

# A roadmap of phylogenomic methods for studying polyploid plant genera

Weixuan Ning<sup>1</sup>  | Heidi M. Meudt<sup>2</sup>  | Jennifer A. Tate<sup>1</sup> 

<sup>1</sup>School of Natural Sciences, Massey University, Palmerston North 4442, New Zealand

<sup>2</sup>Museum of New Zealand Te Papa Tongarewa, Wellington 6011, New Zealand

## Correspondence

Weixuan Ning, School of Natural Sciences, Massey University, Private Bag 11222, Palmerston North 4442, New Zealand.

Email: [wning@massey.ac.nz](mailto:wning@massey.ac.nz)

This article is part of the special issue “Twice as Nice: New Techniques and Discoveries in Polyploid Biology.”

## Abstract

Phylogenetic inference of polyploid species is the first step towards understanding their patterns of diversification. In this paper, we review the challenges and limitations of inferring species relationships of polyploid plants using traditional phylogenetic sequencing approaches, as well as the mischaracterization of the species tree from single or multiple gene trees. We provide a roadmap to infer interspecific relationships among polyploid lineages by comparing and evaluating the application of current phylogenetic, phylogenomic, transcriptomic, and whole-genome approaches using different sequencing platforms. For polyploid species tree reconstruction, we assess the following criteria: (1) the amount of prior information or tools required to capture the genetic region(s) of interest; (2) the probability of recovering homeologs for polyploid species; and (3) the time efficiency of downstream data analysis. Moreover, we discuss bioinformatic pipelines that can reconstruct networks of polyploid species relationships. In summary, although current phylogenomic approaches have improved our understanding of reticulate species relationships in polyploid-rich genera, the difficulties of recovering reliable orthologous genes and sorting all homeologous copies for allopolyploids remain a challenge. In the future, assembled long-read sequencing data will assist the recovery and identification of multiple gene copies, which can be particularly useful for reconstructing the multiple independent origins of polyploids.

## KEYWORDS

allopolyploid, phylogenomics, phylotranscriptomics, polyploidy, reticulation, sequencing, single-copy genes

## POLYPLOIDY AND SPECIES DIVERSIFICATION

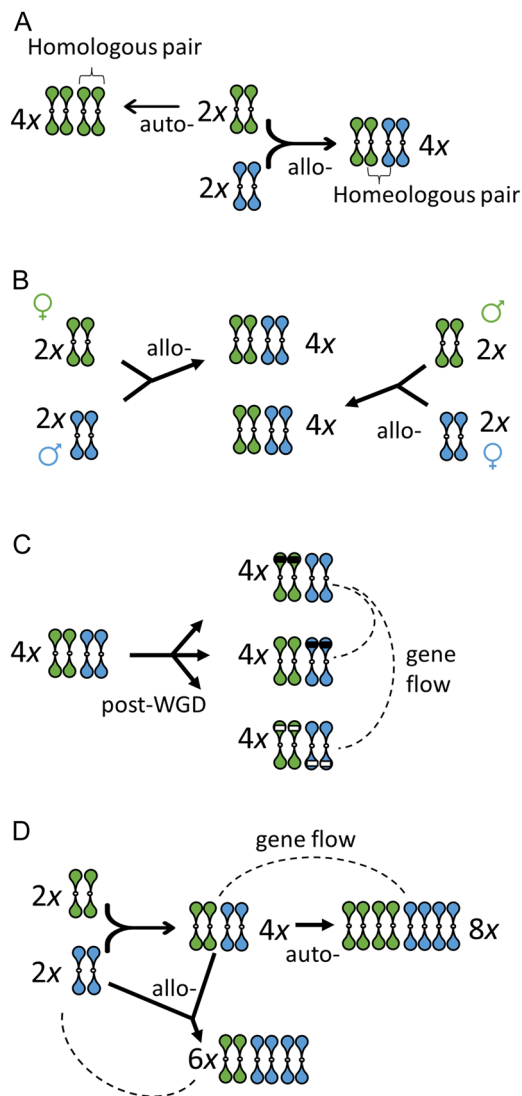
Polyploidy, resulting from one or more whole-genome duplication (WGD) events, describes organisms that have more than two paired sets of chromosomes. Although the long-term evolutionary effects of WGDs remain debatable (Soltis et al., 2014; Mayrose et al., 2015), with some considering polyploidy to be an evolutionary dead end (e.g., higher extinction rates in polyploids) and others suggesting it is ultimately a mechanism to promote speciation, genetic evidence suggests that all flowering plants have undergone at least one WGD event in their evolutionary history (Jiao et al., 2011; One Thousand Plant Transcriptomes Initiative, 2019). Undeniably, WGDs can act as a source of variation and are important for

promoting species diversity. To investigate plant evolution and diversification due to WGD, the origins and relationships of polyploid species must first be reconstructed.

Polyploids are broadly defined by their progenitor genome contributions. An allopolyploid originates from the combination of genomes from different species, while an autopolyploid is formed by genome duplication within a single species (Comai, 2005; Glover et al., 2016). Figure 1A shows an autopolyploid containing multiple homologous chromosome sets with nearly identical or little-diverged sequences, while the allopolyploid contains multiple homeologous (partially homologous) copies of parental chromosomes that often have more diverged sequences. These two groups of defined polyploids occupy the extreme ends of the polyploid-formation spectrum, but it is important to realize

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2024 The Authors. *Applications in Plant Sciences* published by Wiley Periodicals LLC on behalf of Botanical Society of America.



**FIGURE 1** Origins and diversification of polyploid species. The genomes of two diploid parents are represented by green (2x) and blue (2x). (A) An autotetraploid originates from the whole-genome duplication (WGD) of a single diploid parent (autopolyploidization), while an allotetraploid originates from a WGD of two diverged diploid parents that brought homeologous chromosomes together in one cell (allopolyploidization). (B) Reciprocal allopolyploidization forms two allotetraploids with different organellar genomes. (C) Different post-WGD processes, such as DNA deletion (white strips on the chromosomes) or insertion (black strips on the chromosomes), can lead to the diversification of a single allopolyploid lineage. Additional gene flow between tetraploids of different origins (represented by dashed lines) generates further diversity among neopolyploid lineages. (D) Formation of higher polyploids (6x, 8x) via repeated allo- or autopolyploidization events and with possible gene flow among differing ploidal levels indicated by a dashed line.

that there are forms in between that can be difficult to categorize into these strict types. Sometimes, meiotic chromosome pairing behavior can also be used to help elucidate the type of polyploid, whether bivalent (allopolyploids) or multivalent (autopolyploids) pairing occurs (Glover et al., 2016). Regardless of the different origins, the duplicated genomes in both auto- and allopolyploids

can be triggers of genomic, evolutionary, and species diversification in polyploid lineages (Soltis et al., 2015; Michael et al., 2016).

From a genomic point of view, polyploids tend to be more diverse than diploids due to their duplicated genomic content and the amount of initial genetic variation contributed from their parent species (Figure 1A). In general, allopolyploids can have more genetic incompatibility than autopolyploids due to the divergence between the parental genomes (i.e., the differences between duplicated homeologs vs. duplicated homologs). This makes the establishment of allopolyploids more challenging than that of autopolyploids as the coexisting divergent subgenomes must be balanced (Edger et al., 2018). On the other hand, the divergent subgenomes in allopolyploids can provide more opportunities for novel traits to evolve, which often makes it easier to identify allopolyploids than autopolyploids when using morphological traits and genetic sequencing data (Soltis et al., 2007; Qiu et al., 2020).

The cytoplasmic genomes in plants are typically uniparentally inherited (maternally in most flowering plants) (Birky, 1995), which, especially in allopolyploids, can lead to further divergence of the polyploid lineage due to organellar–nuclear genome incompatibility (Sharbrough et al., 2017; Postel and Touzet, 2020). Figure 1B shows the formation of two types of allotetraploids that have the same nuclear genome but different organellar genomes via reciprocal allopolyploidization (e.g., *Tragopogon miscellus* Ownbey; Shan et al., 2020).

Moreover, polyploid species with the same subgenome donors and organellar genomes can have multiple origins (i.e., different populations of a neopolyploid form independently from the same diploid progenitors, such as the *Tragopogon* L. allotetraploids [Soltis and Soltis, 1999]), or can diversify from a single WGD event (Figure 1C; such as allopolyploids in *Nicotiana* L. section *Repandae* Goodsp. [Clarkson et al., 2017; Dodsworth et al., 2017] or *Gossypium* L. [Paterson et al., 2012]). In addition, the genomic complexity of polyploids with multiple origins can be further affected by the divergence of, and variation within, the progenitors (Symonds et al., 2010; Rothfels, 2021). Nevertheless, the divergence of sister polyploid species will be determined eventually by different post-polyploidization processes in each independent lineage (Dodsworth et al., 2016; Li et al., 2021). The molecular-level changes after WGD, including chromosome reorganization, genomic/gene deletions or insertions, transcriptomic regulation, and epigenomic modification, can be affected by various ecological conditions (Chen, 2007; Vicent and Casacuberta, 2017; Li et al., 2021), further increasing the genomic complexity of polyploids. In addition to post-WGD changes, asexual reproduction in polyploids can also lead to species diversification (Pellino et al., 2013), as can sexual reproduction. As a result of reduced recombination between homologs and/or homeologs, mutations can accumulate independently in each set of alleles in a polyploid, resulting in rapid allelic sequence divergence (Hojsgaard and Hörandl, 2015) and therefore speciation (Figure 1C).

The formation of higher-level polyploids from two or more subgenome donors via repeated WGDs (reviewed by Oxelman et al., 2017), as well as additional reticulation events, including hybridization or introgression (Soltis and Soltis, 2009; Twyford and Ennos, 2012; Wagner et al., 2020), may lead to further species complexity and diversity. This might include gene flow between species with the same ploidy level (Figure 1C) or between different ploidy levels (Figure 1D). Moreover, the formation of higher-level polyploids can occur via the autopolyploidization of an existing allopolyploid or via additional hybridization events with a progenitor lineage (Figure 1D), as exemplified by *Fragaria* L. (Wei et al., 2017; Qiao et al., 2021), *Cerastium* L. (Brysting et al., 2007, 2011), *Ranunculus* L. (Paun et al., 2006; Pellino et al., 2013), and *Rosa* L. (Debray et al., 2021).

Given the variable factors that can contribute to the diversity of polyploids, a system with closely related polyploids of variable chromosome numbers (hereafter referred to as polyploid lineages or polyploid-rich genera) can be particularly useful for investigating macroevolutionary patterns (e.g., Meudt et al., 2015; Brittingham et al., 2018; Karbstein et al., 2022; Moraes et al., 2022; Wang et al., 2023a). Due to the initial WGD and additional reticulation events, such genera can have neopolyploid species with comparable and predictable genome sizes (C-value) and ploidy levels when compared to their parental species. The additive genome size or even the ploidy level in neopolyploids can be rapidly altered following each round of post-polyploidization genomic changes and can be mediated by different ecological conditions (Leitch and Bennett, 2004; Wang et al., 2021a). Eventually, neopolyploids or mesopolyploids (later-generation polyploids) within a genus can have diverse genome sizes, morphological characteristics, phenologies, life histories, and geographic distributions (e.g., Qiu et al., 2019; Han et al., 2020; Nieto Feliner et al., 2020; Debray et al., 2021; Lu et al., 2022). Understanding the phylogenetic relationships of species in a polyploid-rich genus that may have gone through multiple rounds of WGD and with reticulate evolutionary histories is therefore both critical and challenging.

## RESOLVING POLYPLOID RELATIONSHIPS USING ORTHOLOGOUS SEQUENCES

### Orthologous genes in polyploids

Over the past few decades, molecular phylogenetic inference of polyploid lineages has provided an affordable approach that has been developed and improved to estimate species relationships and the origins of polyploids (Stebbins, 1980; Hillis and Dixon, 1991; Small et al., 2004; Yang and Rannala, 2012; Guo et al., 2023). In general, phylogenetic inference of species relationships can be estimated from the genealogy of a single genetic marker (i.e., a gene tree) or from genealogies derived from genome-wide markers (i.e., phylogenomic inference). In either case, each set of gene markers is

expected to comprise orthologous sequences that were derived from the same ancestral locus and diverged following speciation, as opposed to paralogous sequences that arise from gene duplication (reviewed by Small et al., 2004). The informative polymorphisms among orthologous genes can be used to infer phylogenetic relationships among species. However, for polyploid lineages with multiple sets of chromosomes (Figure 1D), identifying and selecting informative genetic regions remains challenging (Small et al., 2004; McKain et al., 2018; Rothfels, 2021). This is because identifying orthologous genes for which all homeologous copies are preserved after a WGD can be difficult, as can capturing and recovering all homeologous sequences, as discussed later.

Moreover, not all gene regions or orthologous markers can be used successfully to resolve phylogenetic relationships, as different regions of a locus or different loci can have their own unique evolutionary rates (i.e., nucleotide substitution rates) and characteristics (i.e., inheritance mode) (Small et al., 2004; Soltis et al., 2014). For example, within a single locus, the coding regions are more conserved across different taxa than non-coding regions, the latter of which contain more polymorphisms for phylogenetic inference due to their faster evolutionary rates (Borsch et al., 2009; Pleines et al., 2009). The selection of different genetic markers or even different regions within a locus (e.g., coding vs. non-coding regions) therefore depends on the specific study questions of the plant lineages being investigated.

### Phylogeny of high- and single-copy genes in plants

Here, using polyploid-rich groups as an example, we review the techniques used to investigate polyploid lineages that may have an ancient WGD origin, as well as a recent species radiation involving both polyploidization and hybridization. Below, we discuss the utility and limitations of different genetic markers (i.e., high-copy organellar and nuclear genes vs. single-copy nuclear genes) in phylogenetic reconstruction, especially in identifying the parental species of taxa with hybrid or allopolyploid origins (Figures 1A, 1D).

