Navigating phylogenetic conflict and evolutionary inference in plants with target capture data

E.M. Joyce*[1], A.N. Schmidt-Lebuhn[2], H.K. Orel[3], F.J. Nge[4], B.M. Anderson[5], T.A. Hammer[6,7], and T.G.B. McLay[3,8,9]

[1]Systematik, Biodiversität und Evolution der Pflanzen, Ludwig-Maximilians-Universität München, Menzinger Str. 67, 80638 Munich, Germany

[2]Centre for Australian National Biodiversity Research (a joint venture of Parks Australia and CSIRO), Clunies Ross Street, Canberra ACT 2601, Australia

[3]School of BioSciences, The University of Melbourne, Parkville, Victoria 3010, Australia

[4]National Herbarium of New South Wales, Botanic Gardens of Sydney, Locked Bag 6002, Mount Annan, NSW 2567, Australia

[5]Western Australian Herbarium, Department of Biodiversity, Conservation and Attractions, Locked Bag 104, Bentley Delivery Centre, Bentley, Western Australia 6983, Australia

[6]School of Biological Sciences, The University of Adelaide, Adelaide, South Australia 5005

[7]State Herbarium of South Australia, Botanic Gardens and State Herbarium, Hackney Road, Adelaide, South Australia 5000

[8]National Biodiversity DNA Library, Environomics, CSIRO, Parkville 3010, Victoria, Australia

[9]Royal Botanic Gardens Victoria, Melbourne, Victoria 3004, Australia

*Corresponding author: E.Joyce@lmu.de

Abstract

Target capture has quickly become a preferred approach for plant systematic and evolutionary research, marking a step-change in the generation of data for phylogenetic inference. While this advancement has facilitated the resolution of many phylogenetic relationships, phylogenetic conflict continues to be reported, and often attributed to genome duplication, reticulation, deep coalescence or rapid speciation – processes that are particularly common in plant evolution. The proliferation of methods designed to analyse target capture data in the presence of these processes can be overwhelming for many researchers, especially students. In this review, we guide researchers through the target capture bioinformatic workflow, with a particular focus on robust phylogenetic inference in the presence of conflict. Through the workflow, we highlight key considerations for reducing artefactual conflict, synthesise strategies for managing paralogs, explain the causes and measurement of conflict, and summarise current methods for investigating biological

processes underlying conflict. While we draw from examples in the Australian flora, this review is broadly relevant for any researcher working with target capture data. We conclude that conflict is often inherent and inevitable in plant phylogenetic research, but when properly managed, target capture data can provide unprecedented insight into the extraordinary and complex evolutionary histories of plants.

Keywords: paralogy, polytomy, discordance, incongruence, incomplete lineage sorting, hybridisation, polyploidy, HybSeq, target enrichment, Genomics for Australian Plants, PAFTOL, GAP, Angiosperms353

Introduction

Target capture sequencing (also referred to as target enrichment and HybSeq) has rapidly become a preferred approach for phylogenetic inquiry. In Australian plant systematics, a multitude of data types are used (Nauheimer *et al.* 2019; Fowler *et al.* 2020; Gunn *et al.* 2020, 2024; Orel *et al.* 2023a; Orel *et al.* 2023b), but the ongoing trend points to a greater adoption of target capture sequencing (Fig. 1). This rapid uptake is due to the many advantages that target capture data offers, including the ability to generate large amounts of phylogenetic information, compatibility across datasets using the same bait kits (RNA baits or probes, designed to capture a set of target loci), and the ability to obtain targeted loci from degraded material such as herbarium specimens (Hart *et al.* 2016; Shee *et al.* 2020). It has been further expedited through the establishment of initiatives such as Plant and Fungal Trees of Life (PAFTOL) in 2016 ((Baker *et al.* 2021); https://www.kew.org/science/our-science/projects/plant-and-fungal-trees-of-life) and Genomics for Australian Plants (GAP; https://www.genomicsforaustralianplants.com) in 2017. These ventures coordinated efforts of researchers and institutions to sequence 353 single- or low-copy nuclear loci conserved in angiosperms with the Angiosperms353 (A353) bait kit (Johnson *et al.* 2019), facilitating the generation of the most densely-sampled and data-rich nuclear phylogeny of angiosperms to date (Zuntini *et al.* 2024). Other universal bait kits have also been developed in the past five years, such as the GoFlag bait kit for flagellate plants (Breinholt *et al.* 2021) and the OzBaits kit for Australian plants (Waycott *et al.* 2021), and custom bait kits for particular groups are now commonplace for finer-scale phylogenetic investigation (e.g. Compositae1061, (Siniscalchi *et al.* 2021); (Vatanparast *et al.* 2018)) or when groups have proven challenging to investigate with A353. This has culminated in the production of an unprecedented volume of data for plant phylogenomic research across taxonomic levels within the Australian flora, as detailed in Table 1.

Although target capture has aided the resolution of many previously elusive plant relationships (e.g. Larridon *et al.* 2021; Pillon *et al.* 2021; Schmidt-Lebuhn and Grealy 2024), it has proved not to be the 'silver bullet' for resolving the evolutionary history of many plant groups as well supported bifurcating trees, with 'conflict' continuing to be reported. Conflict, also often referred to as 'discordance' or 'incongruence', refers to when individual loci do not share the same topology with either the species tree or with each other. Such conflict can be the result of contamination during lab work, data artefacts introduced by researchers during analysis, or inherent biases in target capture data (Steenwyk *et al.* 2023; Frost *et al.* 2024). Alternatively, conflict can be the product of real biological processes that cause evolutionary histories of genes and lineages to deviate from each other or from a bifurcating tree. Such processes, like whole-genome duplication (WGD) events, reticulation, and deep coalescence have long been known to be common and important events in the evolution of plants but were difficult to detect in phylogenetic studies prior to high-throughput sequencing techniques. Now, target capture datasets indicate that these processes are pervasive in plants, manifesting as conflict. In the Australian flora, conflict has been attributed to likely WGD events in target capture datasets of *Adenanthos* (Nge, Biffin, *et al.* 2021), *Pomaderris* (Nge, Kellermann, *et al.* 2021), *Calytrix* (Nge *et al.* 2022), *Cryptandra* (Nge *et al.* 2024), *Senecio* (Schmidt-Lebuhn *et al.* 2024), *Celmisiinae* (Nicol *et al.* 2024) and many lineages in Sapindales (Joyce *et al.* 2023). Conflict due to reticulation has been detected in *Adansonia* (Karimi *et al.* 2020) and Thelypteridaceae (Bloesch *et al.* 2022), and reticulation in concert with deep coalescence in *Adenanthos* (Nge, Biffin, *et al.* 2021) and *Eucalyptus* (McLay *et al.* 2023). As illustrated by these examples, conflict in target capture data, if handled carefully, can actually give novel insight into key biological processes in the evolutionary history of plants.

However, the rapid advancement of target capture has also led to a proliferation of software and pipelines for its analysis that can be confusing for those new to the methods, especially students. Faced with this abundance of methods, it can be unclear how to design a bioinformatic pipeline for analysing target capture data in a way that reduces artefactual conflict introduced by the researcher and enables the researcher to test for any biological processes that might underlie any remaining conflict. In this review, we aim to describe key steps in a bioinformatic pipeline with target capture data from the starting point of having raw reads to: 1. Locus extraction; 2. Paralogy reconciliation; 3. Phylogenomic reconstruction of gene trees and species trees; 4. Conflict assessment, and 5. Understanding and investigating patterns an underlying causes of conflict (Fig. 2). While not an exhaustive review of all software available, we highlight key practical considerations for reconstructing and interpreting a phylogeny in the face of conflict. Through the steps of the pipeline we explore what conflict actually is and how to measure it, draw attention to key steps where conflict can be introduced, make recommendations on how to minimise artefactual conflict, and summarise

current approaches for testing for the biological processes of WGD, reticulation, deep coalescence and simultaneous/rapid speciation that may underlie any remaining phylogenetic conflict.

1. Locus extraction

Following the sequencing of enriched libraries and quality control, the targeted loci must be assembled and extracted from the raw reads (Fig. 2). Many methods are now available for this purpose. One of the oldest and most commonly used locus extraction pipelines in plant phylogenomics is HybPiper (Johnson *et al.* 2016), which was developed specifically for the retrieval and assembly of target capture data using A353, and is currently version 2.1.6 (see Table 2). HybPiper uses a read-mapping approach to align raw sequencing reads to reference gene sequences and then assemble those reads into contigs for both exons and their flanking intron regions (Johnson *et al.* 2016). HybPiper was greatly improved through the course of the GAP project, and version 2.0 is much easier to use as either a Python package or container, with improvements in read-mapping, locus assembly, and recovery reporting (Jackson *et al.* 2023). Another program that implements a read-mapping approach is HybPhyloMaker (Fér and Schmickl 2018), which was also written for target capture recovery in plant phylogenomics, but in addition implements phylogenetic reconstruction. An alternative set of methods instead begins by *de novo* assembling all sequencing reads and then retrieving the target loci using reference gene sequences. Such software includes SECAPR (Andermann *et al.* 2020), PHYLUCE (which is more frequently used in animal phylogenomics with ultra-conserved elements; Faircloth 2016), and the recently developed CAPTUS (Ortiz *et al.* 2023). While SECAPR is designed around single-exon targets (one assembled sequence per locus), portions of the pipeline can be adapted for use with gene targets such as A353, which uses probes designed to target multiple exons per locus (which might not all be assembled into a single contiguous piece). Assembly-first methods have the advantage of being able to cluster and extract many off-target loci without a reference target file, facilitating the extraction of additional nuclear and plastid loci for phylogenetic analysis (e.g. Ortiz *et al.* 2023). Both assembly methods require a well designed reference gene sequence file that includes sufficient coverage of the target genes across the phylogenetic scale of interest. For A353, recovery can be improved by expanding the default target file to encompass more phylogenetic breadth (McLay *et al.* 2021). Although comparisons of some locus extraction pipelines have been published (e.g. Zhang *et al.* 2022; Raza *et al.* 2023), a comprehensive comparison of the performance of these methods across lineages and data qualities has not been conducted; as such, multiple methods could be tested on datasets to determine the most optimal and practical pipeline.

Ultimately, the choice of extraction pipeline will depend on the performance (locus recovery), computational efficiency and access to computational resources, as well as the research question. Some questions may require certain downstream analyses that are dependent on the output of particular locus extraction pipelines (such as the extraction and reporting of paralogs, or the raw mapped reads of HybPiper for HybPhaser (Nauheimer *et al.* 2021) (Fig. 2)); in these cases the appropriate extraction pipeline should be used, and extracting loci with multiple pipelines may be warranted. Finally, some locus extraction pipelines offer a workflow for additional steps beyond locus extraction, through to sequence alignment and even tree-estimation (Fér and Schmickl 2018; Ortiz *et al.* 2023). Although these pipelines are user-friendly, we caution against following these workflows without careful consideration and inspection of each step.

Artefactual conflict (as opposed to conflict arising from biological causes) can be introduced by researchers at the locus extraction step in a number of ways. One cause of artefactual conflict can come from the type of short-read assembler implemented in any given locus assembly pipeline. Different assembly pipelines use different short-read assemblers; for example, HybPiper and SECAPR use SPAdes (Bankevich *et al.* 2012), while CAPTUS uses MEGAHIT (Li *et al.* 2015). Each short-read assembler performs differently depending on the pattern of coverage, the presence of highly repetitive regions, GC and AT content, and the structural variation in each dataset (Liao *et al.* 2019). As such, using a suboptimal short-read assembler through the assembly pipeline can contribute to misassemblies and artefactual conflict. Short-read assembling is unavoidable in target capture locus reconstruction (until longer read sequencing becomes more cost effective and efficient for such projects), and currently, the only way to identify short-read assembly problems is to stringently check output sequences and alignments, paying particular attention to alignment gappiness, poorly aligned regions, or assembly error carry-through. Visual inspection of the sequence output for each locus, and each locus alignment can be performed, though this can be time-consuming for a large number of loci and can introduce its own biases and errors. A variety of alignment summary tools are available, such as AMAS (Borowiec 2016) or SEGUL (Handika and Esselstyn 2022). Upon identification of short-read assembly errors, a researcher can try to ameliorate these errors by trying another locus assembly pipeline that uses a different short-read assembler, manually choose sequences unaffected by assembly errors (although this has rarely been done in the literature), or automatically remove errors by cleaning alignments. Many tools are available for alignment cleaning, such as TrimAl (Capella-Gutiérrez *et al.* 2009), ClipKIT (Steenwyk *et al.* 2020) and CIAlign (Tumescheit *et al.* 2022). Determining the correct parameters for these tools requires trial and error. Another option for cleaning alignments is indirectly in downstream phylogenetic analysis, through trimming spuriously long branches in gene trees (potentially indicating erroneous alignments) with tools like TreeShrink (Mai and Mirarab 2018).

Sequence and alignment assessment and cleaning is also important to minimise the amount of missing data, as this can also introduce artefactual conflict. Poor sample input DNA or library preparation can ultimately lead to poor coverage or biased sequence files. This can result in uneven locus recovery across samples, producing alignments with substantial missing data. While there may be a desire to keep all samples in the dataset, samples with missing data (especially with biased recovery across the target genes) can introduce conflict or mislead inference through a lack of information (e.g. Smith *et al.* 2020), and there is some evidence that the impact of missing data is amplified in datasets with high levels of incomplete lineage sorting (ILS) (Xi *et al.* 2016; Nute *et al.* 2018). To avoid potential biases, sample removal thresholds should be high (e.g. remove samples with <75% locus recovery), and inspections of sequences to check for coverage (as well as the percentage of recovered length) should also be conducted using the tools described above. Additionally, the first steps of HybPhaser (e.g. the script 1b_assess_dataset.R), are useful for summarising assembly quality, missingness, and information content of the loci.

2. Paralogy reconciliation

Dealing with paralogs is one of the most important parts of a phylogenomic workflow with target capture data (Smith and Hahn 2021), particularly in plants, where gene or genome duplications resulting in paralogs are common (see section 'Paralogy' below; De Bodt *et al.* 2005; Panchy *et al.* 2016; Ren *et al.* 2018; Landis *et al.* 2018; Almeida-Silva and Van de Peer 2023). There are an increasing number of workflows to handle paralogs in phylogenomic datasets that can be categorised into four main approaches: 1) remove paralogs (or paralogous loci); 2) mask the effects of paralogs; 3) infer ortholog groups, and 4) estimate species trees directly with a paralog-aware method (Fig. 2). Each approach has a different philosophy and set of underlying assumptions, which affects not only species tree estimation, but also the downstream analyses that can be applied to investigate biological processes such as ILS, hybridisation, and WGD events. The choice of approach should therefore be based on the philosophy that is most suitable for the analytical workflow the researcher wants to apply, as well as the biological system and questions at hand.

In the first approach, there are two options for removing paralogs: by filtering paralogs from paralogous loci so that only one copy is retained, or by excluding all paralogous loci detected (Fig. 2). In the first option, paralogs are filtered based on a criterion such as similarity to a reference sequence, pairwise similarity, or length, for example, as implemented in PPD (Zhou *et al.* 2022), ParalogWizard (Ufimov *et al.* 2022) or the filtering steps of CAPTUS (Ortiz *et al.* 2023). As a result, copies of paralogous loci are removed, and only the sequence that is the longest or most similar is retained for each locus. This option

may be suitable for some lineages where minimal paralogy is evident, or for handling plastid loci, where few to no paralogs are expected to be present. However, it should be clear that this approach makes no attempt to infer orthology. As such, analysis of the remaining loci not only violates the assumptions of homology in phylogenetic inference but also runs the risk of estimating erroneous topologies and introducing artefactual conflict into phylogenetic trees, because each retained copy may not share the same evolutionary history. Therefore, for the majority of target capture studies on plants, we do not recommend this approach. However, the second option for removing paralogs (removing any locus determined to be paralogous) is more scientifically defensible. By removing paralogous loci, researchers only include single-copy loci, which are more likely to be orthologous. In effect, this is an attempt to only include orthologs, which then satisfies the assumptions of homology in phylogenetic inference and can be justifiably used for tree inference. The downside of this method, however, is that it can substantially reduce the number of loci and therefore the amount of information for phylogenetic inference, potentially leading to poor resolution in estimated trees in smaller bait kits such as A353, or in lineages that have recent WGD and so all or nearly all loci would be determined to have paralogy. It also removes any signal of biological processes such as hybridisation and WGD events that might be in the evolutionary history of the lineage. If identifying the presence and nature of such processes is of interest to the scientific study, then a different paralog handling approach that uses the information in paralogs is likely to be more appropriate (see paralog reconciliation approaches 2–4 below).

The second approach for dealing with paralogs in target capture datasets is by masking the paralogs with consensus sequences coded with ambiguity codes (Fig. 2). This approach, which can be implemented in pipelines such as HybPhaser (Nauheimer *et al.* 2021) and that of Kates *et al.* (2018), aims to mitigate the effects of paralogs by encoding single nucleotide polymorphisms (SNPs) from different paralogs (and alleles) as ambiguous characters. In doing so, it avoids the pitfall of the first approach whereby non-homologous sequences for the same locus are aligned and used for phylogenetic analysis. Additionally, characterisation of the percentage of SNPs and allele divergence between paralogs through HybPhaser has been shown to be a good indication of ploidy within the phylogenetic tree (Hendriks *et al.* 2023) and can be used to phase paralogs and identify hybridisation events (Nauheimer *et al.* 2021; see section ‘5. Understanding and investigating patterns and underlying causes of tree conflict’). One potential pitfall of this approach is the introduction of ambiguities into the dataset that may eliminate potentially important phylogenetic signal at those sites, which can theoretically decrease the resolution of the tree. Potts *et al.* (2014) found this to be true for a series of short (< c. 1100 bp) single-gene datasets, but Kates *et al.* (2018) did not find the same for target capture data.

The third approach involves the inference of orthologous sequences from all paralogous sequences based on gene tree topologies (Fig. 2). This approach, summarised by Yang and Smith (2014), takes gene trees with paralogs and identifies sub-trees that only contain nodes representing speciation events, rather than nodes that may be a result of gene duplication. These sub-trees therefore include sequences for each gene that are orthologous (i.e., share a common ancestor), which can then be extracted from the original dataset, aligned, and used for species tree estimation. There are several algorithms for pruning trees to ortholog sub-trees. Choice of algorithm depends on the availability and quality of outgroup sequences (which the Maximum Inclusion (MI) algorithm doesn't require) and on the trade-off between retrieving few ortholog groups with good sampling (Monophyletic Outgroups (MO) algorithm) *versus* many ortholog groups with many missing samples (Rooted subTrees (RT) algorithm) (Yang and Smith 2014). For reliable ortholog identification, it is important to carefully clean the initial paralog trees by removing any spurious sequences (e.g. by pruning especially long branches), and reduce any monophyletic tips of the same species (that could represent alleles or neopolyploids) to one representative sequence in order to produce clean homolog trees. Orthology inference (and other downstream analyses such as WGD mapping, GRAMPA and ASTRAL-Pro — see the section '5. Understanding and investigating patterns and underlying causes of tree conflict') can then be performed on the clean homolog trees (also often referred to as 'multi-labelled trees'). Tree-based ortholog inference can be implemented through the scripts developed by Morales-Briones *et al.* (2021), or through the software Paragone (Jackson *et al.* 2023). By identifying orthologous sequences, the underlying assumption of homology in evolutionary models is maintained, and the conflicting signal of paralogs is eliminated, resulting in robust phylogenomic inferences. This approach of identifying homolog and ortholog gene trees also has the advantage of facilitating many options in downstream analyses for meaningful conflict investigation and inferring its underlying biological processes.

The fourth approach to dealing with paralogs entails the estimation of species trees using methods explicitly designed to accommodate paralogs (Fig. 2), as summarised in Smith and Hahn (2021). Instead of assuming a single gene tree topology across all loci, paralog-aware methods explicitly model and account for gene duplication and loss events in the estimation of species trees, though the method of doing so varies between programs. Programs such as ASTRAL-Pro (Zhang *et al.* 2020; Zhang and Mirarab 2022), FastMulRFS (Molloy and Warnow 2020), SpeciesRax (Morel *et al.* 2022), and iGTP (Chaudhary *et al.* 2010) combine homolog trees with two-step coalescent or parsimony-style approaches. Decomposition methods such as DISCO (Willson *et al.* 2022) apply a tree-pruning algorithm to the homolog trees (similar to the third paralog-reconciliation approach), splitting homolog trees into orthogroups prior to species-tree estimation, usually under the coalescent. As with the orthology-inference approach to paralogs, optimal implementation of these methods is dependent on the use of clean homolog trees, rather than raw paralog trees. By

integrating information across paralogous loci while accommodating gene tree discordance, these methods offer a sound option for accurate species tree estimation in complex evolutionary scenarios. However, two-step coalescent-based methods such as ASTRAL-Pro come with some drawbacks, such as treating gene tree topology as fixed even where nodes may be poorly supported because of short individual gene alignments. This potentially misleads species tree inference, and results in undefined branch lengths on the phylogeny (Mirarab *et al.* 2016; Simmons and Gatesy 2021). These issues may be mitigated with the development of new paralog-aware species tree estimation methods such as AleRax (Morel *et al.* 2024), which uses the distribution of homolog tree topologies to inform species tree inference; however, this method is yet to be applied in a plant phylogenetic context. Nevertheless, resolving paralogy *before* phylogenetic analysis gives the researcher more methodological and software options for the latter.

Each of these approaches is based on a distinct philosophy and set of underlying assumptions, influencing not only species tree estimation but also the possibilities for downstream analyses, and their interpretability and robustness. In some cases, taking multiple complementary approaches to dealing with paralogs may give additional insight into any conflict present in the dataset and help to answer research questions pertaining to underlying biological sources of conflict.

3. Phylogenomic reconstruction of gene trees and species trees

Following paralogy reconciliation using approaches 1–3 (i.e. once a sequence copy from each locus has been chosen), there are multiple methods available for species tree inference (Fig. 2). Tree-inference methods have been reviewed comprehensively (see e.g. Leache & Rannala (2011); Simmons and Gatesy (2015); Mirarab *et al.* (2016)), so we will not review these in-depth in this paper. Briefly, the two main approaches currently used for target capture data are Maximum Likelihood analyses conducted on concatenated sequence alignments (such as with IQ-TREE (Nguyen *et al.* 2015; Minh, Schmidt, *et al.* 2020) and RAxML (Stamatakis 2014; Kozlov *et al.* 2019)), and two-step coalescent approaches based on gene tree topologies (such as with ASTRAL (Mirarab *et al.* 2014)). Bayesian analysis of phylogenomic datasets can also be performed using ExaBayes (Aberer *et al.* 2014), and for smaller datasets of fewer than a hundred terminals, Bayesian inference under the multispecies coalescent (e.g. with StarBEAST (Douglas *et al.* 2022)) is another computationally feasible option. Each method comes with its own set of assumptions that may be more or less suitable depending on the scale of taxonomic sampling, size of study group, and lineage. Furthermore, the degree of gene-tree topology error and ILS present in a dataset can also influence the choice of tree-inference method, as conflict can increase the computation effort required (e.g. (Tea *et*

al. 2022)). We recommend using multiple tree-estimation methods for target capture datasets, especially because any conflict may give insight into artefactual issues or biological processes (see section '5. Understanding and investigating patterns and underlying causes of tree conflict').

Should a researcher want to go on to produce a dated phylogeny with node-dating, special considerations need to be made to deal with target capture datasets. Bayesian approaches to obtaining a dated phylogeny (e.g. BEAST (Bouckaert *et al.* 2014), MCMCTree in PAML (Yang 2007)) are computationally demanding, and with large datasets become intractable (Barba-Montoya *et al.* 2021). This can be solved by subsampling genes to choose the most clocklike, as implemented in SortaDate (Smith *et al.* 2018), or by using more computationally efficient phylogenetic dating methods such as penalised likelihood (Sanderson 2002), as implemented in TreePL (Smith and O'Meara 2012), the R package ape (chronos) (Paradis 2013; Paradis and Schliep 2019), and r8s (Sanderson 2003), or the relative rate framework (RRF) as implemented in RelTime (Tamura *et al.* 2012, 2018). However, should extensive conflict be identified in phylogenetic reconstruction, and if there is evidence to suggest that reticulation, deep coalescence or simultaneous speciation is the cause (see section '5. Understanding and investigating patterns and underlying causes of tree conflict'), extreme caution should be used in dating analyses. Currently, there is no method that can date a phylogeny that deviates from a bifurcating tree, and as such, trying to apply a molecular clock to such trees could lead to erroneous results (see section 'Conclusions and future perspectives').

It is important to note that artefactual conflict can arise during phylogenetic tree reconstruction through inappropriate choice of evolutionary models (such as substitution model), and gene tree estimation error (Cai *et al.* 2021). As such, it is good practice to conduct phylogenetic tree reconstruction with multiple approaches, carefully consider the assumptions of all choices made in the tree estimation models used, and to inspect gene trees for signs of error. Error in gene tree topologies can be caused by a number of factors, such as the inclusion of erroneous sequences, uninformative loci (due to slow mutation rates or short loci), or loci with extremely high rates of mutation prone to saturation and homoplasy. Filtering gene trees, or phylogenomic subsampling, can reduce the chance of artefactual conflict occurring by selecting of a subset of genes that are considered reliable. Tools such as GeneSortR (Mongiardino Koch 2021) and PhylteR (Comte *et al.* 2023) perform comparative analyses to identify a set of gene trees that have higher phylogenetic utility and accuracy, as well as removing potential outlier gene trees. GeneSortR is particularly extensive in the comparisons it performs, including average pairwise distance, compositional heterogeneity, level of saturation, root-to-tip variance, Robinson-Foulds distance to a reference topology, average bootstrap support, and proportion of variable sites. It also has the added benefit of producing easy to interpret and publication-ready images of the summarised outputs. TreeShrink (Mai and Mirarab 2018) is

another useful tool to reduce artefactual gene tree conflict, by identifying and pruning outlier long branches, thereby removing spurious samples. In combination with locus assembly and alignment assessment, gene tree assessment and phylogenomic subsampling can reduce the impact of non-biological conflict in the dataset and allow for more clear inferences of the true biological cause of conflict (see section '5. Understanding and investigating patterns and underlying causes of tree conflict').

4. Conflict assessment

With the increasing amount of genetic information available, phylogenetic conflict in plants — whereby individual gene trees do not share the same topology with either the species tree or with each other (Pamilo and Nei 1988; Maddison 1997) — is becoming increasingly reported. Ultimately, conflict is the result of either artefactual data issues or biological processes (see below), but before being able to identify the cause of it, one must first be able to pinpoint where the conflict occurs, and to what degree. Conflict may manifest as discordance between the topologies of species trees estimated with different methods or different data types, or between gene trees. Conflict in topology across species trees inferred with different data types or methods (e.g. discrepancies in the topologies of plastid and nuclear phylogenies, or discordance between coalescent and concatenated phylogenies) is usually identified visually and qualitatively described (Fig. 2). Conflict between the topologies of gene trees can be quantified on the resulting species tree in three main ways: through support values, through concordance vectors (*sensu* Lanfear and Hahn (2024)), and through internode certainty (IC) (Fig. 2).

Support values, such as bootstrap values or posterior probabilities, are statistical measures of confidence for the existence of any given branch, akin to standard errors (Lanfear and Hahn 2024). While important measures, the increasing amount of data from high-throughput sequencing datasets means that support values tend towards their maximum, often giving inflated measures of confidence (Kumar *et al.* 2012; Thomson and Brown 2022). Concordance vectors, on the other hand, are statistical measures of the variation in the relationships of any given branch, analogous to standard deviation. Unlike support values, they are more robust to the effects of larger datasets, giving an informative summary of the variation in the topology of each node independent of the size of the dataset. Concordance vectors can be calculated in three ways: as gene concordance factors, as quartet concordance factors, and as site concordance factors. These are reviewed in depth in Lanfear and Hahn (2024), and here we provide only a brief summary of the major differences between the three measures. In short, gene concordance factors (gCFs) compare the topology for each node of each gene tree to the topology of the species tree, and summarise the proportion of gene

trees that have a topology concordant with the species tree (Ané *et al.* 2007; Baum 2007; Smith *et al.* 2015; Lanfear and Hahn 2024). gCFs can be calculated in a number of ways, and the exact measures of concordance differ slightly depending on the method used. The most computationally feasible and popular methods for large datasets are with IQ-TREE2 (Minh, Hahn, *et al.* 2020), BUCKy (Larget *et al.* 2010), and PhyParts (Smith *et al.* 2015), which can also calculate concordance based on homolog trees (i.e. can account for duplications). Quartet concordance factors (qCFs) are estimated by subsampling all (or many) sets of four taxa for each locus ('quartets'), estimating the unrooted topology for each quartet, and then counting the proportion of quartets that are congruent with the species tree. Tools available to calculate qCFs include the program ASTRAL and its subsequent versions (e.g. (Mirarab *et al.* 2014; Sayyari and Mirarab 2016) and Quartet Sampling (Pease *et al.* 2018). Site concordance factors (sCFs) sample quartets of taxa for each node of the species tree, and use parsimony or maximum likelihood to count the number of informative sites (of a single locus or concatenated loci) that support each of three possible topologies for those taxa (Minh, Hahn, *et al.* 2020). Currently, this method is only implemented in IQ-TREE2 (Minh, Hahn, *et al.* 2020; Mo *et al.* 2023); however, sCFs are more susceptible to the effects of homoplasy than other concordance vectors, and so may overestimate discordance (Kück *et al.* 2022). Another way to measure conflict within a species tree is by calculating internode certainty, which can be seen as a summary of the aforementioned concordance vectors that compares the support for a given branch to the support for the best-supported alternative resolution of that branch (Salichos and Rokas 2013; Zhou *et al.* 2020). Internode certainty can also be compared to branch length to gain an indication of potential factors that may be causing conflict. Visualising these quantified conflicts and the relative frequencies of different topological combinations can also be conducted through DiscoVista (Sayyari *et al.* 2018). Each measure of conflict has nuanced meaning, interpretation, and pitfalls (Lanfear and Hahn 2024), so it is always good practice to characterise conflict through a number of methods.

5. Understanding and investigating patterns and underlying causes of conflict

Once conflict in phylogenetic trees is identified, located and measured, the source of the conflict can be investigated. Conflict can be attributed to two main sources: data artefacts, and biological processes that result in a deviation from a bifurcating pattern of evolution.

Artefactual conflict

Artefactual conflict refers to discordance between gene or species trees that arise from inappropriate bioinformatic choices, leading to errors, anomalies, biases and/or noise in phylogenetic results (e.g. Frost

et al. (2024)). Artefactual conflict can be introduced at any step of the bioinformatic pipeline, but is especially common during locus extraction, paralog reconciliation and phylogenetic tree inference. As such, it is important that the assumptions underlying each method used during these steps are carefully considered, and that the output (especially alignments, homolog trees, and gene trees) are checked and cleaned (see also the quality control steps marked with an asterisk in Fig. 2). For our summary of options for reducing the impact of data artefacts at these stages, see sections '1. Locus extraction', '2. Paralog reconciliation' and '3. Phylogenomic reconstruction of gene trees and species trees'.

