Skeels, A., Sauquet, H., Mast, A., Weston, P. H., Olde, P. M., Reynolds, Z. K. M., Fenker, J., Lemmon, A. R., Lemmon, E. M., & Cardillo, M. (2025). Sources of gene tree discordance and their implications for systematics and evolution of a megadiverse Australian plant radiation (subtribe Hakeinae, Proteaceae). *Bulletin of the Society of Systematic Biologists*. https://doi.org/10.18061/bssb.v3i2.9823

Investigations

Sources of Gene Tree Discordance and Their Implications for Systematics and Evolution of a Megadiverse Australian Plant Radiation (Subtribe Hakeinae, Proteaceae)

Alexander Skeels[1], Hervé Sauquet[2,3], Austin Mast[4], Peter H. Weston[2], Peter M. Olde[2], Zoe K. M. Reynolds[1], Jéssica Fenker[1,5], Alan R. Lemmon[6], Emily Moriarty Lemmon[4], Marcel Cardillo[1]

[1] Research School of Biology, Australian National University, [2] National Herbarium of NSW, Botanic Gardens of Sydney, [3] Evolution and Ecology Research Centre, School of Biological, Earth and Environmental Sciences, University of New South Wales, [4] Department of Biological Science, Florida State University, [5] Sciences Department, Museums Victoria, [6] Department of Scientific Computing, Florida State University

Keywords: Grevillea, phylogenomics, multispecies coalescent, concatenation, incomplete lineage sorting, Anchored hybrid enrichment

https://doi.org/10.18061/bssb.v3i2.9823

Bulletin of the Society of Systematic Biologists

Abstract

Resolving phylogenetic relationships in the presence of conflicting signal across genes is one of the major challenges of the phylogenomic era. Conflicting signal can emerge from biological processes, such as incomplete lineage sorting or introgression, or have technical origins, such as from misaligned sequences. Decisions made in the process of estimating species trees may therefore result in alternative tree topologies and large variation in branch support values with important taxonomic consequences. To explore how these methodological choices affect the estimation of relationships, we compare alternative strategies for alignment cleaning, loci filtering, and phylogenetic estimation for 551 taxa in the Proteaceae subtribe Hakeinae. We found that node support values across gene trees were generally low and gene discordance was high in the Hakeinae, and that the degree of discordance varied across the phylogeny, as well as geographically across Australia. Higher stringency of alignment cleaning tended to decrease node support, while removing undesirable loci tended to increase gene concordance. Cleaning, filtering, and phylogenetic estimation method (short-cut coalescent vs. concatenation) have significant effects on tree topologies with distinct clusters of similar topologies detected in tree space. In some cases, this has important systematic consequences for two of Australia's largest and best-known plant genera, with concatenated approaches resolving *Hakea* and *Grevillea* as reciprocally monophyletic, but coalescent approaches showing that *Hakea* is nested within *Grevillea*. Our results suggest that widespread gene discordance may be the result of rapid radiation and incomplete lineage sorting, demonstrating the importance of assessing the drivers of discordance to understand phylogenetic relationships.

Introduction

Synthesis

It has long been recognised that different genomic regions can have independent evolutionary histories (Goodman et al., 1979; Maddison, 1997), some of which may be in direct conflict, known as gene discordance (Rokas et al., 2003; Steenwyk et al., 2023). A major insight that has emerged in the phylogenomic era is that widespread discordance is the rule, not the exception, in large datasets (e.g. Roycroft et al., 2020; Singhal et al., 2021; B. T. Smith et al., 2023). Gene discordance can arise from homoplasy or technical error in data processing and analysis, for example in sequence alignment, orthology assessment, or model specification. However, it can also reflect real, biologically interesting phenomena in the history of a clade, such as horizontal gene transfer, incomplete lineage sorting (ILS) or introgression (Hahn & Nakhleh, 2016). Both technical error and biological processes can lead to non-random signal in phylogenetic data, and therefore determining the sources of discordance by distinguishing technical error from biological processes, and distinguishing between different processes, is a key, but often overlooked, step (Som, 2015; Steenwyk et al., 2023; Xi et al., 2014). Minimizing the likelihood that discordance is the result of technical errors may increase confidence that processes such as ILS or introgression have been at play, thus improving understanding of the evolutionary history of the group of interest. Changes in phylogenetic branching patterns that result from different ways of handling discordance could also have implications for systematics and classification, or for downstream analyses based on phylogenetic information such as diversification rates or spatial patterns of phylodiversity. Here, we investigate spatial and phylogenetic patterns of gene discordance in a large radiation of Australasian plants (Fam. Proteaceae: subtribe Hakeinae) and the taxonomic consequences of different analytical decisions around managing discordance.

The sources of discordance are relevant for the approach chosen to estimate species trees, which may have important systematic consequences. Supermatrix (concatenation) approaches do not explicitly consider gene tree discordance, and instead rely on the strength of the collective signal of the species tree across different loci. On the other hand, methods designed to accommodate discordance use (or are consistent with) the multi-species coalescent model. These methods explicitly consider gene tree histories but are typically slower to run and may not be tractable for large datasets. Shortcut coalescent species tree approaches, such as ASTRAL-III, have been shown to be good estimators of species-level relationships if ILS is the main source of discordance and levels of discordance are moderate to high (Mirarab & Warnow, 2015), gene tree estimation error is low (Roch & Warnow, 2015), and reticulate evolution is low (Solís-Lemus et al., 2016). When discordance is low or gene tree estimation error is high, supermatrix approaches may be more appropriate (Mirarab, 2019; Mirarab & Warnow, 2015), and when reticulate evolution is common, network methods may be preferred (Mirarab, 2019; Solís-Lemus et al., 2016). As such, estimated tree topologies may be heavily dependent on the presence or absence of discordance in the data and the sources of discordance. So how do we know where discordance comes from?

Different sources of discordance are expected to leave diagnostic signals in the topologies and branch lengths of gene trees, and are expected to lead to differing patterns of discordance across the branches of species trees (Meleshko et al., 2021; B. T. Smith et al., 2023; Zwickl et al., 2014). Sequencing or alignment error can lead to incorrect assessment of primary homology (Bromham, 2016) and is a major source of gene tree estimation error. Alternatively, some loci may have low phylogenetic signal in the absence of sequence error due to the evolution of few parsimony informative characters, so that 'gene shopping' for loci with high species coverage and a large number of parsimony informative sites may improve phylogenetic inference (Molloy & Warnow, 2018; S. A. Smith et al., 2018). Hence, a marked improvement in node support values of gene trees, or a marked decrease in gene tree discordance, after trimming and masking alignments and filtering sequences and loci, would be indicative of technical sources of discordance in the original data (Aberer et al., 2013; Salichos & Rokas, 2013; M. R. Smith, 2022b; Talavera & Castresana, 2007; Wilkinson, 1996; Zhang et al., 2021). On the other hand, discordance that results from biological processes, such as ILS or introgression, is unlikely to be completely resolved by alignment cleaning or locus and sequence filtering, and instead may be evidenced by non-random patterns in space and phylogeny.

The spatial and phylogenetic context of discordance could give insights into the drivers of diversification dynamics. Technical sources of discordance should be associated with phylogenetically or spatially random patterns of discordance, whereas biological sources might be non-randomly distributed in space or among the branches of the phylogeny. For example, a prediction of ILS is that discordance should be most prevalent in shorter branches in the phylogeny, which may have resulted from rapid radiations with little genetic differentiation and insufficient time for complete sorting of gene variants (Degnan & Rosenberg, 2006; Pease et al., 2016; Whitfield & Lockhart, 2007). As such, geographic regions with rapidly diversifying lineages should be characterized by short branch lengths and high discordance. If discordance is driven primarily by introgression, then spatial 'hotspots' of gene discordance could highlight ecological or environmental factors that promote hybridisation and gene-flow such as dynamic environments (Singhal et al., 2021) or speciation mode. In either case, locating regions with elevated rates of discordance compared to expected, could help to pinpoint the underlying ecological and evolutionary drivers of discordance.

Focal system

We explore the sources of, and patterns in gene tree discordance and their implications for phylogenetic inference and understanding the diversification history of the subtribe Hakeinae (Proteaceae; > 525 species). Hakeinae is a large group of shrubs and trees that are mostly endemic to Australia, with a handful of species found in New Guinea, New Caledonia, and Indonesia (Weston & Barker 2006). The subtribe contains five genera, including two of Australia's largest and most well-known genera, *Grevillea* R.Br. ex Knight (>360 species) and *Hakea* Schrad. & J.C.Wendl. (>150 species), together with three small genera: *Finschia* Warb. (3 species), *Buckinghamia* F.Muell. (2 species), and *Opisthiolepis* L.S.Sm (1 species). A recent study based on four plastid and one nuclear gene for a third of the species concluded that *Grevillea* is paraphyletic with respect to *Hakea* and *Finschia* (Mast et al., 2015). Monophyly of *Hakea* was supported by this study and is supported by a well-defined morphological synapomorphy – the fruit being a woody follicle with secondary thickening by a continuous cambium (Johnson & Briggs, 1975). On the other hand, *Grevillea* currently lacks an unambiguous synapomorphy and further investigation of its status is warranted (Mast et al., 2012). We expand the genomic coverage of the study of Mast et al. (2015) from five to 216 loci, and the coverage of Hakeinae species from 150 to 469 species (~90% of Hakeinae species). We test alternative bioinformatic and phylogenetic strategies, and estimate their effects on node support values, gene tree discordance, and topology, both across the phylogeny and spatially across Australasia.

Objectives

Specifically, we ask: (i) how do alternative data cleaning, loci filtering, and phylogenetic inference methods impact node support values and tree topology? (ii) What are the spatial and phylogenetic patterns of gene discordance and are these associated with biomes or branches of the phylogeny under the expectations of ILS or introgression? (iii) Under what treatments is the paraphyly of *Grevillea* supported by phylogenomic data? We predict that technical sources of discordance would be supported if alignment cleaning and loci filtering increases node support values and decreases discordance. On the other hand, these should have little effect if the major causes of discordance are biological, and instead we expect non-random phylogenetic and spatial patterns. Differentiating these sources of discordance should help determine which species tree topologies are most likely to represent the true divergence history of the group and can help us weigh up relative support for alternative generic relationships. If supported, *Grevillea* paraphyly would have obvious taxonomic consequences for the generic status of the major clades within Hakeinae (Mast et al., 2015). Because Hakeinae comprises an iconic and diverse component of Australia's temperate and arid environments, spatial and phylogenetic patterns of discordance would also have implications for understanding the tempo of diversification of the Australian flora and the origins of Australia's temperate biodiversity hotspots in the context of large-scale biome turnover during the Neogene (Cardillo et al., 2017; Mast et al., 2015; Sauquet et al., 2009).

Methods

In this study, we produced phylogenomic data using Anchored Hybrid Enrichment (AHE; Lemmon et al., 2012), with different alignment cleaning, loci filtering, and phylogenetic inference techniques. Our pipeline resulted in 12 species trees: for each alignment set, we obtained a single concatenated tree using IQ-TREE and three short-cut coalescent trees using ASTRAL-III with different data filtering steps (Fig. 1). We compared node support values and tree topologies between treatments and explored the resulting spatial and phylogenetic patterns of discordance.

Sample collection and DNA extraction

This study used 551 sequences for 482 Proteaceae taxa, of which 186 sequences were obtained from an earlier study (Cardillo et al., 2017). These 186 sequences include 151 species of *Hakea,* four species of *Grevillea* (*G. dimorpha*, *G. evanescens*, *G. hookeriana*, and *G. batrachioides*), and *Opisthiolepis heterophylla* of the Hakeinae, together with eight Proteaceae outgroup taxa: *Lomatia silaifolia*, *Stenocarpus davallioides*, *Alloxylon pinnatum*, *A. flammeum*, *Telopea speciosissima*, *Banksia paludosa*, *B. rufa*, and *Lambertia formosa*. In addition, 368 samples were collected for novel DNA extraction and sequencing, sourced from fresh plant tissue obtained from wild or cultivated plants, or from dried herbarium specimens (see Table S1 for a list of samples). These included 320 species of *Grevillea* and *Finschia chloroxantha* from the Hakeinae, and additional Proteaceae outgroups: *Stenocarpus milnei*, *Oreocallis grandiflora*, *Macadamia integrifolia*, *Adenanthos glabrescens*, *Lambertia formosa*, *Isopogon formosus*. Thirty-five *Grevillea* taxa were represented by multiple samples representing different recognised or putative subspecies.

Plant samples were prepared for DNA extraction by first being finely ground into powder using beads in conjunction with the TissueLyser II machine (Qiagen). For particularly tough samples, liquid nitrogen was employed to facilitate the grinding process. Total genomic DNA was subsequently extracted from the powdered samples using the Qiagen DNEasy Plant Mini Kit, following the manufacturer's protocol (Qiagen Inc., California, USA). The concentration of extracted DNA was measured using the High Sensitivity Qbit (Invitrogen) at the Environmental Biology Laboratory (EBL) at the Australian National University. In instances where further purification was necessary, DNA was subjected to an isopropanol precipitation method post-extraction to ensure optimal purity and quality.

Grevilleoideae-specific probe design

In order to obtain efficient sequencing of a diversity of loci in Grevilleoideae subfamily of Proteaceae (which contains the Hakeinae), we extended an existing AHE probe design (Bogarín et al., 2018; Lemmon et al., 2012). We followed the approach of (Hamilton et al., 2016) and (Bogarín et al., 2018), using the code and scripts described therein. To extend the AHE-specific design, our general approach was to develop Grevilleoideae-specific probes using newly collected WGS. More specifically, we collected low-coverage genome sequence data from four Proteaceae samples, *Banksia rufa* (sample ID=MC41, Table S1), *Lambertia formosa* (ID=ZR6), *Alloxylon pinnatum* (ID=ZR145), and *Macadamia integrifolia* (ID=ANBG_681026.3). We prepared Illumina libraries from DNA extracts of these four samples following (Prum et al., 2015). We then sequenced these libraries (the insert size of which was 150-350bp) on an XP flow cell of an Illumina NovaSeq 6000 sequencer using a PE150 bp protocol with dual indexing. After filtering low-quality read pairs using the Casava high-chastity filter, we merged overlapping read pairs following (Rokyta et al., 2012). The sequencing effort generated 20m-40m merged reads per sample. We subsequently mapped these merged reads to the Theobroma cacao reference sequence from the AHE V3 design (Bogarín et al., 2018) and used the merged reads to extend the resulting sequences into flanking regions. We then aligned the extended sequences to the Theobroma cacao reference sequences, inspected the alignments in Geneious (Biomatters Ltd.,(Kearse et al., 2012)), and trimmed alignment ends to remove misaligned sequences regions. Realizing that some of the original AHE loci may be derived from neighboring exons, we compared these extended sequences across loci and identified overlapping sequences. At this point we removed 41 alignments to ensure that we did not target redundant AHE loci. Finally, we identified and masked 9 repetitive regions.

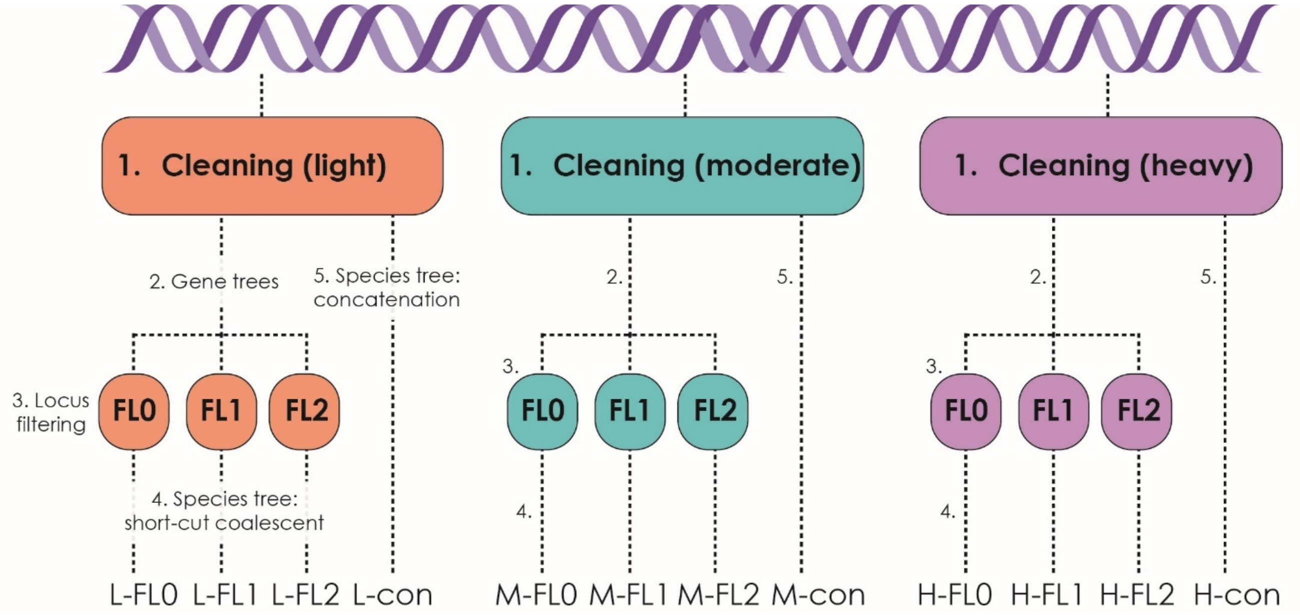

Figure 1. Alignment cleaning and filtering pipeline. 216 Anchored Hybrid Enrichment loci are (1) cleaned by trimming and masking sites under three scenarios (light, L; moderate, M; heavy, H). (2) Separate gene trees are estimated for each locus using IQ-TREE for each alignment set. (3) Loci were either completely retained (FL0) or filtered using two thresholds (moderate, FL1; heavy, FL2) based on taxon sampling, parsimony informative characters, and variation in branch lengths. (4) Species trees were estimated using a short-cut coalescent approach with ASTRAL-III. (5) We also concatenated loci to estimate a single species tree with IQ-Tree. This pipeline produced 12 alternative species trees (L-FL0, L-FL1, L-FL2, L-con, etc.).

Library preparation, read assembly, sequencing and orthology detection

Libraries for the 368 new samples were prepared at the FSU Center for Anchored Phylogenomics. In brief, we followed Lemmon et al (2012) and Prum et al. (2015) to prepare dual-indexed libraries with insert size 150-350. The libraries were then pooled in groups of ~24 and enriched using the probe set described above (The SureDesign XT kit was produced by Agilent Technologies). We performed sequencing at the FSU Translational Laboratory on an Illumina NovaSeq6000 sequencer with a PE150bp sequencing protocol (dual indexing). On average we collected 5.7 million reads per sample. Following sequencing, we removed low-quality reads using the Illumina Casava high-chastity filter.

The reads were processed following Hamilton et al. (2016; detailed methods are contained therein). In brief, we merged overlapping reads (Rokyta et al., 2012) then assembled reads using, as references, sequences used to develop the AHE Angiosperm Grev2 kit (see above). Finally, we established orthology using alignment-free distance estimates followed by clustering via neighbor-joining.

Alignment cleaning, loci filtering, and phylogenetic inference

Sequences in each orthologous cluster were first machine-aligned using MAFFT v7.023b (Katoh & Standley, 2013). Alignment error, through miscalled variants or mistaken alignment of non-homologous regions, may be reduced by masking ambiguously aligned sites (Talavera & Castresana, 2007; Zhang et al., 2021), which is increasingly common in phylogenomic analyses. However, testing the effect of alternative alignment cleaning strategies is rarely done (but see, for example, Helmstetter et al., 2024). We applied three alignment cleaning and trimming thresholds to each of the 216 “raw” machine alignments, as follows.

1. Light cleaning (Figure 1). The alignments were trimmed and masked using a common procedure for AHE data (Prum et al., 2015). Sites with the same character in >50% of sequences were considered “conserved”. A 20-base pair (bp) sliding window was then moved across the alignment, and regions with <14 characters matching the common base at the corresponding conserved site were masked.
2. Moderate cleaning (Figure 1). Starting with the “light” cleaned sequences, we removed sequences with <50% of sites containing unambiguous base pairs. A majority consensus sequence was estimated from the Hakeinae in-group taxa to detect major deviations resulting from poor alignment. To this end, a 7 bp sliding window was moved across each ingroup sequence separately and masked if a majority of bases (4 bp) differed from the consensus. We trimmed the alignments from each end until 66% (2/3) sequences had an unambiguous base using the function “msaTrim” in the R package *microseq* (Snipen & Liland, 2018). Similarly, using the 7 bp sliding window, we also masked segments with *3* ambiguous bases (~40%).

3. Heavy cleaning (Figure 1). Starting with the "light" cleaned sequences, we removed sequences with <50% of sites containing unambiguous base pairs. A majority consensus sequence was estimated from the Hakeinae in-group taxa. A 7 bp sliding window was moved across each ingroup sequence separately and masked if a majority of bases (4 bp) differed from the consensus. We trimmed the alignments from each end until 80% sequences had an unambiguous base. Similarly, using the 7 bp sliding window, we also masked segments with *5* ambiguous bases (~70%).

For each of the three trimmed and masked sets of 216 loci (light, *L;* moderate, *M*; heavy, *H*) we estimated separate unrooted gene trees using a maximum-likelihood approach in the software IQ-TREE (v. 2.2.2.6, (Minh, Schmidt, et al., 2020)). We first estimated the best fitting substitution model for each locus in each alignment set using a model selection procedure based on Bayesian information criterion in ModelFinder (Kalyaanamoorthy et al., 2017) before estimating the tree topology and branch lengths. We estimated node support values using ultrafast bootstrap (UFBoot; (Hoang et al., 2018)) and Shimodaira–Hasegawa-like approximate likelihood ratio tests (SH-aLRT; (Guindon et al., 2010)). In addition to estimating separate gene trees, we also estimated a single species tree for each alignment set with IQ-TREE using concatenation. Under this approach we first partitioned loci and used the best fitting substitution model from ModelFinder for each partition, then estimated a single topology and branch lengths and node support using UFBoot and SH-aLRT.

Taxa that have had errors in orthology detection (Springer & Gatesy, 2018) or have low phylogenetic information content (Salichos & Rokas, 2013) may be prone to shifting phylogenetic position. The removal of such 'rogue' taxa (Wilkinson, 1996) can improve phylogenetic inference (Aberer et al., 2013; M. R. Smith, 2022b). To estimate species trees using short-cut coalescent methods which can account for ILS, we first removed outlier sequences from the gene trees using the R package *PhylteR* (Comte et al., 2023). This method uses multidimensional scaling to detect outliers from multiple phylogenetic distance matrices simultaneously. Phylter detected 203 outliers across 31 loci in alignment set L, 108 outliers across 21 loci in set M, and 218 outliers across 32 loci in set H. After removing outliers, we estimated a species tree for each alignment set using ASTRAL-III (Zhang et al., 2018), which searches for the tree which maximises the number of shared species quartets at each node. ASTRAL-III is statistically consistent with the multi-species coalescent, can account for ILS, and is also equipped to handle the effect of hidden paralogy (Yan et al., 2022).

ASTRAL-III can be biased by gene tree estimation error. To account for this, in addition to producing a species-tree from all 216 loci (FL0; standing for filtered loci set 0), we applied two filtering strategies. Filtered loci sets 1 [FL1] and 2 [FL2] were devised to remove undesirable loci by simultaneously maximising the number of sequences present in each locus (FL1 = 90% of sequences, FL2 = 95% of sequences), maximising the number of parsimony informative characters (FL1 = 200, FL2 = 300), and minimising the coefficient of variation of root-to-tip distances in the gene trees as a measure of the evenness of evolutionary rates across lineages (FL1 = 0.75, FL2 = 0.5). We performed this filtering procedure on all three cleaned datasets (Figure 1). Between 133-140 loci were retained under FL1 across the three cleaning treatments and between 13-32 loci under FL2.

Poorly supported nodes in gene trees may bias the estimate of accurate species trees, so we followed the recommendation of (Mirarab, 2019) to collapse nodes with UFBoot < 10% into polytomies. This was done on gene trees estimated from all three cleaned datasets, before estimating species trees in ASTRAL-III with the three different loci filtering strategies. Shortcut coalescent trees estimated using ASTRAL-III have branch length in coalescent units and we estimated node support values using local posterior probabilities (LPP), a measure of the probability of the species tree branch representing the true branch given variation in the gene trees (Sayyari & Mirarab, 2016).

Phylogenetic and spatial patterns of discordance

To estimate the topological agreements between loci we estimated gene concordance factors (gCF) and site concordance factors (sCF) at each branch in the species tree with IQ-TREE (Minh, Hahn, et al., 2020; Mo et al., 2023). gCF is a measure of the percentage of gene trees which contain a given branch in the species tree, while sCF is a measure of the number of sites in the alignments which support a given branch in the species tree. We investigated any bias in phylogenetic structure of discordance to identify whether discordance was associated with distance from the root (in number of nodes) or with branch length (in expected number of substitutions). We did this for concatenated trees only as short-cut coalescent tree branch length are measured in coalescent units which are *a priori* associated with concordance. We fitted nested models gCF and sCF with both node depth and branch length and compared model fit with likelihood ratio tests. Both gCF and sCF are properties measured at nodes and therefore may represent non-independent data due to phylogenetic relatedness; we acknowledge that the non-phylogenetic models used here may conflate a statistical association with phylogenetic inertia, however here we are primarily interested in the degree of association between variables rather than inference.

To determine whether discordance is spatially non-random, which might be predicted if driven by biological rather than technical causes, we estimated an integrated species-level metric of discordance following (Singhal et al., 2021). Following their approach, we estimated a weighted average of gCF at all nodes subtending each tip in the species tree, to produce tip concordance factors (TCF). We then obtained occurrence records from the Atlas of Living Australia (ala.org.au; accessed July 2023) and reduced this data set to 188,442 records after filtering for geographic outliers using the R package CoordinateCleaner (Zizka et al., 2019), then examining each species by eye and comparing remaining records to the assumed distribution of each taxon from

FloraBase (florabase.dbca.wa.gov.au) and the Flora of Australia (Barker et al., 1999; Makinson, 2000). We matched the TCF to the occurrence records of each taxon and then found the average TCF value in 1x1 degree resolution grid cells across the distribution of Australasian Hakeinae in this study. To see if spatial patterns of TCF differ from what we would expect if TCF was randomly distributed across the branches of the phylogeny, we compared the observed TCF to a null distribution in which we randomly shuffled values of TCF across the phylogeny and re-estimated spatial patterns. We estimated standardised effect sizes (SES) of TCF from this null distribution following:

$$TCF_{SES} = \frac{TCF_{observed} - \ mean(TCF_{null})}{standard\ deviation(TCF_{null})}$$

To investigate the effect of different trimming and masking, loci filtering, and phylogenetic estimation methods on tree topologies, we estimated multi-dimensional distances between each species-tree using phylogenetic information distances (PID) and clustering information distances (CID) in the R package "treespace" (M. R. Smith, 2020, 2022a). We also estimated the support for the reciprocal monophyly of *Grevillea* and *Hakea* from our species trees and gene trees using the proportion of trees which support this topology.

To estimate the possibility of unidirectional introgression, we used Patterson's D statistic (ABBA-BABA test; (Durand et al., 2011; Green et al., 2010)) using a recently modified version of the ABBA-BABA test applied to gene trees (Rancilhac et al., 2021). For a given triad of species plus outgroup, under the assumptions of complete lineage sorting all gene trees should share a common allele topology (A, (A, (B, B))) at each node. Under ILS, the two alternative topologies (A, (B, (B, A))) and (B, (A, (B, A))) should occur in equal frequency (hence, ABBA-BABA). However, under introgression, one alternative topology should occur at greater frequency. Following Singhal et al. (2021), we estimated the effect of introgression in recently diverged monophyletic species triads. Across each of the 12 species trees, we removed subspecific taxa by randomly sampling one subspecies as a representative. This reduced the number of species in the dataset to 468. We coded all instances of fully inclusive, monophyletic species triads and identified their most closely related outgroup taxa. When multiple outgroup species existed we randomly selected a single species from the outgroup clade. For each triad we estimated the D-statistic and its statistical significance using 100 bootstrap replicates (Rancilhac et al., 2021). For each of the 12 species-trees, we used all gene trees (FL0) from the matching cleaning set (L/M/H) to estimate the D-statistic, rather than filtering the gene trees themselves which would vary the sample sizes across treatments. We also look at introgression more deeply in the tree by selecting paraphyletic species triads from across the tree. As it was unfeasible to analyse every combination of taxa, we selected 10,000 random triads and repeated the ABBA-BABA analysis. We did this for two datasets which showed the most different patterns in the recent-divergence ABBA-BABA tests (M-FL0 and H-FL2). We note that the ABBA-BABA test tests for the presence of directional introgression but cannot infer the direction itself.

Results

Anchored Hybrid Enrichment

We obtained contig assemblies for 551 taxa (537 Hakeinae and 14 outgroups), with a maximum of 1169 loci captured per individual, averaging 783 bp in length across the samples. The mean number of loci >250 bp captured per taxon was 580 and the mean number of loci >500 bp was 492. After removing loci with missing data for >50% of samples we were left with 216 loci for which we could obtain good alignments.

Alignment cleaning, loci filtering, and phylogenetic inference

The light (L), moderate (M), and heavily (H) trimmed and masked datasets differed in key attributes (Fig. S1): the proportion of 553 retained sequences (*L*: mean = 0.93, range = 0.52-1.0; *M*: mean=0.92, range = 0.52-1.0; *H*: mean = 0.92, range = 0.52-1.0), the number of sites per locus (*L*: mean = 706, range = 214-1190; *M*: mean=651, range = 214-1126; *H*: mean = 538, range = 214-1056), the average proportion of missing bp per sequence (*L*: mean = 0.10, range = 0.00-0.22; *M*: mean=0.06, range = 0.00-0.12; *H*: mean = 0.03, range = 0.00-0.08), and the number of parsimony informative sites per locus (PIS; *L*: mean = 373, range = 50-639; *M*: mean=334, range = 50-586; *H*: mean = 267, range = 50-557).

Node support values across gene trees, measured using Ultrafast bootstrapping (UFBoot) and the Shimodaira-Hasegawa approximate likelihood ratio test (SH-aLRT), were generally low in short-cut coalescent trees (Fig. 2), with mean UFBoot values across gene trees of 61% and a mean SH-aLRT of 30.01%. Increased alignment cleaning led to a decrease in node support values across gene trees (Fig 2). On average, only 24.6% of nodes were strongly supported with SH-aLRT values greater than 80% (L=26%, M=25%, H=22%) and 29.1% of nodes were strongly supported with UFBoot values greater than 95% (L=31%, M=30%, H=26%). Filtered subsets of genes had marginally higher values using both node support measures, with the average number of nodes supported by UFBoot being 31.1% for FL1 and 33.6% for FL2. The average number of nodes supported by SH-aLRT was 25.6% for FL1 and 28.1% for FL2. Similarly, in the filtered subsets, increased alignment cleaning led to decreased node support with both metrics. Using ordinary least squares (OLS) regression, we found that node support values were predicted by the number of sites, proportion of parsimony informative sites, proportion of missing data, the variance in root-to-tip distance of the gene trees, and the number of sequences in the alignment (Fig. S2). Together, these predictors explained 66% of variation in UFBoot values (R^2=0.66, F(5,641) = 244.11, P < 0.0001) and 65% of variation in SH-aLRT values (R^2=0.65, F(5,641) = 245.19, P < 0.0001).

When concatenating loci and estimating species-trees using IQ-tree, we found node support values were much higher than for individual gene trees. For all three alignment sets, mean SH-aLRT values were greater than 92% (L-

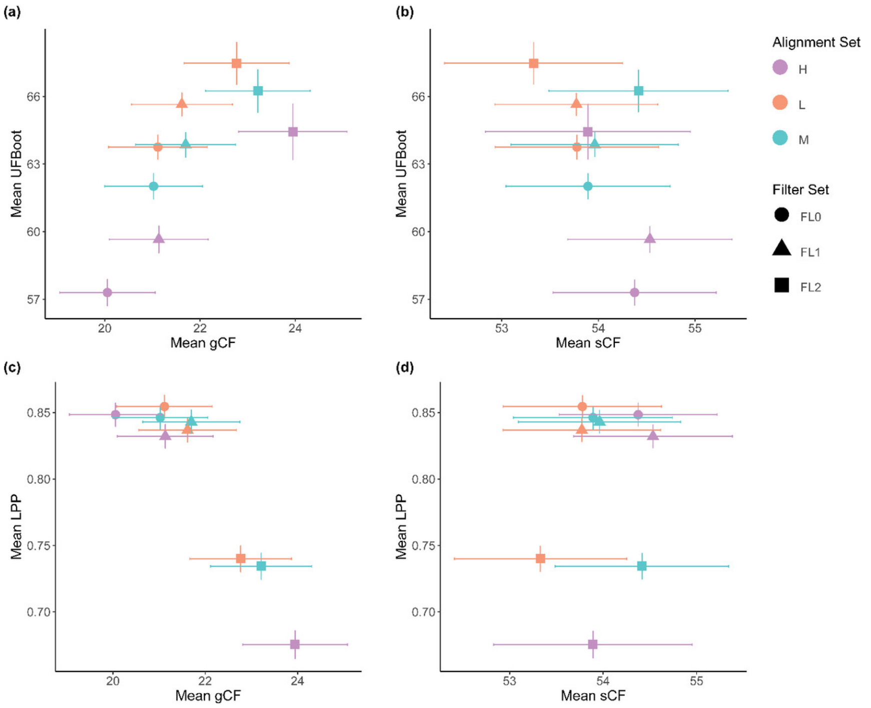

Figure 2. relationship between mean node support values and concordance factors in short-cut coalescent species trees for alternative alignment cleaning sets (colour) and filtering sets (shape). (a) ultrafast bootstrap (UFBoot) and gene concordance factors (GCF), (b) UFBoot and site concordance factors (sCF), (c) Local posterior probabilities and gCF, and (d) LPP and sCF. Error bars represent 1 standard error for each axis. Values for UFBoot were very similar to SH-aLRT which are shown in Figure S3.

con=93.2%, M-con=92.7%, H-con=92.2%) and mean UFBoot values were greater than 94% (L-con=95.2%, M-con=94.7%, H-con=95.5%). In all concatenated trees, more than 99% of nodes were considered strongly supported based on SH-aLRT values greater than 80% and UFBoot values greater than 95%.

Mean local posterior probabilities (LPP) on short-cut coalescent species-trees measures node support across gene trees. We found that mean LPP values ranged between 0.68-0.86 across treatments, with higher average values found in unfiltered sets of loci (FL0=0.85, FL1=0.83, FL2=0.72) and in lighter cleaned alignment sets (L=0.81, M=0.81, H=0.76; Fig. 2). Overall, the proportion of nodes considered well supported (LPP>=0.9) was < 0.4 in high filtering subsets (FL2) and > 0.6 in low filtering subsets (FL0). We estimated very high levels of discordance among loci with low average gene concordance factors (gCF) and site concordance factors (sCF). Across all cleaning and filtering scenarios, average gCF values ranged between 20 and 24 (Fig. 2), meaning that nodes in the species tree were present in between 20-24% of loci. In contrast, sCF values were higher and ranged between 53-55 across all cleaning and filtering strategies (Fig. 2). Although there was little variation in concordance factors, the highest sCF values were observed in the concatenated trees, while the highest gCF values were observed in the ASTRAL trees which were more heavily filtered (FL2), suggesting that concatenation maximises the congruence of signal across all sites in the genome, while coalescent based approaches maximise congruence between alternative gene trees.

Phylogenetic and spatial patterns of discordance

In the concatenated species trees, using likelihood ratio tests we found that branch length and node depth together explained significant variation in gCF and sCF but not in UFBoot values, which were best predicted by node depth independently. This result held across all three cleaning treatments. Overall, the effect size of node depth was considerably smaller than that of branch length (Fig. S4). Together, these results suggest branch length is the strongest predictor of concordance and that there is a tendency for more concordance in relationships estimated in more recently divergent lineages (further from the root).

Following Singhal et al. (2021), we estimated a weighted tip-metric of gCF for each species in the phylogeny (tip concordance factors; TCF). We then estimated the standardised effect size of TCF for species found within each one-degree grid cell across Australia and southern New Guinea (Fig 3a). Patterns of TCF (SES) are spatially structured at the continental scale, with species belonging to well supported lineages accounting for a greater diversity in the Tropical Grasslands biome of northern Australia (Fig 3b-c), as well parts of the Mediterranean and southern Arid biomes (Fig 3a). In contrast, the Pilbara and Great Sandy Desert regions in the northwest of the Arid biome as well as the Temperate Forests and Montane biomes of southeastern Australia had the lowest average TCF values (Fig 3c).

Across the 12 species trees, there were between 61 and 74 monophyletic, inclusive species triads, representing

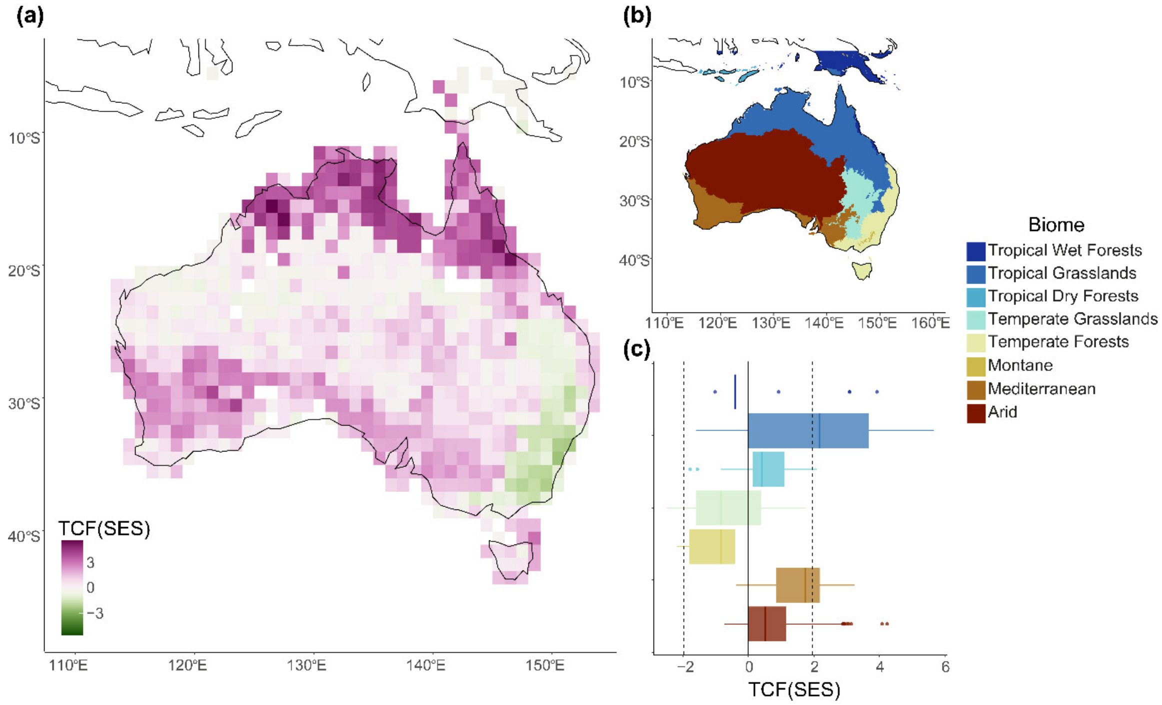

Figure 3. Standardised effect size (SES) of Tip Concordance Factor (TCF) (a) and its distribution within major biomes in Australia (b-c). TCF values above or below two standard deviations (dashed lines) from the mean, centred on zero (bold line) indicate significant deviation from the null model.

more recently diverged sets of species. Of these sets of triads, between 9.8% and 37.5% of triads showed patterns of elevated alternative topologies from ABBA-BABA tests, suggesting possible introgression. The highest proportion of significant D-statistics was found in the alignment sets with the highest cleaning and filtering (H-FL2 = 37.5%, M-FL2 = 27.2%). On the other hand, the lowest levels were seen in the moderately cleaned dataset (M-FL0 = 9.8%, M-FL1 = 10.1%). Across the 12 species-trees, the average value was 20.5%. When sampling triads broadly across the tree, including polyphyletic triads, to look for a signal of historical introgression, we found significant D-statistics in 25.5% of the 10,000 triads in H-FL2 and 5.84% in M-FL1, which are each slightly lower than their estimates from only monophyletic triads.

Topological differences

We found that alignment cleaning strategies, locus filtering, and concatenation influenced tree topology in Hakeinae. We performed K-means clustering on species tree distances estimated with CID (Fig 4a). We found five independent clusters of trees best explained partitioning of the data. These five clusters separated trees inferred from concatenated loci (L-con, M-con, H-con; Cl. 1) from short-cut coalescent trees with low data filtering (FL0, FL1; Cl. 2) while each highly filtered tree (FL2) was assigned its own cluster (Cl. 3-5). This suggests that phylogenetic estimation method (concatenation vs. coalescent) and locus filtering strategies can both impact branching relationships in the inferred phylogeny. We further explored the distribution of trees in tree space using the minimally cleaned alignments.

The subtribe Hakeinae and genus *Hakea* were both reconstructed as monophyletic in all treatments. Further, in all treatments we found that *Grevillea* was paraphyletic with respect to *Finschia* (Fig. 4; Fig. S5[a-l]). When using concatenation, *Grevillea* + *Finschia* (hereafter *Grevillea*) and *Hakea* were reciprocally monophyletic (Topology-1, Figure 4b). A consistent difference between phylogenetic inference strategies and filtering methods was the reconstruction of a monophyletic *Grevillea*: we identified three subclades of *Grevillea* whose relationships to each other, and to *Hakea*, varied between different cleaning and filtering strategies (Fig 4b-e): "*Grevillea* A" contained a pair of species (*G. acaciodes*, *G. endlicheriana*), "*Grevillea* B" contained eight species (*G. gordoniana*, *G. althoferorum*, *G. rudis*, *G. paradoxa*, *G. petrophiloides*, *G. rogersoniana*, *G. tenuiflora*, *G. pulchella*), "*Grevillea* C" contained the remaining *Grevillea* species and *Finschia*. A single alignment set shared Topology-1 with the concatenated trees (FL2-L; Fig. 4b; ((*Grevillea* B, *Grevillea* C), *Grevillea* A), *Hakea*). Three alignment sets (FL1-L, FL1-H, FL2-M) displayed Topology-2 in which *Hakea* is nested within *Grevillea* (Fig. 4c; ((*Grevillea* B, *Grevillea* C), *Hakea*), *Grevillea* A))). A further three alignment sets (FL0-M, FL0-H, FL1-M) displayed Topology-3, which shows a different pattern of paraphylly for *Grevillea* (Fig. 4d; ((*Grevillea* B, *Hakea*), *Grevillea* C), *Grevillea* A). Two further topologies were found in only single alignment sets: FL0-L is the only instance showing Topology-4 (Fig. 4e; (((*Grevillea* C, *Hakea*), *Grevillea* B), *Grevillea* A)), while in FL2-H shows Topology-5 (Fig. 4f) in which *Grevillea* C was paraphyletic with respect to *Hakea* and *Grevillea* B, and together those clades were sister to *Grevillea* A (Fig. S5[i]).

