

Research paper

Evaluating the utility of deep genome skimming for phylogenomic analyses: A case study in the species-rich genus *Rhododendron*

Zhi-Qiong Mo (莫智琼)^{a, b, c}, Chao-Nan Fu (付超男)^{b, d}, Alex D. Twyford^{e, f},
Pete M. Hollingsworth^f, Ting Zhang (张挺)^b, Jun-Bo Yang (杨俊波)^b,
De-Zhu Li (李德铢)^{b, d, g}, Lian-Ming Gao (高连明)^{a, d, g, *}

^a State Key Laboratory of Plant Diversity and Specialty Crops, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650201, Yunnan, China

^b Germplasm Bank of Wild Species & Yunnan Key Laboratory of Crop Wild Relatives Omits, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650201, Yunnan, China

^c University of Chinese Academy of Sciences, Beijing 100049, China

^d Center for Interdisciplinary Biodiversity Research & College of Forestry, Shandong Agricultural University, Tai'an 271018, Shandong, China

^e Institute of Ecology and Evolution, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3FL, Scotland, UK

^f Royal Botanic Garden Edinburgh, Edinburgh EH3 5LR, Scotland, UK

^g Lijiang Forest Biodiversity National Observation and Research Station, Kunming Institute of Botany, Chinese Academy of Sciences, Lijiang 674100, Yunnan, China

ARTICLE INFO

Article history:

Received 25 February 2025

Received in revised form

17 April 2025

Accepted 26 April 2025

Available online 2 May 2025

Keywords:

Herbarium specimens

Degraded DNA

Deep genome skimming

Target enrichment

Non-targeted exon-flanking regions

Phylogenomics

ABSTRACT

Deep genome skimming (DGS) has emerged as a promising approach to recover orthologous nuclear genes for large-scale phylogenomic analyses. However, its reliability with low DNA quality and quantity typical of archival specimens, such as herbarium material, remains largely unexplored. We used *Rhododendron* as a case study to evaluate best practices for DGS in phylogenetic analyses at both deep and shallow scales. We first investigated locus recovery variation with sequencing depth, before evaluating the phylogenetic utility of different sets of loci, including Angiosperms353, target nuclear exons, and extended exon-flanking regions. We found DGS effectively recovered nuclear genes from herbarium specimens, with $\sim 15\times$ coverage performing similarly to deeper sequencing. The recovery of target exon and flanking regions was improved by using supercontigs as a reference, offering a potential solution to limited sequencing depth. The high-integrity nuclear sequences recovered robust phylogenetic relationships within *Rhododendron*. Notably, exon-flanking regions showed significant potential for resolving relationships at shallow scales. Genes recovered with taxon-specific references had less missing data than those recovered by Angiosperms353 and generated higher-resolution phylogenetic trees. This study demonstrates the utility of DGS data for obtaining numerous nuclear genes from herbarium specimens for phylogenetic studies, and makes recommendations for best practices regarding sequencing coverage, locus selection, and bioinformatic approaches.

Copyright © 2025 Kunming Institute of Botany, Chinese Academy of Sciences. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co., Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

* Corresponding author. State Key Laboratory of Plant Diversity and Specialty Crops, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650201, Yunnan, China.

E-mail addresses: mozhiqiong@mail.kib.ac.cn (Z.-Q. Mo), fuchaonan@mail.kib.ac.cn (C.-N. Fu), Alex.Twyford@ed.ac.uk (A.D. Twyford), PHollingsworth@rbge.org.uk (P.M. Hollingsworth), zhangting@mail.kib.ac.cn (T. Zhang), jbyang@mail.kib.ac.cn (J.-B. Yang), dzl@mail.kib.ac.cn (D.-Z. Li), gaolm@mail.kib.ac.cn (L.-M. Gao).

Peer review under the responsibility of Editorial Office of Plant Diversity.

1. Introduction

Plant phylogenetics has expanded through the use of next-generation sequencing technologies and bioinformatic methods (Yu et al., 2018; Guo et al., 2023; Zhang and Ma, 2024), with genome skimming, transcriptome sequencing (RNA-seq), target enrichment, and restriction-site associated DNA sequencing (RAD-seq) proving popular for different applications (Lemmon and Lemmon, 2013; Dodsworth et al., 2019; Guo et al., 2023). However, questions remain regarding the best approach for recovering nuclear

genomic data for resolving close species-level relationships. Specifically, the acquisition of nuclear genomic data should be affordable, scalable, and applicable to preserved material represented in herbarium collections. One approach that meets these criteria is whole genome sequencing, especially if flanking regions can be reliably recovered, though this approach has rarely been tested on herbarium material (Liu et al., 2019, 2020, 2022; Su et al., 2023; Xue et al., 2023). It is now increasingly possible to not only perform low-coverage genome skimming, but to increase the amount of data generated to approximately 10–20× coverage. This deep genome skimming (DGS, also known as whole genome resequencing) enhances the potential of genome skimming to obtain hundreds of single/low-copy nuclear gene sequences (Liu et al., 2021), which may in turn improve phylogenetic inference (Liu et al., 2022; Yang et al., 2023; Yu et al., 2023; Jin et al., 2023, 2024). DGS can utilize not only fresh and silica-gel dried leaves but also effectively recover valuable data from low-quality degraded DNA in herbarium specimens (Bakker et al., 2016). However, little practical exploration has been conducted to evaluate the number and integrity of target nuclear genes recovered from herbarium specimens using DGS, and whether these data will improve phylogenetic resolution.

Various regions of the genome are potential targets for plant phylogenetics. For example, exons are particularly appealing and commonly used as a reference for target sequence assembly, as these regions are easily aligned across taxa. In addition, non-coding sequences that flank exons are thought to have significant potential for resolving shallow-scale phylogenetic and intraspecific relationships (Villaverde et al., 2018; Johnson et al., 2019; Maurin et al., 2021). However, the recovery of exon-flanking sequence is typically limited and in general this approach has been underutilized (Villaverde et al., 2018). Intron-containing sequences (i.e., supercontigs, comprising both exons and introns) as a reference may provide better recovery and allow the full assembly of target exons and their flanking regions. This would be expected to be the case especially for intergenic regions or introns that are significantly longer than the read length, resulting in a considerable reduction in missing data and a significant increase in the number of sites and the degree of sequence variability. Extending sequence recovery beyond the coding sequence into more variable flanking regions is, therefore, of particular importance for addressing shallow-scale phylogenetic relationships.

Rhododendron is an ideal genus for evaluating the utility of different genomic regions in phylogenomic analyses. It is a species-rich and taxonomically challenging genus with over 1000 species that frequently co-occur and hybridize (Chamberlain et al., 1996). Evolutionary relationships of *Rhododendron* have been recently reconstructed using plastomes (Mo et al., 2022) and various types of nuclear genomic data (Ma et al., 2022; Xia et al., 2022), with the latter providing particularly high phylogenetic resolution. Efficiently acquiring nuclear genomic data is crucial for future phylogenetic studies of this genus, especially when taxonomic sampling expands. Recent studies have resolved relationships across the whole angiosperm phylogeny (Zuntini et al., 2024), as well as in focal groups such as Celastrales (Simmons et al., 2022), Bromeliaceae (Yardeni et al., 2022), and *Puya* (Aguirre-Santoro et al., 2024) by using the Angiosperms353 gene set, comprising 353 universal low-copy nuclear genes across angiosperms (Johnson et al., 2019). However, studies have indicated conserved universal gene sets may provide poorer phylogenetic resolution than taxon-specific sequence data, especially for shallow-scale phylogenetic relationships and rapidly radiating groups (Kadlec et al., 2017; Yardeni et al., 2022; Wang et al., 2024). The recent publication of several *Rhododendron* genomes may facilitate the use of taxon-specific references, allowing DGS data to retrieve single- or low-copy nuclear loci (e.g., Zhang et al., 2017; Soza et al., 2019; Shirasawa et al.,

2021; Chang et al., 2023; Xia et al., 2024), and also raising the prospect that introns could be incorporated into the assembly process.

In this study, we investigate the utility of DGS for phylogenomic analyses in the species-rich genus *Rhododendron*. We address a range of practical concerns, including the type of material, the amount of sequencing data (target coverage), the bioinformatic approaches, and the utility of different types and combinations of loci. We first compare sequence recovery rates for both silica-gel dried material and herbarium specimens, with collection ages of approximately 40 and 80 years. We then subsampled sequencing data from a set of samples to explore the optimal minimum sequencing depth for high-quality target gene assembly from herbarium material. In addition to the common practice of using only exons as reference for read sorting in the assembly process, we also used supercontigs as reference to evaluate whether sequence recovery can be extended, particularly for non-targeted exon-flanking regions. We then assessed the phylogenetic resolution and availability of the recovered target loci and flanking regions from herbarium specimens across different sequencing depths and references. Additionally, we identified Angiosperms353-related genes from the target exons obtained here (i.e., taxon-specific sequences) and evaluated the phylogenetic resolution of these genes recovered using both taxon-specific sequences and Angiosperms353 genes as a reference. This allowed us to explore the potential benefits of the transition from Angiosperms353 to taxon-specific reference sequences.