Biological sources of conflict

Once artefactual conflict has been minimised, remaining phylogenetic conflict can give key insight into biological processes that have caused the evolutionary history of that lineage to deviate from one that can be represented in a bifurcating tree. The four main patterns that can be observed are (1) paralogy, (2) reticulation, (3) deep coalescence, and (4) simultaneous speciation or rapid radiation (Fig. 3). Given 'ideal' data, it would be possible to differentiate between these patterns and to reliably infer the underlying evolutionary process in each case; however, sufficiently clear evidence may be unavailable with reduced-representation sequencing methods such as target-capture sequencing, because of secondary loss of gene copies or failure to capture or assemble all existing copies. Further complicating matters is that there is no one method that can satisfactorily model and test for paralogy, reticulation and deep coalescence simultaneously. Therefore it is important to carefully select a suite of methods to test for and tease apart the effect of each of these processes if conflict is detected.

Paralogy

Paralogy is caused either by gene duplication or whole genome duplication (WGD) followed by lineage diversification. WGD events involve the doubling of an organism's entire genetic material within the same species (autopolyploidy), or after inter-species hybridisation (allopolyploidy, see also below) (del Pozo and Ramirez-Parra 2015). They are known to be common and important sources of diversity in the evolution of land plants but present challenges for phylogenomic analysis (Clark and Donoghue 2018; Morales-Briones *et al.* 2021).

Divergent evolution of the resulting gene copies will lead to differences that are inherited by descendent species and can cause retained gene copies in the same individual to group in separate clades in phylogenetic analyses. Copies from the same clade (or ortholog group) of the resulting gene family phylogeny represent orthologs (descendants of the same copy), but copies in separate clades are paralogs (descendants of different copies). Treating paralogs as orthologs can mislead phylogenetic analysis (Struck 2013). In the

ideal case, paralogy would be easily recognised by observing sister clades in a gene tree that both contain the same complement of samples (Fig. 3b), and for WGD, this pattern would be replicated across all genes. However, while duplicated gene copies can be retained (and often specialise in function — a major source of evolutionary novelty (Flagel and Wendel 2009)), more often the locus will re-diploidise over time, leading to a more ambiguous pattern of gene duplications and losses (Fig. 3c) (Mason and Wendel 2020; Bomblies 2020). As such, even when orthology inference is attempted, or only single-copy loci are retained in a dataset (see section on "Paralogy reconciliation" above), loss of paralogs following a polyploidization event can mean some (*c*. 10% in yeasts; Scannel *et al*. (2006)) of these single-copy loci are not orthologs, but 'hidden paralogs'. Therefore, even with careful handling of paralogs in a target-capture dataset, hidden paralogy may be an unavoidable and undetectable source of conflict in a dataset, though more studies are needed to understand the extent of this issue in plants.

Hidden paralogy notwithstanding, the presence of WGD events can be identified through target enrichment data in a number of ways. Locus extraction software such as HybPiper and CAPTUS can infer the presence and number of paralogs for each locus (Johnson *et al.* 2016; Ortiz *et al.* 2023). These are useful for extracting all sequence copies, and for gaining an indication of the amount of paralogy present in a dataset; however, these detected 'paralogs' are also likely to comprise divergent alleles and contigs with sequencing errors, and so further processing is required to identify paralogs that are the result of gene/genome duplication events. One method to more accurately characterise the degree of paralogy is through HybPhaser (Nauheimer *et al.* 2021). HybPhaser enables the user to define the threshold of heterozygosity that most likely represents true paralogs (rather than sequencing errors or alleles) so that they can be quantified. Further, heterozygosity has been shown to be correlated with ploidy level, and can therefore be used to characterise lineages that have a history of genome duplication (Hendriks *et al.* 2023). Alternatively, the paralog output of locus extraction software can be processed by first building clean homolog trees from all paralogs, extracting orthogroups from each homolog tree, and mapping those to the species tree to count the number of gene duplication events that occurred at each node (e.g. (Yang *et al.* 2018; Morales-Briones *et al.* 2021)). While these gene duplication mapping approaches have been shown to be useful for large (transcriptomic and genomic) datasets, their application in smaller target capture datasets (particularly A353 datasets) has not been extensively tested. Homolog trees can also be reconciled with species trees using programs such as GRAMPA (Thomas *et al.* 2017). GRAMPA uses a modified duplication-loss (DL) reconciliation algorithm (e.g. (Goodman *et al.* 1979; Page 1994) to determine whether hypothesised genome duplication events are best explained by allo- or autopolyploidisation events. However, like any DL-based method, GRAMPA does not account for deep coalescence (see section 'Deep coalescence' below), and can only investigate genome duplication at one node at a time; therefore use of such methods

requires careful consideration and interpretation of results. The development of new reconciliation algorithms that can account for the coalescent process are an important area of future research to disentangle WGD and ILS in phylogenomic datasets (Boussau and Scornavacca 2020; Mishra *et al.* 2023). When possible, any of these analyses can be combined with additional sources of evidence, such as Ks plots from transcriptomic and genomic data or karyological data, to pinpoint WGD events in a species tree (e.g. Yang *et al*. 2018).

Reticulation

Reticulation is caused by a variety of processes such as introgression and allopolyploid speciation which are often colloquially lumped together as 'hybridisation'. There are a number of approaches available for testing the presence of reticulation, and several comprehensive reviews already widely address the issue of hybridisation and introgression in phylogenomic datasets (e.g. Hibbins and Hahn 2022; Stull *et al.* 2023; Steenwyk *et al.* 2023). Here we focus on methods that are applicable to target capture data.

Introgression at the same ploidy level occurs when partially fertile hybrids between two species back-cross into one of the parental species. Introgression is increasingly recognised as a major driver of plant evolution. It can act as a source of additional genetic variation in species, and potentially even facilitate adaptation to novel stresses or habitats (Suarez-Gonzalez *et al.* 2018; Edelman and Mallet 2021). Conversely, introgression of maladaptive alleles during incipient speciation can lead to strong selective pressure towards reproductive isolation (Ostevik *et al.* 2016). In a phylogenetic context, introgression leads to incongruence between gene tree and species tree because of the transfer of alleles between species (Fig. 3d). Many methods for the detection of introgression work by comparing the depth of coalescence between estimated gene trees and the species tree, and infer introgression if the coalescence of gene lineages is too recent to be plausibly explained by deep coalescence (e.g. Joly *et al*. (2009); see 'Deep coalescence' below). Programs such as JML (Joly *et al*. 2012), QuIBL (Edelman *et al.* 2019) and Aphid (Galtier 2024) compare branch lengths of taxon triplets from ortholog gene trees, and by examining differences in branch lengths — shorter for gene flow and longer for deep coalescence — provide estimates of speciation times, ancestral population sizes, and quantify the impact of each process on phylogenetic conflict.

A special case of introgression is organelle capture, with chloroplast capture particularly relevant to plant phylogenetics because of the field's traditional reliance on chloroplast loci. Evidence across many taxa indicates that organelles are more easily transferred across lineages than nuclear genes (Stegemann *et al.* 2012), resulting in cases where plastid phylogenies are incongruent with morphological, nuclear ribosomal, and low copy nuclear data (e.g., Schmidt-Lebuhn and Bovill 2021). Perhaps the best-known Australian

example where phylogenetic inference has been confounded by frequent chloroplast capture is in the eucalypts (McKinnon *et al.* 1999; Nevill *et al.* 2014). Organelle capture can sometimes be inferred from conflict in the topology of species trees generated with plastid, morphological and nuclear data, in combination with tests for introgression in the nuclear dataset (McLay *et al.* 2023).

Allopolyploid speciation occurs when a hybrid between two species that is normally sterile due to an inability to produce functional gametes undergoes WGD, often through the production of unreduced gametes (Fig. 3e). This allows meiosis to be successful in their offspring, as the chromosomes can now pair with their duplicates. Allopolyploid speciation is a major factor in plant evolution (Soltis *et al.* 2009; Aïnouche and Wendel 2014; Alix *et al.* 2017; Clark and Donoghue 2018). As in WGD or gene duplication without hybridisation, the resulting gene lineages can specialise or, more frequently, re-diploidise through gene losses (Bomblies 2020). In an ideal case, an allopolyploid lineage could be recognised because the two gene clades in each locus are consistently nested in the two ancestral species clades, and could be recognised with the same tests used to test for paralogy (such as GRAMPA, see 'Paralogy' above). However, as with paralogy, gene losses during re-diploidisation, failure to capture or amplify all existing gene copies, and/or additional confounding factors such as deep coalescence could potentially make it difficult to reliably recognise ancient polyploidy.

Further approaches available for detection of reticulation (both via introgression and allopolyploid speciation) in target-capture datasets include phasing methods, whereby raw reads of a putative hybrids are phased into subgenomes and placed separately into the species tree to identify putative parental lineages. This is commonly achieved in target-capture data through HybPhaser (Nauheimer *et al.* 2021), and has been shown to be highly effective in cases of neoallopolyploidy (e.g. Bloesch *et al.* 2022; Bradican *et al.* 2023); however the method requires careful selection of the presence of diploid references for putative parental clades, and is often unsuitable for groups with complex or ancient reticulation (e.g. McLay *et al.* 2023). The need for a diploid reference is overcome through the Bayesian implementation of phasing in homologizer (Freyman *et al.* 2023), but may require subsampling of target-capture datasets to reduce computational demands.

Other available methods derive from population genetics ABBA-BABA (or 'D-statistic') tests, whereby any deviation in site pattern probabilities from what would be expected in a bifurcating tree indicates reticulation or deep coalescence. Such tests can be implemented in programs such as HyDe (Blischak *et al.* 2018). However, the site pattern probabilities expected in ABBA-BABA methods are calculated under a suite of assumptions, including symmetrical gene-flow between populations and constant substitution rate

across lineages and genes, which may be unrealistic, and could lead to inaccurate results (Frankel and Ané 2023). Therefore, these methods should be applied and interpreted with care.

Finally, network-based methods can be used to explore and depict reticulate evolutionary relationships, but these methods are still in their infancy and present a much-needed area for development. Distance-based methods such as Neighbor-Net (Bryant and Moulton 2004) and split decomposition methods such as SplitsTree (Huson 1998) are computationally feasible for phylogenomic datasets, but do not explicitly incorporate models of evolution, nor do they account for biological processes such as deep coalescence. Other phylogenetic network packages can implement more complex models, including the likelihood methods used in PhyloNet (Than *et al.* 2008) and more recently-developed Bayesian and coalescent methods (e.g. Yu and Nakhleh 2015; Solís-Lemus and Ané 2016; Wen *et al.* 2016; Zhang *et al.* 2018), such as those applied in PhyloNetworks (Solıs-Lemus *et al.* 2017). These can give robust estimations of phylogenetic networks but remain computationally intensive (often prohibitively so), restricting analysis to datasets of very few terminals. As phylogenetic network methods develop, they will be powerful tools to model and understand evolution in the presence of reticulation. However, networks are models that are more parameter-rich than bifurcating trees, so complex, reticulate scenarios will tend to be more statistically probable, even when they may not be true (Blair and Ané 2020). Therefore, results of phylogenetic network analyses should be evaluated critically, and they are usually most useful as a complement to bifurcating trees, rather than as a replacement for them.

Deep coalescence

Under coalescent theory, incongruence between gene trees and between an individual gene tree and the species tree is expected even in the absence of paralogy or reticulation. The underlying process has been understood for decades (Pamilo and Nei 1988; Maddison 1997).

Ancestral species with large effective population sizes are able to maintain a high diversity of alleles. At a lineage split in an ancestral species, both daughter species inherit a random sample of this diversity. If a gene lineage splits simultaneously with the species split, over time genetic drift will lead to the extinction of relictual ancestral alleles, and the remaining alleles in each species will be monophyletic (i.e. completely sorted), and the gene tree will be concordant with the species tree. However, if effective population sizes remain large, and species lineage splits follow quickly upon each other, then ancestral alleles that diverged prior to the species splits will not yet have been lost (i.e. incompletely sorted), and may be inherited. In this case, the gene tree will not reflect the species phylogeny (Fig. 3f). This pattern is known as deep coalescence, because the gene lineages coalesce deeper in the phylogeny than the species lineages they are evolving in. The stochastic inheritance of persistent ancestral alleles is referred to as ILS.

Despite the expected incongruence between gene trees due to ILS, the underlying species tree can still be reliably inferred under the assumption that ILS is the process causing the incongruence, i.e. under the multispecies coalescent (see section '3. Phylogenomic reconstruction of gene trees and species trees'). While there are multiple methods for estimating the species tree, the most relevant to target capture data that accounts for deep coalescence are summary approaches like ASTRAL that take gene trees as input. If deep coalescence is the main cause of conflict in the data, target capture data are particularly promising for resolving the species tree because individual loci may be long enough to produce relatively resolved gene trees and there are many loci. However, depending on the biological system and bait set, if most loci have little phylogenetic signal, this can mislead methods like ASTRAL and make it harder to estimate the species tree (Molloy and Warnow 2018), underscoring the importance of evaluating phylogenetic signal and gene trees earlier (see section '3. Phylogenomic reconstruction of gene trees and species trees').

Given that deep coalescence can leave a genetic signature similar to that of reticulation, most tests of deep coalescence also test for reticulation to differentiate the effect of these processes. Such tests include ABBA-BABA tests, and branch-length based methods like Aphid (Galtier 2024) and QuIBL (Edelman *et al.* 2019) (see 'Reticulation' above).

Simultaneous speciation and rapid radiations

Simultaneous speciation, whereby multiple species evolve at the same time rather than in a bifurcating manner, can also be a source of conflict in a phylogenetic tree. Simultaneous speciation is thought to occur rarely, but most commonly through allopatric, non-adaptive speciation, whereby a population is separated into more than two isolated geographic areas (e.g. through vicariance, mountain building or glaciation), and the individuals in each area evolve into separate lineages (Matsubayashi and Yamaguchi 2022, e.g. Dillenberger and Kadereit 2017); however, it is also theoretically possible for multiple species to evolve simultaneously through sympatric adaptive radiation events (Bolnick 2006), and combinatorial mechanisms (Marques *et al.* 2019). In phylogenetic terms, the simultaneous evolution of multiple lineages is a 'hard polytomy', with multifurcating branches, rather than bifurcating branches (Maddison 1989; Hoelzer and Meinick 1994). Hard polytomies can manifest as conflict (high levels of discordance across gene tree topologies or low support values) at nodes between short branches in a bifurcating species tree. When forced to be represented as bifurcations, each gene tree may have random, conflicting topologies between lineages originating by simultaneous speciation simply because the pattern of mutation does not comply with a bifurcating pattern. However, the challenge with inferring simultaneous speciation is differentiating it from cases of rapid radiations that do follow a bifurcating pattern of evolution. In these cases, little time and few

mutations may separate divergent lineages, and the lack of information makes these relationships particularly difficult to reconstruct. These are often referred to as 'soft polytomies' (Maddison 1989), where it is unclear if conflict is a result of a lack of information to resolve the true, bifurcating relationships of a rapid radiation, or a genuine case of simultaneous speciation (DeSalle *et al.* 1994; Whitfield and Lockhart 2007; Orel, McLay, Neal, *et al.* 2023; Zhang *et al.* 2023).

Most methods available to test for simultaneous speciation and rapid radiations are statistical tests based on the idea of treating a polytomy as the null hypothesis (whereby the branch length is zero), and rejecting it based on data (Swofford *et al.* 1996; Anisimova and Gascuel 2006). Some versions also incorporate a power test to facilitate the differentiation of soft and hard polytomies (Walsh *et al.* 1999). Alternative Bayesian approaches such as that described by Lewis *et al.* (2005) are also available. However, these methods can only be applied to single-locus data. As such, the most popular method currently used for phylogenomic data is the polytomy test available through ASTRAL (Sayyari and Mirarab 2018). This method is also based on the concept of rejecting the null hypothesis of zero-branch-length polytomies in the tree, can test across multiple gene trees, and also accounts for ILS, but by nature is sensitive to errors in gene tree topology (Sayyari and Mirarab 2018). Regardless of the analysis conducted, conclusively inferring simultaneous speciation, and differentiating it from a rapid radiation or a deficit of data is usually very difficult unless the study is conducted with large amounts of phylogenetic data, at a shallow phylogenetic scale, and in conjunction with a great deal of ecological and biological knowledge of the group in question. Realistically, the exact mode of evolution in most cases of rapid radiation and simultaneous speciation is therefore unknowable.