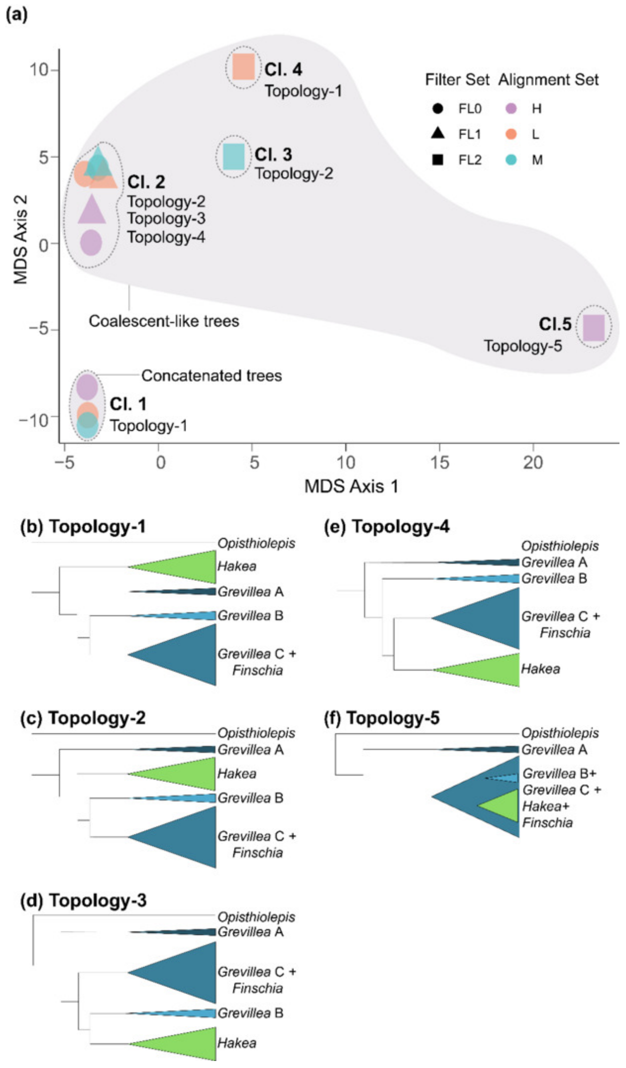

Figure 4. Multidimensional scaling (MDS) of clustering information distances representing topological differences in phylogenetic tree space (a). Five distinct clusters (Cl. 1-5) were identified in tree space: Cluster 1 contains the three concatenated trees; Cluster 2 contains the less filtered short-cut coalescent species trees from ASTRAL-III (FL0-FL1); Cluster 3-5 are each a separate alignment cleaning set with high degree of loci filtering (FL2). Panels b-e show the four most common generic level relationships within the Hakeinae.

We explored the position of *Hakea* across all gene trees which had information on outgroups to root the tree (181 loci, 84%) and found that in 162-164 gene trees in each alignment set (>90%), *Hakea* was nested within *Grevillea*. On the other hand, *Grevillea* was nested within *Hakea* in between 16-30 gene trees (9%-17%). *Grevillea* and *Hakea* were resolved as sister taxa in only 11-14 gene trees (6%-7%). These ratios were roughly similar when only considering filtered loci. This suggests that a topology with *Hakea* nested within *Grevillea* is the most common across gene trees, however these relationships are more variable depending on choice of alignment cleaning, loci filtering, and choice of phylogenetic inference method.

Discussion

Gene tree discordance is commonly inferred in phylogenomic studies of disparate taxa from across the tree of life. We found that this is also true in the case of Proteaceae subtribe Hakeinae, with on average only about 20% of genes supporting any given node in the phylogeny. In our data, several lines of evidence (discussed below) suggest that extensive gene discordance is most likely the result of low phylogenetic signal and ILS during a rapid radiation, rather than alignment error or widespread introgression. Taken together, our results also point towards a paraphyletic *Grevillea*, and we show that failing to take gene discordance into account can lead to alternative inference of reciprocal monophyly of *Grevillea* and *Hakea*. Distinguishing between these two alternatives topologies has important consequences for how we consider the evolution of Australia's diverse sclerophyllous flora and important hotspot genera.

How does alignment cleaning and loci filtering affect phylogenetic inference?

We found that additional masking of sites with missing data or clearly erroneous windows of sequences under our three alternative cleaning strategies made distinguishable differences to node support values, the inference of introgression, and species tree topologies, but did very little to affect detected rates of gene concordance. Additional cleaning led to a reduction in average node support values across gene trees as measured by UFBoot and SH-aLRT. This conforms with previous studies that have shown that 'over' cleaning tends to decrease support (Aagesen, 2004; Portik & Wiens, 2021; Tan et al., 2015), and which may be explained by the removal of informative signal alongside potential error when cleaning heavily. We also found that filtering the subset of gene trees with the most phylogenetic information tended to increase both node support and concordance factors, which has also been found in previous work (Hutter & Duellman, 2023), although the effect on concordance factors in our study was weak (1-2% difference). Interestingly, while filtering trees improved gCF, it had little effect on sCF, and in fact decreased node support values on the species trees (LPP) considerably. As such, choices about how much cleaning and filtering of alignments should be done will depend on what estimate of tree support is maximised, e.g., maximising average UFBoot across gene trees versus LPP on the species tree require different cleaning and filtering strategies (Fig. 1).

In addition to the lowest LPP, the most filtered datasets (FL2) each had uniquely divergent topologies and the highest estimated rates of introgression. Previous studies have shown that excessive filtering can reduce the accuracy of species-tree inference (Portik & Wiens, 2021) and that a minimum number of loci are required to resolve relationships in large clades (Shekhar et al., 2017) and we believe these issues make overfiltering problematic for the Hakeinae. Instead, light to moderate cleaning (L, M) and up to moderate amounts of filtering (FL0-FL1) tend to have similar LPP, UFBoot, gCF, and sCF values (Fig. 2), as well as more similar topologies (Fig. 4). Our results highlight that resolving a single species-tree with confidence is difficult even with genome-scale data, and that cleaning and filtering data does not necessarily improve concordance and node support values across all metrics, which is becoming clear from phylogenomic exploration of other recalcitrant radiations (Thomas et al., 2021). However, overall, we found that less stringent alignment cleaning (L treatment) and the less stringent locus filtering strategy (FL1 treatment) optimises both concordance factors (sCF, gCF), gene tree node support values (UFBoot, SHaLRT), and species tree node support values (LPP) and is recommended relative to high amounts of cleaning and filtering or no filtering at all.

Patterns of gene discordance across phylogeny and geography

Characterising patterns of gene discordance can help identify its ultimate causes. In Hakeinae, gene discordance was strongly associated with short internode distances and more weakly with the number of nodes separating the root (node depth) in concatenated trees. Short internodes are associated with rapidly diverging taxa (Pease et al., 2016) and can lead to high levels of discordance due to the limited time for genes to sort into lineages (Whitfield & Lockhart, 2007) or for parsimony informative characters to evolve between lineages (Rokas & Carroll, 2006). It is predicted that deeper nodes in the tree may also have higher discordance because of low phylogenetic signal in a dataset. The deepest nodes should be the hardest to resolve as they require slowly evolving regions of the genome with enough informative sites to characterise differences between taxa. The AHE approach employed in this study is well positioned to resolve these nodes as it targets genomic regions with both conserved and divergent sites and is aimed at estimating deep and shallow divergences (Lemmon et al., 2012). In fact, here we find the opposite pattern, in which node depth is negatively related to concordance factors, such that the nodes most separated from the root tend to be the most poorly resolved (Fig. S4). Our ability to resolve deeper nodes in the phylogeny more confidently compared to other recent studies may be due to the shallower phylogenetic timescales of this study. For example, discordance was positively related to node depth in analysis of deeply divergent metazoan taxa compared to shallower vertebrate taxa (Salichos & Rokas, 2013). It may also be the case that if rapid radiation is a key cause of discordance, more discordance in recently diverged taxa may be the result of recent diversification of the group, for example in response to major Miocene and Pliocene climate change in Australia, but this remains to be tested.

Not only is discordance non-randomly distributed across the phylogenetic tree, but it is also spatially non-random. Using a metric which integrates concordance factors along the branches leading to each tip (Singhal et al., 2021), we found that the lineages with the highest levels of concordance were found in the Tropical Grasslands biome of northern Australia and the Mediterranean biome of southwestern Australia – a centre of diversity for the clade and likely distribution of the common ancestor of extant *Hakea* (Cardillo et al., 2017). This matches expectations for a biological origin of discordance, as we can think of no reasons to expect such strong phylogenetic and spatial patterns if the primary causes of discordance are technical, analytical, or linked with data issues. We show this directly using a null model of tip-discordance which shuffled values randomly across the phylogeny and this generated a more spatially even distribution of tip discordance values (Fig S6). Based on predictions from our results showing greater concordance in longer, deeper branches of the tree, the prevalence of greater concordance in the Tropical Grasslands biome of the northern monsoonal tropics and the southwest Mediterranean biome might be related to the older presence of lineages in these regions, relative to the more environmentally recent and dynamic arid zones and temperate biomes of south-eastern Australia. There is some evidence for this in *Hakea* as the crown node of this group was inferred as having a Mediterranean-biome origin about 30 Ma, and the three main lineages from the monsoonal tropics having occupied that biome between 25-15 Ma, compared to the temperate forest lineages which occupied that biome < 15 Ma (Cardillo et al., 2017). The temperate forests biome in southeastern Australia is also the most topographically complex region of the continent and includes Australia's small area of alpine ecosystems. This complex topography, together with increasing aridity during the Miocene (Byrne et al., 2011), may have promoted recent and rapid diversification in this biome leading to greater discordance among its lineages. Although this region is not a major centre of diversity for most plant lineages in Australia, the Temperate Forests biome of southeastern Australia have been estimated to have high rates of diversification in a number of lineages and to be typically younger than southwestern counterparts due to greater rates of extinction during the Cenozoic (Nge et al., 2020). This hypothesis could be further tested by estimating the temporal biogeographic and diversification history of the Hakeinae and their close relatives across Australia.

Introgression in recently diverged taxa

Introgression between divergent taxa is one possible driver of gene tree discordance. In our sample of taxa, we detected introgression in between 10-35% (20% on average) of recently diverged monophyletic species triads, and between 5-25% of triads more broadly sampled from across the tree (including polyphyletic ones) depending on the alignment set. This result suggests some degree of introgression which could be widespread, particularly in recently diverged taxa. The highest proportions of introgression were detected when using the FL2 species-tree topologies, which are the most different to the remaining species-trees (Fig. 4), which may inflate the detection of introgression if these topologies are incorrect. Hakeinae, and particularly *Grevillea*, are commonly hybridised for cultivation (Pharmawati & Macfarlane, 2013). There is also evidence of speciation by natural hybridisation in a species complex of *Grevillea* in New Caledonia (Pillon et al., 2023). The average rates of introgression of recently diverged taxa across alignment sets (~20%) is towards the lower range of values detected in many other groups, including sub-oscine passerines (37%, Singhal et al., 2021), Andean Asteraceae (40%, (Vargas et al., 2017)), and wild tomatoes (88%, (Hamlin et al., 2020)). The Reduced Hybridization Hypothesis (RHH) suggests that natural hybridisation should be rare in the Mediterranean-biome region of Southwest Australia - the centre of diversity for the Hakeinae (Hopper, 2018). The RHH proposes that climatically disturbed or dynamic regions have high rates of hybridisation, because lineages may be shuffled around and be more likely to come into secondary contact following speciation, or because weak reproductive isolation barriers are selected for to promote genetic diversity and adaptive variation from introgression. Old, climatically buffered infertile landscapes (OCBILS) such as Southwest Australia, on the other hand, should have less hybridisation because populations may be more likely to be narrow range endemics which persist in isolation for long timespans and have genomic mechanisms to retain genetic diversity which might prevent natural hybridisation (The James Effect, (Hopper, 2009)). Supporting this, Hopper (2018) found that only 0.29% of over 14,000 Grevillea herbarium specimens from southwest Australia were from naturally hybridised taxa. However, in one of the analyses with the most identified introgression events (M-FL1), we found that 10 out of 16 introgressed species-pairs were from southwestern Australia, suggesting that although introgression may be modest in the group overall, it is not rarer in the southwest.

Paraphyly of *Grevillea*

We found that node support values estimated using UFBoot and SH-aLRT were much higher in the concatenated analysis compared with those in individual gene trees. Taken at face value this would suggest that concatenation does a better job of estimating robust phylogenetic relationships. However, high bootstrap support in concatenation analysis is expected as a larger number of informative characters can increase support without an increase in phylogenetic signal (Rokas & Carroll, 2006) and this may be misleading in the presence of heterogeneous evolutionary histories between genes (Salichos & Rokas, 2013). High bootstrap support despite widespread conflict among genes has been shown in taxa from across the tree of life, including murid rodents (Roycroft et al., 2020), yeast (Salichos & Rokas, 2013), and wild tomatoes (Pease et al., 2016; Salichos & Rokas, 2013). The reciprocal monophyly of *Hakea* and *Grevillea*, which was recovered in the concatenated analyses (Topology-1; Fig. 4), was present in only 6-7% of loci and only a single short-cut coalescent species tree using a highly filtered dataset (L-FL2), which suggests that the signal for this topology is uncommon across the Hakeinae genome. As proposed by Mast et al. (2015), a paraphyletic *Grevillea* would be consistent with a lack of any well-defined morphological synapomorphy of the genus (although this is not a requisite for monophyly). Our results strengthen the case for revising the generic classification of Hakeinae, although uncertainty is retained through the concatenated analyses and the very low confidence around the placement of the major clades in *Grevillea* (Topologies-2-4). This may be the result of very rapid divergences early in the clade's history with very low signal to resolve these higher-level relationships in our dataset. The necessary taxonomic changes to recircumscribe Grevillea as a monophyletic genus, must consider alternative topologies and their ultimate causes, and will be explored in a separate paper, considering further sampling of Hakeinae and its outgroups in the Grevilleoideae.

Conclusions

Our results are consistent with a complex scenario of incomplete lineage sorting and potential gene tree estimation error in the presence of low phylogenetic signal. These factors are illustrated by widespread discordance that is concentrated in short branches towards the tips of the tree and is spatially clustered in younger and more topographically complex biomes which may have undergone rapid radiation in the recent past. Together, these factors make estimating well resolved relationships in the Hakeinae difficult. Ways of addressing these factors including choice of phylogenetic inference method, alignment cleaning and 'gene shopping' strategies all have a large effect on the resulting node support values and tree topologies, with important consequences for the taxonomy of some of Australia's largest and most well-known plant genera. Our results are consistent with other recent studies showing that too much cleaning and filtering may lead to worse inference of species-level relationships, and that concatenation versus short-cut coalescent approaches can lead to alternative topologies, but small to moderate amounts of cleaning generally do not make huge differences to the outcomes. Our results highlight that some important nodes in the phylogeny can remain recalcitrant despite larger data sets and more advanced computational methods for phylogenetic inference, and that assessing the effect of different alignment cleaning, filtering, and phylogenetic inference strategies is important to survey uncertainties and limitations in our current bioinformatics and phylogenetic pipelines.

Additional

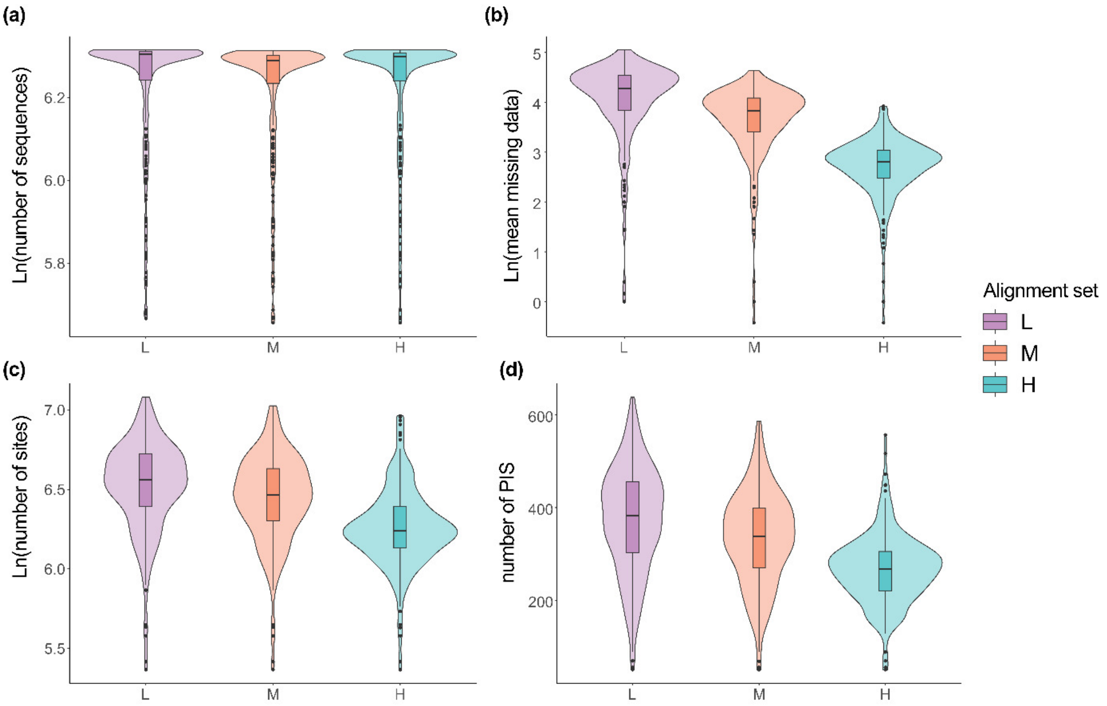

Figure S1. Violin plots showing distribution and density of the natural logarithm (Ln) of the number of sequences per locus, number of sites per locus, mean number of missing base pairs per sequence, and the number of phylogenetically information sites (PIS) across the three alternative alignment cleaning treatments.

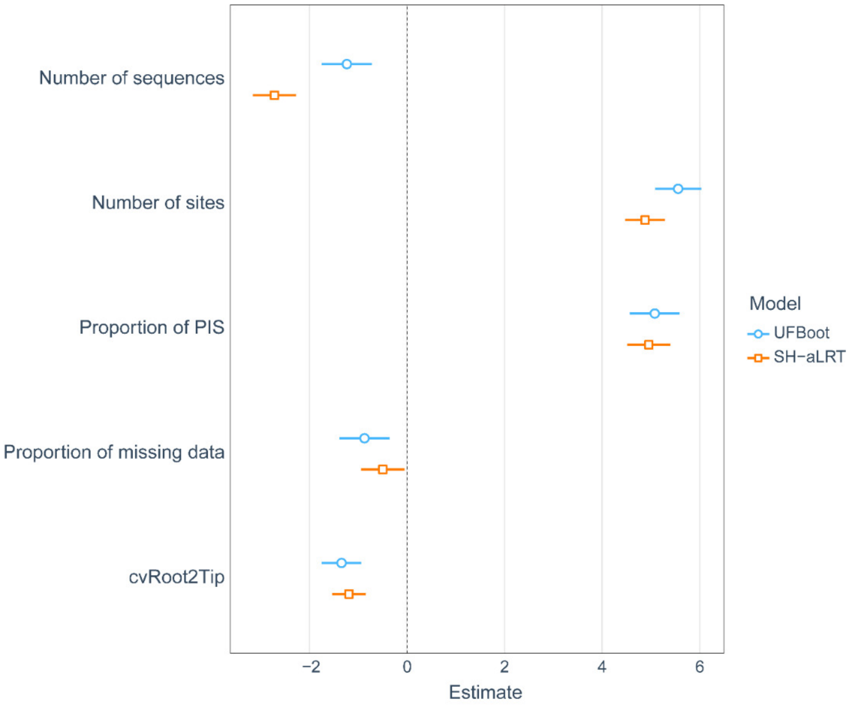

Figure S2. Coefficient estimates from a multiple regression of ultrafast bootstrap (UFBoot) and Shimodaira–Hasegawa-like approximate likelihood ratio test (SH-aLRT) values of node support support and five measures of alignment information and quality: the number of sequences present, the number of sites, the proportion of parsimony informative sites (PIS), the proportion of missing data, and the coefficient of variation of root-to-tip distance (cvRoot2Tip) as a measure of clocklikeness.

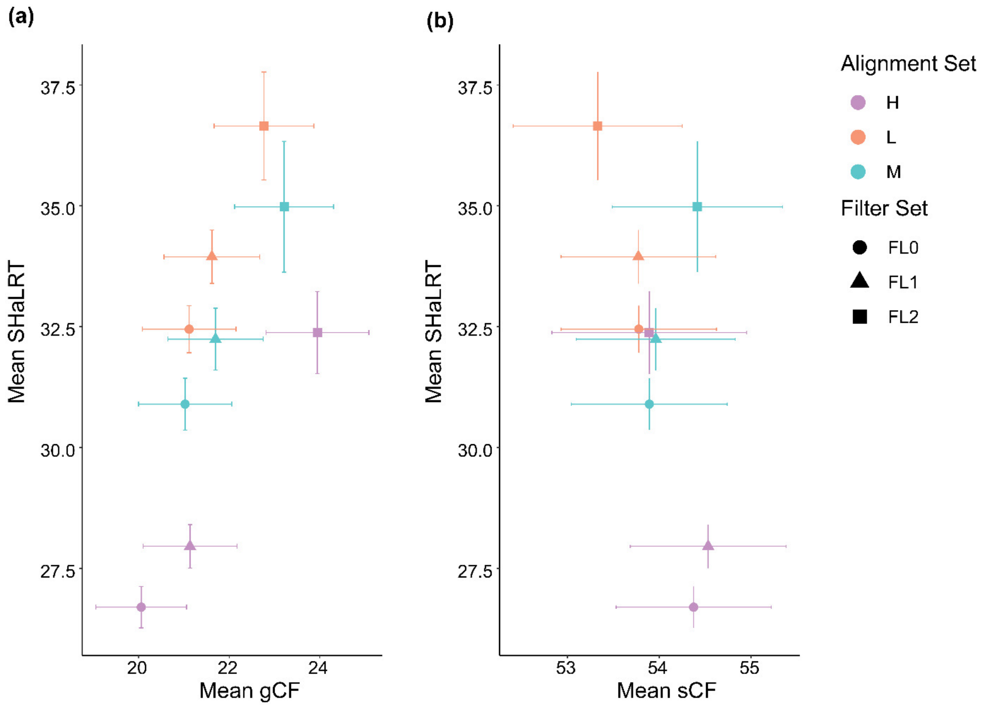

Figure S3. Relationship between mean Shimodaira–Hasegawa-like approximate likelihood ratio test (SH-aLRT) and concordance factors in short-cut coalescent species trees for alternative alignment cleaning sets (colour) and filtering sets (shape). (a) SH-aLRT and gene concordance factors (GCF), (b) SH-aLRT and site concordance factors (sCF). Error bars represent 1 standard error for each axis.

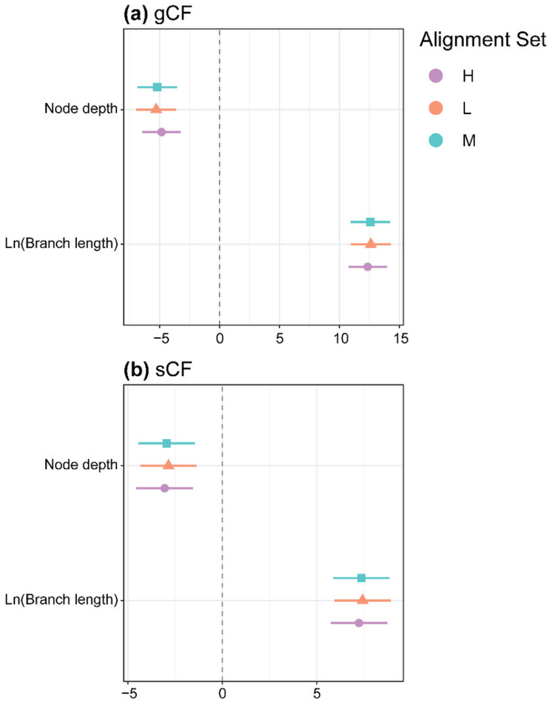

Figure S4. Coefficient estimates from a multiple regression of site concordance factors (sCF) and gene concordance factors (gCF) and node depth (number of nodes separating the branch from the root) and the natural logarithm of branch length (estimated number of substitutions per site).

Figure S5. Please view at the DRYAD repository https://doi.org/10.5061/dryad.6t1g1jx6b.

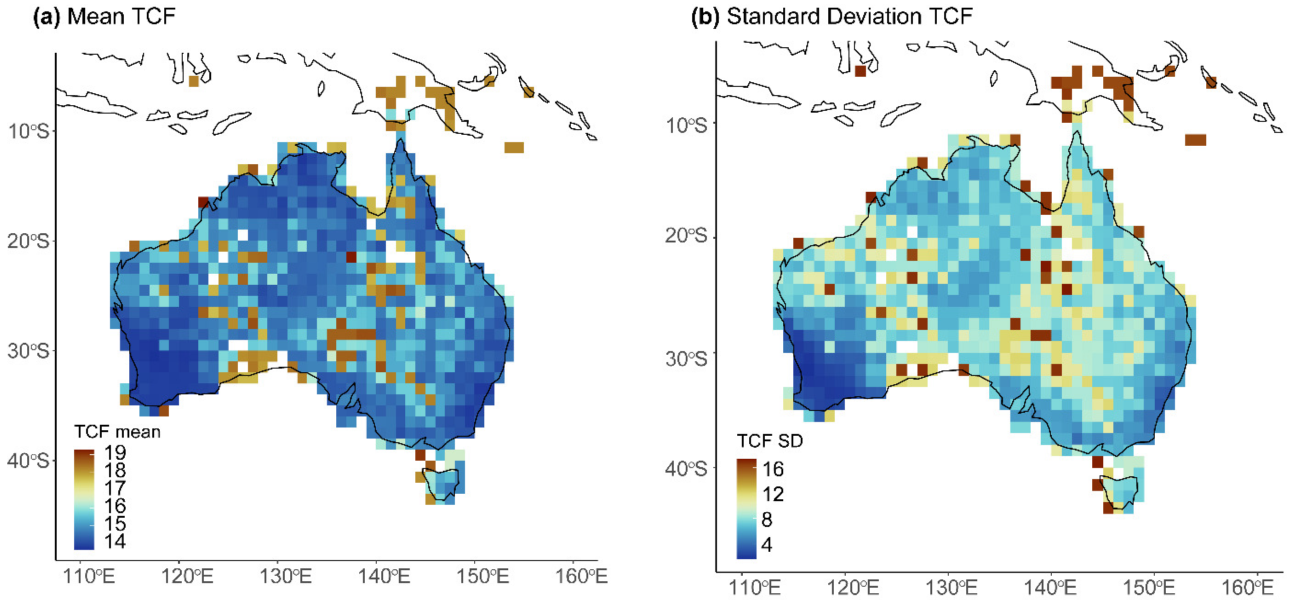

Figure S6. Distribution of the mean and standard deviation of tip concordance factors (TCF) from a tip shuffle randomisation across 1 degree x 1 degree grid cells in Australia

Funding

This research was supported by the Australian Research Council grant DP200102927. We would further like to thank the Grevillea Study Group of the Australian Plant Society (ANPSA) for covering expeditionary expenses in relation to the collection of *Grevillea* and *Hakea* specimens. We appreciate early feedback and discussion about our research with Sarin Tiatragul, Keaghan Yaxley, Ryan O'Donnell, and Rob Lanfear at the Australian National University.

Data availability

All sequence data, sequence information including accession and location information (Table S1), Figure S5, cleaned spatial occurrence data, and data processing scripts in the R language are available in the supplementary information in the DRYAD repository https://doi.org/10.5061/dryad.6t1g1jx6b. Raw occurrence records for the Proteaceae were obtained from the Atlas of Living Australia on 24 July 2023: (https://doi.org/10.26197/ala.f0b551b1-fff7-4e18-a17b-31d6b1719c1b).

Submitted: November 04, 2024 EST. Accepted: December 16, 2024 EST. Published: February 17, 2025 EST.

References

Aagesen, L. (2004). The information content of an ambiguously alignable region, a case study of the trnL intron from the Rhamnaceae. *Organisms Diversity & Evolution*, *4*(1–2), 35–49. https://doi.org/10.1016/j.ode.2003.11.003

Aberer, A. J., Krompass, D., & Stamatakis, A. (2013). Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. *Systematic Biology*, *62*(1), 162–166. https://doi.org/10.1093/sysbio/sys078

Barker, R. M., Haegi, L., & Barker, W. R. (1999). Hakea. *Flora of Australia*, *17b*, 31–170.

Bogarín, D., Pérez-Escobar, O. A., Groenenberg, D., Holland, S. D., Karremans, A. P., Lemmon, E. M., Lemmon, A. R., Pupulin, F., Smets, E., & Gravendeel, B. (2018). Anchored hybrid enrichment generated nuclear, plastid and mitochondrial markers resolve the Lepanthes horrida (Orchidaceae: Pleurothallidinae) species complex. *Molecular Phylogenetics and Evolution*, *129*, 27–47. https://doi.org/10.1016/j.ympev.2018.07.014

Bromham, L. (2016). *An introduction to molecular evolution and phylogenetics*. Oxford University Press. https://doi.org/10.1093/hesc/9780198736363.001.0001

Byrne, M., Steane, D. A., Joseph, L., Yeates, D. K., Jordan, G. J., Crayn, D., Aplin, K., Cantrill, D. J., Cook, L. G., Crisp, M. D., Keogh, J. S., Melville, J., Moritz, C., Porch, N., Sniderman, J. M. K., & Sunnucks, P. (2011). Decline of a biome: evolution, contraction, fragmentation, extinction and invasion of the Australian mesic zone biota. *Journal of Biogeography*, *38*, 1635–1656. https://doi.org/10.1111/j.1365-2699.2011.02535.x

Cardillo, M., Weston, P. H., Reynolds, Z. K. M., Olde, P. M., Mast, A. R., Lemmon, E. M., Lemmon, A. R., & Bromham, L. (2017). The phylogeny and biogeography of Hakea (Proteaceae) reveals the role of biome shifts in a continental plant radiation. *Evolution*, *71*(8), 1928–1943. https://doi.org/10.1111/evo.13276

Comte, A., Tricou, T., Tannier, E., Joseph, J., Siberchicot, A., Penel, S., Allio, R., Delsuc, F., Dray, S., & de Vienne, D. M. (2023). PhylteR: efficient identification of outlier sequences in phylogenomic datasets. *Molecular Biology and Evolution*, *40*(11), msad234. https://doi.org/10.1093/molbev/msad234

Degnan, J. H., & Rosenberg, N. A. (2006). Discordance of species trees with their most likely gene trees. *PLoS Genetics*, *2*(5), e68. https://doi.org/10.1371/journal.pgen.0020068

Durand, E. Y., Patterson, N., Reich, D., & Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, *28*(8), 2239–2252. https://doi.org/10.1093/molbev/msr048

Goodman, M., Czelusniak, J., Moore, G. W., Romero-Herrera, A. E., & Matsuda, G. (1979). Fitting the Gene Lineage into its Species Lineage, a Parsimony Strategy Illustrated by Cladograms Constructed from Globin Sequences. *Systematic Zoology*, *28*(2), 132–163. https://doi.org/10.2307/2412519

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., & Fritz, M. H.-Y. (2010). A draft sequence of the Neandertal genome. *Science*, *328*(5979), 710–722. https://doi.org/10.1126/science.1188021

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, *59*(3), 307–321. https://doi.org/10.1093/sysbio/syq010

Hahn, M. W., & Nakhleh, L. (2016). Irrational exuberance for resolved species trees. *Evolution*, *70*(1), 7–17. https://doi.org/10.1111/evo.12832

Hamilton, C. A., Lemmon, A. R., Lemmon, E. M., & Bond, J. E. (2016). Expanding anchored hybrid enrichment to resolve both deep and shallow relationships within the spider tree of life. *BMC Evolutionary Biology*, *16*(1), 212. https://doi.org/10.1186/s12862-016-0769-y

Hamlin, J. A. P., Hibbins, M. S., & Moyle, L. C. (2020). Assessing biological factors affecting postspeciation introgression. *Evolution Letters*, *4*(2), 137–154. https://doi.org/10.1002/evl3.159

Helmstetter, A. J., Ezedin, Z., Lírio, E. J. de, Oliveira, S. M. de, Chatrou, L. W., Erkens, R. H. J., Larridon, I., Leempoel, K., Maurin, O., Roy, S., Zuntini, A. R., Baker, W. J., Couvreur, T. L. P., Forest, F., & Sauquet, H. (2024). Towards a phylogenomic classification of Magnoliidae. *BioRxiv*, *2024*, 2024.01.09.574948. https://doi.org/10.1101/2024.01.09.574948

Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*, *35*(2), 518–522. https://doi.org/10.1093/molbev/msx281

Hopper, S. D. (2009). OCBIL theory: towards an integrated understanding of the evolution, ecology and conservation of biodiversity on old, climatically buffered, infertile landscapes. *Plant and Soil*, *322*, 49–86. https://doi.org/10.1007/s11104-009-0068-0

Hopper, S. D. (2018). Natural hybridization in the context of Ocbil theory. *South African Journal of Botany*, *118*, 284–289. https://doi.org/10.1016/j.sajb.2018.02.410

Hutter, C. R., & Duellman, W. (2023). Filtration of Gene Trees From 9,000 Exons, Introns, and UCEs Disentangles Conflicting Phylogenomic Relationships in Tree Frogs (Hylidae). *Genome Biology and Evolution*, *15*(5), evad070. https://doi.org/10.1093/gbe/evad070

Johnson, L. A. S., & Briggs, B. G. (1975). On the Proteaceae—the evolution and classification of a southern family. *Botanical Journal of the Linnean Society*, *70*(2), 83–182. https://doi.org/10.1111/j.1095-8339.1975.tb01644.x

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*(6), 587–589. https://doi.org/10.1038/nmeth.4285

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., & Duran, C. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, *28*(12), 1647–1649. https://doi.org/10.1093/bioinformatics/bts199

Lemmon, A. R., Emme, S. A., & Lemmon, E. M. (2012). Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology*, *61*(5), 727–744. https://doi.org/10.1093/sysbio/sys049

Maddison, W. P. (1997). Gene Trees in Species Trees. *Systematic Biology*, *46*(3), 523–536. https://doi.org/10.1093/sysbio/46.3.523

Makinson, R. O. (2000). Grevillea. In *Flora of Australia* (Vol. 17A, pp. 20–506).

Mast, A. R., Milton, E. F., Jones, E. H., Barker, R. M., Barker, W. R., & Weston, P. H. (2012). Time-calibrated phylogeny of the woody Australian genus Hakea (Proteaceae) supports multiple origins of insect-pollination among bird-pollinated ancestors. *American Journal of Botany*, *99*(3), 472–487. https://doi.org/10.3732/ajb.1100420

Mast, A. R., Olde, P. M., Makinson, R. O., Jones, E., Kubes, A., Miller, E. T., & Weston, P. H. (2015). Paraphyly changes understanding of timing and tempo of diversification in subtribe Hakeinae (Proteaceae), a giant Australian plant radiation. *American Journal of Botany*, *102*(10), 1634–1646. https://doi.org/10.3732/ajb.1500195

Meleshko, O., Martin, M. D., Korneliussen, T. S., Schröck, C., Lamkowski, P., Schmutz, J., Healey, A., Piatkowski, B. T., Shaw, A. J., & Weston, D. J. (2021). Extensive genome-wide phylogenetic discordance is due to incomplete lineage sorting and not ongoing introgression in a rapidly radiated bryophyte genus. *Molecular Biology and Evolution*, *38*(7), 2750–2766. https://doi.org/10.1093/molbev/msab063

Minh, B. Q., Hahn, M. W., & Lanfear, R. (2020). New methods to calculate concordance factors for phylogenomic datasets. *Molecular Biology and Evolution*, *37*(9), 2727–2733. https://doi.org/10.1093/molbev/msaa106

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., 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*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Mirarab, S. (2019). Species tree estimation using ASTRAL: practical considerations. *ArXiv Preprint ArXiv:1904.03826*. https://doi.org/10.48550/arXiv.1904.03826

Mirarab, S., & Warnow, T. (2015). ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, *31*(12), i44–i52. https://doi.org/10.1093/bioinformatics/btv234

Mo, Y. K., Lanfear, R., Hahn, M. W., & Minh, B. Q. (2023). Updated site concordance factors minimize effects of homoplasy and taxon sampling. *Bioinformatics*, *39*(1), btac741. https://doi.org/10.1093/bioinformatics/btac741

Molloy, E. K., & Warnow, T. (2018). To include or not to include: the impact of gene filtering on species tree estimation methods. *Systematic Biology*, *67*(2), 285–303. https://doi.org/10.1093/sysbio/syx077

Nge, F. J., Biffin, E., Thiele, K. R., & Waycott, M. (2020). Extinction pulse at Eocene–Oligocene boundary drives diversification dynamics of two Australian temperate floras. *Proceedings of the Royal Society B*, *287*(1919), 20192546. https://doi.org/10.1098/rspb.2019.2546

Pease, J. B., Haak, D. C., Hahn, M. W., & Moyle, L. C. (2016). Phylogenomics Reveals Three Sources of Adaptive Variation during a Rapid Radiation. *PLOS Biology*, *14*(2), e1002379. https://doi.org/10.1371/journal.pbio.1002379

Pharmawati, M., & Macfarlane, I. J. (2013). The Genetic Relationships of Grevillea Hybrids Determined by RAPD Marker. *HAYATI Journal of Biosciences*, *20*(4), 196–200. https://doi.org/10.4308/hjb.20.4.196

Pillon, Y., Majourau, P., Gotty, K., Isnard, S., Fogliani, B., Chase, M. W., & Kergoat, G. J. (2023). The allopolyploid origin(s) and diversification of New Caledonian Grevillea (Proteaceae). *Botany Letters*, *170*(3), 425–438. https://doi.org/10.1080/23818107.2023.2187454

Portik, D. M., & Wiens, J. J. (2021). Do alignment and trimming methods matter for phylogenomic (UCE) analyses? *Systematic Biology*, *70*(3), 440–462. https://doi.org/10.1093/sysbio/syaa064

Prum, R. O., Berv, J. S., Dornburg, A., Field, D. J., Townsend, J. P., Lemmon, E. M., & Lemmon, A. R. (2015). A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature*, *526*(7574), 569–573. https://doi.org/10.1038/nature15697

Rancilhac, L., Irisarri, I., Angelini, C., Arntzen, J. W., Babik, W., Bossuyt, F., Künzel, S., Lüddecke, T., Pasmans, F., & Sanchez, E. (2021). Phylotranscriptomic evidence for pervasive ancient hybridization among Old World salamanders. *Molecular Phylogenetics and Evolution*, *155*, 106967. https://doi.org/10.1016/j.ympev.2020.106967

Roch, S., & Warnow, T. (2015). On the Robustness to Gene Tree Estimation Error (or lack thereof) of Coalescent-Based Species Tree Methods. *Systematic Biology*, *64*(4), 663–676. https://doi.org/10.1093/sysbio/syv016

Rokas, A., & Carroll, S. B. (2006). Bushes in the tree of life. *PLoS Biology*, *4*(11), e352. https://doi.org/10.1371/journal.pbio.0040352

Rokas, A., Williams, B. L., King, N., & Carroll, S. B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, *425*(6960), 798–804. https://doi.org/10.1038/nature02053

Rokyta, D. R., Lemmon, A. R., Margres, M. J., & Aronow, K. (2012). The venom-gland transcriptome of the eastern diamondback rattlesnake (Crotalus adamanteus). *BMC Genomics*, *13*(1), 312. https://doi.org/10.1186/1471-2164-13-312

Roycroft, E. J., Moussalli, A., & Rowe, K. C. (2020). Phylogenomics Uncovers Confidence and Conflict in the Rapid Radiation of Australo-Papuan Rodents. *Systematic Biology*, *69*(3), 431–444. https://doi.org/10.1093/sysbio/syz044

Salichos, L., & Rokas, A. (2013). Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*, *497*(7449), 327–331. https://doi.org/10.1038/nature12130

Sauquet, H., Weston, P. H., Anderson, C. L., Barker, N. P., Cantrill, D. J., Mast, A. R., & Savolainen, V. (2009). Contrasted patterns of hyperdiversification in Mediterranean hotspots. *Proc Natl Acad Sci U S A*, *106*(1), 221–225. https://doi.org/10.1073/pnas.0805607106

Sayyari, E., & Mirarab, S. (2016). Fast coalescent-based computation of local branch support from quartet frequencies. *Molecular Biology and Evolution*, *33*(7), 1654–1668. https://doi.org/10.1093/molbev/msw079

Shekhar, S., Roch, S., & Mirarab, S. (2017). Species tree estimation using ASTRAL: how many genes are enough? *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *15*(5), 1738–1747. https://doi.org/10.1109/TCBB.2017.2757930

Singhal, S., Derryberry, G. E., Bravo, G. A., Derryberry, E. P., Brumfield, R. T., & Harvey, M. G. (2021). The dynamics of introgression across an avian radiation. *Evolution Letters*. https://doi.org/10.1002/evl3.256

Smith, B. T., Merwin, J., Provost, K. L., Thom, G., Brumfield, R. T., Ferreira, M., Mauck, W. M., III, Moyle, R. G., Wright, T. F., & Joseph, L. (2023). Phylogenomic analysis of the parrots of the world distinguishes artifactual from biological sources of gene tree discordance. *Systematic Biology*, *72*(1), 228–241. https://doi.org/10.1093/sysbio/syac055

Smith, M. R. (2020). TreeDist: distances between phylogenetic trees. *R Package Version*, *2*(3).

Smith, M. R. (2022a). Robust analysis of phylogenetic tree space. *Systematic Biology*, *71*(5), 1255–1270. https://doi.org/10.1093/sysbio/syab100

Smith, M. R. (2022b). Using information theory to detect rogue taxa and improve consensus trees. *Systematic Biology*, *71*(5), 1088–1094. https://doi.org/10.1093/sysbio/syab099

Smith, S. A., Brown, J. W., & Walker, J. F. (2018). So many genes, so little time: a practical approach to divergence-time estimation in the genomic era. *PloS One*, *13*(5), e0197433. https://doi.org/10.1371/journal.pone.0197433

Snipen, L., & Liland, H. (2018). microseq: Basic Biological Sequence Handling. *R Package Version*, *2*(5).

Solís-Lemus, C., Yang, M., & Ané, C. (2016). Inconsistency of species tree methods under gene flow. *Systematic Biology*, *65*(5), 843–851. https://doi.org/10.1093/sysbio/syw030

Som, A. (2015). Causes, consequences and solutions of phylogenetic incongruence. *Briefings in Bioinformatics*, *16*(3), 536–548. https://doi.org/10.1093/bib/bbu015

Springer, M. S., & Gatesy, J. (2018). On the importance of homology in the age of phylogenomics. *Systematics and Biodiversity*, *16*(3), 210–228. https://doi.org/10.1080/14772000.2017.1401016

Steenwyk, J. L., Li, Y., Zhou, X., Shen, X.-X., & Rokas, A. (2023). Incongruence in the phylogenomics era. *Nature Reviews Genetics*, 1–17. https://doi.org/10.1038/s41576-023-00620-x

Talavera, G., & