Silla-Martinez, and T. Gabaldon. 2009. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25: 1972–1973.

Czech, L., J. Huerta-Cepas, and A. Stamatakis. 2017. A critical review on the use of support values in tree viewers and bioinformatics toolkits. *Molecular Biology and Evolution* 34: 1535–1542.

Darriba, D., D. Posada, A. M. Kozlov, A. Stamatakis, B. Morel, and T. Flouri. 2019. ModelTest-NG: A new and scalable tool for the selection of DNA and protein evolutionary models. *Molecular Biology and Evolution* 37: 291–294.

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.

Dodsworth, S., L. Pokorny, M. G. Johnson, J. T. Kim, O. Maurin, N. J. Wickett, F. Forest, and W. J. Baker. 2019. Hyb-Seq for flowering plant systematics. *Trends in Plant Science* 24: 887–891.

Edwards, D. 2022. Plant bioinformatics: Methods and protocols, 1374. Humana Press, New York, New York, USA.

Eserman, L. A., S. K. Thomas, E. E. D. Coffey, and J. H. Leebens-Mack. 2021. Target sequence capture in orchids: Developing a kit to sequence hundreds of single-copy loci. *Applications in Plant Sciences* 9: e11416.

Faircloth, B. C. 2015. PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics* 32: 786–788.

Fér, T., and R. E. Schmickl. 2018. HybPhyloMaker: Target enrichment data analysis from raw reads to species trees. *Evolutionary Bioinformatics* 14: e1176934317742613.

Folk, R. A., H. R. Kates, R. LaFrance, D. E. Soltis, P. S. Soltis, and R. P. Guralnick. 2021. High-throughput methods for efficiently building massive phylogenies from natural history collections. *Applications in Plant Sciences* 9: e11410.

Fu, X., S. Liu, R. van Velzen, G. W. Stull, Q. Tian, Y. Li, R. A. Folk, et al. 2022. Phylogenomic analysis of the hemp family (Cannabaceae) reveals deep cyto-nuclear discordance and provides new insights into generic relationships. *Journal of Systematics and Evolution* 61: 806–826.

Gu, W., T. Zhang, S.-Y. Liu, Q. Tian, C.-X. Yang, Q. Lu, X.-G. Fu, et al. 2024. Phylogenomics, reticulation, and biogeographical history of Elaeagnaceae. *Plant Diversity* 46: 683–697.

Guindon, S., J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology* 59: 307–321.

Guo, C., Y. Luo, L. Gao, T. Yi, H. Li, J. Yang, and D. Li. 2022. Phylogenomics and the flowering plant tree of life. *Journal of Integrative Plant Biology* 65: 299–323.

Hendriks, K. P., T. Mandáková, N. M. Hay, E. Ly, A. Hooft van Huysduynen, R. Tamrakar, S. K. Thomas, et al. 2021. The best of both worlds: Combining lineage-specific and universal bait sets in target-enrichment hybridization reactions. *Applications in Plant Sciences* 9: e11438.

Herrando-Moraira, S. 2018. Exploring data processing strategies in NGS target enrichment to disentangle radiations in the tribe Cardueae (Compositae). *Molecular Phylogenetics and Evolution* 128: 69–87.

Hoang, D. T., O. Chernomor, A. von Haeseler, B. Q. Minh, and L. S. Vinh. 2017. UFBoot2: Improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution* 35: 518–522.

Hosner, P. A., B. C. Faircloth, T. C. Glenn, E. L. Braun, and R. T. Kimball. 2015. Avoiding missing data biases in phylogenomic inference: An empirical study in the landfowl (Aves: Galliformes). *Molecular Biology and Evolution* 33: 1110–1125.

Jackson, C., T. McLay, and A. N. Schmidt-Lebuhn. 2023. hybpiper-nf and paragone-nf: Containerization and additional options for target capture assembly and paralog resolution. *Applications in Plant Sciences* 11: e11532.

Jia, D.-R., and I. V. Bartish. 2018. Climatic changes and orogeneses in the Late Miocene of Eurasia: The main triggers of an expansion at a continental scale? *Frontiers in Plant Science* 9: e1400.

Johnson, M. G., E. M. Gardner, Y. Liu, R. Medina, B. Goffinet, A. J. Shaw, N. J. C. Zerega, and N. J. Wickett. 2016. HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in Plant Sciences* 4: e1600016.

Johnson, M. G., L. Pokorny, S. Dodsworth, L. R. Botigué, R. S. Cowan, A. Devault, W. L. Eiserhardt, et al. 2019. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *Systematic Biology* 68: 594–606.

Joyce, E. M., A. N. Schmidt-Lebuhn, H. K. Orel, F. J. Nge, B. M. Anderson, T. A. Hammer, and T. G. B. McLay. 2025. Navigating phylogenetic conflict and evolutionary inference in plants with target-capture data. *Australian Systematic Botany* 38: SB24011.

Junier, T., and E. M. Zdobnov. 2010. The Newick utilities: High-throughput phylogenetic tree processing in the Unix shell. *Bioinformatics* 26: 1669–1670.

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.

Kocot, K. M., M. R. Citarella, L. L. Moroz, and K. M. Halanych. 2013. PhyloTreePruner: A phylogenetic tree-based approach for selection of orthologous sequences for phylogenomics. *Evolutionary Bioinformatics* 9: EBO.S12813.

Kozlov, A. M., D. Darriba, T. Flouri, B. Morel, and A. Stamatakis. 2019. RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35: 4453–4455.

Li, H., and R. Durbin. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26: 589–595.

Mandel, J. R., R. B. Dikow, V. A. Funk, R. R. Masalia, S. E. Staton, A. Kozik, R. W. Michelmore, et al. 2014. A target enrichment method for gathering phylogenetic information from hundreds of loci: An example from the Compositae. *Applications in Plant Sciences* 2: e1300085.