Conclusions and future perspectives

Given the recent and rapid advancement of target capture data for plant phylogenetic studies, there are many areas of the bioinformatic workflow that need improvement and research to further reduce artefactual conflict. One important area is locus extraction and assembly; the accurate and complete detection of paralogs resulting from gene duplication remains a challenge for plant phylogenomics, especially in groups currently without reference genomes. Aside from the unavoidable issue of hidden paralogy (see 'Biological sources of conflict – Paralogy'), and mis-assembly issues that can arise from short-read assembler errors (see '1. Locus extraction'), there are indications that current assembly approaches may be underestimating real paralogy in A353 datasets based on comparisons with reference genomes (Theodore Allnut, pers. comm.). While the extent of such issues are unknown at present, further refinement of locus extraction and

assembly programs, along with more affordable reference genomes will greatly assist future studies. Further, a better understanding of hidden paralogy in plants, and how different bait sets and different targeted loci (e.g. introns *vs.* exons) perform is needed to further develop best practices for reducing artefactual noise and investigating the causes of conflict.

Even when phylogenomic analysis is carefully conducted so that data artefacts are minimised, phylogenetic conflict is often inevitable. Although this conflict can often be carefully investigated to identify processes such as reticulation, paralogy, deep coalescence or polytomies as its cause, many methods to detect these processes are still being developed. The inability of many conflict interrogation analyses to account for more than two processes at once can make it difficult to differentiate their effects on phylogenetic conflict, and their influence on evolution. Phylogenetic network methods also have much need for improvement so that they can be applied meaningfully to large datasets. Currently, one of the main hindrances to the further development of these analyses is modelling the complex interactions between ILS, reticulation, paralogy and polytomies in such large datasets. It is likely that machine learning will play a large role in overcoming these obstacles in the future, although this would also come with its own set of caveats and limitations (Mo *et al.* 2024). In the meantime, the limitations of the data and assumptions of models used should always be acknowledged and taken into consideration while these methods develop.

Although identification of conflict and its underlying biological processes offers interesting insights into the mode of plant evolution, it also presents challenges for downstream evolutionary analyses such as dating, diversification analyses, ancestral area reconstruction and ancestral trait reconstruction. In cases where conflict is caused by noise from paralogs or deep coalescence, researchers may opt for downstream analyses that can account for the topological uncertainty at nodes with high degrees of conflict. In cases of reticulation and simultaneous speciation, however, it is more difficult to reasonably conduct these analyses, as most cannot yet account for evolution that is not modelled by a bifurcating tree. In these cases, we encourage researchers to be realistic about what analyses can be justifiably conducted, and when they are conducted, to be transparent about assumption violations and uncertainty in results. It is possible that creative solutions can be found in these scenarios by, for example, conducting analyses on subsets of taxa or trees. On the whole though, development of dating, diversification and ancestral state reconstruction models that can account for these processes is another area of much needed research, especially for plant evolutionary research.

We have long known that plant evolution is complex, with reticulation, whole genome duplication events, ILS and rapid radiations commonly reported, and so it should be unsurprising that phylogenetic conflict is

inherent within many plant lineages. Although target capture has made conflict more obvious, in some cases it can also give unprecedented capacity to empirically test for the underlying biological processes causing it, giving new insights into the extraordinary complexities of plant evolution. As large target capture phylogenies are generated, it would be fascinating to empirically quantify the extent of these processes across plants. Nevertheless, as with all models, researchers should always understand and be realistic about the limitations of the data and assumptions of analyses used.

For the Australian flora, this means answering long-standing questions that have been hampered by the 'lack of resolution' intractable from previously available technologies. To date, target capture studies have greatly enhanced our understanding of the timing and tempo of radiations (Joyce *et al.* 2023; Nge *et al.* 2024), the role of hybridisation and introgression in evolution (Bloesch *et al.* 2022; McLay *et al.* 2023; Nge *et al.* 2021 – *Adenanthos*), polyploidy and WGD events (Nge, Kellermann, *et al.* 2021; Schmidt-Lebuhn *et al.* 2024), and the evolution of diverse and important ecological groups in Australia (Crisp *et al.* 2024; McLay *et al.* 2023; Peakall *et al*. 2021; Schmidt-Lebuhn and Bovill 2021). They have also shed light on biogeography within Australia (Nge *et al.* 2021a; 2021b – *Calytrix* and *Pomaderris*), as well as Australia's biotic connection with other land-masses (Joyce *et al.* 2023; Nge *et al.* 2021 – *Pomaderris*), demonstrating Australia's role as both a source and sink of global plant diversity (Pillon *et al.* 2021; Van Dijk *et al.* 2023). Further, they have aided in taxonomic classification and the description of new species (Cooper *et al.* 2023; Crisp *et al*. 2024; Schmidt-Lebuhn and Grealy 2024; Simpson *et al.* 2022). These studies are only scratching the surface, but clearly have been an extraordinary advancement for our understanding of the Australian flora. We envisage that greater adoption of target capture approaches through collective (GAP and the Australian Angiosperm Tree of Life – Schmidt-Lebuhn *et al.* in prep) and group-specific studies (e.g. Stage 2 GAP phylogenomics – https://www.genomicsforaustralianplants.com/phylogenomics/, accessed May 2024) will spearhead research on the evolution and systematics of the Australian flora.

The difficulties dealing with the conflict within datasets and the vast array of methods involved in analysing this type of data offer new challenges to overcome and complexity to decipher. Rather than being seen as an obstacle, this will ultimately provide a more comprehensive understanding and more realistic and accurate evolutionary reconstruction of our interest groups. By addressing these challenges within the context of the Australian flora, we have a great opportunity to spotlight it on the global stage.

Declaration of funding

EMJ is supported by funding from the Prinzessin Therese von Bayern Stiftung.

Conflict of interest statement

The authors declare on conflicts of interest.

Data availability statement

No new data was generated for this review article.

References

Aberer AJ, Kobert K, Stamatakis A (2014) ExaBayes: Massively Parallel Bayesian Tree Inference for the Whole-Genome Era. *Molecular Biology and Evolution* **31**, 2553–2556. doi:10.1093/molbev/msu236.

Aïnouche M, Wendel J (2014) 'Polyploid Speciation and Genome Evolution: Lessons from Recent Allopolyploids.' doi:10.1007/978-3-319-07623-2_5.

Alix K, Gérard PR, Schwarzacher T, Heslop-Harrison JS (Pat) (2017) Polyploidy and interspecific hybridization: partners for adaptation, speciation and evolution in plants. *Annals of Botany* **120**, 183–194. doi:10.1093/aob/mcx079.

Almeida-Silva F, Van de Peer Y (2023) Whole-genome Duplications and the Long-term Evolution of Gene Regulatory Networks in Angiosperms. *Molecular Biology and Evolution* **40**, msad141. doi:10.1093/molbev/msad141.

Andermann T, Torres Jiménez MF, Matos-Maraví P, Batista R, Blanco-Pastor JL, Gustafsson ALS, Kistler L, Liberal IM, Oxelman B, Bacon CD, Antonelli A (2020) A Guide to Carrying Out a Phylogenomic Target Sequence Capture Project. *Frontiers in Genetics* **10**, 1407. doi:10.3389/fgene.2019.01407.

Ané C, Larget B, Baum DA, Smith SD, Rokas A (2007) Bayesian estimation of concordance among gene trees. *Molecular Biology and Evolution* **24**, 412–426.

Anisimova M, Gascuel O (2006) Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative. *Systematic Biology* **55**, 539–552. doi:10.1080/10635150600755453.

718 Baker WJ, Dodsworth S, Forest F, Graham SW, Johnson MG, McDonnell A, Pokorny L, Tate JA, Wicke
719 S, Wickett NJ (2021) Exploring Angiosperms353: An open, community toolkit for collaborative
720 phylogenomic research on flowering plants. *American Journal of Botany* **108**, 1059–1065.
721 doi:10.1002/ajb2.1703.
722 Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham
723 S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA
724 (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.
725 *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* **19**, 455–
726 477. doi:10.1089/cmb.2012.0021.
727 Barba-Montoya J, Tao Q, Kumar S (2021) Assessing Rapid Relaxed-Clock Methods for Phylogenomic
728 Dating. *Genome Biology and Evolution* **13**, evab251. doi:10.1093/gbe/evab251.
729 Baum DA (2007) Concordance trees, concordance factors, and the exploration of reticulate genealogy.
730 *TAXON* **56**, 417–426. doi:10.1002/tax.562013.
731 Blair C, Ané C (2020) Phylogenetic Trees and Networks Can Serve as Powerful and Complementary
732 Approaches for Analysis of Genomic Data. *Systematic Biology* **69**, 593–601.
733 doi:10.1093/sysbio/syz056.
734 Blischak PD, Chifman J, Wolfe AD, Kubatko LS (2018) HyDe: A Python Package for Genome-Scale
735 Hybridization Detection. *Systematic Biology* **67**, 821–829. doi:10.1093/sysbio/syy023.
736 Bloesch Z, Nauheimer L, Elias Almeida T, Crayn D, Field AR (2022) HybPhaser identifies hybrid evolution
737 in Australian Thelypteridaceae. *Molecular Phylogenetics and Evolution* **173**, 107526.
738 doi:10.1016/j.ympev.2022.107526.
739 Bolnick DI (2006) Multi-species outcomes in a common model of sympatric speciation. *Journal of*
740 *Theoretical Biology* **241**, 734–744. doi:10.1016/j.jtbi.2006.01.009.
741 Bomblies K (2020) When everything changes at once: finding a new normal after genome duplication.
742 *Proceedings of the Royal Society B: Biological Sciences* **287**, 20202154.
743 doi:10.1098/rspb.2020.2154.
744 Borowiec ML (2016) AMAS: a fast tool for alignment manipulation and computing of summary statistics.
745 *PeerJ* **4**,. doi:10.7717/peerj.1660.
746 Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A, Drummond AJ
747 (2014) BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Computational*
748 *Biology* **10**, e1003537.
749 Boussau B, Scornavacca C (2020) Reconciling Gene trees with Species Trees. 'Phylogenetics in the
750 Genomic Era'. (Eds C Scornavacca, F Delsuc, N Galtier) p. 3.2:1-3.2:23. (No commercial publisher
751 | Authors open access book) https://hal.science/hal-02535529.

Bradican JP, Tomasello S, Boscutti F, Karbstein K, Hörandl E (2023) Phylogenomics of Southern European Taxa in the Ranunculus auricomus Species Complex: The Apple Doesn't Fall Far from the Tree. *Plants* **12**, 3664. doi:10.3390/plants12213664.

Breinholt JW, Carey SB, Tiley GP, Davis EC, Endara L, McDaniel SF, Neves LG, Sessa EB, Von Konrat M, Chantanaorrapint S, Fawcett S, Ickert-Bond SM, Labiak PH, Larraín J, Lehnert M, Lewis LR, Nagalingum NS, Patel N, Rensing SA, Testo W, Vasco A, Villarreal JC, Williams EW, Burleigh JG (2021) A target enrichment probe set for resolving the flagellate land plant tree of life. *Applications in Plant Sciences* **9**, e11406. doi:10.1002/aps3.11406.

Bryant D, Moulton V (2004) Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks. *Molecular Biology and Evolution* **21**, 255–265. doi:10.1093/molbev/msh018.

Cai L, Xi Z, Lemmon EM, Lemmon AR, Mast A, Buddenhagen CE, Liu L, Davis CC (2021) The Perfect Storm: Gene Tree Estimation Error, Incomplete Lineage Sorting, and Ancient Gene Flow Explain the Most Recalcitrant Ancient Angiosperm Clade, Malpighiales (M Fishbein, Ed.). *Systematic Biology* **70**, 491–507. doi:10.1093/sysbio/syaa083.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973.

Cardillo M, Weston PH, Reynolds ZKM, Olde PM, Mast AR, Lemmon EM, Lemmon AR, Bromham L (2017) The phylogeny and biogeography of *Hakea* (Proteaceae) reveals the role of biome shifts in a continental plant radiation. *Evolution* **71**, 1928–1943. doi:10.1111/evo.13276.

Chaudhary R, Bansal MS, Wehe A, Fernández-Baca D, Eulenstein O (2010) iGTP: A software package for large-scale gene tree parsimony analysis. *BMC Bioinformatics* **11**, 574. doi:10.1186/1471-2105-11-574.

Clark JW, Donoghue PCJ (2018) Whole-Genome Duplication and Plant Macroevolution. *Trends in Plant Science* **23**, 933–945. doi:10.1016/j.tplants.2018.07.006.

Comte A, Tricou T, Tannier E, Joseph J, Siberchicot A, Penel S, Allio R, Delsuc F, Dray S, De Vienne DM (2023) PhylteR: Efficient Identification of Outlier Sequences in Phylogenomic Datasets (K Tamura, Ed.). *Molecular Biology and Evolution* **40**, msad234. doi:10.1093/molbev/msad234.

Cooper WE, Crayn DM, Joyce EM (2023) Aglaia fellii W.E.Cooper & Joyce (Meliaceae), a new species for Cape York Peninsula. *Australian Journal of Taxonomy* **16**, 1–9. doi:https://doi.org/10.54102/ajt.p8to6.

Crisp MD, Minh BQ, Choi B, Edwards RD, Hereward J, Kulheim C, Lin YP, Meusemann K, Thornhill AH, Toon A, Cook LG (2024) Perianth evolution and implications for generic delimitation in the

785 eucalypts (Myrtaceae), including the description of the new genus, *Blakella*. *Journal of Systematics*
786 *and Evolution* jse.13047. doi:10.1111/jse.13047.
787 De Bodt S, Maere S, Van de Peer Y (2005) Genome duplication and the origin of angiosperms. *Trends in*
788 *Ecology & Evolution* **20**, 591–597.
789 DeSalle R, Absher R, Amato G (1994) Speciation and phylogenetic resolution. *Trends in Ecology &*
790 *Evolution* **9**, 297–298. doi:10.1016/0169-5347(94)90034-5.
791 Dillenberger MS, Kadereit JW (2017) Simultaneous speciation in the European high mountain flowering
792 plant genus Facchinia (Minuartia sl, Caryophyllaceae) revealed by genotyping-by-sequencing.
793 *Molecular Phylogenetics and Evolution* **112**, 23–35.
794 Douglas J, Jiménez-Silva CL, Bouckaert R (2022) StarBeast3: Adaptive Parallelized Bayesian Inference
795 under the Multispecies Coalescent. *Systematic Biology* **71**, 901–916. doi:10.1093/sysbio/syac010.
796 Edelman NB, Frandsen PB, Miyagi M, Clavijo B, Davey J, Dikow RB, García-Accinelli G, Van Belleghem
797 SM, Patterson N, Neafsey DE, Challis R, Kumar S, Moreira GRP, Salazar C, Chouteau M,
798 Counterman BA, Papa R, Blaxter M, Reed RD, Dasmahapatra KK, Kronforst M, Joron M, Jiggins
799 CD, McMillan WO, Di Palma F, Blumberg AJ, Wakeley J, Jaffe D, Mallet J (2019) Genomic
800 architecture and introgression shape a butterfly radiation. *Science* **366**, 594–599.
801 doi:10.1126/science.aaw2090.
802 Edelman NB, Mallet J (2021) Prevalence and Adaptive Impact of Introgression. *Annual Review of Genetics*
803 **55**, 265–283. doi:10.1146/annurev-genet-021821-020805.
804 Faircloth BC (2016) PHYLUCE is a software package for the analysis of conserved genomic loci.
805 *Bioinformatics* **32**, 786–788. doi:10.1093/bioinformatics/btv646.
806 Fér T, Schmickl RE (2018) HybPhyloMaker: Target Enrichment Data Analysis From Raw Reads to Species
807 Trees. *Evolutionary Bioinformatics Online* **14**, 1176934317742613.
808 doi:10.1177/1176934317742613.
809 Flagel LE, Wendel JF (2009) Gene duplication and evolutionary novelty in plants. *The New Phytologist*
810 **183**, 557–564. doi:10.1111/j.1469-8137.2009.02923.x.
811 Fowler RM, McLay TGB, Schuster TM, Buirchell BJ, Murphy DJ, Bayly MJ (2020) Plastid phylogenomic
812 analysis of tribe Myoporeae (Scrophulariaceae). *Plant Systematics and Evolution* **306**, 52.
813 doi:10.1007/s00606-020-01678-4.
814 Frankel LE, Ané C (2023) Summary tests of introgression are highly sensitive to rate variation across
815 lineages. 2023.01.26.525396. doi:10.1101/2023.01.26.525396.
816 Freyman WA, Johnson MG, Rothfels CJ (2023) homologizer: Phylogenetic phasing of gene copies into
817 polyploid subgenomes. *Methods in Ecology and Evolution*. doi:https://doi.org/10.1111/2041-
818 210X.14072.