2. Materials and methods

2.1. Sampling, DNA extraction, and sequencing

Silica-gel dried leaf samples from 13 *Rhododendron* species, and leaf material from seven herbarium specimens of five *Rhododendron* species were sampled (Table S1). These specimens were chosen to represent various intrageneric ranks, including subgenera, sections, and subsections. We used three accessions, *R. delavayi*, *R. heliopsis*, and *R. oreotrephes*, that were collected approximately 40 years ago, and four accessions of two species, *R. rubiginosum* and *R. racemosum*, that were collected around 40 and 80 years ago, to evaluate the potential of DGS for older natural history collections. These two time periods (~40 and 80 years) were chosen as at these intervals *Rhododendron* collections are abundant and well-represented in the herbarium. Silica-dried material of *Diplarche multiflora* was used as the outgroup for phylogenetic analysis.

For the silica-gel dried material, total genomic DNA was extracted using a DNeasy Plant kit (QIAGEN), and its quality and quantity were assessed with a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). Genomic sequencing libraries were prepared using the NEBNext Ultra™ DNA Library Prep Kit (NEB, MA, USA) according to the manufacturer's protocol. Libraries were sequenced with 150 bp paired-end reads on the Illumina HiSeq X Ten platform (San Diego, CA, USA) at Novogene Bioinformatics Technology Co., Ltd. (Beijing, China). The sequencing depth ranged from 10.7× to 15.7×, with an average of 12.1×, assuming an estimated genome size of ~700 Mb based on the genome of the diploid species *R. griersonianum* (677 Mb; Ma et al., 2021) and *R. delavayi* (695 Mb; Zhang et al., 2017). For the herbarium specimens, DNA extraction, library preparation, and sequencing followed the methods described by Zeng et al. (2018), with the sequencing depths ranging from 20.7× to 33.2×, with an average of 26.6× (Table S1). Despite the polyploid nature of certain species, sequencing depth was calculated as the diploid genome size.

2.2. Data processing and sequence subsampling

Adapters and low-quality sequences from raw sequencing data were trimmed and filtered using fastp v.0.21.0 (Chen et al., 2018). To explore the optimal minimum depth for herbarium specimens, a total of 21 subsets were generated using seqtk v.1.3-r106 (available from <https://github.com/lh3/seqtk>) by randomly subsampling reads with approximately 10×, 15×, and 20× coverage from the clean data files of the seven herbarium accessions (Table S1). Several plastid genomes (*R. delavayi* NC_047438, *R. simsii* MT239364, *R. molle* MZ073672, and *Vaccinium macrocarpon* NC_019616) and a mitochondrial genome (*R. simsii* NC_053763) were downloaded from GenBank (<http://www.ncbi.nlm.nih.gov/genbank/>) to construct an organellar database. Reads mapping to this organellar database were then filtered out from the sequencing data of silica-gel dried material and subsets generated from herbarium specimens using Bowtie2 v.2.3.4 (Langmead and Salzberg, 2012).

2.3. Single-copy nuclear marker development

Two methods, OrthoFinder (Emms and Kelly, 2015) and Hyb-Seq (Weitemier et al., 2014), were employed to identify single-copy orthologous genes. Genes identified by both methods, with a minimum length of 600 bp, were retained as the final reference for sequence capture. Protein-coding sequences from the available chromosome-level genomes of diploid *Vaccinium macrocarpon* (Díaz-García et al., 2021) and three *Rhododendron* species, *R. simsii* (Yang et al., 2020), *R. griersonianum* (Ma et al., 2021), and *R. ovatum* (Wang et al., 2021), were clustered into homologous gene families using OrthoFinder v.2.3.12 with default settings. Alternative splicing isoforms and duplicated or incorrect annotations often result in inaccurate estimation of single-copy orthologous genes. To address this issue, we retained orthogroups that should be classified as single copy but were initially misidentified as multiple copies, and removed orthogroups containing genes that were clustered into multiple orthogroups. Furthermore, only those single-copy orthologous genes present in all species were retained.

To identify single-copy genes with the Hyb-Seq method, we used the nuclear genome and protein-coding sequences of *Rhododendron simsii*, as the exon(s) of each gene of this species were stored as a single protein-coding sequence. The genome sequences were matched against protein-coding sequences with a minimum sequence identity threshold of 99% using BLAT v.35 (Kent, 2002), and only loci with a single match against the genome were retained. Additionally, genes with ≥ 90% similarity were removed from the remaining genes using CD-HIT-EST v.4.6.1 (Li and Godzik, 2006).

2.4. Capturing target nuclear loci and Angiosperms353-related genes in silico

We used HybPiper pipeline v.1.3.1 (Johnson et al., 2016) to capture specified single-copy nuclear genes and their flanking regions from DGS data, using default settings. This analysis was performed on the cleaned raw reads without organelle-matching reads. In this pipeline, BWA v.0.7.17 (Li and Durbin, 2009) was employed to map reads to specified exon sequences. SPAdes v.3.12.0 (Bankevich et al., 2012) was employed for *de novo* assembly of reads into contigs. Exonerate v.2.2.0 (Slater and Birney, 2005) was used to align contigs to target exons and determine exon-intron boundaries. The script *intronerate.py* extracted flanking regions from assemblies, and *retrieve_sequences.py* was used to extract the assembled target exons and flanking sequences. Furthermore, we added supercontigs (containing both exon and

intron regions) to the reference for the read sorting step in the assembly process, and specified to only use exons for the Exonerate step, instead of following common practice of using only exons as reference. Both exon and flanking sequences were also extracted in this novel analysis.

Each recovered exon and flanking region was aligned using MAFFT v.7.407 (Katoh and Standley, 2013) with the “-maxiterate 1000 -globalpair” option. TAPER v.1.0.0 (Zhang et al., 2021) was then used to identify and mask divergent outlier regions in each multiple alignment using default parameter values together with the command “-m N” to mask identified nucleotides with the “N” ambiguity code. To assess sequence recovery, we summarized the count and integrity of target exons and flanking regions for each accession post TAPER processing, excluding “N” nucleotides. The efficiency of sequence recovery was visualized using the R script *gene_recovery_heatmap_ggplot.R* included in the HybPiper pipeline.

OrthoFinder was employed to identify the *Rhododendron*-specific orthologs of the Angiosperms353 gene set. If genes were previously identified by both OrthoFinder and Hyb-Seq in *R. simsii* (above), then these were retained as references for the HybPiper assembly. This gene set recovered from silica-gel dried material and herbarium specimens with 20× coverage was designated as the ‘TaxonSpecific-1’ dataset. Additionally, the Angiosperms353 gene set was also used as a reference for target gene assembly, and the results were designated as the ‘Angiosperms353-1’ dataset. Genes in these two datasets were aligned using MAFFT and filtered with TAPER.

2.5. Nuclear sequence quality-control and phylogenetic analyses

We performed a series of filtering steps on the assembled sequences of target loci from *Rhododendron* prior to phylogenetic analysis. Initially, we retained exons and exon-flanking regions from the reference that fell within a length range of 900–3000 bp, resulting in 3988 exons and 1257 flanking regions. Recovered sequences shorter than half of the reference length were excluded, and the outgroup sequence was removed from the TAPER outputs to avoid potential artifacts caused by its long branch length. Alignments were conducted using MAFFT, followed by gene tree construction using RAXML v.8.2.12 (Stamatakis, 2014) with a GTR + Γ substitution model and 500 bootstrap (BS) replicates. To address assembly errors and paralogs, we applied scripts *trim_tips.py* and *cut_long_internal_branches.py* (available from https://bitbucket.org/yanglab/phylogenomic_dataset_construction/src/master/scripts/) from Yang and Smith (2014) to trim spurious terminal branches and cut deep paralogs in gene trees. Based on the trimmed trees, we combined sequences from the subsetted sequence reads from the herbarium specimens at 10×, 15×, and 20× depths separately with the silica-gel dried material to create 12 datasets, with the outgroup sequence added. Exons or flanking regions were removed from these datasets if they contained fewer than 19 accessions (< 90%; 19/21). We adopted a custom naming convention (E/S) (10/15/20)×-(exon/flank) for these 12 datasets, where components in parentheses indicate the reference sequence used (exon or supercontig), the depth of sequencing data for the herbarium specimens (10×, 15×, or 20×), and the type of assemblies (exon or flanking region). Finally, each exon and its flanking region were aligned using MAFFT and trimmed using trimAl v.1.4 (Capella-Gutierrez et al., 2009) with “-automated1” parameter.