Minh, B. Q., H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, and R. Lanfear. 2020. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution* 37: 1530–1534.

Moore-Pollard, E. R., D. S. Jones, and J. R. Mandel. 2024. Compositae-ParaLoss-1272: A complementary sunflower-specific probe set reduces paralogs in phylogenomic analyses of complex systems. *Applications in Plant Sciences* 12: e11568.

Nauheimer, L., N. Weigner, E. Joyce, D. Crayn, C. Clarke, and K. Nargar. 2021. HybPhaser: A workflow for the detection and phasing of hybrids in target capture data sets. *Applications in Plant Sciences* 9: e11441.

Ortiz, E. M., A. Höwener, G. Shigita, M. Raza, O. Maurin, A. Zuntini, F. Forest, et al. 2023. A novel phylogenomics pipeline reveals complex pattern of reticulate evolution in Cucurbitales. *bioRxiv* [preprint]. Available at: https://doi.org/10.1101/2023.10.27.564367 [posted 31 January 2026; accessed 21 April 2026].

Price, M. N., P. S. Dehal, and A. P. Arkin. 2010. FastTree 2: Approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5: e9490.

Rincón-Barrado, M., T. Villaverde, M. F. Perez, I. Sanmartín, and R. Riina. 2024. The sweet tabaiba or there and back again: Phylogeographical history of the Macaronesian *Euphorbia balsamifera*. *Annals of Botany* 133: 883–904.

Slater, G. S. C., and E. Birney. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6: e31.

Smith, S. A., M. J. Moore, J. W. Brown, and Y. Yang. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evolutionary Biology* 15: e150.

Stamatakis, A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312–1313.

Sun, M., and Q. Lin. 2010. A revision of *Elaeagnus* L. (Elaeagnaceae) in mainland China. *Journal of Systematics and Evolution* 48: 356–390.

Sun, M., R. Naeem, J.-X. Su, Z.-Y. Cao, J. Burleigh, P. Soltis, and Z. Chen. 2016. Phylogeny of the *Rosidae*: A dense taxon sampling analysis. *Journal of Systematics and Evolution* 54: 363–391.

Tria, F. D. K., G. Landan, and T. Dagan. 2017. Phylogenetic rooting using minimal ancestor deviation. *Nature Ecology & Evolution* 1: 0193.

Villaverde, T., L. Pokorny, S. Olsson, M. Rincón-Barrado, M. G. Johnson, E. M. Gardner, N. J. Wickett, et al. 2018. Bridging the micro- and macroevolutionary levels in phylogenomics: Hyb-Seq solves relationships from populations to species and above. *New Phytologist* 220: 636–650.

Wang, X., T. Xiong, Y. Wang, X. Zhang, and M. Sun. 2024. Integrating genomic sequencing resources: An innovative perspective on recycling with universal Angiosperms353 probe sets. *Horticulture Advances* 2: e4.

Weitemier, K., S. C. K. Straub, R. C. Cronn, M. Fishbein, R. Schmickl, A. McDonnell, and A. Liston. 2014. Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences* 2: e1400042.

Wu, Z. Y., P. H. Raven, and D. Y. Hong. 2007. Flora of China. Science Press, Beijing, China, and Missouri Botanical Garden Press, St. Louis, Missouri, USA.

Xu, L., Z. Song, T. Li, Z. Jin, B. Zhang, S. Du, S. Liao, et al. 2024. New insights into the phylogeny and infrageneric taxonomy of *Saussurea* based on hybrid capture phylogenomics (Hyb-Seq). *Plant Diversity* 47: 21–33.

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.

Zhang, C., M. Rabiee, E. Sayyari, and S. Mirarab. 2018. ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* 19: 153.

Zhang, C., and S. Mirarab. 2022. Weighting by gene tree uncertainty improves accuracy of quartet-based species trees. *Molecular Biology and Evolution* 39: msac215.

Zhang, C., R. Nielsen, and S. Mirarab. 2025. ASTER: A package for large-scale phylogenomic reconstructions. *Molecular Biology and Evolution* 42: msaf172.

Zuntini, A. R., L. P. Frankel, L. Pokorny, F. Forest, and W. J. Baker. 2021. A comprehensive phylogenomic study of the monocot order Commelinales, with a new classification of Commelinaceae. *American Journal of Botany* 108: 1066–1086.

Zuntini, A. R., T. Carruthers, O. Maurin, P. C. Bailey, K. Leempoel, G. E. Brewer, N. Epitawalage, et al. 2024. Phylogenomics and the rise of the angiosperms. *Nature* 629: 843–850.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Appendix S1: Step-by-step description of the HybSuite workflow.

Appendix S2: Required software dependencies for running HybSuite.

Appendix S3: Sampling information for the two example datasets used in this study.

Appendix S4: Gene recovery and paralog distribution across the two example datasets used in this study.

Appendix S5: Summary of recovered genes from HybPiper assemblies for the two example datasets in this study, generated using the command *hybpiper stats* in HybPiper.

Appendix S6: Results of statistical analyses comparing seven paralog-handling methods applied to the two example datasets.

Appendix S7: Feature comparison between HybSuite and other Hyb-Seq-based software.

How to cite this article: Liu, Y.-X., Z.-J. Lu, W. O. Omollo, M.-M. Wang, L.-G. Zhang, T. Xiong, X.-Q. Wang, Y.-Y. Wang, and M. Sun. 2026. HybSuite: An integrated pipeline for hybrid capture phylogenomics from reads to trees. *Applications in Plant Sciences* 14(3): e70059. https://doi.org/10.1002/aps3.70059