819 Frost LA, Bedoya AM, Lagomarsino LP (2024) Artifactual Orthologs and the Need for Diligent Data
820 Exploration in Complex Phylogenomic Datasets: A Museomic Case Study from the Andean Flora.
821 *Systematic Biology* syad076. doi:10.1093/sysbio/syad076.
822 Galtier N (2024) An approximate likelihood method reveals ancient gene flow between human, chimpanzee
823 and gorilla. *Peer Community Journal* **4**,. doi:10.24072/pcjournal.359.
824 Goodman M, Czelusniak J, Moore GW, Romero-Herrera AE, Matsuda G (1979) Fitting the gene lineage
825 into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin
826 sequences. *Systematic Biology* **28**, 132–163.
827 Gunn BF, Murphy DJ, Walsh NG, Conran JG, Pires JC, Macfarlane TD, Birch JL (2020) Evolution of
828 Lomandroideae: Multiple origins of polyploidy and biome occupancy in Australia. *Molecular*
829 *Phylogenetics and Evolution* **149**, 106836. doi:10.1016/j.ympev.2020.106836.
830 Gunn BF, Murphy DJ, Walsh NG, Conran JG, Pires JC, Macfarlane TD, Crisp MD, Cook LG, Birch JL
831 (2024) Genomic data resolve phylogenetic relationships of Australian mat-rushes, *Lomandra*
832 (Asparagaceae: Lomandroideae). *Botanical Journal of the Linnean Society* **204**, 1–22.
833 doi:10.1093/botlinnean/boad034.
834 Hancock LP, Obbens F, Moore AJ, Thiele K, De Vos JM, West J, Holtum JAM, Edwards EJ (2018)
835 Phylogeny, evolution, and biogeographic history of *Calandrinia* (Montiaceae). *American Journal*
836 *of Botany* **105**, 1021–1034. doi:10.1002/ajb2.1110.
837 Handika H, Esselstyn J (2022) SEGUL: An ultrafast, memory-efficient alignment manipulation and
838 summary tool for phylogenomics. doi:10.22541/au.165167823.30911834/v1.
839 Hart ML, Forrest LL, Nicholls JA, Kidner CA (2016) Retrieval of hundreds of nuclear loci from herbarium
840 specimens. *TAXON* **65**, 1081–1092. doi:10.12705/655.9.
841 Hendriks KP, Kiefer C, Al-Shehbaz IA, Bailey CD, van Huysduynen AH, Nikolov LA, Nauheimer L,
842 Zuntini AR, German DA, Franzke A (2023) Global Brassicaceae phylogeny based on filtering of
843 1,000-gene dataset. *Current Biology* **33**, 4052–4068.
844 Hibbins MS, Hahn MW (2022) Phylogenomic approaches to detecting and characterizing introgression.
845 *Genetics* **220**, iyab173. doi:10.1093/genetics/iyab173.
846 Hoelzer GA, Meinick DJ (1994) Patterns of speciation and limits to phylogenetic resolution. *Trends in*
847 *Ecology & Evolution* **9**, 104–107. doi:10.1016/0169-5347(94)90207-0.
848 Huson DH (1998) SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* **14**, 68–73.
849 doi:10.1093/bioinformatics/14.1.68.
850 Jackson C, McLay T, Schmidt-Lebuhn AN (2023) hybpiper-nf and paragone-nf: Containerization and
851 additional options for target capture assembly and paralog resolution. *Applications in Plant*
852 *Sciences* **11**, e11532. doi:10.1002/aps3.11532.

Johnson MG, Gardner EM, Liu Y, Medina R, Goffinet B, Shaw AJ, Zerega NJC, Wickett NJ (2016) HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in Plant Sciences* **4**,. doi:10.3732/apps.1600016.

Johnson MG, Pokorny L, Dodsworth S, Botigué LR, Cowan RS, Devault A, Eiserhardt WL, Epitawalage N, Forest F, Kim JT, Leebens-Mack JH, Leitch IJ, Maurin O, Soltis DE, Soltis PS, Wong GK, Baker WJ, Wickett NJ (2019) A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using k-Medoids Clustering (S Renner, Ed.). *Systematic Biology* **68**, 594–606. doi:10.1093/sysbio/syy086.

Joly S, McLenachan PA, Lockhart PJ (2009) A Statistical Approach for Distinguishing Hybridization and Incomplete Lineage Sorting. *The American Naturalist* **174**, E54–E70. doi:10.1086/600082.

Joly S (2012) JML: testing hybridization from species trees. Molecular Ecology Resources 12, 179–184. doi:10.1111/j.1755-0998.2011.03065.x.

Joyce EM, Appelhans MS, Buerki S, Cheek M, de Vos JM, Pirani JR, Zuntini AR, Bachelier JB, Bayly MJ, Callmander MW, Devecchi MF, Pell SK, Groppo M, Lowry PP, Mitchell J, Siniscalchi CM, Munzinger J, Orel HK, Pannell CM, Nauheimer L, Sauquet H, Weeks A, Muellner-Riehl AN, Leitch IJ, Maurin O, Forest F, Nargar K, Thiele KR, Baker WJ, Crayn DM (2023) Phylogenomic analyses of Sapindales support new family relationships, rapid Mid-Cretaceous Hothouse diversification, and heterogeneous histories of gene duplication. *Frontiers in Plant Science* **14**,. https://www.frontiersin.org/articles/10.3389/fpls.2023.1063174.

Karimi N, Grover CE, Gallagher JP, Wendel JF, Ané C, Baum DA (2020) Reticulate Evolution Helps Explain Apparent Homoplasy in Floral Biology and Pollination in Baobabs (Adansonia; Bombacoideae; Malvaceae) (E Susko, Ed.). *Systematic Biology* **69**, 462–478. doi:10.1093/sysbio/syz073.

Kates HR, Johnson MG, Gardner EM, Zerega NJC, Wickett NJ (2018) Allele phasing has minimal impact on phylogenetic reconstruction from targeted nuclear gene sequences in a case study of Artocarpus. *American Journal of Botany* **105**, 404–416. doi:10.1002/ajb2.1068.

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

Kück P, Romahn J, Meusemann K (2022) Pitfalls of the site-concordance factor (sCF) as measure of phylogenetic branch support. *NAR Genomics and Bioinformatics* **4**, lqac064. doi:10.1093/nargab/lqac064.

Kumar S, Filipski AJ, Battistuzzi FU, Kosakovsky Pond SL, Tamura K (2012) Statistics and Truth in Phylogenomics. *Molecular Biology and Evolution* **29**, 457–472. doi:10.1093/molbev/msr202.

887 Landis JB, Soltis DE, Li Z, Marx HE, Barker MS, Tank DC, Soltis PS (2018) Impact of whole-genome
888 duplication events on diversification rates in angiosperms. *American Journal of Botany* **105**, 348–
889 363. doi:10.1002/ajb2.1060.
890 Lanfear R, Hahn M (2024) The meaning and measure of concordance factors in phylogenomics.
891 https://ecoevorxiv.org/repository/view/6484/.
892 Larget BR, Kotha SK, Dewey CN, Ané C (2010) BUCKy: Gene tree/species tree reconciliation with
893 Bayesian concordance analysis. *Bioinformatics* **26**, 2910–2911.
894 doi:10.1093/bioinformatics/btq539.
895 Larridon I, Zuntini AR, Barrett RL, Wilson KL, Bruhl JJ, Goetghebeur P, Baker WJ, Brewer GE,
896 Epitawalage N, Fairlie I, Forest F, Sabino Kikuchi IAB, Pokorny L, Semmouri I, Spalink D,
897 Simpson DA, Muasya AM, Roalson EH (2021) Resolving generic limits in Cyperaceae tribe
898 Abildgaardieae using targeted sequencing. *Botanical Journal of the Linnean Society* **196**, 163–187.
899 doi:10.1093/botlinnean/boaa099.
900 Leaché AD, Rannala B (2011) The Accuracy of Species Tree Estimation under Simulation: A Comparison
901 of Methods. *Systematic Biology* **60**, 126–137. doi:10.1093/sysbio/syq073.
902 Lewis PO, Holder MT, Holsinger KE (2005) Polytomies and Bayesian Phylogenetic Inference. *Systematic*
903 *Biology* **54**, 241–253. doi:10.1080/10635150590924208.
904 Li D, Liu C-M, Luo R, Sadakane K, Lam T-W (2015) MEGAHIT: an ultra-fast single-node solution for
905 large and complex metagenomics assembly via succinct *de Bruijn* graph. *Bioinformatics* **31**, 1674–
906 1676. doi:10.1093/bioinformatics/btv033.
907 Liao X, Li M, Zou Y, Wu F, Yi-Pan, Wang J (2019) Current challenges and solutions of ***de novo*** assembly.
908 *Quantitative Biology* **7**, 90–109. doi:10.1007/s40484-019-0166-9.
909 Maddison W (1989) Reconstructing Character Evolution on Polytomous Cladograms. *Cladistics* **5**, 365–
910 377. doi:10.1111/j.1096-0031.1989.tb00569.x.
911 Maddison WP (1997) Gene Trees in Species Trees. *Systematic Biology* **46**, 523–536.
912 doi:10.1093/sysbio/46.3.523.
913 Mai U, Mirarab S (2018) TreeShrink: fast and accurate detection of outlier long branches in collections of
914 phylogenetic trees. *BMC Genomics* **19**, 23–40.
915 Marques DA, Meier JI, Seehausen O (2019) A Combinatorial View on Speciation and Adaptive Radiation.
916 *Trends in Ecology & Evolution* **34**, 531–544. doi:10.1016/j.tree.2019.02.008.
917 Mason AS, Wendel JF (2020) Homoeologous Exchanges, Segmental Allopolyploidy, and Polyploid
918 Genome Evolution. *Frontiers in Genetics* **11**,. doi:10.3389/fgene.2020.01014.

Matsubayashi KW, Yamaguchi R (2022) The speciation view: Disentangling multiple causes of adaptive and non-adaptive radiation in terms of speciation. *Population Ecology* **64**, 95–107. doi:10.1002/1438-390X.12103.

McKinnon GE, Steane DA, Potts BM, Vaillancourt RE (1999) Incongruence between chloroplast and species phylogenies in Eucalyptus subgenus Monocalyptus (Myrtaceae). *American Journal of Botany* **86**, 1038–1046. doi:10.2307/2656621.

McLay TGB, Birch JL, Gunn BF, Ning W, Tate JA, Nauheimer L, Joyce EM, Simpson L, Schmidt-Lebuhn AN, Baker WJ, Forest F, Jackson CJ (2021) New targets acquired: Improving locus recovery from the Angiosperms353 probe set. *Applications in Plant Sciences* **9**,. doi:10.1002/aps3.11420.

McLay TGB, Fowler RM, Fahey PS, Murphy DJ, Udovicic F, Cantrill DJ, Bayly MJ (2023) Phylogenomics reveals extreme gene tree discordance in a lineage of dominant trees: hybridization, introgression, and incomplete lineage sorting blur deep evolutionary relationships despite clear species groupings in *Eucalyptus* subgenus *Eudesmia*. *Molecular Phylogenetics and Evolution* **187**, 107869. doi:10.1016/j.ympev.2023.107869.

Minh BQ, Hahn MW, Lanfear R (2020) New Methods to Calculate Concordance Factors for Phylogenomic Datasets. *Molecular Biology and Evolution* **37**, 2727–2733. doi:10.1093/molbev/msaa106.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R (2020) IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* **37**, 1530–1534. doi:10.1093/molbev/msaa015.

Mirarab S, Bayzid MS, Warnow T (2016) Evaluating Summary Methods for Multilocus Species Tree Estimation in the Presence of Incomplete Lineage Sorting. *Systematic Biology* **65**, 366–380. doi:10.1093/sysbio/syu063.

Mirarab S, Reaz R, Bayzid MdS, Zimmermann T, Swenson MS, Warnow T (2014) ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30**, i541–i548. doi:10.1093/bioinformatics/btu462.

Mishra S, Smith ML, Hahn MW (2023) reconcILS: A gene tree-species tree reconciliation algorithm that allows for incomplete lineage sorting. 2023.11.03.565544. doi:10.1101/2023.11.03.565544.

Mo YK, Hahn MW, Smith ML (2024) Applications of machine learning in phylogenetics. *Molecular Phylogenetics and Evolution* **196**, 108066. doi:10.1016/j.ympev.2024.108066.

Mo YK, Lanfear R, Hahn MW, Minh BQ (2023) Updated site concordance factors minimize effects of homoplasy and taxon sampling. *Bioinformatics* **39**, btac741. doi:10.1093/bioinformatics/btac741.

Molloy EK, Warnow T (2018) To Include or Not to Include: The Impact of Gene Filtering on Species Tree Estimation Methods. *Systematic Biology* **67**, 285–303. doi:10.1093/sysbio/syx077.

Molloy EK, Warnow T (2020) FastMulRFS: fast and accurate species tree estimation under generic gene duplication and loss models. *Bioinformatics* **36**, i57–i65. doi:10.1093/bioinformatics/btaa444.

Mongiardino Koch N (2021) Phylogenomic Subsampling and the Search for Phylogenetically Reliable Loci (Y Satta, Ed.). *Molecular Biology and Evolution* **38**, 4025–4038. doi:10.1093/molbev/msab151.

Morales-Briones DF, Gehrke B, Huang C-H, Liston A, Ma H, Marx HE, Tank DC, Yang Y (2021) Analysis of Paralogs in Target Enrichment Data Pinpoints Multiple Ancient Polyploidy Events in *Alchemilla* s.l. (Rosaceae) (J Mandel, Ed.). *Systematic Biology* **71**, 190–207. doi:10.1093/sysbio/syab032.

Morel B, Schade P, Lutteropp S, Williams TA, Szöllősi GJ, Stamatakis A (2022) SpeciesRax: A Tool for Maximum Likelihood Species Tree Inference from Gene Family Trees under Duplication, Transfer, and Loss. *Molecular Biology and Evolution* **39**, msab365. doi:10.1093/molbev/msab365.

Morel B, Williams TA, Stamatakis A, Szöllősi GJ (2024) AleRax: a tool for gene and species tree co-estimation and reconciliation under a probabilistic model of gene duplication, transfer, and loss. *Bioinformatics* **40**, btae162. doi:10.1093/bioinformatics/btae162.

Nauheimer L, Cui L, Clarke C, Crayn DM, Bourke G, Nargar K (2019) Genome skimming provides well resolved plastid and nuclear phylogenies, showing patterns of deep reticulate evolution in the tropical carnivorous plant genus Nepenthes (Caryophyllales). *Australian Systematic Botany*. doi:10.1071/SB18057.

Nauheimer L, Weigner N, Joyce E, Crayn D, Clarke C, Nargar K (2021) HybPhaser: A workflow for the detection and phasing of hybrids in target capture data sets. *Applications in Plant Sciences* **9**,. doi:10.1002/aps3.11441.