In flowering plants, organellar genes from the chloroplast (plastid DNA or cpDNA) or mitochondria (mtDNA) are often only uniparentally inherited from the maternal lineage (Birky, 1995). Because of this, they can be particularly useful to identify the maternal lineage of allopolyploids (Figure 1B) and to resolve complex species relationships, especially in conjunction with nuclear genes (e.g., Debray et al., 2021; Šlenker et al., 2021; de Lima Ferreira et al., 2022). These organellar genomes are present in high copy numbers in the cell, and their gene regions are often conserved across taxa, two attributes that make them easy to capture and sequence (Small et al., 2004; McKain et al., 2018). The mitochondrial genome in plants has more structural variation (i.e., high intramolecular recombination) and often less sequence

variation (i.e., low evolutionary rates from low nucleotide substitution rates) than the chloroplast genome, which lacks recombination and has a relatively faster evolutionary rate (Small et al., 2004; Ravi et al., 2008); therefore, chloroplast markers generally have been preferred for species-level phylogenetic inference. Nevertheless, phylogenies reconstructed from organellar markers may provide researchers with a limited ability to correctly infer polyploid relationships due to WGDs and reticulate evolutionary histories. In particular, local introgression events between closely related polyploid species will lead to a plastid phylogeny that shows local geographical structure instead of species relationships (Tsitroni et al., 2003).

Biparentally inherited nuclear genes (Small et al., 2004) often have faster evolutionary rates than cpDNA or mtDNA markers (e.g., Gaut, 1998; Huang et al., 2012). Nuclear genes can be more informative about reticulate polyploid species relationships, given the recent formation of neopolyploids that often contain largely intact subgenomes (Li et al., 2021). The orthologous nuclear genes can be further divided into two categories, high-copy and single-copy nuclear genes, both of which are useful for resolving the phylogenetic relationships of polyploid lineages (Small et al., 2004). High-copy nuclear genes (e.g., the internal transcribed spacers [ITS] of the nuclear ribosomal DNA [nrDNA]), similar to organellar genes, are also often conserved across taxa and are easy to capture and sequence (Baldwin et al., 1995). However, these genes may be of limited use for phylogenetic inference because they have insufficient polymorphic sites due to a recent radiation among species, or because the homogenization process (concerted evolution) of homologous loci can result in a single copy remaining in the genome (reviewed by Álvarez and Wendel, 2003; Soltis et al., 2014).

By contrast, single-copy nuclear genes are less likely to be subjected to concerted evolution and tend to have faster evolutionary rates with more informative sites (Small et al., 2004). Therefore, phylogenies reconstructed from single-copy nuclear genes can be used to characterize species relationships more precisely, especially when high-copy genes alone, such as plastid markers and nrDNA, cannot provide resolved phylogenies for polyploid-rich genera whose species have gone through rapid species radiation, WGD, and reticulation (Sang, 2002; Soltis et al., 2014).

## Chromosome-level whole-genome comparisons

To date, the most advanced and comprehensive approach to reveal the origins and evolution of polyploid lineages, especially for allopolyploids (Figure 1A), is to compare the similarities and variations within each complete set of duplicated genomes (e.g., Chen et al., 2020; Gordon et al., 2020; Qiao et al., 2021; Peng et al., 2022). The comparison of the orthologous sequences between entire genomes of polyploids, also referred to as a comparative whole genomic analysis, utilizes the chromosome-level

assembled scaffolds. This necessitates the generation of a reference genome for each polyploid lineage, as well as their diploid parents or relatives, mostly via a de novo assembly method (Bayer et al., 2020; Lei et al., 2021).

The whole-genome sequence comparison will not only show the origins of each subgenome set, but can also reveal the evolutionary changes after polyploidization events (Deb et al., 2023), such as large genomic reorganization (Chen et al., 2020; Cerca et al., 2022), genomic content evolution (Bozan et al., 2023), and genome-wide signatures of selection (Qiao et al., 2021) or introgression (Wang et al., 2023b). However, in addition to the technical difficulties of whole-genome assembly (discussed later), when a polyploid-rich genus is rapidly evolved or extensively diverged (e.g., Meudt et al., 2021), the comparative genomic analysis approach can be highly cost-prohibitive. Especially for non-model polyploid lineages that contain multiple ploidy levels, the steps required to generate and annotate the subgenomes are, in most cases, time consuming and not cost effective (Kyriakidou et al., 2018). As the cost of genome sequencing continues to decline, this may be a more tractable option for non-model polyploid groups in the future.

## AN OVERVIEW OF SEQUENCING APPROACHES FOR PHYLOGENETIC INFERENCE

Phylogenetic approaches can capitalize on next-generation high-throughput sequencers to capture genome-wide markers, including whole or partial organellar genomes and numerous biparentally inherited nuclear loci, to infer the WGD histories as well as the phylogeny of polyploid species (McKain et al., 2018; Kapli et al., 2020). These sequencing platforms include first-generation Sanger sequencing (e.g., the Applied Biosystems sequencer [Thermo Fisher Scientific, Waltham, Massachusetts, USA]), second-generation high-throughput sequencing of short paired-end reads (e.g., Illumina [San Diego, California, USA] sequencer), and third-generation single-molecule real-time sequencing of long reads (e.g., Pacific Biosciences [PacBio, Menlo Park, California, USA] or Oxford Nanopore Technologies [Oxford, United Kingdom] sequencers).

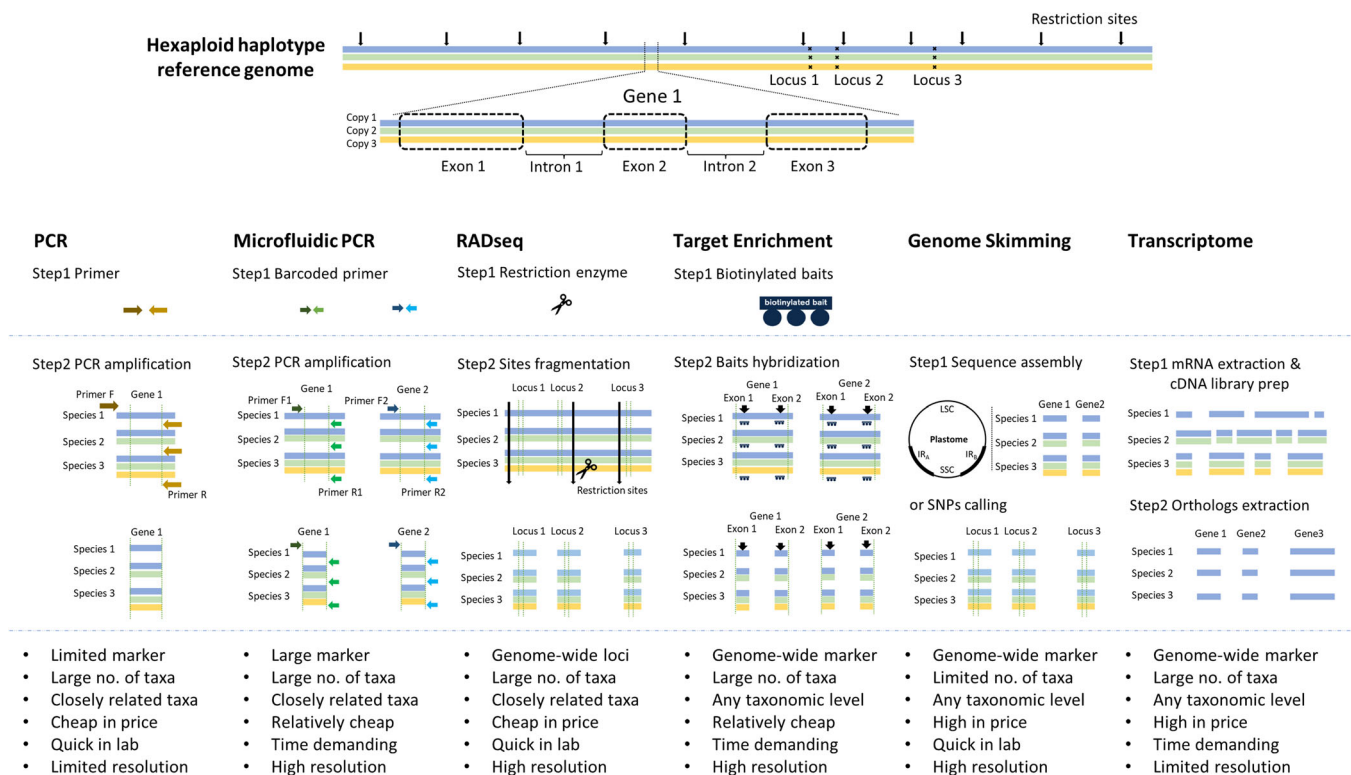
In addition to the comparison of whole chromosome-level genome sequences, there are currently six popular phylogenetic sequencing approaches (including phylogenomics and phylotranscriptomics) that can be divided into four categories based on their required investment (e.g., cost and whether reference data are required) and the efficiency of the downstream data analysis (Figure 2) (reviewed by McKain et al., 2018). The groups are as follows: (1) designing primer pairs for capturing an individual locus is time consuming, but downstream analysis is efficient, e.g., PCR and microfluidic PCR; (2) restriction enzyme-based approaches that can reduce genomic complexity (usually little prior work is needed to select restriction enzymes) for downstream analysis, e.g., restriction site-associated DNA sequencing (RADseq);

(3) target enrichment sequencing (Hyb-Seq) that can capture specific genomic regions simultaneously using pre-designed biotinylated RNA baits, including complex taxon-specific (requires prior selection of single-copy genes using transcriptome or genome-skimming data) or simplified universal (no prior information required) bait sets (Johnson et al., 2018), with reduced downstream computational analysis; (4) no prior information is required but downstream computational methods are demanding, e.g., genome-skimming sequencing with depth or transcriptome sequencing (i.e., phylotranscriptomics), as well as whole-genome comparisons.

As outlined above, there are different approaches to resolving polyploid species origins and relationships. Selecting the most efficient sequencing platform for each phylogenetic approach plays an important role in recovering informative orthologous gene sequences and their

duplicated homeologous gene copies, which is essential for the phylogenetic inference of polyploid species. Depending on the availability of plant material and finances, some of these approaches will be more tractable than others for a particular study. The “gold standard” may be whole-genome sequencing, but this also may not be necessary to answer the questions of interest for a particular polyploid group and/or it may exceed budgets. Whether the study focuses on all species within a genus (a large number of taxa with various divergence rates), a closely related group of species (low interspecific divergence rates between a small number of taxa), or a species complex (multiple WGD and reticulation events) will factor into decisions regarding what approach is most appropriate.

For better or worse, polyploid plant species are not a “one size fits all,” so what works for one group may not



**FIGURE 2** A comparison of five phylogenomic approaches (PCR, microfluidic PCR, restriction site-associated DNA sequencing [RADseq], target enrichment, and genome skimming) and one phylotranscriptomic approach (transcriptome sequencing) for capturing targeted markers between a diploid (2x: Species1) vs. two allopolyploids (4x: Species2, 6x: Species3). A hexaploid haplotype reference genome is shown at the top and illustrates the targeted gene sequence (e.g., exons and introns in Gene1 with three homeologous copies) and genome-wide restriction sites (black arrows) for extracting SNPs (black crosses represent SNPs at loci 1–3). Each method is compared by their required prior preparation and conceptual laboratory workflow, as well as the general suitability for studying complex polyploid plant groups. (1) PCR can capture and amplify the targeted locus (e.g., the sequence of Gene1 in Step2) with a pair of prior designed primers (Step1). (2) Microfluidic PCR also starts with primer design (Step1) for each targeted locus (Gene1 and Gene2 in Step2). The amplicons can be individually barcoded for each sample, allowing the amplicons from different samples in one array to be multiplexed and sequenced. (3) RADseq utilizes restriction enzymes (indicated by scissors in Step1) that can recognize specific restriction sites (black arrows in Step2) and shred the genome into random, simplified, and comparative DNA fragments that contain informative DNA sites for sequencing. The sequenced reads are often used to extract the SNP variation (Step2 loci 1–3). (4) Target enrichment sequencing or Hyb-Seq uses pre-designed RNA biotinylated baits (Step1) to hybridize with individually barcoded genomic libraries in one pool (Step2). The baits bind to conserved exons of the targeted genes (e.g., exons in Gene1), which can be sequenced on a high-throughput sequencer to recover the exons. (5) Genome skimming requires no prior information. The output of genome skimming depends on the sequencing depth (Step1; see main context) but is generally sufficient to capture high-copy regions (e.g., plastome, nrDNA). (6) Transcriptome sequencing also does not require any prior information; however, due to the sequencing reads being only from the exons and post-transcriptomic modification processes, mRNA data often cannot be used to fully identify the genes with homeologous copies (Step1; see the main text for details).

directly correlate to another. Nevertheless, below we offer suggestions for selecting genetic- or genomic-scale data that are broadly suitable for studies of polyploid species to suit different budgets and timelines. By comparing the efficiency of sequencing genome-wide markers in diploids vs. allopolyploids ( $2x$  vs.  $4x$  and  $6x$ ), and the recovery of homeologs in polyploids (Figure 2), we discuss the application of three generations of sequencing platforms combined with six sequencing approaches.

## PCR and microfluidic PCR

PCR amplification of a target region requires a pair of primers that can bind to the specific sequence of interest (Figure 2). The commonly used regions, such as plastid DNA or the nuclear ribosomal ITS region, which are conserved between taxa and present in high copy numbers in the genome, are easy to amplify via universal primers using PCR and can be informative about the relationships of diploid species without reticulate evolutionary histories (Hillis and Dixon, 1991; Shaw et al., 2014). For homoploid hybrids or polyploid lineages, PCR-amplified biparentally inherited nuclear genes, such as the ITS homeologs that have not been affected by concerted evolution (i.e., concerted evolution has not occurred to the extent as to render homeologous copies uniform) or single-copy nuclear gene homeologs, can be particularly useful to infer the reticulate species relationships within polyploid-rich genera (Rothfels et al., 2017; Xu et al., 2017; Osuna-Mascaró et al., 2022).