For the Angiosperms353-1 and TaxonSpecific-1 datasets, sequences shorter than half of the reference length were removed from the TAPER results, and genes with more than 50% of total accessions (i.e., > 10 accessions) were retained. A MAFFT analysis

was re-run for each remaining gene, and the aligned matrix was trimmed using trimAl.

We employed both coalescent- and concatenation-based methods to construct phylogenetic trees for these 14 nuclear datasets. For the coalescent-based phylogeny, a ML tree based on each trimmed matrix was generated using RAxML with a GTR + Γ substitution model and 500 BS replicates. Gene trees for each dataset were used to infer a species tree with ASTRAL-III v.5.7.8 (Zhang et al., 2018). For the concatenation-based analysis, AMAS (Borowiec, 2016) was used to concatenate the trimmed matrices of each dataset, and IQ-TREE v.2.2.2.6 (Nguyen et al., 2015) was used to construct a ML tree for each concatenated supermatrix with the best-fitting model selected by ModelFinder (Kalyaanamoorthy et al., 2017) and 1000 ultrafast bootstrap (UFBS) replicates (Hoang et al., 2018).

PhyParts v.0.0.1 (Smith et al., 2015) was employed to calculate the amount of conflict and concordance by comparing gene trees against the corresponding species tree for the dataset S15 $\times$ -exon. Gene trees containing the outgroup (*Diplarche multiflora*) and/or the early-diverging *Rhododendron* species (*R. redowskianum*) were retained for PhyParts analysis and rooted using the nw_reroot program in Newick utilities (Junier and Zdobnov, 2010). Ten gene trees that lacked these two species were excluded from the analysis. Branches in gene trees with BS $\leq$ 50% were considered uninformative. PhyParts results were summarized and visualized with phyPartspiecharts.py (available from <https://github.com/mossmatters/MJPythonNotebooks/>).

3. Results

3.1. Single-copy nuclear loci

OrthoFinder initially identified 4566 nuclear protein-coding loci as single-copy orthologous gene clusters. An additional 2085 clusters initially identified as multi-copy were reclassified as single copy, likely due to multiple splicing isoforms, duplicated annotations, and/or incorrect annotations. Of the 6651 clusters, 82 were removed because sequences from the same gene were clustered into two or more gene families, and 636 were excluded due to not meeting the minimum length threshold of 600 bp. This resulted in 5933 genes remaining from the OrthoFinder output. Of these, 5358 genes overlapped with the Hyb-Seq results and were retained as references for assembly (Table S2).

3.2. Capturing single-copy nuclear orthologous genes

Across samples, the number of target exons and their flanking regions recovered with 15 $\times$ and 20 $\times$ depth was higher than those with 10 $\times$ depth (Figs. 1 and 2). For silica-gel dried material, the combined numbers of exons and flanking regions was no less than 4763 of the total 5358 loci (exceeding 88% recovery), regardless of whether exons or supercontigs were used as a reference (Fig. 2; Tables S1 and S3–S6). Herbarium specimen subsampled to 10 $\times$ depth led to the recovery of both exons and flanking regions between 74% and 96%. At 15 $\times$ depth, the recovery of both significantly increased, surpassing 97% ($\geq$ 5201). At 20 $\times$ depth, both exons and flanking regions exceeded 99% recovery ($\geq$ 5335) (Fig. 2; Tables S1 and S3–S6).

The integrity of recovered sequences, especially flanking regions, improved significantly using supercontigs as a reference compared to using exons as a reference (Fig. 3). When using exons as reference, recovered exon integrity was superior to that of flanking regions, and both increased with sequencing depth in herbarium specimens, albeit with a small increase in flanking regions (Fig. 3). Most exons (> 75%) recovered from silica-gel dried

materials exceeded 50% of the reference length, with certain genes nearly as long as the reference, whereas most flanking regions were less than 50% of the reference length (Fig. 4A). For the same specimen accession, exons recovered from deeper sequencing depth (15 $\times$ and 20 $\times$) were closer in length to the reference and significantly longer than those from 10 $\times$ depth (Fig. 4B). Although the integrity of flanking regions increased with sequencing depth, most flanking regions from 10 $\times$ depth were less than 50% of the reference length, and from 15 $\times$ and 20 $\times$ depths were still less than 75% of the reference length (Fig. 4B).

Using supercontigs as a reference, most exons and flanking regions recovered from silica-gel dried materials and most herbarium specimen subsets at 15 $\times$ and 20 $\times$ depths were over 75% or close to the reference length (Fig. 4). At 10 $\times$ depth, recovery was also improved compared to using exons as a reference, but only a few accessions had sequences close to the reference length, with certain accessions showing low integrity in a large part of the flanking region (Fig. 4). For the same specimen accession, the integrity of flanking regions increased with sequencing depth increase. The flanking regions of herbarium specimens 80 years old were well recovered but shorter than those from specimens 40 years old (Fig. 4B).

The number of exon alignments passing the filtering process, conducted prior to phylogenetic analysis, increased with sequencing depth for herbarium specimens (Table 1). Notably, using supercontigs as a reference produced significantly more filtered exon alignments than using exons as a reference. Specifically, the S20 $\times$ -exon dataset yielded 1.7 times more alignments than did the E20 $\times$ -exon (3547 vs. 2053), the S15 $\times$ -exon dataset yielded 1.9 times more than did the E15 $\times$ -exon (3391 vs. 1751), and the S10 $\times$ -exon dataset yielded 2.6 more alignments than did the E10 $\times$ -exon (1619 vs. 614). When using exons as a reference, only two recovered flanking regions met the filtering criteria in dataset E10 $\times$ -flank, even though higher sequencing depths of 15 $\times$ and 20 $\times$ (datasets E15 $\times$ -flank and E20 $\times$ -flank) yielded 36 and 62 flanking regions, respectively. In contrast, when using supercontigs as a reference, 183 flanking regions met the filtering criteria in dataset S10 $\times$ -flank, and a much larger number in datasets with herbarium specimens at sequencing depths of 15 $\times$ and 20 $\times$, accounting for 42% (528/1257) and 47% (591/1257) of the total, respectively.

A total of 255 genes from the Angiosperms353 gene set were identified from the 5358 single-copy *Rhododendron* genes, of which 77.3% (197/255) were retained in the TaxonSpecific-1 dataset for subsequent phylogenetic inference. Additionally, using the Angiosperms353 gene set as a reference, only 40.2% (142/353) were retained in the Angiosperms353-1 dataset.

3.3. Phylogeny inferences

Datasets using supercontigs as a reference exhibited higher phylogenetic resolution than those using exons as reference (Figs. 5 and S1). Coalescent phylogenetic trees constructed from all six datasets using supercontigs as a reference (i.e., datasets S10 $\times$ -exon, S15 $\times$ -exon, S20 $\times$ -exon, S10 $\times$ -flank, S15 $\times$ -flank, and S20 $\times$ -flank) showed a consistent topology with high local posterior probabilities (LPP) across all nodes (Figs. 5 and S1), aligning with a previous nuclear gene-based phylogeny (Xia et al., 2022). Among these datasets, four showed the identical topology with the maximum UFBS from IQ-TREE. However, datasets S10 $\times$ -exon and S10 $\times$ -flank displayed different topologies due to the position changes of *R. heliolepis* and *R. oreotrephes*. When using exons as a reference, phylogenetic trees constructed with ASTRAL and IQ-TREE from datasets E10 $\times$ -exon, E15 $\times$ -exon, and E20 $\times$ -exon all exhibited the same topology (Fig. 5), but datasets E10 $\times$ -flank, E15 $\times$ -flank, and E20 $\times$ -flank exhibited multiple conflicting relationships with low