Nevill PG, Després T, Bayly MJ, Bossinger G, Ades PK (2014) Shared phylogeographic patterns and widespread chloroplast haplotype sharing in Eucalyptus species with different ecological tolerances. *Tree Genetics & Genomes* **10**, 1079–1092. doi:10.1007/s11295-014-0744-y.

Nge FJ, Biffin E, Thiele KR, Waycott M (2021) Reticulate Evolution, Ancient Chloroplast Haplotypes, and Rapid Radiation of the Australian Plant Genus Adenanthos (Proteaceae). *Frontiers in Ecology and Evolution* **8**,. doi:10.3389/fevo.2020.616741.

Nge FJ, Biffin E, Waycott M, Thiele KR (2022) Phylogenomics and continental biogeographic disjunctions: insight from the Australian starflowers (*Calytrix*). *American Journal of Botany* **109**, 291–308. doi:10.1002/ajb2.1790.

Nge FJ, Kellermann J, Biffin E, Thiele KR, Waycott M (2024) Rise and fall of a continental mesic radiation in Australia: spine evolution, biogeography, and diversification of *Cryptandra* (Rhamnaceae: Pomaderreae). *Botanical Journal of the Linnean Society* **204**, 327–342. doi:10.1093/botlinnean/boad051.

Nge FJ, Kellermann J, Biffin E, Waycott M, Thiele KR (2021) Historical biogeography of Pomaderris (Rhamnaceae): Continental vicariance in Australia and repeated independent dispersals to New Zealand. *Molecular Phylogenetics and Evolution* **158**, 107085. doi:10.1016/j.ympev.2021.107085.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ (2015) IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* **32**, 268–274. doi:10.1093/molbev/msu300.

Nicol DA, Saldivia P, Summerfield TC, Heads M, Lord JM, Khaing EP, Larcombe MJ (2024) Phylogenomics and morphology of Celmisiinae (Asteraceae: Astereae): Taxonomic and evolutionary implications. *Molecular Phylogenetics and Evolution* **195**, 108064. doi:10.1016/j.ympev.2024.108064.

Nute M, Chou J, Molloy EK, Warnow T (2018) The performance of coalescent-based species tree estimation methods under models of missing data. *BMC Genomics* **19**, 286. doi:10.1186/s12864-018-4619-8.

Orel HK, McLay TG, Guja LK, Duretto MF, Bayly MJ (2023a) Genomic data inform taxonomy and conservation of critically endangered shrubs: a case study of Zieria (Rutaceae) species from eastern Australia. *Botanical Journal of the Linnean Society* boad069.

Orel HK, McLay TGB, Neal WC, Forster PI, Bayly MJ (2023b) Plastid phylogenomics of the Eriostemon group (Rutaceae; Zanthoxyloideae): support for major clades and investigation of a backbone polytomy. *Australian Systematic Botany* **36**, 355–385. doi:10.1071/SB23011.

Ortiz EM, Höwener A, Shigita G, Raza M, Maurin O, Zuntini A, Forest F, Baker WJ, Schaefer H (2023) A novel phylogenomics pipeline reveals complex pattern of reticulate evolution in Cucurbitales. 2023.10.27.564367. doi:10.1101/2023.10.27.564367.

Ostevik KL, Andrew RL, Otto SP, Rieseberg LH (2016) Multiple reproductive barriers separate recently diverged sunflower ecotypes. *Evolution; International Journal of Organic Evolution* **70**, 2322–2335. doi:10.1111/evo.13027.

Page RD (1994) Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Systematic Biology* **43**, 58–77.

Pamilo P, Nei M (1988) Relationships between gene trees and species trees. *Molecular Biology and Evolution* **5**, 568–583. doi:10.1093/oxfordjournals.molbev.a040517.

Panchy N, Lehti-Shiu M, Shiu S-H (2016) Evolution of Gene Duplication in Plants. *Plant Physiology* **171**, 2294–2316. doi:10.1104/pp.16.00523.

Paradis E (2013) Molecular dating of phylogenies by likelihood methods: A comparison of models and a new information criterion. *Molecular Phylogenetics and Evolution* **67**, 436–444. doi:10.1016/j.ympev.2013.02.008.

Paradis E, Schliep K (2019) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R (R Schwartz, Ed.). *Bioinformatics* **35**, 526–528. doi:10.1093/bioinformatics/bty633.

Peakall R, Wong DCJ, Phillips RD, Ruibal M, Eyles R, Rodriguez-Delgado C, Linde CC (2021) A multitiered sequence capture strategy spanning broad evolutionary scales: Application for phylogenetic and phylogeographic studies of orchids. *Molecular Ecology Resources* **21**, 1118–1140. doi:10.1111/1755-0998.13327.

Pease JB, Brown JW, Walker JF, Hinchliff CE, Smith SA (2018) Quartet Sampling distinguishes lack of support from conflicting support in the green plant tree of life. *American Journal of Botany* **105**, 385–403. doi:10.1002/ajb2.1016.

Pillon Y, Hopkins HCF, Maurin O, Epitawalage N, Bradford J, Rogers ZS, Baker WJ, Forest F (2021) Phylogenomics and biogeography of Cunoniaceae (Oxalidales) with complete generic sampling and taxonomic realignments. *American Journal of Botany* **108**, 1181–1200. doi:10.1002/ajb2.1688.

Potts AJ, Hedderson TA, Grimm GW (2014) Constructing phylogenies in the presence of intra-individual site polymorphisms (2ISPs) with a focus on the nuclear ribosomal cistron. *Systematic Biology* **63**, 1–16.

del Pozo JC, Ramirez-Parra E (2015) Whole genome duplications in plants: an overview from Arabidopsis. *Journal of Experimental Botany* **66**, 6991–7003. doi:10.1093/jxb/erv432.

Raza M, Ortiz EM, Schwung L, Shigita G, Schaefer H (2023) Resolving the phylogeny of Thladiantha (Cucurbitaceae) with three different target capture pipelines. *BMC Ecology and Evolution* **23**, 75. doi:10.1186/s12862-023-02185-z.

Ren R, Wang H, Guo C, Zhang N, Zeng L, Chen Y, Ma H, Qi J (2018) Widespread whole genome duplications contribute to genome complexity and species diversity in angiosperms. *Molecular Plant* **11**, 414–428.

Salichos L, Rokas A (2013) Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* **497**, 327–331.

Sanderson MJ (2002) Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Molecular Biology and Evolution* **19**, 101–109.

Sanderson MJ (2003) r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19**, 301–302. doi:10.1093/bioinformatics/19.2.301.

Sayyari E, Mirarab S (2016) Fast Coalescent-Based Computation of Local Branch Support from Quartet Frequencies. *Molecular Biology and Evolution* **33**, 1654–1668. doi:10.1093/molbev/msw079.

Sayyari E, Mirarab S (2018) Testing for Polytomies in Phylogenetic Species Trees Using Quartet Frequencies. *Genes* **9**, 132. doi:10.3390/genes9030132.

Sayyari E, Whitfield JB, Mirarab S (2018) DiscoVista: Interpretable visualizations of gene tree discordance. *Molecular Phylogenetics and Evolution* **122**, 110–115. doi:10.1016/j.ympev.2018.01.019.

Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH (2006) Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* **440**, 341–345. doi:10.1038/nature04562.

Schmidt-Lebuhn AN, Bovill J (2021) Phylogenomic data reveal four major clades of Australian Gnaphalieae (Asteraceae). *TAXON* **70**, 1020–1034. doi:10.1002/tax.12510.

Schmidt-Lebuhn AN, Egli D, Grealy A, Nicholls JA, Zwick A, Dymock JJ, Gooden B (2024) Genetic data confirm the presence of Senecio madagascariensis in New Zealand. *New Zealand Journal of Botany* **62**, 1–13. doi:10.1080/0028825X.2022.2148544.

Schmidt-Lebuhn AN, Grealy A (2024) Transfer of Cotula alpina to the genus Leptinella (Asteraceae: Anthemideae) (J Strijk, Ed.). *Australian Systematic Botany* **37**,. doi:10.1071/SB23012.

Shee ZQ, Frodin DG, Cámara-Leret R, Pokorny L (2020) Reconstructing the Complex Evolutionary History of the Papuasian Schefflera Radiation Through Herbariomics. *Frontiers in Plant Science* **11**,. doi:10.3389/fpls.2020.00258.

Simmons MP, Gatesy J (2015) Coalescence vs. concatenation: Sophisticated analyses vs. first principles applied to rooting the angiosperms. *Molecular Phylogenetics and Evolution* **91**, 98–122. doi:10.1016/j.ympev.2015.05.011.

Simmons MP, Gatesy J (2021) Collapsing dubiously resolved gene-tree branches in phylogenomic coalescent analyses. *Molecular Phylogenetics and Evolution* **158**, 107092. doi:10.1016/j.ympev.2021.107092.

Simpson J, Conran JG, Biffin E, Van Dijk K, Waycott M (2022) The Crinum flaccidum (Amaryllidaceae) species complex in Australia (J Bruhl, Ed.). *Australian Systematic Botany* **35**, 395–402. doi:10.1071/SB21038.

Siniscalchi CM, Hidalgo O, Palazzesi L, Pellicer J, Pokorny L, Maurin O, Leitch IJ, Forest F, Baker WJ, Mandel JR (2021) Lineage-specific vs. universal: A comparison of the Compositae1061 and Angiosperms353 enrichment panels in the sunflower family. *Applications in Plant Sciences* **9**, 10.1002/aps3.11422. doi:10.1002/aps3.11422.

Smith SA, Brown JW, Walker JF (2018) So many genes, so little time: A practical approach to divergence-time estimation in the genomic era. *PLOS ONE* **13**, e0197433. doi:10.1371/journal.pone.0197433.

Smith ML, Hahn MW (2021) New Approaches for Inferring Phylogenies in the Presence of Paralogs. *Trends in Genetics* **37**, 174–187. doi:10.1016/j.tig.2020.08.012.

Smith BT, Mauck WM, Benz BW, Andersen MJ (2020) Uneven Missing Data Skew Phylogenomic Relationships within the Lories and Lorikeets (B Holland, Ed.). *Genome Biology and Evolution* **12**, 1131–1147. doi:10.1093/gbe/evaa113.

Smith SA, Moore MJ, Brown JW, Yang Y (2015) Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evolutionary Biology* **15**, 150. doi:10.1186/s12862-015-0423-0.

Smith SA, O'Meara BC (2012) treePL: divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics* **28**, 2689–2690. doi:10.1093/bioinformatics/bts492.

Solís-Lemus C, Ané C (2016) Inferring Phylogenetic Networks with Maximum Pseudolikelihood under Incomplete Lineage Sorting. *PLOS Genetics* **12**, e1005896. doi:10.1371/journal.pgen.1005896.

Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, Sankoff D, de Pamphilis CW, Wall PK, Soltis PS (2009) Polyploidy and angiosperm diversification. *American Journal of Botany* **96**, 336–348. doi:10.3732/ajb.0800079.

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

Steenwyk JL, Buida TJ, Li Y, Shen X-X, Rokas A (2020) ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference (A Hejnol, Ed.). *PLOS Biology* **18**, e3001007. doi:10.1371/journal.pbio.3001007.

Steenwyk JL, Li Y, Zhou X, Shen X-X, Rokas A (2023) Incongruence in the phylogenomics era. *Nature Reviews Genetics* **24**, 834–850. doi:10.1038/s41576-023-00620-x.

Stegemann S, Keuthe M, Greiner S, Bock R (2012) Horizontal transfer of chloroplast genomes between plant species. *Proceedings of the National Academy of Sciences* **109**, 2434–2438. doi:10.1073/pnas.1114076109.

Strobel V (2018) Pold87/academic-keyword-occurrence: First release. doi:10.5281/ZENODO.1218409.

Struck TH (2013) The Impact of Paralogy on Phylogenomic Studies – A Case Study on Annelid Relationships. *PLOS ONE* **8**, e62892. doi:10.1371/journal.pone.0062892.

Stull GW, Pham KK, Soltis PS, Soltis DE (2023) Deep reticulation: the long legacy of hybridization in vascular plant evolution. *The Plant Journal* **114**, 743–766. doi:10.1111/tpj.16142.

Suarez-Gonzalez A, Lexer C, Cronk QCB (2018) Adaptive introgression: a plant perspective. *Biology Letters* **14**, 20170688. doi:10.1098/rsbl.2017.0688.

Swofford DL, Olsen GJ, Waddell PJ, Hillis DM (1996) Phylogenetic inference. 'Molecular Systematics'. (Eds DM Hillis, C Moritz, BK Mable) pp. 407–514. (Sinauer Associates: Sunderland, MA, USA)

Tamura K, Battistuzzi FU, Billing-Ross P, Murillo O, Filipski A, Kumar S (2012) Estimating divergence times in large molecular phylogenies. *Proceedings of the National Academy of Sciences* **109**, 19333–19338. doi:10.1073/pnas.1213199109.

Tamura K, Tao Q, Kumar S (2018) Theoretical Foundation of the RelTime Method for Estimating Divergence Times from Variable Evolutionary Rates (C Russo, Ed.). *Molecular Biology and Evolution* **35**, 1770–1782. doi:10.1093/molbev/msy044.

Tea Y-K, Xu X, DiBattista JD, Lo N, Cowman PF, Ho SYW (2022) Phylogenomic Analysis of Concatenated Ultraconserved Elements Reveals the Recent Evolutionary Radiation of the Fairy Wrasses (Teleostei: Labridae: Cirrhilabrus). *Systematic Biology* **71**, 1–12. doi:10.1093/sysbio/syab012.

Than C, Ruths D, Nakhleh L (2008) PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics* **9**, 322. doi:10.1186/1471-2105-9-322.

Thomas GWC, Ather SH, Hahn MW (2017) Gene-Tree Reconciliation with MUL-Trees to Resolve Polyploidy Events. *Systematic Biology* **66**, 1007–1018. doi:10.1093/sysbio/syx044.

Thomson RC, Brown JM (2022) On the Need for New Measures of Phylogenomic Support (B Carstens, Ed.). *Systematic Biology* **71**, 917–920. doi:10.1093/sysbio/syac002.

Tumescheit C, Firth AE, Brown K (2022) CIAlign: A highly customisable command line tool to clean, interpret and visualise multiple sequence alignments. *PeerJ* **10**, e12983. doi:10.7717/peerj.12983.

Ufimov R, Gorospe JM, Fer T, Kandziora M, Salomon L, Loo M, Schmickl R (2022) Utilizing paralogs for phylogenetic reconstruction has the potential to increase species tree support and reduce gene tree discordance in target enrichment data. *Molecular Ecology Resources* **22**,. doi:10.1111/1755-0998.13684.

Van Dijk K, Waycott M, Biffin E, Creed JC, Albertazzi FJ, Samper-Villarreal J (2023) Phylogenomic Insights into the Phylogeography of Halophila baillonii Asch. *Diversity* **15**, 111. doi:10.3390/d15010111.

Vatanparast M, Powell A, Doyle JJ, Egan AN (2018) Targeting legume loci: A comparison of three methods for target enrichment bait design in Leguminosae phylogenomics. *Applications in Plant Sciences* **6**, e1036. doi:10.1002/aps3.1036.

Walsh HE, Kidd MG, Moum T, Friesen VL (1999) Polytomies and the Power of Phylogenetic Inference. *Evolution* **53**, 932–937. doi:10.1111/j.1558-5646.1999.tb05386.x.

Waycott M, Van Dijk K, Biffin E (2021) A hybrid capture RNA bait set for resolving genetic and evolutionary relationships in angiosperms from deep phylogeny to intraspecific lineage hybridization. doi:10.1101/2021.09.06.456727.

Wen D, Yu Y, Nakhleh L (2016) Bayesian Inference of Reticulate Phylogenies under the Multispecies Network Coalescent. *PLOS Genetics* **12**, e1006006. doi:10.1371/journal.pgen.1006006.