Separating PCR-amplified homeologous gene copies is challenging. Traditional Sanger sequencing often requires additional cloning or homeolog-specific primers to separate the homeologous gene copies (Brysting et al., 2011). Combining PCR with second- or third-generation sequencers provides an efficient way to amplify and separate different alleles or homeologous loci. For example, Rothfels et al. (2017) sequenced PCR amplicons (1 kbp) of four single-copy nuclear genes using PacBio; by adding ambiguous sites among primers, they captured multiple phased (separated) individual homeolog sequences in the resulting long reads.

Microfluidic PCR enhances the parallel amplification of multiple loci by combining PCR with microarray technology, which enables the amplification of thousands of targets per array (Zhang and Ozdemir, 2009). The combination of microfluidic PCR with second-generation Illumina sequencers is a powerful tool that can capture and sequence hundreds of single-copy nuclear genes, with additional homeologous sequences recovered from the sequenced reads (Uribe-Convers et al., 2016; Debray et al., 2021; Frost et al., 2021). Utilizing unique barcodes and paired-end sequence information, the individual locus sequences can be demultiplexed and assembled from the sequenced reads for each input taxon via bioinformatic tools such as the Pipeline for Untangling Reticulate Complexes (PURC) (Rothfels et al., 2017; Schafran et al., 2023) or Fluidigm2PURC (Blischak et al., 2018). After filtering the chimeric sequences

from PCR amplifications, the haplotype sequences across all input taxa at each locus can be clustered based on their sequence similarities to continue the downstream gene alignments (Figure 2). However, Illumina sequencers can return a maximum of 300 bp for a single-end read, which means the targeted gene length is often required to be shorter than 1 kbp to recover the whole gene sequence without any gaps present (Debray et al., 2021).

For non-model lineages, generating additional reference genomes or transcriptome sequences is required before targeted locus selection. In addition, testing whether the orthologous genes are conserved between all taxa can be time consuming. Selecting the amplified polyploid loci for which all homeologous copies are present, designing the individual primer pairs for each locus, and optimizing the PCR conditions for each gene amplification will require additional effort.

When targeting commonly used nuclear (Small et al., 2004) and plastid loci (Shaw et al., 2014), PCR-based approaches would suit studies aimed at larger groups of species within a genus, especially groups for which no prior phylogenetic data have been generated. Moreover, using PCR-based methods to test for specific allelic sequence differences or loci related to functional divergence that resulted from WGD can be particularly useful when combined with a third-generation sequencer (Suissa et al., 2022; Joshi et al., 2023).

## RADseq

The use of RADseq is common for population genetic studies and it has become more popular in phylogenomic studies (reviewed by Leaché and Oaks, 2017). This approach uses restriction enzymes that can recognize specific genome sites (e.g., 4-bp cutter, 6-bp cutter, or 8-bp cutter) to shear DNA into simplified yet comparable fragments that contain informative sites for quantitative genetic and population genetic studies of individuals from different populations (reviewed by Davey et al., 2011). Similarly, enzyme-digested DNA fragments can also be used to compare the genetic divergence of closely related taxa or reconstruct their phylogenetic relationships (Figure 2) (Wang et al., 2021b; Karbstein et al., 2022; Suissa et al., 2022).

Andrews et al. (2016) reviewed RADseq-related approaches, such as double-digest RAD (ddRAD) and genotyping-by-sequencing (GBS), which often rely on single nuclear polymorphisms (SNPs) extracted from Illumina-sequenced reads (maximum single-end 300 bp length) for downstream analysis. After acquiring the data from RADseq reads, the next step is to identify the orthologous DNA fragments based on their sequence similarities (Figure 2). Bioinformatic tools such as ipyrad (Eaton and Overcast, 2020) can demultiplex the sequence reads into each individual using sequencing barcode information, after which the similar reads can be sorted into each DNA fragment cluster (locus) with or without a reference genome. After assembling each DNA cluster block with

overlapped reads into consensus sequences and comparing all input taxa, a variant call format (VCF) file that contains all SNPs for each individual can be generated for downstream analysis.

Allopolyploids that contain diverged subgenome donors are also expected to have more heterozygous SNP sites. Assigning or phasing the genome-wide SNPs of polyploids to each subgenome donor using RADseq DNA fragments is not feasible without the reference genome of each subgenome donor, unless the RADseq data include both the digested DNA fragments of polyploid species as well as their potential diploid subgenome donors (Wang et al., 2021b). Conversely, inferring reticulate polyploid species relationships using RADseq data often has lower requirements for SNP phasing, but is also limited by the necessity of including parental or related diploid species (Wang et al., 2021b; Karbstein et al., 2022), as discussed later.

The RADseq method often requires a low degree of divergence (or a low substitution rate) among the investigated taxa. This method can also work for diverged taxa (Guo et al., 2023); however, the efficiency of the RADseq approach for studying diverged taxa may depend on the proportion of missing SNPs because fewer common DNA cut sites between taxa would be expected (Eaton et al., 2017; Wagner et al., 2020, 2023). On the other hand, the random distribution of the enzymatic sites means that no specific gene sequence (e.g., single-copy genes) can be recovered (McKain et al., 2018). Using RADseq for highly degraded DNA samples (e.g., herbarium specimens) will yield low output (few loci) due to strong locus dropout in degraded material.

Although RADseq can be limited in polyploids with higher ploidy levels (6x and above) or complex genomic content, with proper taxon sampling it can be suitable for studying lower-level (4x) polyploids in closely related species complexes (Wagner et al., 2020). With detailed individual sampling for each polyploid species, RADseq could be useful for addressing the population genetic variation within each lineage (Andrews et al., 2016), as well as providing further insight into genomic adaptation after WGD (Bürki et al., 2023).

## Target enrichment sequencing

Target enrichment sequencing (Hyb-Seq) provides a straightforward way to capture desired genomic regions via pre-designed biotinylated RNA baits (Weitemier et al., 2014; Andermann et al., 2020). These baits are often at least 100 bp in length and can capture specific genomic regions based on their sequence similarities (Figure 2). The bait set captures DNA fragments from the targeted genes, which can be enriched via PCR and sequenced using high-throughput sequencing platforms (Weitemier et al., 2014). For phylogenetic inference, the exons of targeted genes are more conserved than introns or flanking regions across

taxa, making them ideal regions for bait design (Weitemier et al., 2014; Schmickl et al., 2016).

Using a pipeline such as HybPiper (Johnson et al., 2016), the single-copy genes can be extracted by mapping the sequence capture reads to a reference file containing the reference sequences of the desired loci (e.g., McLay et al., 2021), and the mapping step can be done using the DNA nucleotide sequences or with the transcribed amino acid sequences. Each reads cluster can then be de novo assembled into contigs and correctly ordered by comparing to the exon sequences in the reference file, enabling the final extraction of exons or even the gene sequences (including exons, some or partial introns, and flanking regions). Furthermore, the limited number of targeted loci increases the opportunity to pool hundreds of individuals in one batch for second-generation sequencing.

The read depth of Hyb-Seq data can also provide an additional opportunity for the extraction of each homeologous gene copy. A few bioinformatic tools have been developed to detect and extract allelic variation from these data. ParalogWizard (Ufimov et al., 2022) or the putative paralogs detection (PPD) pipeline (Zhou et al., 2022) can increase the performance of the detection of “paralogs” (or different homeologous gene copies for allopolyploids) by comparing the divergence level among the assembled gene sequences. Gardner et al. (2021) added each copy of the pre-identified homeologous sequences into the reference file as the “reference sequence” to assist homeolog phasing using HybPiper. HybPhaser (Nauheimer et al., 2021) can phase (separate) the Hyb-Seq data of hybrid or polyploid species by splitting and mapping the sequenced reads via BSplit (BMAP; Bushnell, 2023) to the predefined pseudo-subgenome donors (e.g., a diploid that relates to the polyploid and does not contain a large number of SNPs). PATÉ (Tiley et al., 2021) utilizes the ploidy level and Hyb-Seq sequencing depth (i.e., overlapped reads) information of polyploid species to phase the haplotype blocks for each targeted gene via GATK (DePristo et al., 2011) and HpoPG (Xie et al., 2016). Similarly, the Hyb-Seq allele phasing pipeline that combines GATK and WhatsHap (Patterson et al., 2015) can also extract the haplotypes of each targeted gene using the depth of the sequenced reads (e.g., Šlenker et al., 2021). The SORTER pipeline (Jonas et al., 2023) also increases the ability for filtering paralogous gene sequences and phasing the Hyb-Seq reads using SAMtools (Li et al., 2009).

Johnson et al. (2018) identified 353 conserved single-copy nuclear genes among the angiosperms and published the Angiosperms353 bait set for phylogenomic inference of any flowering plant group. Breinholt et al. (2021) subsequently published the GoFlag451 bait set, which can capture up to 248 single-copy nuclear genes across the flagellate plant lineages, including bryophytes, ferns, and all gymnosperms. These universal bait sets improved the flexibility of phylogenetic inference of the investigated group even with divergent evolutionary histories, such as in the Tree of Life project (Baker et al., 2022). The off-target reads can also be

used for high-copy gene extraction, which provides additional phylogenetic signal for the inference of species relationships (Karimi et al., 2020; de Lima Ferreira et al., 2022). Moreover, pipelines such as HybSeq-SNP-Extraction (Sлимп et al., 2021) can extract SNPs from Hyb-Seq data and the output can be analyzed using methods similar to those used for RADseq.

Although the universal bait sets significantly reduce the amount of time needed for Hyb-Seq experimental design for non-model plant groups, building taxon-specific bait sets from genome-skimming or transcriptome reads (discussed later) should still be prioritized whenever the resources are available (Siniscalchi et al., 2021; Yardeni et al., 2022). Many published taxon-specific bait sets (at the family level) have also reduced the time required for prior bait design using reference sequences and increased the opportunities to extract allelic variation from the Hyb-Seq reads due to higher rates of recovering genes with introns and flanking regions (Šlenker et al., 2021; Crowl et al., 2022). These more specific bait sets are expected to yield more targeted loci, which can lead to phylogenetic inference of closely related polyploid groups with greater resolution, or even provide further insight for future population genetic analysis of each taxon (Phang et al., 2023).

For polyploid lineages, the bioinformatic pipelines are still limited to extracting the complete homeologous sequences of the targeted single-copy nuclear genes, due to lack of ploidy-level information or limitations of the phasing program, which may not be able to handle polyploids with multiple subgenome donors. Moreover, phasing can be even more challenging when polyploids have subgenomes with low divergence rates or when a universal bait set (e.g., Johnson et al., 2018) is applied, which will often yield similar exon sequences and a lower recovery rate of introns and flanking regions (i.e., discontinuous exons) compared with the lineage-specific bait sets (Hendriks et al., 2021; Yardeni et al., 2022). Eventually, the targeted loci may result in chimeric gene sequences that contain exons from different homeologous copies.

Nonetheless, target enrichment sequencing can be a useful tool for phylogenomic reconstruction of rapidly diverged or more distantly related species, but may benefit from taxon-specific baits rather than universal baits for resolving the complex of closely related polyploids. With sufficient sequencing depth, efficient phasing of homeologous loci, and a well-annotated reference genome, Hyb-Seq may also be used to address the post-WGD locus evolution of polyploids.

## Genome skimming by shotgun sequencing

Genome skimming by shotgun sequencing can generate sequences that represent the whole genome of the sequenced individual with shallow coverage of sequencing reads, and often uses second-generation (short-read) sequencers (Figure 2). This method requires no prior primer design

and is flexible with respect to the divergence rate of the studied groups. It also provides an opportunity to recover the homeologs of extracted loci for polyploid lineages.

Depending on the sequencing coverage, shallow genome skimming can be an efficient approach to recover high-copy-number organellar and nuclear genes, such as nrDNA and whole plastid genomes. High-copy-number genes can be de novo assembled from genome-skimming reads due to their abundance in the genome, even with low sequencing coverage (e.g., 1× of whole-genome sequencing coverage) (Straub et al., 2012). Taking the assembly of a plastome as an example, the seed-and-extend-based de novo assembly (e.g., assembler compared in Freudenthal et al., 2020) or the reference mapping-based haplotype calling methods (e.g., Takamatsu et al., 2018) via GTAK HaplotypeCaller (DePristo et al., 2011) or SAMtools mpileup (Li et al., 2009) can produce a complete plastome. For repetitively occurring nrDNA, this may also be treated as a “circular” genome like the plastome to be de novo assembled with different initial seeds, e.g., using the assembler GetOrganelle (Jin et al., 2020). Phasing the nrDNA allele or homeolog variation from the depth of sequence reads would require additional ploidy-level information or the number of subgenome donors of the input taxa for use with bioinformatic tools such as GTAK HaplotypeCaller.

In addition, genome skimming with depth can also be useful for extracting SNPs or single-copy nuclear genes; for example, SNPs among genome-skimming reads between low-diverged input taxa can be extracted using the reference mapping-based method via joint SNP calling in GATK (DePristo et al., 2011), and the SNP data can be analyzed using sliding window-based methods or used to infer the species phylogeny based on independent SNPs (Meleshko et al., 2021). In addition, similar to the assembly of single-copy nuclear genes using Hyb-Seq data, user-friendly bioinformatic tools (methods compared by Michel et al., 2022) such as HybPiper (Johnson et al., 2016; Jackson et al., 2023) and HybPhyloMaker (Fér and Schmickl, 2018) can also de novo assemble the deep genome-skimming reads into exons by mapping them to a reference file that contains the exons of a related species, and can correct the order of exons within a locus by comparing them to the selected reference gene. Moreover, the reference file can be custom designed using published genome or transcriptome data via bioinformatic tools such as MarkerMiner (Chamala et al., 2015) or a universal angiosperm reference file that contains the reference gene sequences for the Angiosperms353 loci (McLay et al., 2021).