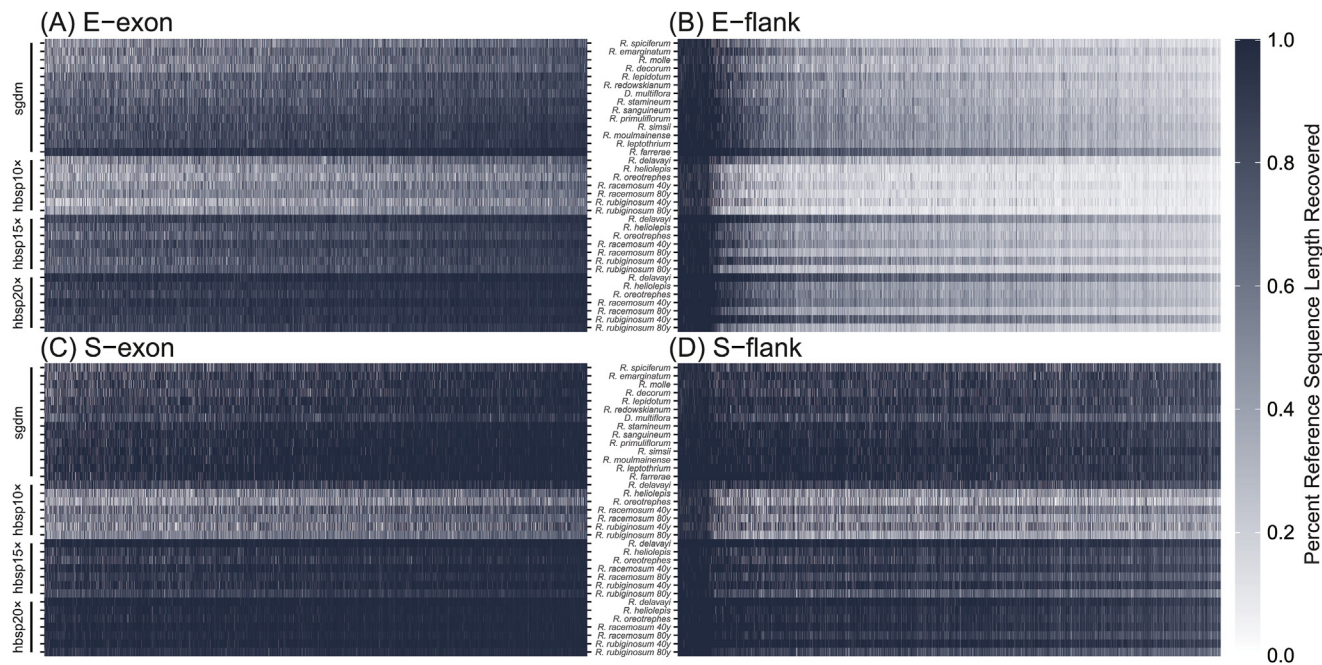


Fig. 1. Sequence recovery efficiency for 5358 genes using HybPiper. E-exon (A) and E-flank (B) represent the exons and flanking regions recovered using exon as reference, respectively. S-exon (C) and S-flank (D) represent the exons and flanking regions recovered using intron-containing sequences (supercontigs) as reference, respectively. Each column corresponds to a gene, and each row corresponds to a sample. The shading in each cell indicates the proportion of the recovered sequence length relative to the reference length, with 1.0 representing complete recovery. sgdm represents silica-gel dried material and hbsp represents herbarium specimen. The labels 10×, 15× and 20× represent the sequencing depths. Full data are provided in Tables S3–S6.

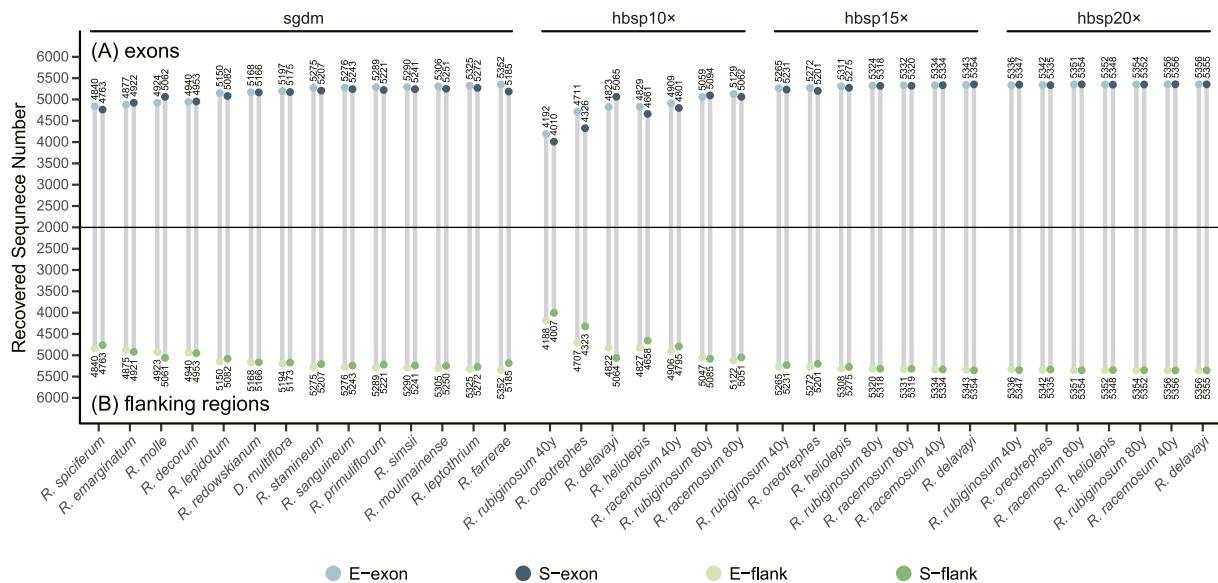


Fig. 2. Number of recovered exons (A) and exon-flanking regions (B) of 5358 genes. sgdm represents silica-gel dried material and hbsp represents herbarium specimen. The labels 10×, 15× and 20× represent the sequencing depths. E/S-exon/flank indicates the exons/flanking regions recovered using exons/supercontigs as reference for read sorting.

phylogenetic resolution (Fig. S1), likely due to a limited number of informative sites. The TaxonSpecific-1 dataset showed higher phylogenetic resolution compared to the Angiosperms353-1 dataset (Fig. S2). For TaxonSpecific-1, phylogenetic trees had relatively high LPP values, with three accessions (*Rhododendron leptothrium*, *R. rubiginosum* 40y, and *R. spiciferum*) showing different phylogenetic placements. However, certain samples or nodes (e.g., *R. heliolepis*,

R. spiciferum, *R. leptodotum*, *R. molle*, *R. oreotrephes*, and *R. rubiginosum* 40y) in the trees from dataset Angiosperms353-1 were unresolved. Phyparts analysis of the 3381 exon gene trees revealed that most topologies were conflicting (red) or uninformative (grey) compared to the internal nodes of the species tree, except for the nodes of *R. sect. Choniastrum*, subgenera *Hymenanthus* and *Tsutsusi*, which showed more congruence (blue) (Fig. S3).

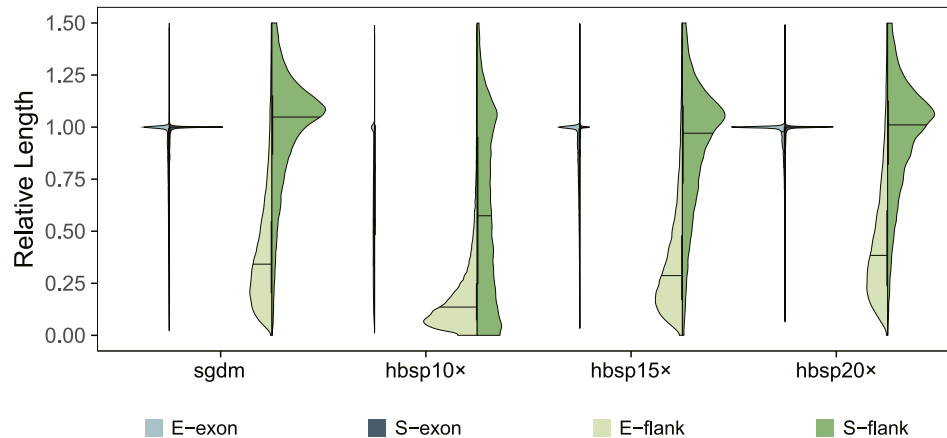


Fig. 3. Relative lengths of sequences recovered from silica-gel dried materials and herbarium specimens at sequencing depths of 10 $\times$, 15 $\times$ and 20 $\times$. sgdm represents silica-gel dried material and hbsp represents herbarium specimen. E/S-exon/flank represents the exons/flanking regions recovered using exons/supercontigs as reference for read sorting.

4. Discussion

4.1. DGS is an effective strategy for obtaining nuclear loci from herbarium specimens

Phylogenomics studies have demonstrated that genomic data, particularly extensive nuclear orthologous gene data, can reveal robust and reliable phylogenetic relationships, especially for taxa that have undergone recent rapid radiation (Xia et al., 2022; Yu et al., 2023; Chen et al., 2025). This highlights the benefit of recovering nuclear sequences from DGS data for phylogenetic research (Liu et al., 2021, 2022; Yu et al., 2023). Recent efforts have increasingly aimed at utilizing plant nuclear genetic resources from DGS data (Vargas et al., 2019; Liu et al., 2021). Consistent with this, Wang et al. (2024) reported that increased sequencing depth generally leads to better recovery of Angiosperms353 loci across genome skimming datasets, reinforcing the importance of adequate sequencing coverage. Similarly, Liu et al. (2021) suggested that a minimum of 10 $\times$ coverage is optimal for recovering high-quality single-copy nuclear genes. Our assembly results from silica-gel dried materials (10.7–15.7 $\times$, mean coverage 12.1 $\times$) corroborated this recommendation, as evidenced by the large number and high integrity of recovered target loci; however, further coverage beyond 10 $\times$ can still be beneficial to recover the most complete loci (Figs. 1A, 2, 3 and 4A).

We demonstrated that DGS can also efficiently recover many nuclear loci from herbarium specimen, comparable to those obtained from silica-gel dried material. While DGS has been shown to be useful for low-quality DNA materials (Bakker et al., 2016; Zeng et al., 2018), such as herbarium specimens up to 146 years old, these studies primarily focused on obtaining organelle genomes and nrDNA sequences without evaluating the availability of nuclear orthologous genes. Unveiling the extensive genomic information preserved in herbarium specimens could create unprecedented opportunities for biological research, including studies of phylogeny, domestication, and population genomics (Staats et al., 2013; Burbano and Gutaker, 2023). Using herbarium collections can substantially reduce the necessity for fieldwork, thereby increasing sample availability. More importantly, this facilitates historical studies comparing modern samples to historical ones, allowing for an examination of evolutionary changes over time.

Choosing the optimal sequencing depth can be challenging, and this is particularly difficult for historical samples where the extent of DNA degradation is unknown. A sequencing depth of 15 $\times$ performed similarly to 20 $\times$ in term of the recovered sequence integrity

for herbarium specimens (Figs. 1–4), suggesting high coverage might be not routinely required. However, 15 $\times$ outperformed 10 $\times$ depth by yielding more and longer target loci, and approached the performance of silica-gel dried material. The phylogenetic topologies constructed from the 15 $\times$ data of herbarium specimens was consistent with those from the 20 $\times$ data (Fig. 5F, H, J, and L). Although the herbarium specimen subsets with 10 $\times$ data recovered the same topology, certain nodes showed low support values (Fig. 5B and D). Therefore, we propose 15 $\times$ as a rule of thumb for the optimal minimum sequencing depth for herbarium specimens; however, further work must be done to identify if different threshold coverages are required under certain suboptimal conditions (e.g., age, method of drying, preservation condition, or presence of contamination).

4.2. Improved sequence recovery of exons and their flanking regions

In this study, we found notable improvements in the integrity of exons and their flanking regions, especially the latter, when using intron-containing sequences (supercontigs) as references during the assembly process, compared to the common practice of using exons alone (Figs. 1, 3 and 4). Exon-flanking regions are useful for resolving close-species level phylogenetic relationships, even at the intraspecific level (Johnson et al., 2019). Improved reference information provides possibilities for assembling longer non-coding regions and capturing various combinations of single- or low-copy nuclear genes (e.g., with or without introns) for phylogenetic inference (Liu et al., 2021). However, current studies use only a small portion of the flanking region adjacent to target exons (Villaverde et al., 2018), which is not surprising given the challenge of obtaining them from contigs assembled using short reads (e.g., 150 bp pair-end reads) mapped to target exons.

When using exons as a reference, the recovered exons exhibited high integrity and closely matched the reference length, even in the herbarium specimen (Figs. 1, 3 and 4), making them sufficient for high-resolution phylogenetic analysis (e.g., Liu et al., 2021; Liu et al., 2022; Yu et al., 2023). The high integrity of exons retrieved at higher coverage (15 $\times$ and 20 $\times$ depths) indicates that increasing sequencing depth can compensate for the effects of degraded DNA fragmentation on sequence recovery. Notably, a substantial proportion of exons and flanking regions recovered with 10 $\times$ sequencing depth from herbarium specimens exhibited low integrity (Figs. 3 and 4). Nevertheless, integrity can be improved when supercontigs are used as a reference. This suggests that using supercontigs as a reference can effectively alleviate the issue of

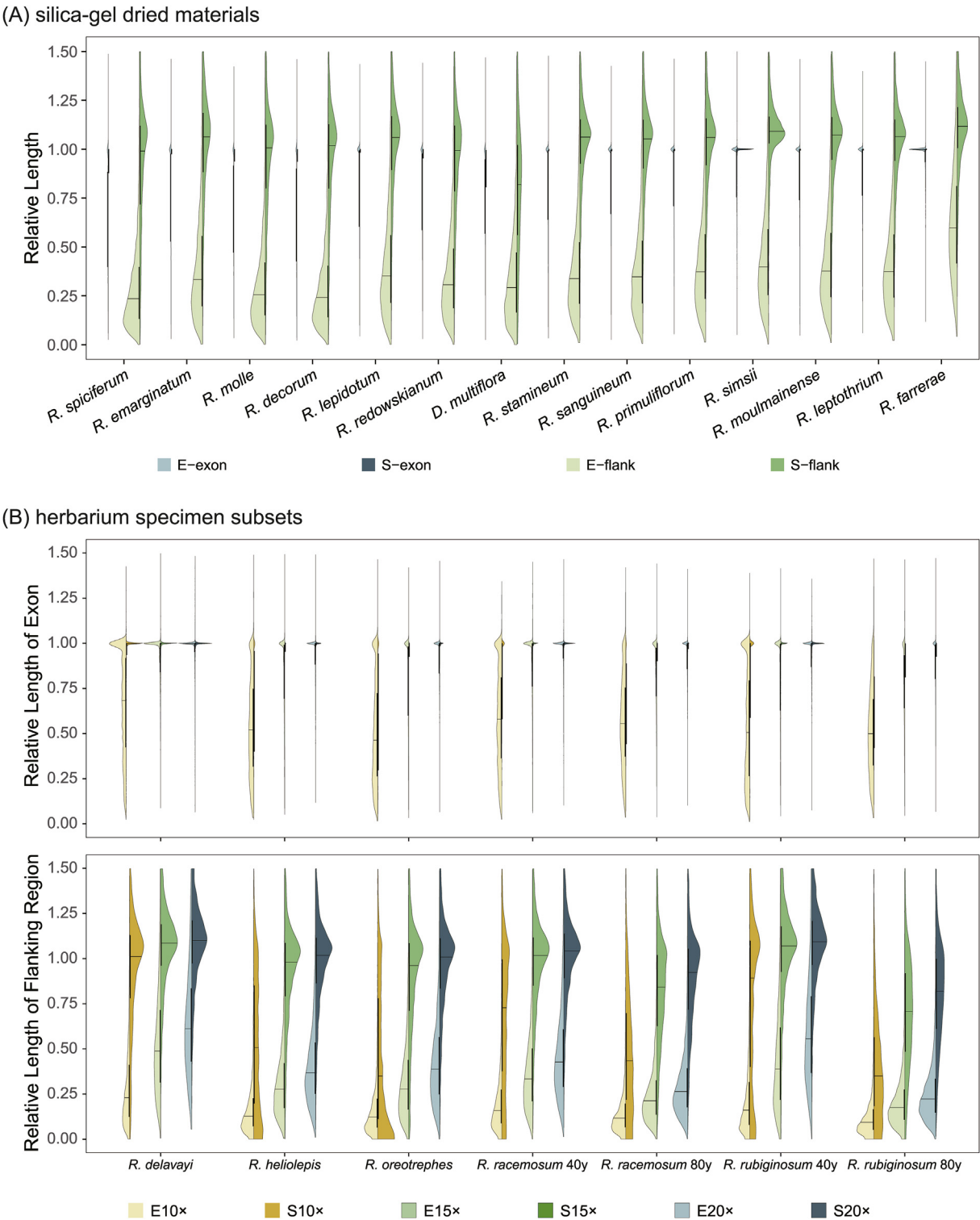


Fig. 4. Relative lengths of sequences recovered from all silica-gel dried materials (A) and herbarium specimen subsets (B). E/S-exon/flank represents the exons/flanking regions recovered using exons/supercontigs as reference for read sorting. Combinations of E/S and 10/15/20× represent the recovered sequence from herbarium specimens with sequencing depths of 10×, 15× and 20×, using exons/supercontigs as reference for read sorting.

limited recovery, even with lower sequencing depth. The growing availability of plant genomes and whole genome sequencing data makes it feasible and promising to use these intron-containing sequences as references for target loci assembly, even with only ~15× sequencing data from degraded DNA.

The assembly results from herbarium specimens at 15× and 20× depths (Fig. 4B) indicate that the age of herbarium specimens in this study had little influence on the acquisition of exon sequences. However, the integrity of flanking regions varied with specimen age at the same depth. Herbarium specimens

Table 1
Summary information of 12 datasets obtained for *Rhododendron* from silica-gel dried material and herbarium specimen subset to different sequencing depths, including total loci number, trimmed multiple sequence alignment (MSA) length, and parsimony informative (PI) and variable sites number.

Dataset	No. of loci	Trimmed MSA length (bp)	No. of PI sites	No. of variable sites
E10×-exon	614	1,074,167	52,781	171,134
S10×-exon	1619	2,760,236	124,374	412,146
E10×-flank	2	3852	257	1183
S10×-flank	183	419,221	30,828	113,893
E15×-exon	1751	2,906,046	147,201	455,663
S15×-exon	3391	5,429,981	252,260	810,689
E15×-flank	36	67,663	4776	18,449
S15×-flank	528	1,223,350	98,230	341,419
E20×-exon	2053	3,360,854	168,074	519,520
S20×-exon	3547	5,639,252	263,481	839,930
E20×-flank	62	105,794	7458	28,052
S20×-flank	591	1,378,150	113,041	387,055

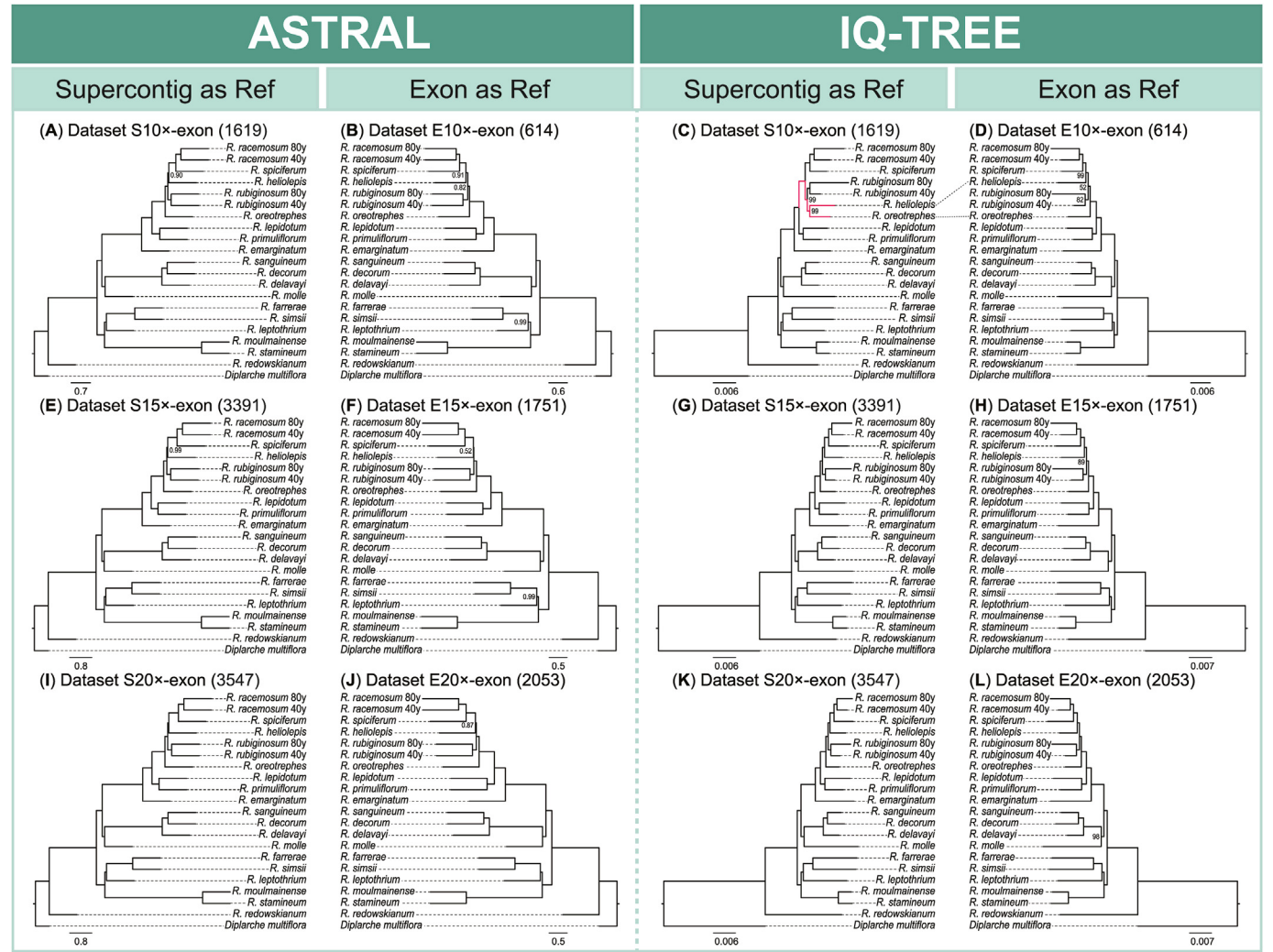


Fig. 5. Phylogenetic trees inferred using ASTRAL and IQ-TREE from exon datasets E10×-exon, S10×-exon, E15×-exon, S15×-exon, E20×-exon and S20×-exon (A–L). Numbers in parentheses represent the gene count in the dataset. Comparisons of phylogenetic trees constructed from datasets using exons or supercontigs as reference for read sorting were made. Accessions with conflicting placements between the trees are connected with dashed lines, and accessions or clades with conflicts in the primary topology are highlighted in red. Branch support values are shown near the nodes, where values are omitted when fully supported (LPP = 1.0 and UFBS = 100%).

approximately 40 years old yielded high-integrity sequences compared to older specimens around 80 years old. Increasing sequencing depth can improve the integrity of flanking sequences when using exons as reference for older specimens, although this improvement is less pronounced when compared to using

supercontigs as a reference. It is important to note that these findings were based on a limited number of samples. Future research should consider expanding the sample size to further validate these results and explore the impact of specimen age on sequencing quality.

We demonstrated that using supercontigs as a reference can yield a large number of nuclear loci (targeted exons and their flanking regions) with high integrity from both silica-gel dried material and herbarium specimen. This approach requires a genomic reference or a large number of targeted genes assembled from transcriptome and genome data (Zhang et al., 2019; Liu et al., 2021). In the genomic era, with numerous plant genomes either published or in progress, it is now feasible to obtain numerous single- or low-copy nuclear genes as references, enabling the assembly of these genes from DGS data for phylogenomic applications.

4.3. Phylogenetic utility of high-integrity nuclear genes

DGS data at $\sim 12.1\times$ coverage for silica-gel dried material and $15\times$ coverage for herbarium specimen, coupled with the use of supercontigs as a reference, represents an effective approach of recovering a large amount of useful comparative nuclear single-copy orthologous loci for phylogenetic analysis (3391 genes and 528 flanking regions; Table 1). The resulting data from this approach gave comparable levels of phylogenetic resolution and node support as our reference phylogeny, which was based on $20\times$ sequence coverage of herbarium specimens (Figs. 5 and S1). Specifically, our ‘ $10\times$ -silica $15\times$ -herbarium supercontig’ approach recovered 100% of species relationships congruent with the reference phylogeny, and all nodes showing UFBS = 100%/LPP = 1, except for one node with LPP = 0.99 support. This is markedly higher than other methods, e.g., ‘ $10\times$ -silica $15\times$ -herbarium exon’. In this approach, the flanking regions only recovered 66.7%–88.9% (12–14/18) species relationships with 5–6 nodes showing UFBS $\leq$ 72%/LPP $\leq$ 0.65 support, and the exons recovered all species relationships but with one node showing UFBS = 89% and another node with LPP = 0.52.

In this study, the utilization of taxon-specific genes as a reference recovered a higher number of genes and improved sequence integrity, and enhanced resolution capability compared to the Angiosperms353 gene set (Fig. S2). Our findings are in line with those of Yardeni et al. (2022), who also reported that a taxon-specific kit yielded more segregating sites, which showed a comparable phylogenetic resolution between universal and taxon-specific set. However, our results indicate that the latter offers a distinct advantage, particularly in the case of the rapid radiating group. Moreover, gene recovery is not only affected by the types of loci being recovered, but the presence or absence of a reference genome, and the genetic distance of the reference genome to the focal group of interest. While the use of a divergent reference genome will impact sequence recovery, we observed high sequence integrity and recovery of *Diplarche multiflora* (outgroup) even with *Rhododendron* as the reference genome. However, reduced integrity due to the use of a distant reference genome may compromise phylogenetic resolution or even produce inaccurate relationships, as seen when using the Angiosperms353 gene set, where low sequence recovery yielded a poorly resolved tree. We recommend assessing the number and integrity of assembled loci in detail when a close reference genome is not available to ensure phylogenetic relationships are not inferred based on large amounts of missing data. Further exploration across a broader sampling is needed to fully assess the potential of such sets in other groups.