Whitfield JB, Lockhart PJ (2007) Deciphering ancient rapid radiations. *Trends in Ecology & Evolution* **22**, 258–265. doi:10.1016/j.tree.2007.01.012.

Willson J, Roddur MS, Liu B, Zaharias P, Warnow T (2022) DISCO: Species Tree Inference using Multicopy Gene Family Tree Decomposition. *Systematic Biology* **71**, 610–629. doi:10.1093/sysbio/syab070.

Xi Z, Liu L, Davis CC (2016) The Impact of Missing Data on Species Tree Estimation. *Molecular Biology and Evolution* **33**, 838–860. doi:10.1093/molbev/msv266.

Yang Z (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* **24**, 1586–1591. doi:10.1093/molbev/msm088.

Yang Y, Moore MJ, Brockington SF, Mikenas J, Olivieri J, Walker JF, Smith SA (2018) Improved transcriptome sampling pinpoints 26 ancient and more recent polyploidy events in Caryophyllales, including two allopolyploidy events. *New Phytologist* **217**, 855–870. doi:10.1111/nph.14812.

Yang Y, Smith SA (2014) Orthology Inference in Nonmodel Organisms Using Transcriptomes and Low-Coverage Genomes: Improving Accuracy and Matrix Occupancy for Phylogenomics. *Molecular Biology and Evolution* **31**, 3081–3092. doi:10.1093/molbev/msu245.

Yu Y, Nakhleh L (2015) A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics* **16**, S10. doi:10.1186/1471-2164-16-S10-S10.

Zhang Q, Folk RA, Mo Z-Q, Ye H, Zhang Z-Y, Peng H, Zhao J-L, Yang S-X, Yu X-Q (2023) Phylotranscriptomic analyses reveal deep gene tree discordance in *Camellia* (Theaceae). *Molecular Phylogenetics and Evolution* **188**, 107912. doi:10.1016/j.ympev.2023.107912.

Zhang C, Mirarab S (2022) ASTRAL-Pro 2: ultrafast species tree reconstruction from multi-copy gene family trees. *Bioinformatics* **38**, 4949–4950.

Zhang C, Ogilvie HA, Drummond AJ, Stadler T (2018a) Bayesian Inference of Species Networks from Multilocus Sequence Data. *Molecular Biology and Evolution* **35**, 504–517. doi:10.1093/molbev/msx307.

Zhang C, Rabiee M, Sayyari E, Mirarab S (2018b) ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19, 153. doi:10.1186/s12859-018-2129-y.

Zhang C, Scornavacca C, Molloy EK, Mirarab S (2020) ASTRAL-Pro: quartet-based species-tree inference despite paralogy. Molecular *Biology and Evolution* **37**, 3292–3307.

Zhang Z, Xie P, Guo Y, Zhou W, Liu E, Yu Y (2022) Easy353: A Tool to Get Angiosperms353 Genes for Phylogenomic Research. *Molecular Biology and Evolution* **39**, msac261. doi:10.1093/molbev/msac261.

Zhou X, Lutteropp S, Czech L, Stamatakis A, Looz MV, Rokas A (2020) Quartet-Based Computations of Internode Certainty Provide Robust Measures of Phylogenetic Incongruence. *Systematic Biology* **69**, 308–324. doi:10.1093/sysbio/syz058.

Zhou W, Soghigian J, Xiang Q-Y (Jenny) (2022) A New Pipeline for Removing Paralogs in Target Enrichment Data. *Systematic Biology* **71**, 410–425. doi:10.1093/sysbio/syab044.

Zuntini AR, Carruthers T, Maurin O, Bailey PC, Leempoel K, Brewer GE, Epitawalage N, Françoso E, Gallego-Paramo B, McGinnie C, Negrão R, Roy SR, Simpson L, Toledo Romero E, Barber VMA, Botigué L, Clarkson JJ, Cowan RS, Dodsworth S, Johnson MG, Kim JT, Pokorny L, Wickett NJ, Antar GM, DeBolt L, Gutierrez K, Hendriks KP, Hoewener A, Hu A-Q, Joyce EM, Kikuchi IABS, Larridon I, Larson DA, de Lírio EJ, Liu J-X, Malakasi P, Przelomska NAS, Shah T, Viruel J, Allnutt TR, Ameka GK, Andrew RL, Appelhans MS, Arista M, Ariza MJ, Arroyo J, Arthan W, Bachelier JB, Bailey CD, Barnes HF, Barrett MD, Barrett RL, Bayer RJ, Bayly MJ, Biffin E, Biggs N, Birch JL, Bogarín D, Borosova R, Bowles AMC, Boyce PC, Bramley GLC, Briggs M, Broadhurst L, Brown GK, Bruhl JJ, Bruneau A, Buerki S, Burns E, Byrne M, Cable S, Calladine A, Callmander MW, Cano Á, Cantrill DJ, Cardinal-McTeague WM, Carlsen MM, Carruthers AJA, de Castro Mateo A, Chase MW, Chatrou LW, Cheek M, Chen S, Christenhusz MJM, Christin P-A, Clements MA, Coffey SC, Conran JG, Cornejo X, Couvreur TLP, Cowie ID, Csiba L, Darbyshire I, Davidse G, Davies NMJ, Davis AP, van Dijk K, Downie SR, Duretto MF, Duvall MR, Edwards SL, Eggli U, Erkens RHJ, Escudero M, de la Estrella M, Fabriani F, Fay MF, Ferreira P de L, Ficinski SZ, Fowler RM, Frisby S, Fu L, Fulcher T, Galbany-Casals M, Gardner EM, German DA, Giaretta A, Gibernau M, Gillespie LJ, González CC, Goyder DJ, Graham SW, Grall A, Green L, Gunn BF, Gutiérrez DG, Hackel J, Haevermans T, Haigh A, Hall JC, Hall T, Harrison MJ, Hatt SA, Hidalgo O, Hodkinson TR, Holmes GD, Hopkins HCF, Jackson CJ, James SA, Jobson RW, Kadereit G, Kahandawala IM, Kainulainen K, Kato M, Kellogg EA, King GJ, Klejevskaja B, Klitgaard BB, Klopper RR, Knapp S, Koch MA, Leebens-Mack JH, Lens F, Leon CJ, Léveillé-Bourret É, Lewis GP, Li D-Z, Li L, Liede-Schumann S, Livshultz T, Lorence D, Lu M, Lu-Irving P, Luber J, Lucas EJ, Luján M, Lum M, Macfarlane TD, Magdalena C, Mansano VF, Masters LE, Mayo SJ, McColl K, McDonnell AJ, McDougall AE, McLay TGB, McPherson H, Meneses RI, Merckx VSFT, Michelangeli FA, Mitchell JD, Monro AK, Moore MJ, Mueller TL, Mummenhoff K, Munzinger J, Muriel P, Murphy DJ, Nargar K, Nauheimer L, Nge FJ, Nyffeler R, Orejuela A, Ortiz EM, Palazzesi L, Peixoto AL, Pell SK, Pellicer J, Penneys DS, Perez-Escobar OA, Persson C, Pignal M, Pillon Y, Pirani JR, Plunkett GM, Powell RF, Prance GT, Puglisi C, Qin M, Rabeler RK, Rees PEJ, Renner M, Roalson EH, Rodda M, Rogers ZS, Rokni S, Rutishauser R, de Salas MF, Schaefer H, Schley RJ, Schmidt-Lebuhn A, Shapcott A, Al-Shehbaz I, Shepherd KA, Simmons MP, Simões AO, Simões ARG, Siros M, Smidt EC, Smith JF, Snow N, Soltis DE, Soltis PS, Soreng RJ, Sothers CA, Starr JR, Stevens PF, Straub SCK, Struwe L, Taylor JM, Telford IRH, Thornhill AH, Tooth I, Trias-Blasi A, Udovicic F, Utteridge TMA, Del Valle JC, Verboom GA, Vonow HP, Vorontsova MS, de

Vos JM, Al-Wattar N, Waycott M, Welker CAD, White AJ, Wieringa JJ, Williamson LT, Wilson TC, Wong SY, Woods LA, Woods R, Worboys S, Xanthos M, Yang Y, Zhang Y-X, Zhou M-Y, Zmarzty S, Zuloaga FO, Antonelli A, Bellot S, Crayn DM, Grace OM, Kersey PJ, Leitch IJ, Sauquet H, Smith SA, Eiserhardt WL, Forest F, Baker WJ (2024) Phylogenomics and the rise of the angiosperms. *Nature* 1–8. doi:10.1038/s41586-024-07324-0.

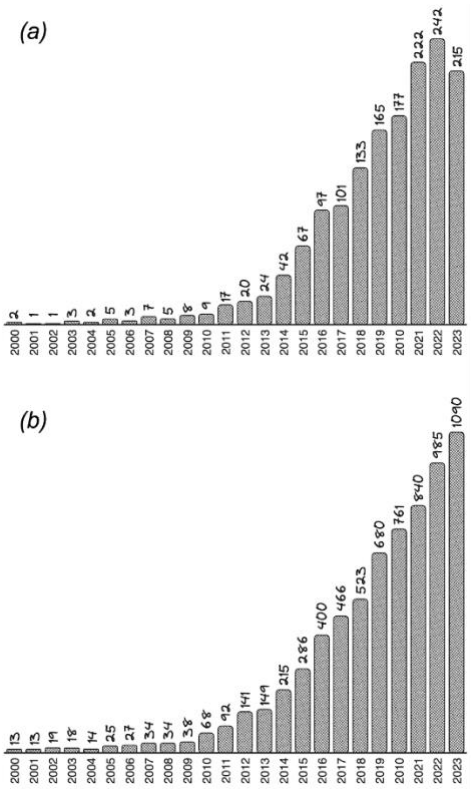

Fig. 1. Number of academic papers in Google Scholar published each year from 2000–2023 matching the search terms ‘"target capture" OR "target enrichment" OR "Hyb-Seq" AND "DNA" AND "plant"’: *(a)* matches also including the search term ‘Australia’, *(b)* matches not including the search term ‘Australia’. Obtained using the Python script of Strobel (2018).

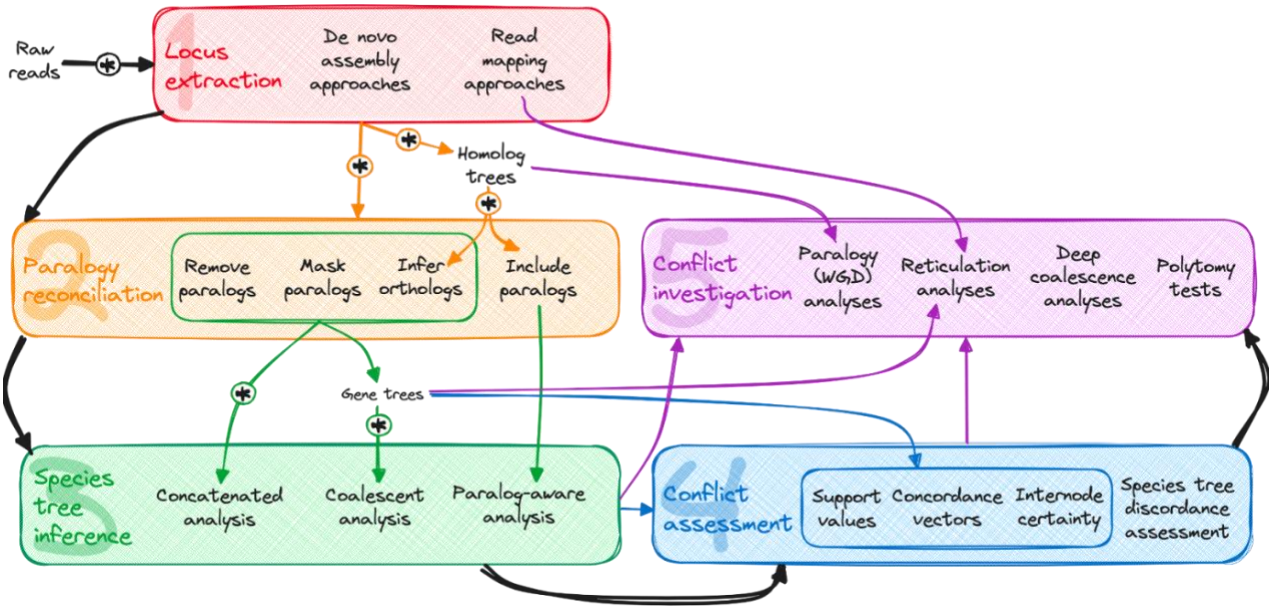

Fig. 2. Overview of the five major steps for a phylogenomic workflow with target capture data outlined in this review, from raw reads to 1) Locus extraction, 2) Paralogy reconciliation, 3) Species tree inference, 4) Conflict assessment and 5) Conflict investigation. Black arrows indicate the general direction of the workflow. Within each step of the workflow, the main approaches are summarised, and coloured arrows indicate which approaches are compatible from each step. Circles with asterisks (*) indicate particularly important points where quality control should be conducted on the output of the previous step to avoid the introduction of artefactual conflict (e.g. by checking and cleaning alignments and gene tree topologies). For more details on each step, see the relevant section of this review.

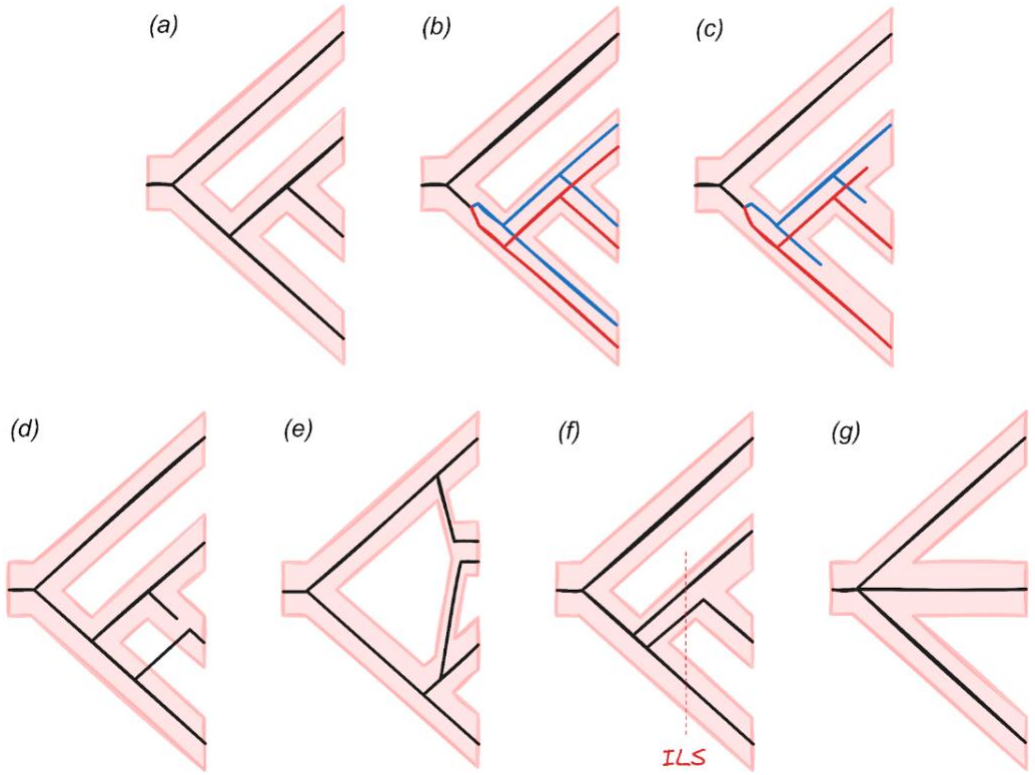

Fig. 3. Possible scenarios for gene evolution during species diversification. (*a*) Congruence between the species tree (pink bars) and gene tree (narrow lines). (*b*) Paralogy with one gene duplication and no gene losses. Red and blue indicate two ortholog groups. (*c*) Paralogy with one gene duplication followed by gene losses (or failure to capture or assemble gene copies) that left no evidence of paralogy. (*d*) Introgression (reticulation). (*e*) Allopolyploid hybridogenic speciation (reticulation). (*f*) Deep coalescence. The dotted line marked 'ILS' indicates transient incomplete lineage sorting in an ancestral lineage, i.e., the two alleles present in the middle lineage (large population) are not monophyletic (one is more closely related to an allele in the sister lineage). *(g)* True multifurcation due to simultaneous speciation.

Table 1 List of studies using target capture sequencing that have included members of the Australian flora. Ongoing work on several other Australian plant groups as part of GAP Stage 2 using A353 baits can be found here: https://www.genomicsforaustralianplants.com/phylogenomics/. 'Nuclear' is abbreviated as 'nuc'; 'chloroplast' is abbreviated as 'cp'.

Plant group	**Baits kit**	**Assembly method**	**Tree inference method (concatenated or coalescent)**	**Authors**	**DOI**
Caladenia and Diurideae (Orchidaceae)	Custom baits (up to 1000+ loci)	Custom pipeline, Hybpiper	Both	Peakall *et al.* (2021)	10.1111/1755-0998.13327
Eucalypts (Myrtaceae)	Custom baits (101 low-copy nuc exons)	Custom pipeline	Both	Crisp *et al.* (2024)	10.1111/jse.13047
Eucalyptus (Myrtaceae)	Custom baits (568 nuc genes, including A353 and OzBaits)	Hybpiper-nf, HybPhaser	Both	McLay *et al.* (2023)	10.1016/j.ympev.2023.107869
Calandrinia (Montiaceae)	Custom baits for Caryophyllales	Custom pipeline	Both	Hancock *et al.* (2018)	10.1002/ajb2.1110
Cryptandra (Rhamnaceae)	OzBaits	Custom pipeline	Concatenated	Nge *et al.* (2024)	10.1093/botlinnean/boad051
Pomaderris (Rhamnaceae)	OzBaits	Custom pipeline	Concatenated	Nge *et al.* (2021)	10.1016/j.ympev.2021.107085
Calytrix (Myrtaceae)	OzBaits	Custom pipeline	Both	Nge *et al.* (2022)	10.1002/ajb2.1790
Adenanthos (Proteaceae)	OzBaits	Custom pipeline	Both	Nge *et al.* (2021)	10.3389/fevo.2020.616741
Crinum (Amaryllidaceae)	OzBaits	Custom pipeline	Concatenated	Simpson *et al.* (2022)	10.1071/SB21038

Halophila (Hydrocharitaceae)	OzBaits	Custom pipeline	Concatenated	Van Dijk *et al.* (2023)	10.3390/d15010111
Pogonolepis (Asteraceae)	A353	Hybpiper	Concatenated	Schmidt-Lebuhn (2022)	10.1071/SB22010
Anthemideae tribe (Asteraceae)	A353	Hybpiper-nf	Concatenated	Schmidt-Lebuhn & Grealy (2024)	10.1071/SB23012
Gnaphalieae tribe (Asteraceae)	Custom baits (Compositae 1061)	Hybpiper	Both	Schmidt-Lebuhn & Bovill (2021)	10.1002/tax.12510
Hakea (Proteaceae)	Custom bait kit (450 nuc loci)	Custom pipeline	Both	Cardillo *et al.* (2017)	10.1111/evo.13276
Thelypteridaceae	GoFlag (451 nuc loci)	Hybpiper, HybPhaser	Both	Bloesch *et al.* (2022)	10.1016/j.ympev.2022.107526
Cunoniaceae	A353	Hybpiper	Coalescent	Pillon *et al.* (2021)	10.1002/ajb2.1688
Zanthoxyloideae subfamily (Rutaceae; Sapindales)	A353	Hybpiper, Hybphaser	Both	Joyce *et al.* (2023)	10.3389/fpls.2023.1063174
Aglaia (Meliaceae)	A353	Hybpiper, Hybphaser	Concatenated	Cooper *et al*. (2023)	10.54102/ajt.p8to6
Celmisiinae	A353	Hybpiper-nf	Both	Nicol *et al*. (2024)	10.1016/j.ympev.2024.108064
Hibbertia (Dilleniaceae)*	A353, OzBaits (nuc), OzBaits (cp)	CAPTUS	Both	Hammer *et al.*	GAP special issue
Drosera (Droseraceae)	A353, OzBaits (nuc), OzBaits (cp)	CAPTUS	Both	Williamson *et al.*	GAP special issue
Alismatales	A353, OzBaits (nuc), OzBaits (cp)	CAPTUS	Both	Waycott *et al.*	GAP special issue

Chamelaucieae tribe (Myrtaceae)	A353	SECAPR	Both	Nge *et al.*	GAP special issue
Minuria (Asteraceae)	A353	HybPiper-nf	Both	Schmidt-Lebuhn *et al.*	GAP special issue

Table 2 Summary of softwares and tools mentioned in this paper.

Tool	Use	Output	Citation and URL
1. Locus extraction			
HybPiper	Locus assembly and extraction	Sequence files for each locus, assembly reporting, paralogy reporting, exons and intron sequences	Johnson *et al.* 2016, Jackson *et al.* 2023; https://github.com/mossmatters/HybPiper; https://github.com/chrisjackson-pellicle/hybpiper-nf
HybPhyloMaker	Locus assembly and extraction, plus alignment, trees	Sequence files for each locus, assembly reporting, paralogy reporting, exons and intron sequences, alignments, gene trees, species trees	Fér and Schmickl 2018; https://github.com/tomas-fer/HybPhyloMaker
SECAPR	Locus assembly and extraction	Sequence files for each locus, assembly reporting, paralogy reporting, exons and intron sequences, phased loci	Andermann *et al.* 2018; https://github.com/AntonelliLab/seqcap_processor

PHYLUCE	Locus assembly and extraction, typically UCE's	Sequence files for each locus, assembly reporting, alignments	Faircloth 2016; https://github.com/faircloth-lab/phyluce
CAPTUS	Locus assembly and extraction	Sequence files for each locus, assembly reporting, paralogy reporting, exons and intron sequences, plus organellar sequences, alignments	Ortiz *et al.* 2023; https://github.com/edgardomortiz/Captus
NewTargets	Expanding target file phylogenetic breadth using available genomic resources	An expanded target file, curated to end-user needs for improved recovery.	McLay *et al.* 2021; https://github.com/chrisjackson-pellicle/NewTargets
1. Locus extraction: Post-assembly assessment and alignment filtering			
AMAS	Alignment assessment and manipulation (e.g. sample removal)	Various alignment formats, individual locus or concatenated alignments plus partition files	Borowiec 2016; https://github.com/marekborowiec/AMAS/
TrimAL	Alignment trimming based on several parameters (e.g. gappiness, informativeness, overlap)	Trimmed alignments in user specified format	Capella-Gutierrez *et al.* 2009; https://vicfero.github.io/trimal/
ClipKit	Alignment trimming based on several parameters (e.g. gappiness, informativeness, codon position)	Trimmed alignments in user specified format	Steenwyk 2020; https://github.com/JLSteenwyk/ClipKIT
CIAlign	Alignment trimming based on several parameters (e.g. gappiness, informativeness, length, alignment quality)	Trimmed alignments in user specified format	Tumescheit *et al.* 2022; https://github.com/KatyBrown/CIAlign
2. Paralogy reconciliation			

PPD	Uses sequence identity and heterozygous sites to identify and remove paralogs	Alignments with detected paralogs removed	Zhou *et al.* 2022; https://github.com/Bean061/putative_paralog
ParalogWizard	Refined reassembly of loci to identifiy divergent sequences (i.e. paralogs or alleles) and perform orthology inference	Alignments sorted into orthologous groups	Ufimov *et al.* 2022; https://github.com/rufimov/ParalogWizard
CAPTUS	Identification of paralogous sequences during pipeline by comparing to a reference sequence	Paralogous sequences can be sorted into different alignments with user-defined parameters, including 'best' and 'similarity', or all copies can be kept, or removed	Ortiz *et al.* 2023; https://github.com/edgardomortiz/Captus
HybPhaser	Infer parental lineages of putatively hybridogenic lineages	Phased alignments with paralogs removed	Nauheimer *et al.* 2021; https://github.com/LarsNauheimer/HybPhaser
PARAGONE	Implements Yang and Smith's (2014) collection of methods for resolving paralogy using gene tree topologies	Paralogy resolved alignments and gene trees from each Y&S algorithm	Jackson *et al.* 2023; https://github.com/chrisjackson-pellicle/paragone-nf
3. Species tree inference: Paralog-aware phylogenetic tree reconstruction			
ASTRAL-PRO	Two-step coalescent phylogenetics from multi-labelled trees (i.e including paralogous sequences)	Coalescent species tree with paralogous tips reconciled to species	Zhang and Mirarab 2022; https://github.com/chaoszhang/ASTER; https://github.com/chaoszhang/A-pro
FastMulRFS	Two-step coalescent phylogenetics from multi-labelled trees (i.e including paralogous sequences), using Robinson-	Coalescent species tree with paralogous tips reconciled to species	Molloy and Warnow 2020; https://github.com/ekmolloy/fastmulrfs

	Fould's distances to summarise paralogous sequences		
SpeciesRax	Likelihood inference of species tree from gene alignments or gene family trees	Species tree, and gene trees if starting with alignments	Morel *et al.* 2022; https://github.com/BenoitMorel/GeneRax
DISCO	Performs orthology inference of each gene tree to preserve orthologous sequences and discard paralogs	Coalescent species tree with paralogous tips reconciled to species	Willson *et al.* 2022; https://github.com/JSdoubleL/DISCO
AleRax	Likelihood inference of species tree from samples of estimated gene family trees	Species tree, reconciled and consensus gene trees, number of events	Morel *et al.* 2024; https://github.com/BenoitMorel/AleRax

3. Species tree inference: Phylogenetic tree reconstruction on single-sequence-per-species alignments

IQ-TREE	Likelihood phylogenetics on concatenated data	Phylogenetic tree (and other outputs depending on analysis)	Minh *et al.* 2020; http://www.iqtree.org
RAxML	Likelihood phylogenetics on concatenated data	Phylogenetic tree (and other outputs depending on analysis)	Stamatakis 2014,, Kozlov *et al.* 2019; https://github.com/stamatak/standard-RAxML; https://github.com/amkozlov/raxml-ng
ASTRAL	Two-step coalescent phylogenetics	Species tree	Mirarab *et al.* 2014, Zhang *et al.* 2018b; https://github.com/smirarab/ASTRAL
SplitsTree	Implements a range of network analyses, including the popular NeighbourNet and Consensus Network algorithms	Phylogenetic network	Huson and Bryant 2006; https://uni-tuebingen.de/en/fakultaeten/mathematisch-naturwissenschaftliche-fakultaet/fachbereiche/informatik/lehrstuehle/algorithms-in-bioinformatics/software/splitstree/; https://github.com/husonlab/splitstree6

ExaBayes	Bayesian phylogenetics on concatenated data	Phylogenetic tree (and other outputs depending on analysis)	Aberer *et al.* 2014; https://cme.h-its.org/exelixis/web/software/exabayes/
StarBeast	Bayesian inference of gene trees and species tree under the multispecies coalescent.	Posterior distribution and summary tree for species tree and gene trees	Douglas *et al.* 2022; https://github.com/rbouckaert/starbeast3
3. Species tree inference: Gene tree assessment and phylogenomic subsampling			
GeneSortR	Sorting and subsampling phylogenomic datasets to quantify phylogenetic usefulness	Sorted alignment, partition file, gene tree file and a plot of sorted genes by estimated properties, graphical summary of metrics employed to subsample	Mongiardino Koch 2021; https://github.com/mongiardino/genesortR
PhylteR	Identify outlier loci in phylogenomic datasets	Visualised output of outlier loci for removal	Comte *et al.* 2023; https://github.com/damiendevienne/phylter
TreeShrink	Pruning long, likely erroneous long branches from sets of phylogenetic trees	Pruned phylogenetic trees and corresponding alignments	Mai and Mirarab 2018; https://github.com/uym2/TreeShrink
SortaDate	Phylogenomic subsampling to identify most-clock like genes for phylogenetic dating	List of locus alignments of most clock-like genes	Smith *et al.* 2018; https://github.com/FePhyFoFum/sortadate
4. Conflict assessment			
IQ-TREE	Likelihood phylogenetics on concatenated data and locus alignments or partition file	Gene concordance factors (gCFs) and site concordance factors (sCFs) on phylogeny as branch labels	Minh *et al.* 2020; http://www.iqtree.org

PhyParts	Identification of concordant and conflicting bipartitions	Species phylogeny with concordance and conflict as branch labels	Smith *et al.* 2015; https://bitbucket.org/blackrim/phyparts/src/master/
ASTRAL	Measuring concordance/discordance by percentages of supporting quartets used to produce species tree	Species phylogeny with quartet concordance and conflict as branch labels	Mirarab *et al.* 2014; https://github.com/smirarab/ASTRAL
BUCKy	Estimating concordance factors from Bayesian MCMC trees of many loci	Species phylogeny with concordance and discordance scores as branch labels	Larget *et al.* 2010; https://pages.stat.wisc.edu/~ane/bucky/

5. Conflict investigation

HyDe	Detects hybridization in phylogenomic data sets	Values including identification of species and population level hybrids with ABBA-BABA tests	Blischak *et al.* 2018; https://github.com/pblischak/HyDe
JML	Detects hybridisation on time-calibrated trees, with information about population sizes	Distances between sequences for species pairs with *P*-values for hybridisation according to the posterior predictive distributions	Joly *et al*. 2012; https://github.com/simjoly/jml
Aphid	Estimating the contributions of gene flow and incomplete lineage sorting to phylogenetic conflict	Per gene tree output of conflict and the estimated cause (e.g. ILS or gene flow) in a csv file	Galtier 2024; https://gitlab.mbb.cnrs.fr/ibonnici/aphid
GRAMPA	Use homolog gene tree topologies (i..e. MULtrees) to identify placement and types of WGD	Tree and txt files detailing the estimated ploidy placement and type of polyploid	Thomas *et al.* 2017; https://github.com/gwct/grampa

QuIBL	Uses gene tree internal branch lengths to distinguish between hybridisation and deep coalescence	For each triplet in the species tree, an estimate of the relative contribution of the locus set to ILS or gene flow	Edelman *et al.* 2019; https://github.com/miriammiyagi/QuIBL
PhyloNet	Infer phylogenetic networks from sets of loci while accounting for both reticulation and ILS, using mostly maximum likelihood-based algorithms	Networks as Nexus files	Than *et al.* 2008, Wen *et al.* 2018; https://phylogenomics.rice.edu/html/phylonet.html
PhyloNetworks	Infer phylogenetic networks from sets of loci while accounting for both reticulation and ILS, under a coalescent model	Networks as Newick files	Solís-Lemus *et al.* 2017; https://github.com/crsl4/PhyloNetworks.jl
Quartet Sampling	Repeated sampling of quartets to analyze branch support on molecular phylogenies	Newick tree files of various scores, a FigTree file containing all scores, and statistics files	Pease *et al.* 2018; https://github.com/FePhyFoFum/quartetsampling
DiscoVista	Quantify and visualise a range of phylogenomic metrics including species tree and gene tree compatibility, branch quartet frequencies and GC content.	Figures showing gene tree discordance and relative frequency of different topologies, species tree discordance and taxon occupancy	Sayyari *et al.* 2018; https://github.com/esayyari/DiscoVista
ASTRAL	Perform a polytomy test to determine if the polyomy is ‘hard’ or ‘soft’	Species phylogeny with significance values that indicate the presence of a hard polytomy	Mirarab *et al.* 2014; Sayyari and Mirarab 2018; https://github.com/smirarab/ASTRAL