Genome skimming with sufficient depth is an efficient way to recover genome-wide single-copy orthologous genes and possibly all homeologous sequences for nuclear genes that are biparentally inherited. Nevertheless, determining the proper coverage for lineages with multiple ploidy levels is challenging, given that each homeologous copy will require sufficient sequencing coverage to be recovered. Liu et al. (2021) compared the number of successfully recovered single-copy nuclear genes and the depth of

genome-skimming reads (2× to 20× coverage) in Vitaceae. Their results suggested that at least 10× coverage was required to extract over 800 single-copy nuclear genes for phylogenomic inference. Moreover, the extensive genomic data (e.g., the high number of reads for sequencing coverage and depth) required for each sequenced sample also limits the number of input samples that can be added to one sequencing pool, which increases the total cost of sequencing. In addition, the downstream filtering and extracting of the informative loci from massive amounts of genome-skimming sequenced reads increases the computational analysis time.

Genome skimming, as the most comprehensive phylogenomic method, is now a popular approach to capture both high- and single-copy genetic markers without requiring a reference genome. It is useful for studying polyploid complexes at any taxonomic level. Although recovering all homeologous copies (gene sequences or SNPs) for allopolyploids will depend on the genome size and the sequencing depth of coverage (which can be an extra financial cost due to the limited number of individuals in one sequencing pool), in general, genome skimming is a promising approach for divergent species within a genus or closely related species.

## Transcriptome sequencing

Transcriptome sequencing (mRNA-Seq) can also be used to capture genome-wide markers (including high- and single-copy genes) without the requirement for any prior primers or restriction enzyme selection steps (Cheon et al., 2020). The transcriptomic (or genomic) data can be used to estimate ancient WGD events, using  $K_s$ -based methods ( $K_s$  plot) to calculate and visualize the distribution of synonymous substitutions per site ( $K_s$ ) among paralogous genes within the genome of a single species (Cui et al., 2006; Li and Barker, 2020; Tiley et al., 2021). These methods have been widely used for detecting ancient WGD events and estimating the age of WGD events in flowering plants (De Bodt et al., 2005; Pelosi et al., 2022; Zhao et al., 2023). In addition, when a specific experimental design targeting specific tissues or collection times is used, the resulting mRNA comparisons can further address the functional divergence of the loci under investigation (McKain et al., 2018).

In comparison to the DNA-based phylogenomic approaches, the phylotranscriptomic approach compares orthologous variation extracted from mRNA (Figure 2). Moreover, annotated reference transcriptomes are increasingly being added to online databases (One Thousand Plant Transcriptomes Initiative, 2019), and the pipelines for identifying orthologs using mRNA are continuously improving (Cheon et al., 2020). After acquiring the mRNA-Seq data and cleaning the reads (removing the adapters, poly-N sequences, and low-quality bases), extracting orthologous gene sequences often starts by distributing the mRNA-Seq reads into each protein-coding gene cluster for each sequenced individual, using bioinformatic tools such as CD-HIT (Li and Godzik, 2006).

The homology identification of each cluster across all input taxa can be done in several ways (e.g., compared by Cheon et al., 2020), such as using a “core ortholog” (i.e., a predefined set of orthologs) to identify the ortholog clusters in HaMStR (Ebersberger et al., 2009), the gene tree-based homology searching method in Yang and Smith (2014), or the ortholog group-based and gene tree-based approaches in OrthoFinder (Emms and Kelly, 2015).

To resolve the multiple origins of polyploids using mRNA data, extracting the haplotype isoforms using short-read Illumina sequencing can be particularly challenging without reference genomes, because the discontinuous mRNA sequences are transcribed from the exons only. Although the haplotypes of gene isoforms can also be phased using third-generation long-read sequencers to determine the origins of different subgenome donors (e.g., Leung et al., 2021; Cerca et al., 2022), mRNA data are often affected by post-polyploidization genomic and transcriptomic modifications (e.g., tissue-specific or homeolog-specific expression, gene loss, gene silencing, pseudo/sub-functionalization, and mRNA alternative splicing) (Adams et al., 2003; Zhou et al., 2011; Soltis et al., 2014; Van de Peer et al., 2017; Pelosi et al., 2022), which can result in biased sequencing depths for homeologs or only one gene copy retained. In addition, collected mRNA samples are more difficult to preserve, typically requiring flash-freezing in liquid nitrogen and storage at  $-80^{\circ}\text{C}$  or below. The extraction of mRNA and preparation of transcriptome libraries are both more complex and costly than extracting DNA and preparing DNA genomic libraries (McKain et al., 2018), which makes phylotranscriptomics a less attractive option than phylogenomics.

A phylotranscriptomic approach can be used to access the ancient polyploidization events (Morales-Briones et al., 2021), but may not be the most appropriate method for resolving the relationships of a complex with closely related neopolyploids that have more recent divergent histories, given the additional difficulties in recovering the homeologous sequences. Nonetheless, with precise experimental design, transcriptome comparison between sister polyploids and/or ancestor diploids remains an important method for investigating potential post-WGD functionally related divergence or further species adaptation (Wang et al., 2021c).

## Whole-genome long-scaffold sequencing

In comparison to phylogenomic or phylotranscriptomic approaches that mostly rely on second-generation short-read sequencers, whole-genome long-scaffold sequencing primarily uses third-generation sequencers to aid the haplotype de novo assembly of different subgenomes in polyploids (reviewed by Kyriakidou et al., 2018). Each sequenced long read (maximum length of over 100–300 kbp via Nanopore ultra-long-read sequencing) is from each haplotype of a subgenome (Koo et al., 2023), and based on

the sequence similarity of overlapped reads, assembled long reads are expected to resolve the haplotype genome complexity of polyploids (Jiao and Schneeberger, 2017). The assembly of the entire genome of a plant species is generally limited by the sequenced read length (Kyriakidou et al., 2018), the large repetitive genome content (Lei et al., 2021), and the excessive amount of heterozygosity within each haplotype genome (e.g., a heterozygous diploid; Zhou et al., 2020a). Additionally, various ploidy levels and different origins among polyploids increase the difficulty of assembling the haplotypes of each subgenome sequence, especially for homologous pairs in polyploids that are expected to have low divergence rates (Figure 1A).

Comparative genomic analyses using long-read sequences are the most promising method for investigating the genomic origins and evolution of polyploids (Jiao and Schneeberger, 2017; Kyriakidou et al., 2018). Building on this, additional chromosome-range scaffold-building technologies, such as optical genome mapping (Bionano Genomics, San Diego, California, USA) or chromosome conformation capturing (Hi-C; Belton et al., 2012), have been used successfully to assist the haplotype genome assembly of complex polyploids (Yang et al., 2016; Zhang et al., 2018a). Moreover, various bioinformatics programs have been developed to improve the accuracy of read assembly (reviewed in Kong et al., 2023), as have long-read haplotype phasing programs (summarized by Zhang et al., 2020a; Saada et al., 2022); however, discussing these methods is outside the scope of this review of the phylogenetic inference of polyploids.

Nevertheless, given the technical difficulties of whole-genome assembly, comparative genome analysis of a polyploid-rich genus is not currently a realistic choice for the study of most plant groups with a large number of taxa or variable ploidy levels. Compared with other phylogenomic approaches, the overall costs to generate and analyze the reference genomes for each polyploid lineage is not feasible yet, despite the continuously falling prices of third-generation sequencing technologies (Pucker et al., 2022). Finally, whole-genome long-scaffold sequencing requires a large quantity of high-quality (undegraded) extracted DNA (Hon et al., 2020; Wang et al., 2021d), thus limiting the use of herbarium specimens from endangered or extinct species, which could otherwise be crucial for drawing a complete picture of polyploid origins.

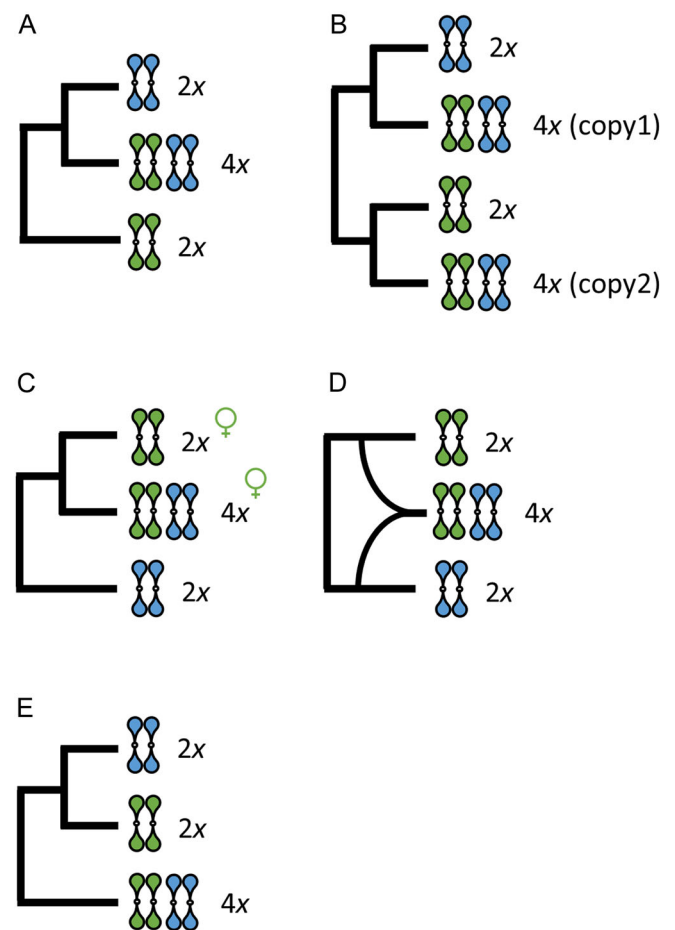
## PHYLOGENETIC INFERENCE OF SPECIES RELATIONSHIPS

### Phylogenetic inference of an individual locus (gene trees)

After assembling the targeted loci (e.g., SNPs, cpDNA, nrDNA, or single-copy nuclear genes) from the sequenced reads, the next step is to reconstruct the phylogeny for each individual locus. There are four main methods of phylogenetic

inference for an individual gene tree, namely neighbor joining, maximum parsimony, maximum likelihood, and Bayesian inference (reviewed by Holder and Lewis, 2003; Yang and Rannala, 2012; Kapli et al., 2020).

An allopolyploid species is expected to have only one copy of each uniparentally inherited locus (e.g., plastid loci) and multiple homeologous gene copies or heterozygous SNPs at each biparentally inherited nuclear locus (Figure 3). The homeologs would be expected to have greater sequence-level differences than the homologs. For methods that can produce individual gene sequences, aligning the gene sequences is the first step to identify the homology of a



**FIGURE 3** Phylogenetic inference of an allopolyploid species using different tree reconstruction methods. (A) A traditional bifurcating phylogenetic tree based on a nuclear marker reconstructs a polyploid as sister to one parent. The other parent may be observed to be more phylogenetically distant, depending on whether gene loss has occurred, divergence between the two parents, and overall sampling in the tree. (B) A multi-labeled nuclear gene tree shows two homeologous copies in an allotetraploid, each derived from one diploid parent. (C) A bifurcating phylogenetic tree based on an organellar marker reflects the maternal progenitor of an allopolyploid. As with the nuclear-based bifurcating tree, the other parent may be more distantly related based on divergence between the two parents and overall sampling. (D) A network based on a nuclear marker shows both parents that contributed to the allopolyploid genome. (E) A bifurcating gene tree inferred from chimeric assembled gene sequence of a polyploid species or possible recombination between homeologs.

sequenced locus prior to phylogenetic tree construction by any of the four methods mentioned above (bioinformatic tools reviewed by Kapli et al., 2020; Guo et al., 2023).

Taking the cytoplasmic marker as an example (Figure 3C), a traditional bifurcating gene tree can be used to infer only the maternal lineage of a polyploid species, given that these markers often only have one copy present, with exceptions such as biparentally inherited plastid DNA (e.g., Barnard-Kubow et al., 2017). By contrast, a multi-labeled bifurcating gene tree (MUL-tree) can show multiple origins for divergent homeologous gene copies of a biparentally inherited locus (Figure 3B) (Huber et al., 2008; Czabarka et al., 2013). A MUL-tree can be informative about reticulate species relationships, when the terminal branches come from the same individual that can be further merged together via Dendroscope (Huson and Scornavacca, 2012) to visualize the network relationships of polyploids (Figure 3D) (Morrison, 2014; Hibbins and Hahn, 2022).

By contrast, SNP data from known allopolyploids can be mapped back to the potential subgenome donors (e.g., the putative  $2x$  ancestors) and the proportion of mapped reads may also indicate the origins of the species (e.g., Wang et al., 2021b). Similarly, using the mapped read depth and genotypes of each SNP, the SNIploid pipeline (Peralta et al., 2013; Wagner et al., 2020) can detect the origins of allopolyploids and their parentage, and can separate homeologous SNPs (from WGDs) from post-origin SNPs (heterozygous in diploids). Moreover, the longer shredded DNA fragments (e.g., assembled contigs larger than 200 bp) that contain multiple SNPs can also be treated as an individual “locus” for gene phylogeny inference (e.g., Wang et al., 2021b). SNPs from the same DNA fragments can be phased and analyzed together to infer species relationships (e.g., Karbstein et al., 2022) via RADpainter and fineRADstructure (Malinsky et al., 2018). Alternatively, the individual biallelic SNPs (i.e., SNPs pruned in linkage disequilibrium) can often be analyzed as individual loci to reconstruct the “gene trees” via quartet-inference using SVDquartets (Chifman and Kubatko, 2014) or Bayesian inference-based SNAPP (Bryant et al., 2012) under the coalescent model. Eventually, a consensus tree can be generated by considering the concordance levels among all individual SNP trees.