Accurate species delimitation is essential for reliable phylogenetic inference, particularly in taxa with cryptic hybridization or introgression (Wang et al., 2019). In our study, species were selected based on distinct morphology and prior DNA barcoding results (Fu et al., 2022), with careful identification to ensure accurate representation of species. This approach minimized the risk of misidentifying hybrid or introgressed individual, thereby ensuring

robust phylogenetic analyses. Despite these checks, there is inherent complexity within *Rhododendron* that is unavoidable, and we note that several polyploid species, such as *R. helirolepis* (hexaploid), *R. oreotrephes* (octaploid or decaploid), and *R. racemosum* (tetraploid), were included here. Interestingly, high-integrity target loci and flanking regions can be recovered from these accessions at similar rates as those of diploid species with similar DGS data. Furthermore, under the current sampling framework, the phylogenetic relationships of these species were reliably resolved using all exon datasets and the datasets of flanking sequences with supercontigs serving as reference (Figs. 5 and S1). As the number of sampled species increases, the results may not fully capture the complicated reticulate relationships that characterize polyploid species, possibly contributing to phylogenetic conflicts. The presence of polyploids, especially allopolyploids, poses significant challenges in phylogenetic studies, complicating the reconstruction of evolutionary relationships. The difficulty in identifying reliable orthologous genes and accurately distinguishing between multiple homeologous copies in allopolyploids remains a substantial obstacle in the field (Ning et al., 2024). Therefore, in genera such as *Rhododendron* and other polyploid-rich groups, it is crucial to accurately identify and handle multiple gene copies in future studies. Fortunately, recently developed tools that phase gene copies into polyploid subgenomes using phylogenetic and similarity approaches (e.g., Nauheimer et al., 2021; Freyman et al., 2023) offer promising solutions for addressing these complex challenges.

5. Conclusions

With the decreasing cost of genome sequencing, DGS has become feasible for large-scale phylogenetic analyses across a wide range of taxa. By increasing sequencing depth, it is practical to successfully recover high-integrity target genes from DGS data, whether obtained from silica-gel dried material as demonstrated in previous studies, or from herbarium specimen where the DNA is typically degraded into short fragments as shown in this study. Our study shows that, for herbarium specimens, DGS at a depth of $15\times$ recovered extensive comparative nuclear genome data for phylogenetic analysis. While greater sequencing depth can further enhance data recovery, $15\times$ can be considered the optimal minimum coverage for herbarium specimens, and acts as a cost-effective means to recover a large number of target genes with high integrity. Additionally, using supercontigs as a reference significantly improved the recovery of both target exons and flanking regions compared to an exon-only reference, underscoring its value in advancing phylogenomic insights. Furthermore, sequences assembled using taxon-specific genes as a reference exhibited higher sequence recovery and phylogenetic resolution than those using a universal angiosperm gene set (the Angiosperms353), highlighting their potential for phylogenetic studies, especially for groups undergoing evolutionary radiations. These insights, demonstrated using *Rhododendron* as a case study, offer valuable guidelines for applying DGS in herbarium genomics for phylogenetic research, especially in groups experiencing rapid diversification.

CRedit authorship contribution statement

Zhi-Qiong Mo: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis. **Chao-Nan Fu:** Resources, Investigation. **Alex D. Twyford:** Writing – review & editing, Methodology. **Pete M. Hollingsworth:** Writing – review & editing, Methodology. **Ting Zhang:** Resources, Investigation. **Jun-Bo Yang:** Resources, Investigation. **De-Zhu Li:** Writing – review & editing, Conceptualization. **Lian-Ming Gao:** Writing –

review & editing, Writing – original draft, Supervision, Project administration, Funding acquisition, Conceptualization.

Data availability statement

Sequencing data generated in this study are deposited in the NCBI SRA database under BioProject PRJNA1177361. Sequence alignments, phylogenetic trees and other data files that support the findings of this study are openly available in the Science Data Bank at <https://doi.org/10.57760/sciencedb.15194>.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study was supported by the National Key Research and Development Program of China (2023YFA0915800), the Science and Technology Basic Resources Investigation Program (2021FY100200), the Key Basic Research program of Yunnan Province, China (202101BC070003), the Large-scale Scientific Facilities of the Chinese Academy of Sciences (2017-LSFGBOWS-02), and the Chinese Academy of Sciences President's International Fellowship Initiative (2025PD0021). We are grateful to curator of KUN to access the herbarium specimens, and Jiayun Zou, Zhirong Zhang, Jing Yang, Hongtao Li and Chunxia Zeng, for their help with sample collection, laboratory work and data analysis. Molecular experiments and data analysis were performed at the Laboratory of Molecular Biology and iFlora High-Performance Computing Center of Germplasm Bank of Wild Species, Kunming Institute of Botany, CAS.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.pld.2025.04.006>.

References

- Aguiñe-Santoro, J., Zuluaga, A., Stonesmyth, E., et al., 2024. Phylogenomics of *Puya* (Bromeliaceae): evolution in the Andean slopes and sky island ecosystems. *J. Syst. Evol.* 62, 257–274.
- Bakker, F.T., Lei, D., Yu, J., et al., 2016. Herbarium genomics: plastome sequence assembly from a range of herbarium specimens using an iterative organelle genome assembly pipeline. *Biol. J. Linn. Soc.* 117, 33–43.
- Bankevich, A., Nurk, S., Antipov, D., et al., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477.
- Borowiec, M.L., 2016. AMAS: a fast tool for alignment manipulation and computing of summary statistics. *PeerJ* 4, e1660.
- Burbano, H.A., Gutaker, R.M., 2023. Ancient DNA genomics and the renaissance of herbaria. *Science* 382, 59–63.
- Capella-Gutierrez, S., Silla-Martinez, J.M., Gabaldon, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973.
- Chamberlain, D., Hyam, R., Argent, G., et al., 1996. The Genus *Rhododendron*: its Classification and Synonymy. The Alden Press, Oxford.
- Chang, Y., Zhang, R., Ma, Y., et al., 2023. A haplotype-resolved genome assembly of *Rhododendron vialii* based on PacBio HiFi reads and Hi-C data. *Sci. Data* 10, 451.
- Chen, S., Zhou, Y., Chen, Y., et al., 2018. fastp: an ultra-fast all-in-one FASTQ pre-processor. *Bioinformatics* 34, 884–890.
- Chen, Y.P., Sunojkumar, P., Spicer, R.A., et al., 2025. Rapid radiation of a plant lineage sheds light on the assembly of dry valley biomes. *Mol. Biol. Evol.* 42, msaf011.
- Diaz-Garcia, L., Garcia-Ortega, L.F., Gonzalez-Rodriguez, M., et al., 2021. Chromosome-level genome assembly of the American cranberry (*Vaccinium macrocarpon* Ait.) and its wild relative *Vaccinium microcarpum*. *Front. Plant Sci.* 12, 633310.
- Dodsworth, S., Pokorny, L., Johnson, M.G., et al., 2019. Hyb-Seq for flowering plant systematics. *Trends Plant Sci.* 24, 887–891.
- Emms, D.M., Kelly, S., 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16, 157.
- Freyman, W.A., Johnson, M.G., Rothfels, C.J., 2023. homologizer: phylogenetic phasing of gene copies into polyploid subgenomes. *Methods Ecol. Evol.* 14, 1230–1244.
- Fu, C.N., Mo, Z.Q., Yang, J.B., et al., 2022. Testing genome skimming for species discrimination in the large and taxonomically difficult genus *Rhododendron*. *Mol. Ecol. Resour.* 22, 404–414.
- Guo, C., Luo, Y., Gao, L.M., et al., 2023. Phylogenomics and the flowering plant tree of life. *J. Integr. Plant Biol.* 65, 299–323.
- Hoang, D.T., Chernomor, O., von Haeseler, A., et al., 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* 35, 518–522.
- Jin, Z.T., Hodel, R.G.J., Ma, D.K., et al., 2023. Nightmare or delight: taxonomic circumscription meets reticulate evolution in the phylogenomic era. *Mol. Phylogenet. Evol.* 189, 107914.
- Jin, Z.T., Ma, D.K., Liu, G.N., et al., 2024. Advancing *Pyrus* phylogeny: deep genome skimming-based inference coupled with paralogy analysis yields a robust phylogenetic backbone and an updated infrageneric classification of the pear genus (Maleae, Rosaceae). *Taxon* 73, 784–799.
- Johnson, M.G., Gardner, E.M., Liu, Y., et al., 2016. HybPiper: extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Appl. Plant Sci.* 4, 1600016.
- Johnson, M.G., Pokorny, L., Dodsworth, S., et al., 2019. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *Syst. Biol.* 68, 594–606.
- Junier, T., Zdobnov, E.M., 2010. The Newick utilities: high-throughput phylogenetic tree processing in the Unix shell. *Bioinformatics* 26, 1669–1670.
- Kadlec, M., Bellstedt, D.U., Le Maitre, N.C., et al., 2017. Targeted NGS for species level phylogenomics: “made to measure” or “one size fits all”? *PeerJ* 5, e3569.
- Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., et al., 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589.
- Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780.
- Kent, W.J., 2002. BLAT — the BLAST-like alignment tool. *Genome Res.* 12, 656–664.
- Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359.
- Lemmon, E.M., Lemmon, A.R., 2013. High-throughput genomic data in systematics and phylogenetics. *Annu. Rev. Ecol. Syst.* 44, 99–121.
- Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760.
- Li, W., Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659.
- Liu, B.B., Hong, D.Y., Zhou, S.L., et al., 2019. Phylogenomic analyses of the *Photinia* complex support the recognition of a new genus *Phippisioemeles* and the resurrection of a redefined *Stranvaesia* in Maleae (Rosaceae). *J. Syst. Evol.* 57, 678–694.
- Liu, B.B., Liu, G.N., Hong, D.Y., Wen, J., 2020. *Eriobotrya* belongs to *Rhaphiolepis* (Maleae, Rosaceae): evidence from chloroplast genome and nuclear ribosomal DNA data. *Front. Plant Sci.* 10, 1731.
- Liu, B.B., Ma, Z.Y., Ren, C., et al., 2021. Capturing single-copy nuclear genes, organellar genomes, and nuclear ribosomal DNA from deep genome skimming data for plant phylogenetics: a case study in Vitaceae. *J. Syst. Evol.* 59, 1124–1138.
- Liu, B.B., Ren, C., Kwak, M., et al., 2022. Phylogenomic conflict analyses in the apple genus *Malus* s.l. reveal widespread hybridization and allopolyploidy driving diversification, with insights into the complex biogeographic history in the Northern Hemisphere. *J. Integr. Plant Biol.* 64, 1020–1043.
- Ma, H., Liu, Y.B., Liu, D.T., et al., 2021. Chromosome-level genome assembly and population genetic analysis of a critically endangered *Rhododendron* provide insights into its conservation. *Plant J.* 107, 1533–1545.
- Ma, Y.Z., Mao, X.X., Wang, J., et al., 2022. Pervasive hybridization during evolutionary radiation of *Rhododendron* subgenus *Hymenanthes* in mountains of southwest China. *Natl. Sci. Rev.* 9, nwac276.
- Maurin, O., Anest, A., Bellot, S., et al., 2021. A nuclear phylogenomic study of the angiosperm order Myrtales, exploring the potential and limitations of the universal Angiosperms353 probe set. *Am. J. Bot.* 108, 1087–1111.
- Mo, Z.Q., Fu, C.N., Zhu, M.S., et al., 2022. Resolution, conflict and rate shifts: insights from a densely sampled plastome phylogeny for *Rhododendron* (Ericaceae). *Ann. Bot.* 130, 687–701.
- Nauheimer, L., Weigner, N., Joyce, E., et al., 2021. HybPhaser: a workflow for the detection and phasing of hybrids in target capture datasets. *Appl. Plant Sci.* 9, e11441.
- Nguyen, L.T., Schmidt, H.A., von Haeseler, A., et al., 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274.
- Ning, W., Meudt, H.M., Tate, J.A., 2024. A roadmap of phylogenomic methods for studying polyploid plant genera. *Appl. Plant Sci.* 12, e11580.
- Shirasawa, K., Kobayashi, N., Nakatsuka, A., et al., 2021. Whole-genome sequencing and analysis of two azaleas, *Rhododendron ripense* and *Rhododendron kiyosumense*. *DNA Res.* 28, dsab010.
- Simmons, M.P., Maurin, O., Bailey, P., et al., 2022. Benefits of alignment quality-control processing steps and an Angiosperms353 phylogenomics pipeline applied to the Celastrales. *Cladistics* 38, 595–611.