However, not all homeologous copies of a targeted locus can be successfully recovered, because the post-polyploidization genomic modification process can also lead to duplicated genes with different fates (Figure 1C); for example, the duplicated gene copies may be either retained or lost, or subfunctionalization or neofunctionalization of duplicated genes can occur (reviewed by Prince and Pickett, 2002; Comai, 2005). Even if all gene copies are retained at one locus and no recombination between homeologs has occurred, phasing the homeologs using short reads can still be challenging, given the limited bioinformatic tools that can handle polyploids with multiple subgenome donors and the large amount of missing ploidy-level information for non-model groups. Finally, the phylogeny of polyploid species may result in incomplete gene tree inference with only

one copy retrieved from one of the subgenome donors (Figure 3A vs. Figure 3B). In addition, for all six phylogenomic or phylotranscriptomic methods, the misidentification of the orthologous gene copy can happen due to the short divergence time between parental genomes or gene loss during post-WGD genomic modification (Unruh et al., 2018). Furthermore, Hyb-Seq and mRNA-Seq, which are the most specific methods to recover only conserved exons (Weitemier et al., 2014; Cheon et al., 2020; Zhou et al., 2022), can be additionally problematic to extract the correct orthologs and their homeologous copies using the short reads. This is because the possibility of merging chimeric concatenated exons from different homeologs may result in incorrect gene tree inference (Figure 3E vs. Figure 3B).

Therefore, phylogenetic relationships can rarely be estimated from a single locus. Most studies now use a combination of genetic markers from different genomes (e.g., organellar vs. nuclear), consider the evolutionary background of each type of marker, and reconcile genome-wide signals to understand the origins and relationships of polyploids to improve their phylogeny reconstruction (Holder and Lewis, 2003; Soltis et al., 2014).

## Inferring a bifurcating species tree from gene trees

In addition to post-WGD genomic modification processes, discordance between different gene trees (i.e., phylogenetic incongruence) may be caused by their independent evolutionary histories. This might include different origins or evolutionary rates of the selected genes or genetic regions as mentioned above, or additional biological factors (e.g., incomplete lineage sorting [ILS], reticulation, horizontal gene transfer, and polyploidization) that can also contribute to the gene trees being discordant with the underlying species phylogeny (reviewed by Maddison, 1997; Degnan and Rosenberg, 2009; Twyford and Ennos, 2012; Mallet et al., 2016). Moreover, the concordance level of genome-wide captured nuclear loci can be further used to resolve the reticulate relationships of species in polyploid-rich genera as discussed below (Than et al., 2008; Solís-Lemus et al., 2017).

Maddison and Knowles (2006) and Yan et al. (2022) showed that a reliable species phylogeny can be generated from a sufficient number of nuclear loci, and that two evolutionary models are typically applied when inferring the species phylogeny from the list of loci. First, the species tree can be estimated by joining all the markers together (concatenation model), such that all markers are considered to have the same evolutionary history. The concatenation method may be more robust for species phylogeny estimation, but it can also result in overconfident node support values (Kubatko and Degnan, 2007). By contrast, each gene can be analyzed individually (multispecies coalescent [MSC] model), and under the coalescent model, the list of gene trees can be used to infer the species tree (Degnan and Rosenberg, 2009). The MSC model assumes the gene trees are independently evolving, and it applies the coalescent-based theory to estimate

the coalescence time (Heled and Drummond, 2010). This model can be more consistent in identifying ILS, and therefore provides a more accurate estimation of species trees (Mirarab et al., 2014).

For polyploid lineages, summarizing gene trees for species phylogenetic inference should be considered in two aspects: (1) if all gene trees of biparentally inherited loci have all homeologous copies from all subgenome donors (e.g., the MUL-tree in Figure 3B), or (2) if each gene tree only contains one set of an orthologous gene copy from only one of the subgenome donors (e.g., the plastid ortholog gene tree in Figure 3C). Inferring the species phylogeny using MUL-trees across genome-wide biparentally inherited loci requires the proper handling of each gene copy. Although a robust method such as Bayesian-based ASTRAL-Pro (Zhang et al., 2020b) under the MSC model allows for multiple alleles of one individual to be present and thus may improve the final species tree inference compared with the similar program ASTRAL (Mirarab et al., 2014), which only allows for a single copy present in one individual, it nevertheless assumes bifurcating species relationships. Studies of polyploid-rich lineages mostly still rely on utilizing orthologs extracted from one subgenome donor, using methods such as Hyb-Seq (Thomas et al., 2021; Karbstein et al., 2022), mRNA-Seq (Morales-Briones et al., 2021; Zhang et al., 2021), or deep genome skimming (Liu et al., 2021). Assuming the gene sequences are correctly assembled, gene trees can be summarized into a bifurcating species tree using a concatenation model such as IQTree (Minh et al., 2020a) or an MSC model such as ASTRAL or StarBEAST2 (Ogilvie et al., 2017).

## Networks in polyploid species

Under the assumption that all recovered gene sequences or SNPs are correctly identified orthologous markers, the conflict between the topologies of gene or SNP trees can be used to infer the allopolyploidization or hybridization histories between polyploids, which can be calculated via gene concordance analysis (Bouckaert, 2010; Smith et al., 2015) or network inference (Than et al., 2008; Solís-Lemus et al., 2017).

Using a traditional bifurcating approach under the multispecies coalescent model, the conflicts among independent biallelic SNP trees generated by SNAPP can be visualized by DensiTree (Bouckaert, 2010). Moreover, the ABBA-BABA or D-statistic test (Patterson et al., 2012; Hibbins and Hahn, 2022) can calculate the overall genomic introgression signals using biallelic SNPs via bioinformatic tools such as Dsuite (Malinsky et al., 2021). In addition, Bayesian-based methods, such as MCMC\_BiMarkers (Zhu et al., 2018) implemented in PhyloNet (Than et al., 2008) or SnappNet (Rabier et al., 2021) implemented in BEAST2 (Bouckaert et al., 2014), are extensions of SNAPP that can infer the network relationships of polyploids using biallelic SNPs. Quartet Sampling (Pease et al., 2018) is a statistical method that can test the incongruences and conflicting patterns in bifurcating tree topologies at each node, which

can be applied to phylogenetic trees derived from any marker set (Wagner et al., 2020; Karbstein et al., 2022) and does not require a priori knowledge of ploidy levels.

Similarly, the concordance levels within a set of orthologous gene trees, or between gene trees and the species tree, can indicate ILS and the complex evolutionary history of the included taxa (e.g., WGD or reticulation) (Minh et al., 2020b). Concordance analysis methods such as DensiTree can also visualize the topological incongruence between a set of gene trees (Zhou et al., 2020b), and PhyParts (Smith et al., 2015) calculates the concordance level between a set of gene trees for each internode of a species tree. By contrast, a network analysis can be performed in SpeciesNetwork (Zhang et al., 2018b), which uses an MSC model to infer each gene tree in BEAST and summarizes the final species tree with reticulations. Conversely, the likelihood programs that infer species relationships with a reticulate model and ILS, such as InferNetwork\_ML and InferNetwork\_MPL as implemented in PhyloNet (Than et al., 2008) or SNaQ as implemented in PhyloNetworks (Solís-Lemus et al., 2017), can produce a network of species relationships by calculating the concordance level from a list of gene trees, which can be useful to identify reticulation or allopolyploidization events. However, neither of these tools considers the subgenome origins of the gene copy and assumes strictly two progenitors per hybrid lineage, which limits analysis of higher ploidy levels (i.e., more than 4x). These also require testing for the number of “hybrid or WGD” nodes based on pseudo-likelihood or maximum likelihood, which can be computationally demanding when inferring polyploid-rich groups with large numbers of input taxa, as well as multiple rounds of hybridization and allopolyploidization events. Furthermore, PhyloNet does not allow polyploids with more than two subgenomes, i.e., a maximum of two alleles per gene. In addition, the robust method SplitsTree (Huson and Bryant, 2006) can be used to show the conflicts between concatenated gene sequences using a “network” approach.

By contrast, a Bayesian-based pipeline such as Homologizer (Freyman et al., 2023) or the gene tree-based method AlleleSorting (<https://github.com/MarekSlenker/AlleleSorting>; Šlenker et al., 2021) can first identify the origins of each gene copy and assign the gene copy to the correct subgenome donor, generating a phased MUL-tree. The phased MUL-trees can be summarized into a species MUL-tree via a program such as ASTRAL, where each tree tip corresponds to each subgenome of a polyploid species (as discussed by Debray et al., 2021). Moreover, phased MUL-trees can be used to infer the species network and ILS via AlloppNET (Jones et al., 2013) with the MSC model implemented in BEAST, which also infers the network of polyploid species that have a maximum of two subgenome donors (e.g., Rothfels et al., 2017; Eriksson et al., 2018). Despite this progress, phasing the allele variation and assigning to the correct subgenomes can be challenging due to bioinformatic difficulties (Ufimov et al., 2022; Zhou et al., 2022) and post-WGD genomic modification uncertainties (Li et al., 2021).

## DISCUSSION AND FUTURE PROSPECTS

The development of phylogenomic approaches has significantly improved our understanding of species diversification in polyploid-rich genera (Rothfels et al., 2017; Johnson et al., 2018; Debray et al., 2021; Rothfels, 2021). The financial cost of capturing genome-wide markers has decreased largely because of the more advanced sequencing technologies, increasingly available universal primers, or reference genome sequences, as well as more commercially available taxon-specific Hyb-Seq baits. Recently developed bioinformatic tools can further increase the accuracy of extracting the orthologous loci from multiple subsets of genomes of polyploid-rich groups.

No one method can be used to resolve the taxonomic relationships for all problematic plant genera. Selecting a suitable approach requires consideration of several factors: the complexity of the targeted group (number of species and ploidy-level variation); the availability of information (reference genome or taxon-specific bait set); the starting materials available for DNA/RNA extraction (freshly collected or mostly degraded specimens); and the evolutionary study questions to be answered after resolving the species relationships (e.g., population genetic variation, post-WGD genomic evolution, or mRNA expression-level differences).

After comparing all the current sequencing approaches, Hyb-Seq showed the most potential for the investigation of polyploid groups that may contain a large number of species with various ploidy levels. This method is also useful for including specimen samples with degraded or low-quality DNA. By combining a universal bait set with a taxon-specific bait set, Hyb-Seq can capture genome-wide loci from distant or closely related polyploid taxa. In addition, SNPs among targeted genes can be extracted and analyzed from Hyb-Seq data. This method also significantly decreases the cost by adding multiple individuals in one high-throughput sequencing pool while acquiring enough sequencing depth for each locus; however, it is worth noting that adding a taxon-specific bait set is crucial for achieving greater resolution among polyploid species that are closely related. Meanwhile, adapting the Hyb-Seq approach with third-generation sequencing, such as PacBio, by increasing the fragment length of each genomic library (Figure 2) will eventually improve orthologous gene assembly and homeolog phasing.

Finally, most bioinformatic tools are still limited by the requirement of having related or parental diploids, a small number of input taxa, lower ploidy levels (e.g., 4x), or strictly only two subgenome donors. In the future, improving bioinformatic pipelines to correctly assemble orthologous genes and perform phasing for homeologous copies, as well as using the phased MUL-trees to infer networks and to tackle multiple origins of polyploids, will all further the phylogenomic inference of polyploid lineages.

## AUTHOR CONTRIBUTIONS

W.N. conceived the review and wrote the paper with conceptual input and editing by H.M.M. and J.A.T. All authors approved the final version of the manuscript.

## ACKNOWLEDGMENTS

The authors thank Carl J. Rothfels (Utah State University) and Rob Smissen (Manaaki Whenua – Landcare Research) for their suggestions on an earlier version of the manuscript. This study was conducted as part of a Ph.D. thesis by W.N., which was funded by the Royal Society of New Zealand Marsden fund (17-LCR-006) to H.M.M. and J.A.T. Open access publishing facilitated by Massey University, as part of the Wiley - Massey University agreement via the Council of Australian University Librarians.

## ORCID

Weixuan Ning  <http://orcid.org/0000-0002-2013-8723>

Heidi M. Meudt  <http://orcid.org/0000-0002-2433-9071>

Jennifer A. Tate  <http://orcid.org/0000-0001-5138-2115>

## REFERENCES

- Adams, K. L., R. Cronn, R. Percifield, and J. F. Wendel. 2003. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proceedings of the National Academy of Sciences, USA* 100: 4649–4654.
- Álvarez, I., and J. F. Wendel. 2003. Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution* 29: 417–434.
- Andermann, T., M. F. Torres Jiménez, P. Matos-Maraví, R. Batista, J. L. Blanco-Pastor, A. L. S. Gustafsson, L. Kistler, et al. 2020. A guide to carrying out a phylogenomic target sequence capture project. *Frontiers in Genetics* 10: 1407.
- Andrews, K. R., J. M. Good, M. R. Miller, G. Luikart, and P. A. Hohenlohe. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics* 17: 81–92.
- 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.
- Baldwin, B. G., M. J. Sanderson, J. M. Porter, M. F. Wojciechowski, C. S. Campbell, and M. J. Donoghue. 1995. The ITS region of nuclear ribosomal DNA: A valuable source of evidence on angiosperm phylogeny. *Annals of the Missouri Botanical Garden* 82(2): 247–277.
- Barnard-Kubow, K. B., M. A. McCoy, and L. F. Galloway. 2017. Biparental chloroplast inheritance leads to rescue from cytonuclear incompatibility. *New Phytologist* 213: 1466–1476.
- Bayer, P. E., A. A. Golicz, A. Scheben, J. Batley, and D. Edwards. 2020. Plant pan-genomes are the new reference. *Nature Plants* 6: 914–920.
- Belton, J.-M., R. P. McCord, J. H. Gibcus, N. Naumova, Y. Zhan, and J. Dekker. 2012. Hi-C: A comprehensive technique to capture the conformation of genomes. *Methods* 58: 268–276.
- Birky, C. W. 1995. Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and evolution. *Proceedings of the National Academy of Sciences, USA* 92: 11331–11338.
- Blischak, P. D., M. Latvis, D. F. Morales-Briones, J. C. Johnson, V. S. Di Stilio, A. D. Wolfe, and D. C. Tank. 2018. Fluidigm2PURC: Automated processing and haplotype inference for double-barcoded PCR amplicons. *Applications in Plant Sciences* 6: e01156.
- Borsch, T., D. Quandt, and M. Koch. 2009. Molecular evolution and phylogenetic utility of non-coding DNA: Applications from species to deep level questions. *Plant Systematics and Evolution* 282: 107–108.
- Bouckaert, R. R. 2010. DensiTree: Making sense of sets of phylogenetic trees. *Bioinformatics* 26: 1372–1373.

- Bouckaert, R., J. Heled, D. Kühnert, T. Vaughan, C.-H. Wu, D. Xie, M. A. Suchard, et al. 2014. BEAST 2: A software platform for Bayesian evolutionary analysis. *PLoS Computational Biology* 10: e1003537.
- Bozan, I., S. R. Achakgari, N. L. Anglin, D. Ellis, H. H. Tai, and M. V. Strömvik. 2023. Pangenome analyses reveal impact of transposable elements and ploidy on the evolution of potato species. *Proceedings of the National Academy of Sciences, USA* 120: e2211117120.
- 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.
- Brittingham, H. A., M. H. Koski, and T.-L. Ashman. 2018. Higher ploidy is associated with reduced range breadth in the Potentilleae tribe. *American Journal of Botany* 105: 700–710.
- Bryant, D., R. Bouckaert, J. Felsenstein, N. A. Rosenberg, and A. RoyChoudhury. 2012. Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution* 29: 1917–1932.
- Brysting, A. K., B. Oxelman, K. T. Huber, V. Moulton, and C. Brochmann. 2007. Untangling complex histories of genome mergings in high polyploids. *Systematic Biology* 56: 467–476.
- Brysting, A. K., C. Mathiesen, and T. Marcussen. 2011. Challenges in polyploid phylogenetic reconstruction: A case story from the arctic-alpine *Cerastium alpinum* complex. *Taxon* 60: 333–347.
- Bürki, T., V. Pulver, S. Grünig, M. Čertner, and C. Parisod. 2023. Adaptive differentiation on serpentine soil in diploid versus autotetraploid populations of *Biscutella laevigata* (Brassicaceae). *Oikos* e09834. <https://doi.org/10.1111/oik.09834>
- Bushnell, B. 2023. BBMap. Website: <https://sourceforge.net/projects/bbmap/> [accessed 7 March 2024].
- Cerca, J., B. Petersen, J. M. Lazaro-Guevara, A. Rivera-Colón, S. Birkeland, J. Vizueta, S. Li, et al. 2022. The genomic basis of the plant island syndrome in Darwin's giant daisies. *Nature Communications* 13: 3729.
- Chamala, S., N. García, G. T. Godden, V. Krishnakumar, I. E. Jordon-Thaden, R. De Smet, W. B. Barbazuk, et al. 2015. MarkerMiner 1.0: A new application for phylogenetic marker development using angiosperm transcriptomes. *Applications in Plant Sciences* 3: 1400115.
- Chen, Z. J. 2007. Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. *Annual Review of Plant Biology* 58: 377–406.
- Chen, Z. J., A. Sreedasyam, A. Ando, Q. Song, L. M. De Santiago, A. M. Hulse-Kemp, M. Ding, et al. 2020. Genomic diversifications of five *Gossypium* allopolyploid species and their impact on cotton improvement. *Nature Genetics* 52: 525–533.
- Cheon, S., J. Zhang, and C. Park. 2020. Is phylotranscriptomics as reliable as phylogenomics? *Molecular Biology and Evolution* 37: 3672–3683.
- Chifman, J., and L. Kubatko. 2014. Quartet inference from SNP data under the coalescent model. *Bioinformatics* 30: 3317–3324.
- Clarkson, J. J., S. Dodsworth, and M. W. Chase. 2017. Time-calibrated phylogenetic trees establish a lag between polyploidisation and diversification in *Nicotiana* (Solanaceae). *Plant Systematics and Evolution* 303: 1001–1012.
- Comai, L. 2005. The advantages and disadvantages of being polyploid. *Nature Reviews Genetics* 6: 836–846.
- Crowl, A. A., P. W. Fritsch, G. P. Tiley, N. P. Lynch, T. G. Ranney, H. Ashrafi, and P. S. Manos. 2022. A first complete phylogenomic hypothesis for diploid blueberries (*Vaccinium* section *Cyanococcus*). *American Journal of Botany* 109: 1596–1606.
- Cui, L., P. K. Wall, J. H. Leebens-Mack, B. G. Lindsay, D. E. Soltis, J. J. Doyle, P. S. Soltis, et al. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Research* 16: 738–749.
- Czabarka, É., P. L. Erdős, V. Johnson, and V. Moulton. 2013. Generating functions for multi-labeled trees. *Discrete Applied Mathematics* 161: 107–117.
- Davey, J. W., P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen, and M. L. Blaxter. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* 12: 499–510.
- De Bodt, S., S. Maere, and Y. Van de Peer. 2005. Genome duplication and the origin of angiosperms. *Trends in Ecology & Evolution* 20: 591–597.
- de Lima Ferreira, P., R. Batista, T. Andermann, M. Groppo, C. D. Bacon, and A. Antonelli. 2022. Target sequence capture of Barnadesioideae (Compositae) demonstrates the utility of low coverage loci in phylogenomic analyses. *Molecular Phylogenetics and Evolution* 169: 107432.
- Deb, S. K., P. P. Edger, J. C. Pires, and M. R. McKain. 2023. Patterns, mechanisms, and consequences of homoeologous exchange in allopolyploid angiosperms: A genomic and epigenomic perspective. *New Phytologist* 238: 2284–2304.
- Debray, K., M.-C. Le Paslier, A. Bérard, T. Thouroude, G. Michel, J. Marie-Magdelaine, A. Bruneau, et al. 2021. Unveiling the patterns of reticulated evolutionary processes with phylogenomics: Hybridization and polyploidy in the genus *Rosa*. *Systematic Biology* 71: 547–569.
- Degnan, J. H., and N. A. Rosenberg. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution* 24: 332–340.
- DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* 43: 491–498.
- Dodsworth, S., M. W. Chase, and A. R. Leitch. 2016. Is post-polyploidization diploidization the key to the evolutionary success of angiosperms? *Botanical Journal of the Linnean Society* 180: 1–5.
- Dodsworth, S., T.-S. Jang, M. Struebig, M. W. Chase, H. Weiss-Schneeweiss, and A. R. Leitch. 2017. Genome-wide repeat dynamics reflect phylogenetic distance in closely related allotetraploid *Nicotiana* (Solanaceae). *Plant Systematics and Evolution* 303: 1013–1020.
- Eaton, D. A. R., and I. Overcast. 2020. ipyrad: Interactive assembly and analysis of RADseq datasets. *Bioinformatics* 36: 2592–2594.
- Eaton, D. A. R., E. L. Spriggs, B. Park, and M. J. Donoghue. 2017. Misconceptions on missing data in RAD-seq phylogenetics with a deep-scale example from flowering plants. *Systematic Biology* 66: 399–412.
- Ebersberger, I., S. Strauss, and A. von Haeseler. 2009. HaMStR: Profile hidden markov model based search for orthologs in ESTs. *BMC Evolutionary Biology* 9: 157.
- Edger, P. P., M. R. McKain, K. A. Bird, and R. VanBuren. 2018. Subgenome assignment in allopolyploids: Challenges and future directions. *Current Opinion in Plant Biology* 42: 76–80.
- Emms, D. M., and S. Kelly. 2015. OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* 16: 157.
- Eriksson, J. S., F. de Sousa, Y. J. K. Bertrand, A. Antonelli, B. Oxelman, and B. E. Pfeil. 2018. Allele phasing is critical to revealing a shared allopolyploid origin of *Medicago arborea* and *M. strasseri* (Fabaceae). *BMC Evolutionary Biology* 18: 9.
- Fér, T., and R. E. Schmickl. 2018. HybPhyloMaker: Target enrichment data analysis from raw reads to species trees. *Evolutionary Bioinformatics* 14: 1176934317742613.
- Freudenthal, J. A., S. Pfaff, N. Terhoeven, A. Korte, M. J. Ankenbrand, and F. Förster. 2020. A systematic comparison of chloroplast genome assembly tools. *Genome Biology* 21: 254.
- Freyman, W. A., M. G. Johnson, and C. J. Rothfels. 2023. Homologizer: Phylogenetic phasing of gene copies into polyploid subgenomes. *Methods in Ecology and Evolution* 14: 1230–1244.
- Frost, L. A., N. O'Leary, L. P. Lagomarsino, D. C. Tank, and R. G. Olmstead. 2021. Phylogeny, classification, and character evolution of tribe Citharexyleae (Verbenaceae). *American Journal of Botany* 108: 1982–2001.
- Gardner, E. M., M. G. Johnson, J. T. Pereira, A. S. A. Puad, D. Arifiani, Sahromi, N. J. Wickett, and N. J. C. Zerega. 2021. Paralogs and off-target sequences improve phylogenetic resolution in a densely sampled study of the breadfruit genus (*Artocarpus*, Moraceae). *Systematic Biology* 70: 558–575.

- Gaut, B. S. 1998. Molecular clocks and nucleotide substitution rates in higher plants. In M. K. Hecht, R. J. Macintyre, and M. T. Clegg [eds.], *Evolutionary Biology*, 93–120. Springer, Boston, Massachusetts, USA.
- Glover, N. M., H. Redestig, and C. Dessimoz. 2016. Homoeologs: What are they and how do we infer them? *Trends in Plant Science* 21: 609–621.
- Gordon, S. P., B. Contreras-Moreira, J. J. Levy, A. Djamei, A. Czedik-Eysenberg, V. S. Tartaglio, A. Session, et al. 2020. Gradual polyploid genome evolution revealed by pan-genomic analysis of *Brachypodium hybridum* and its diploid progenitors. *Nature Communications* 11: 3670.
- Guo, C., Y. Luo, L.-M. Gao, T.-S. Yi, H.-T. Li, J.-B. Yang, and D.-Z. Li. 2023. Phylogenomics and the flowering plant tree of life. *Journal of Integrative Plant Biology* 65: 299–323.
- Han, T.-S., Q.-J. Zheng, R. E. Onstein, B. M. Rojas-Andrés, F. Hauenschild, A. N. Muellner-Riehl, and Y.-W. Xing. 2020. Polyploidy promotes species diversification of *Allium* through ecological shifts. *New Phytologist* 225: 571–583.
- Heled, J., and A. J. Drummond. 2010. Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution* 27: 570–580.
- 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.
- Hibbins, M. S., and M. W. Hahn. 2022. Phylogenomic approaches to detecting and characterizing introgression. *Genetics* 220: iyab173. <https://doi.org/10.1093/genetics/iyab220>
- Hillis, D. M., and M. T. Dixon. 1991. Ribosomal DNA: molecular evolution and phylogenetic inference. *Quarterly Review of Biology* 66: 411–453.
- Hojsgaard, D., and E. Hörandl. 2015. A little bit of sex matters for genome evolution in asexual plants. *Frontiers in Plant Science* 6: 82. <https://doi.org/10.3389/fpls.2015.00082>
- Holder, M., and P. O. Lewis. 2003. Phylogeny estimation: Traditional and Bayesian approaches. *Nature Reviews Genetics* 4: 275–284.
- Hon, T., K. Mars, G. Young, Y.-C. Tsai, J. W. Karalius, J. M. Landolin, N. Maurer, et al. 2020. Highly accurate long-read HiFi sequencing data for five complex genomes. *Scientific Data* 7: 399.
- Huang, C.-C., K.-H. Hung, W.-K. Wang, C.-W. Ho, C.-L. Huang, T.-W. Hsu, N. Osada, et al. 2012. Evolutionary rates of commonly used nuclear and organelle markers of *Arabidopsis* relatives (Brassicaceae). *Gene* 499: 194–201.
- Huber, K. T., M. Lott, V. Moulton, and A. Spillner. 2008. The complexity of deriving multi-labeled trees from bipartitions. *Journal of Computational Biology* 15: 639–651.
- Huson, D. H., and D. Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23: 254–267.
- Huson, D. H., and C. Scornavacca. 2012. Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. *Systematic Biology* 61: 1061–1067.
- 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.
- Jiao, W.-B., and K. Schneeberger. 2017. The impact of third generation genomic technologies on plant genome assembly. *Current Opinion in Plant Biology* 36: 64–70.
- Jiao, Y., N. J. Wickett, S. Ayyampalayam, A. S. Chandrabali, L. Landherr, P. E. Ralph, L. P. Tomsho, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473: 97–100.
- Jin, J.-J., W.-B. Yu, J.-B. Yang, Y. Song, C. W. DePamphilis, T.-S. Yi, and D.-Z. Li. 2020. GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biology* 21: 241.
- 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. Pokorný, S. Dodsworth, L. R. Botigué, R. S. Cowan, A. Devault, W. L. Eiserhardt, et al. 2018. 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.
- Jonas, M.-R., J. G. Burleigh, and M. S. Erin. 2023. Target capture methods offer insight into the evolution of rapidly diverged taxa and resolve allopolyploid homeologs in the fern genus *Polypodium* s.s. *Systematic Botany* 48: 96–109.
- Jones, G., S. Sagitov, and B. Oxelman. 2013. Statistical inference of allopolyploid species networks in the presence of incomplete lineage sorting. *Systematic Biology* 62: 467–478.
- Joshi, P., H. Ansari, R. Dickson, N. W. Ellison, C. Skema, and J. A. Tate. 2023. Polyploidy on islands – concerted evolution and gene loss amid chromosomal stasis. *Annals of Botany* 131: 33–44.
- Kapli, P., Z. Yang, and M. J. Telford. 2020. Phylogenetic tree building in the genomic age. *Nature Reviews Genetics* 21: 428–444.
- Karbstein, K., S. Tomasello, L. Hodač, N. Wagner, P. Marinček, B. H. Barke, C. Paetzold, and E. Hörandl. 2022. Untying Gordian knots: Unraveling reticulate polyploid plant evolution by genomic data using the large *Ranunculus auricomus* species complex. *New Phytologist* 235: 2081–2098.
- Karimi, N., C. E. Grover, J. P. Gallagher, J. F. Wendel, C. Ané, and D. A. Baum. 2020. Reticulate evolution helps explain apparent homoplasy in floral biology and pollination in baobabs (*Adansonia*; Bombacoideae; Malvaceae). *Systematic Biology* 69: 462–478.
- Kong, W., Y. Wang, S. Zhang, J. Yu, and X. Zhang. 2023. Recent advances in assembly of plant complex genomes. *Genomics, Proteomics & Bioinformatics* 21(3): 427–439.
- Koo, H., G.-W. Lee, S.-R. Ko, S. Go, S.-Y. Kwon, Y.-M. Kim, and A.-Y. Shin. 2023. Two long read-based genome assembly and annotation of polyploidy woody plants, *Hibiscus syriacus* L. using PacBio and Nanopore platforms. *Scientific Data* 10: 713.
- Kubatko, L. S., and J. H. Degnan. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology* 56: 17–24.
- Kyriakidou, M., H. H. Tai, N. L. Anglin, D. Ellis, and M. V. Strömviik. 2018. Current strategies of polyploid plant genome sequence assembly. *Frontiers in Plant Science* 9: 1660.
- Leaché, A. D., and J. R. Oaks. 2017. The utility of single nucleotide polymorphism (SNP) data in phylogenetics. *Annual Review of Ecology, Evolution and Systematics* 48: 69–84.
- Lei, L., E. Goltsman, D. Goodstein, G. A. Wu, D. S. Rokhsar, and J. P. Vogel. 2021. Plant pan-genomics comes of age. *Annual Review of Plant Biology* 72: 411–435.
- Leitch, I. J., and M. D. Bennett. 2004. Genome downsizing in polyploid plants. *Biological Journal of the Linnean Society* 82: 651–663.
- Leung, S. K., A. R. Jeffries, I. Castanho, B. T. Jordan, K. Moore, J. P. Davies, E. L. Dempster, et al. 2021. Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing. *Cell Reports* 37: 110022.
- Li, W., and A. Godzik. 2006. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22: 1658–1659.
- Li, Z., and M. S. Barker. 2020. Inferring putative ancient whole-genome duplications in the 1000 Plants (1KP) initiative: Access to gene family phylogenies and age distributions. *GigaScience* 9(2): g1aa004.
- Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, et al. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25: 2078–2079.
- Li, Z., M. T. W. McKibben, G. S. Finch, P. D. Blischak, B. L. Sutherland, and M. S. Barker. 2021. Patterns and processes of diploidization in land plants. *Annual Review of Plant Biology* 72: 387–410.
- Liu, B. B., Z. Y. Ma, C. Ren, R. G. Hodel, M. Sun, X. Q. Liu, G. N. Liu, 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. *Journal of Systematics and Evolution* 59: 1124–1138.
- Lu, W.-X., X.-Y. Hu, Z.-Z. Wang, and G.-Y. Rao. 2022. Hyb-Seq provides new insights into the phylogeny and evolution of the *Chrysanthemum zawadskii* species complex in China. *Cladistics* 38: 663–683.

- Maddison, W. P. 1997. Gene trees in species trees. *Systematic Biology* 46: 523–536.
- Maddison, W. P., and L. L. Knowles. 2006. Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology* 55: 21–30.
- Malinsky, M., E. Trucchi, D. J. Lawson, and D. Falush. 2018. RADpainter and fineRADstructure: Population Inference from RADseq data. *Molecular Biology and Evolution* 35: 1284–1290.
- Malinsky, M., M. Matschiner, and H. Svardal. 2021. Dsuite-fast D-statistics and related admixture evidence from VCF files. *Molecular Ecology Resources* 21: 584–595.
- Mallet, J., N. Besansky, and M. W. Hahn. 2016. How reticulated are species? *BioEssays* 38: 140–149.
- Mayrose, I., S. H. Zhan, C. J. Rothfels, N. Arrigo, M. S. Barker, L. H. Rieseberg, and S. P. Otto. 2015. Methods for studying polyploid diversification and the dead end hypothesis: a reply to Soltis et al. (2014). *New Phytologist* 206: 27–35.
- McKain, M. R., M. G. Johnson, S. Uribe-Convers, D. Eaton, and Y. Yang. 2018. Practical considerations for plant phylogenomics. *Applications in Plant Sciences* 6: e1038.
- McLay, T. G. B., J. L. Birch, B. F. Gunn, W. Ning, J. A. Tate, L. Nauheimer, E. M. Joyce, et al. 2021. New targets acquired: Improving locus recovery from the Angiosperms353 probe set. *Applications in Plant Sciences* 9: e11420.
- Meleshko, O., M. D. Martin, T. S. Korneliussen, C. Schröck, P. Lamkowski, J. Schmutz, A. Healey, et al. 2021. Extensive genome-wide phylogenetic discordance is due to incomplete lineage sorting and not ongoing introgression in a rapidly radiated bryophyte genus. *Molecular Biology and Evolution* 38: 2750–2766.
- Meudt, H. M., B. M. Rojas-Andrés, J. M. Prebble, E. Low, P. J. Garnock-Jones, and D. C. Albach. 2015. Is genome downsizing associated with diversification in polyploid lineages of *Veronica*? *Botanical Journal of the Linnean Society* 178: 243–266.
- Meudt, H. M., D. C. Albach, A. J. Tanentzap, J. Igea, S. C. Newmarch, A. J. Brandt, W. G. Lee, and J. A. Tate. 2021. Polyploidy on islands: Its emergence and importance for diversification. *Frontiers in Plant Science* 12: 637214.
- Michael, S. B., A. Nils, E. B. Anthony, L. Zheng, and A. L. Donald. 2016. On the relative abundance of autopolyploids and allopolyploids. *New Phytologist* 210: 391–398.
- Michel, T., Y.-H. Tseng, H. Wilson, K.-F. Chung, and C. Kidner. 2022. A hybrid capture bait set for *Begonia*. *Edinburgh Journal of Botany* 79: 409.
- Minh, B. Q., H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. Von Haeseler, and R. Lanfear. 2020a. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution* 37: 1530–1534.
- Minh, B. Q., M. W. Hahn, and R. Lanfear. 2020b. New methods to calculate concordance factors for phylogenomic datasets. *Molecular Biology and Evolution* 37: 2727–2733.
- Mirarab, S., R. Reaz, M. S. Bayzid, T. Zimmermann, M. S. Swenson, and T. Warnow. 2014. ASTRAL: Genome-scale coalescent-based species tree estimation. *Bioinformatics* 30: 541–548.
- Moraes, A. P., T. B. J. Engel, E. R. Forni-Martins, F. de Barros, L. P. Felix, and J. S. Cabral. 2022. Are chromosome number and genome size associated with habit and environmental niche variables? Insights from the Neotropical orchids. *Annals of Botany* 130: 11–25.
- Morales-Briones, D. F., G. Kadereit, D. T. Tefarikis, M. J. Moore, S. A. Smith, S. F. Brockington, A. Timoneda, et al. 2021. Disentangling sources of gene tree discordance in phylogenomic data sets: Testing ancient hybridizations in Amaranthaceae sl. *Systematic Biology* 70: 219–235.
- Morrison, D. 2014. Phylogenetic networks: A review of methods to display evolutionary history. *Annual Research & Review in Biology* 4: 1518–1543.
- 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.
- Nieto Feliner, G., J. Casacuberta, and J. F. Wendel. 2020. Genomics of evolutionary novelty in hybrids and polyploids. *Frontiers in Genetics* 11: 792.
- Ogilvie, H. A., R. R. Bouckaert, and A. J. Drummond. 2017. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. *Molecular Biology and Evolution* 34: 2101–2114.
- One Thousand Plant Transcriptomes Initiative. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574: 679–685.
- Osuna-Mascaró, C., R. R. de Casas, M. Berbel, J. M. Gómez, and F. Perfectti. 2022. Lack of ITS sequence homogenization in *Erysimum* species (Brassicaceae) with different ploidy levels. *Scientific Reports* 12: 16907.
- Oxelman, B., A. K. Brysting, G. R. Jones, T. Marcussen, C. Oberprieler, and B. E. Pfeil. 2017. Phylogenetics of allopolyploids. *Annual Review of Ecology, Evolution, and Systematics* 48: 543–557.
- Paterson, A. H., J. F. Wendel, H. Gundlach, H. Guo, J. Jenkins, D. Jin, D. Llewellyn, et al. 2012. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature* 492: 423–427.
- Patterson, N., P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, et al. 2012. Ancient admixture in human history. *Genetics* 192: 1065–1093.
- Patterson, M., T. Marschall, N. Pisanti, L. van Iersel, L. Stougie, G. W. Klau, and A. Schönhuth. 2015. WhatsHap: Weighted haplotype assembly for future-generation sequencing reads. *Journal of Computational Biology* 22: 498–509.
- Paun, O., T. F. Stuessy, and E. Hörandl. 2006. The role of hybridization, polyploidization and glaciation in the origin and evolution of the apomictic *Ranunculus cassubicus* complex. *New Phytologist* 171: 223–236.
- Pease, J. B., J. W. Brown, J. F. Walker, C. E. Hinchliff, and S. A. Smith. 2018. Quartet Sampling distinguishes lack of support from conflicting support in the green plant tree of life. *American Journal of Botany* 105: 385–403.
- Pellino, M., D. Hojsgaard, T. Schmutzer, U. Scholz, E. Hörandl, H. Vogel, and T. F. Sharbel. 2013. Asexual genome evolution in the apomictic *Ranunculus auricomus* complex: Examining the effects of hybridization and mutation accumulation. *Molecular Ecology* 22: 5908–5921.
- Pelosi, J. A., E. H. Kim, W. B. Barbazuk, and E. B. Sessa. 2022. Phylotranscriptomics illuminates the placement of whole genome duplications and gene retention in ferns. *Frontiers in Plant Science* 13: 882441.
- Peng, Y., H. Yan, L. Guo, C. Deng, C. Wang, Y. Wang, L. Kang, et al. 2022. Reference genome assemblies reveal the origin and evolution of allohexaploid oat. *Nature Genetics* 54: 1248–1258.
- Peralta, M., M.-C. Combes, A. Cenci, P. Lashermes, and A. Dereeper. 2013. SnpIoid: A utility to exploit high-throughput SNP data derived from RNA-seq in allopolyploid species. *International Journal of Plant Genomics* 2013: 890123.
- Phang, A., F. F. Pezzini, D. F. R. P. Burslem, G. S. Khew, D. J. Middleton, M. Ruhsam, and P. Wilkie. 2023. Target capture sequencing for phylogenomic and population studies in the Southeast Asian genus *Palaquium* (Sapotaceae). *Botanical Journal of the Linnean Society* 203: 134–147.
- Pleines, T., S. S. Jakob, and F. R. Blattner. 2009. Application of non-coding DNA regions in intraspecific analyses. *Plant Systematics and Evolution* 282: 281–294.
- Postel, Z., and P. Touzet. 2020. Cytonuclear genetic incompatibilities in plant speciation. *Plants* 9: 487.
- Prince, V. E., and F. B. Pickett. 2002. Splitting pairs: The diverging fates of duplicated genes. *Nature Reviews Genetics* 3: 827–837.
- Pucker, B., I. Irisarri, J. de Vries, and B. Xu. 2022. Plant genome sequence assembly in the era of long reads: Progress, challenges and future directions. *Quantitative Plant Biology* 3: e5.
- Qiao, Q., P. P. Edger, L. Xue, L. Qiong, J. Lu, Y. Zhang, Q. Cao, et al. 2021. Evolutionary history and pan-genome dynamics of strawberry

- (*Fragaria* spp.). *Proceedings of the National Academy of Sciences, USA* 118: e2105431118.
- Qiu, F., E. J. Baack, K. D. Whitney, D. G. Bock, H. M. Tetreault, L. H. Rieseberg, and M. C. Ungerer. 2019. Phylogenetic trends and environmental correlates of nuclear genome size variation in *Helianthus* sunflowers. *New Phytologist* 221: 1609–1618.
- Qiu, T., Z. Liu, and B. Liu. 2020. The effects of hybridization and genome doubling in plant evolution via allopolyploidy. *Molecular Biology Reports* 47: 5549–5558.
- Rabier, C.-E., V. Berry, M. Stoltz, J. D. Santos, W. Wang, J.-C. Glaszmann, F. Pardi, and C. Scornavacca. 2021. On the inference of complex phylogenetic networks by Markov Chain Monte-Carlo. *PLoS Computational Biology* 17: e1008380.
- Ravi, V., J. P. Khurana, A. K. Tyagi, and P. Khurana. 2008. An update on chloroplast genomes. *Plant Systematics and Evolution* 271: 101–122.
- Rothfels, C. J. 2021. Polyploid phylogenetics. *New Phytologist* 230: 66–72.
- Rothfels, C. J., K. M. Pryer, and F. W. Li. 2017. Next-generation polyploid phylogenetics: Rapid resolution of hybrid polyploid complexes using PacBio single-molecule sequencing. *New Phytologist* 213: 413–429.
- Saada, O. A., A. Friedrich, and J. Schacherer. 2022. Towards accurate, contiguous and complete alignment-based polyploid phasing algorithms. *Genomics* 114: 110369.
- Sang, T. 2002. Utility of low-copy nuclear gene sequences in plant phylogenetics. *Critical Reviews in Biochemistry and Molecular Biology* 37: 121–147.
- Schafraan, P., F.-W. Li, and C. J. Rothfels. 2023. PURC provides improved sequence inference for polyploid phylogenetics and other manifestations of the multiple-copy problem. In Y. Van de Peer [ed], *Polyploidy: Methods and Protocols*, 189–206. Springer, New York, New York, USA.
- Schmickl, R., A. Liston, V. Zeisek, K. Oberlander, K. Weitemier, S. C. K. Straub, R. C. Cronn, et al. 2016. Phylogenetic marker development for target enrichment from transcriptome and genome skim data: The pipeline and its application in southern African *Oxalis* (Oxalidaceae). *Molecular Ecology Resources* 16: 1124–1135.
- Shan, S., J. L. Boatwright, X. Liu, A. S. Chanderbali, C. Fu, P. S. Soltis, and D. E. Soltis. 2020. Transcriptome dynamics of the inflorescence in reciprocally formed allopolyploid *Tragopogon miscellus* (Asteraceae). *Frontiers in Genetics* 11: 888. Available from: <https://doi.org/10.3389/fgene.2020.00888>
- Sharbrough, J., J. L. Conover, J. A. Tate, J. F. Wendel, and D. B. Sloan. 2017. Cytonuclear responses to genome doubling. *American Journal of Botany* 104: 1277–1280.
- Shaw, J., H. L. Shafer, O. R. Leonard, M. J. Kovach, M. Schorr, and A. B. Morris. 2014. Chloroplast DNA sequence utility for the lowest phylogenetic and phylogeographic inferences in angiosperms: The tortoise and the hare IV. *American Journal of Botany* 101: 1987–2004.
- Siniscalchi, C. M., O. Hidalgo, L. Palazzesi, J. Pellicer, L. Pokorný, O. Maurin, I. J. Leitch, et al. 2021. Lineage-specific vs. universal: A comparison of the Compositae1061 and Angiosperms353 enrichment panels in the sunflower family. *Applications in Plant Sciences* 9: e11422.
- Šlenker, M., A. Kantor, K. Marhold, R. Schmickl, T. Mandáková, M. A. Lysak, M. Perný, et al. 2021. Allele sorting as a novel approach to resolving the origin of allotetraploids using Hyb-Seq data: A case study of the Balkan Mountain endemic *Cardamine barbaraeoides*. *Frontiers in Plant Science* 12: 659275. Available from: <https://doi.org/10.3389/fpls.2021.659275>
- Slimp, M., L. D. Williams, H. Hale, and M. G. Johnson. 2021. On the potential of Angiosperms353 for population genomic studies. *Applications in Plant Sciences* 9: e11419.
- Small, R. L., R. C. Cronn, and J. F. Wendel. 2004. Use of nuclear genes for phylogeny reconstruction in plants. *Australian Systematic Botany* 17: 145–170.
- 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: 150.
- Solís-Lemus, C., P. Bastide, and C. Ané. 2017. PhyloNetworks: A package for phylogenetic networks. *Molecular Biology and Evolution* 34: 3292–3298.
- Soltis, D. E., and P. S. Soltis. 1999. Polyploidy: Recurrent formation and genome evolution. *Trends in Ecology & Evolution* 14: 348–352.
- Soltis, P. S., and D. E. Soltis. 2009. The role of hybridization in plant speciation. *Annual Review of Plant Biology* 60: 561–588.
- Soltis, D. E., P. S. Soltis, D. W. Schemske, J. F. Hancock, J. N. Thompson, B. C. Husband, and W. S. Judd. 2007. Autopolyploidy in angiosperms: Have we grossly underestimated the number of species? *Taxon* 56: 13–30.
- Soltis, D. E., C. J. Visger, and P. S. Soltis. 2014. The polyploidy revolution then...and now: Stebbins revisited. *American Journal of Botany* 101: 1057–1078.
- Soltis, P. S., D. B. Marchant, Y. Van de Peer, and D. E. Soltis. 2015. Polyploidy and genome evolution in plants. *Current Opinion in Genetics & Development* 35: 119–125.
- Stebbins, G. L. 1980. Polyploidy in plants: Unsolved problems and prospects. In W. H. Lewis [ed.], *Polyploidy: Biological Relevance*, 495–520. Springer, Boston, Massachusetts, USA.
- Straub, S. C., M. Parks, K. Weitemier, M. Fishbein, R. C. Cronn, and A. Liston. 2012. Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. *American Journal of Botany* 99: 349–364.
- Suissa, J. S., S. P. Kinosian, P. W. Schafran, J. F. Bolin, W. C. Taylor, and E. A. Zimmer. 2022. Homoploid hybrids, allopolyploids, and high ploidy levels characterize the evolutionary history of a western North American quillwort (*Isoetes*) complex. *Molecular Phylogenetics and Evolution* 166: 107332.
- Symonds, V. V., P. S. Soltis, and D. E. Soltis. 2010. Dynamics of polyploid formation in *Tragopogon* (Asteraceae): Recurrent formation, gene flow, and population structure. *Evolution* 64: 1984–2003.
- Takamatsu, T., M. Baslam, T. Inomata, K. Oikawa, K. Itoh, T. Ohnishi, T. Kinoshita, and T. Mitsui. 2018. Optimized method of extracting rice chloroplast DNA for high-quality plastome resequencing and *de novo* assembly. *Frontiers in Plant Science* 9: 266. Available from: <https://doi.org/10.3389/fpls.2018.00266>
- Than, C., D. Ruths, and L. Nakhleh. 2008. PhyloNet: A software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics* 9: 322.
- Thomas, A. E., J. Igea, H. M. Meudt, D. C. Albach, W. G. Lee, and A. J. Tanentzap. 2021. Using target sequence capture to improve the phylogenetic resolution of a rapid radiation in New Zealand *Veronica*. *American Journal of Botany* 108: 1289–1306.
- Tiley, G. P., A. A. Crowl, P. S. Manos, E. B. Sessa, C. Solís-Lemus, A. D. Yoder, and J. G. Burleigh. 2021. Benefits and limits of phasing alleles for network inference of allopolyploid complexes. *bioRxiv* 2021.05.04.442457 [preprint]. Published 5 November 2023 [accessed 4 March 2024]. Available from: <https://doi.org/10.1101/2021.05.04.442457>
- Tsitrona, A., M. Kirkpatrick, and D. A. Levin. 2003. A model for chloroplast capture. *Evolution* 57: 1776–1782.
- Twyford, A. D., and R. A. Ennos. 2012. Next-generation hybridization and introgression. *Heredity* 108: 179–189.
- Ufimov, R., J. M. Gorospe, T. Fér, M. Kandziara, L. Salomon, M. van Loo, and R. Schmickl. 2022. Utilizing paralogues for phylogenetic reconstruction has the potential to increase species tree support and reduce gene tree discordance in target enrichment data. *Molecular Ecology Resources* 22: 3018–3034.
- Unruh, S. A., M. R. McKain, Y.-I. Lee, T. Yukawa, M. K. McCormick, R. P. Shefferson, A. Smithson, et al. 2018. Phylotranscriptomic analysis and genome evolution of the Cypripedioideae (Orchidaceae). *American Journal of Botany* 105: 631–640.
- Uribe-Convers, S., M. L. Settles, and D. C. Tank. 2016. A phylogenomic approach based on PCR target enrichment and high throughput sequencing: Resolving the diversity within the South American species of *Bartsia* L. (Orobanchaceae). *PLoS ONE* 11: e0148203.

- Van de Peer, Y., E. Mizrahi, and K. Marchal. 2017. The evolutionary significance of polyploidy. *Nature Reviews Genetics* 18: 411–424.
- Vicient, C. M., and J. M. Casacuberta. 2017. Impact of transposable elements on polyploid plant genomes. *Annals of Botany* 120: 195–207.
- Wagner, N. D., L. He, and E. Hörandl. 2020. Phylogenomic relationships and evolution of polyploid *Salix* species revealed by RAD sequencing data. *Frontiers in Plant Science* 11: 36–41.
- Wagner, N. D., P. Marinček, L. Pittet, and E. Hörandl. 2023. Insights into the taxonomically challenging hexaploid alpine shrub willows of *Salix* sections *Phylicifoliae* and *Nigricantes* (Salicaceae). *Plants* 12: 1144.
- Wang, X., J. A. Morton, J. Pellicer, I. J. Leitch, and A. R. Leitch. 2021a. Genome downsizing after polyploidy: Mechanisms, rates and selection pressures. *The Plant Journal* 107: 1003–1015.
- Wang, N., L. J. Kelly, H. A. McAllister, J. Zohren, and R. J. Buggs. 2021b. Resolving phylogeny and polyploid parentage using genus-wide genome-wide sequence data from birch trees. *Molecular Phylogenetics and Evolution* 160: 107126.
- Wang, C., T. Wang, M. Yin, F. Eller, L. Liu, H. Brix, and W. Guo. 2021c. Transcriptome analysis of tetraploid and octoploid common reed (*Phragmites australis*). *Frontiers in Plant Science* 12: 653183. Available from: <https://doi.org/10.3389/fpls.2021.653183>
- Wang, Y., Y. Zhao, A. Bollas, Y. Wang, and K. F. Au. 2021d. Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology* 39: 1348–1365.
- Wang, G., N. Zhou, Q. Chen, Y. Yang, Y. Yang, and Y. Duan. 2023a. Gradual genome size evolution and polyploidy in *Allium* from the Qinghai–Tibetan Plateau. *Annals of Botany* 131: 109–122.
- Wang, T., A. D. J. van Dijk, J. Bucher, J. Liang, J. Wu, G. Bonnema, and X. Wang. 2023b. Interploidy introgression shaped adaptation during the origin and domestication history of *Brassica napus*. *Molecular Biology and Evolution* 40: msad199.
- Wei, N., J. A. Tennesen, A. Liston, and T.-L. Ashman. 2017. Present-day sympatry belies the evolutionary origin of a high-order polyploid. *New Phytologist* 216: 279–290.
- 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: 1400042.
- Xie, M., Q. Wu, J. Wang, and T. Jiang. 2016. H-PoP and H-PoPG: Heuristic partitioning algorithms for single individual haplotyping of polyploids. *Bioinformatics* 32: 3735–3744.
- Xu, B., X.-M. Zeng, X.-F. Gao, D.-P. Jin, and L.-B. Zhang. 2017. ITS non-concerted evolution and rampant hybridization in the legume genus *Lespedeza* (Fabaceae). *Scientific Reports* 7: 40057.
- Yan, Z., M. L. Smith, P. Du, M. W. Hahn, and L. Nakhleh. 2022. Species tree inference methods intended to deal with incomplete lineage sorting are robust to the presence of paralogs. *Systematic Biology* 71: 367–381.
- Yang, Z., and B. Rannala. 2012. Molecular phylogenetics: principles and practice. *Nature Reviews Genetics* 13: 303–314.
- 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.
- Yang, J., D. Liu, X. Wang, C. Ji, F. Cheng, B. Liu, Z. Hu, et al. 2016. The genome sequence of allopolyploid *Brassica juncea* and analysis of differential homoeolog gene expression influencing selection. *Nature Genetics* 48: 1225–1232.
- Yardeni, G., J. Viruel, M. Paris, J. Hess, C. Groot Crego, M. de La Harpe, N. Rivera, et al. 2022. Taxon-specific or universal? Using target capture to study the evolutionary history of rapid radiations. *Molecular Ecology Resources* 22: 927–945.
- Zhang, Y., and P. Ozdemir. 2009. Microfluidic DNA amplification—A review. *Analytica Chimica Acta* 638: 115–125.
- Zhang, J., X. Zhang, H. Tang, Q. Zhang, X. Hua, X. Ma, F. Zhu, et al. 2018a. Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. *Nature Genetics* 50: 1565–1573.
- Zhang, C., H. A. Ogilvie, A. J. Drummond, and T. Stadler. 2018b. Bayesian inference of species networks from multilocus sequence data. *Molecular Biology and Evolution* 35: 504–517.
- Zhang, X., R. Wu, Y. Wang, J. Yu, and H. Tang. 2020a. Unzipping haplotypes in diploid and polyploid genomes. *Computational and Structural Biotechnology Journal* 18: 66–72.
- Zhang, C., C. Scornavacca, E. K. Molloy, and S. Mirarab. 2020b. ASTRAL-Pro: Quartet-based species-tree inference despite paralogy. *Molecular Biology and Evolution* 37: 3292–3307.
- Zhang, C., C.-H. Huang, M. Liu, Y. Hu, J. L. Panero, F. Luebert, T. Gao, and H. Ma. 2021. Phylotranscriptomic insights into Asteraceae diversity, polyploidy, and morphological innovation. *Journal of Integrative Plant Biology* 63: 1273–1293.
- Zhao, L., Y.-Y. Yang, X.-J. Qu, H. Ma, Y. Hu, H.-T. Li, T.-S. Yi, and D.-Z. Li. 2023. Phylotranscriptomic analyses reveal multiple whole-genome duplication events, the history of diversification and adaptations in the Araceae. *Annals of Botany* 131: 199–214.
- Zhou, R., N. Moshgabadi, and K. L. Adams. 2011. Extensive changes to alternative splicing patterns following allopolyploidy in natural and resynthesized polyploids. *Proceedings of the National Academy of Sciences, USA* 108: 16122–16127.
- Zhou, Q., D. Tang, W. Huang, Z. Yang, Y. Zhang, J. P. Hamilton, R. G. F. Visser, et al. 2020a. Haplotype-resolved genome analyses of a heterozygous diploid potato. *Nature Genetics* 52: 1018–1023.
- Zhou, W., Q.-Y. Xiang, and J. Wen. 2020b. Phylogenomics, biogeography, and evolution of morphology and ecological niche of the eastern Asian–eastern North American *Nyssa* (Nyssaceae). *Journal of Systematics and Evolution* 58: 571–603.
- Zhou, W., J. Soghigian, and Q.-Y. Xiang. 2022. A new pipeline for removing paralogs in target enrichment data. *Systematic Biology* 71: 410–425.
- Zhu, J., D. Wen, Y. Yu, H. M. Meudt, and L. Nakhleh. 2018. Bayesian inference of phylogenetic networks from bi-allelic genetic markers. *PLoS Computational Biology* 14: e1005932.

**How to cite this article:** Ning, W., H. M. Meudt, and J. A. Tate. 2024. A roadmap of phylogenomic methods for studying polyploid plant genera. *Applications in Plant Sciences* 12: e11580. <https://doi.org/10.1002/aps3.11580>