- Slater, G.S., Birney, E., 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6, 31.
- Smith, S.A., Moore, M.J., Brown, J.W., et al., 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evol. Biol.* 15, 150.
- Soza, V.L., Lindsley, D., Waalkes, A., et al., 2019. The *Rhododendron* genome and chromosomal organization provide insight into shared whole-genome duplications across the heath family (Ericaceae). *Genome Biol. Evol.* 11, 3353–3371.
- Staats, M., Erkens, R.H.J., van de Vossen, B., et al., 2013. Genomic treasure troves: complete genome sequencing of herbarium and insect museum specimens. *PLoS One* 8, e69189.
- Stamatakis, A., 2014. RAXML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313.
- Su, N., Hodel, R., Wang, X., et al., 2023. Molecular phylogeny and inflorescence evolution of *Prunus* (Rosaceae) based on RAD-seq and genome skimming analyses. *Plant Divers.* 45, 397–408.
- Vargas, O.M., Heuertz, M., Smith, S.A., et al., 2019. Target sequence capture in the Brazil nut family (Lecythidaceae): marker selection and in silico capture from genome skimming data. *Mol. Phylogenet. Evol.* 135, 98–104.
- Villaverde, T., Pokorny, L., Olsson, S., et al., 2018. Bridging the micro- and macro-evolutionary levels in phylogenomics: Hyb-Seq solves relationships from populations to species and above. *New Phytol.* 220, 636–650.
- Wang, J., Luo, J., Ma, Y.Z., et al., 2019. Nuclear simple sequence repeat markers are superior to DNA barcodes for identification of closely related *Rhododendron* species on the same mountain. *J. Syst. Evol.* 57, 278–286.
- Wang, X.Q., Xiong, T., Wang, Y.Y., et al., 2024. Integrating genomic sequencing resources: an innovative perspective on recycling with universal Angiosperms353 probe sets. *Hortic. Adv.* 2, 4.
- Wang, X.Y., Gao, Y., Wu, X.P., et al., 2021. High-quality evergreen azalea genome reveals tandem duplication-facilitated low-altitude adaptability and floral scent evolution. *Plant Biotechnol. J.* 19, 2544–2560.
- Weitemier, K., Straub, S.C.K., Cronn, R.C., et al., 2014. Hyb-Seq: combining target enrichment and genome skimming for plant phylogenomics. *Appl. Plant Sci.* 2, 1400042.
- Xia, X.M., Du, H.L., Hu, X.D., et al., 2024. Genomic insights into adaptive evolution of the species-rich cosmopolitan plant genus *Rhododendron*. *Cell Rep.* 43, 114745.
- Xia, X.M., Yang, M.Q., Li, C.L., et al., 2022. Spatiotemporal evolution of the global species diversity of *Rhododendron*. *Mol. Biol. Evol.* 39, msab314.
- Xue, T.T., Janssens, S.B., Liu, B.B., et al., 2023. Phylogenomic conflict analyses of the plastid and mitochondrial genomes via deep genome skimming highlight their independent evolutionary histories: a case study in the cinquefoil genus *Potentilla* sensu lato (Potentilleae, Rosaceae). *Mol. Phylogenet. Evol.* 190, 107956.
- Yang, F.S., Nie, S., Liu, H., et al., 2020. Chromosome-level genome assembly of a parent species of widely cultivated azaleas. *Nat. Commun.* 11, 5269.
- Yang, L.H., Shi, X.Z., Wen, F., et al., 2023. Phylogenomics reveals widespread hybridization and polyploidization in *Henckelia* (Gesneriaceae). *Ann. Bot.* 131, 953–966.
- Yang, Y., Smith, S.A., 2014. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. *Mol. Biol. Evol.* 31, 3081–3092.
- Yardeni, G., Viruel, J., Paris, M., et al., 2022. Taxon-specific or universal? Using target capture to study the evolutionary history of rapid radiations. *Mol. Ecol. Resour.* 22, 927–945.
- Yu, J.R., Zhao, H., Niu, Y.T., et al., 2023. Distinct hybridization modes in wide- and narrow-ranged lineages of *Causonis* (Vitaceae). *BMC Biology* 21, 209.
- Yu, X.Q., Yang, D., Guo, C., et al., 2018. Plant phylogenomics based on genome-partitioning strategies: progress and prospects. *Plant Divers.* 40, 158–164.
- Zeng, C.X., Hollingsworth, P.M., Yang, J., et al., 2018. Genome skimming herbarium specimens for DNA barcoding and phylogenomics. *Plant Methods* 14, 43.
- Zhang, C., Rabiie, M., Sayyari, E., et al., 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* 19, 153.
- Zhang, C., Zhao, Y., Braun, E.L., et al., 2021. TAPER: pinpointing errors in multiple sequence alignments despite varying rates of evolution. *Methods Ecol. Evol.* 12, 2145–2158.
- Zhang, F., Ding, Y.H., Zhu, C.D., et al., 2019. Phylogenomics from low-coverage whole-genome sequencing. *Methods Ecol. Evol.* 10, 507–517.
- Zhang, G.J., Ma, H., 2024. Nuclear phylogenomics of angiosperms and insights into their relationships and evolution. *J. Integr. Plant Biol.* 66, 546–578.
- Zhang, L., Xu, P., Cai, Y., et al., 2017. The draft genome assembly of *Rhododendron delavayi* Franch. var. *delavayi*. *GigaScience* 6, gix076.
- Zuntini, A.R., Carruthers, T., Maurin, O., et al., 2024. Phylogenomics and the rise of the angiosperms. *Nature* 629, 843–850.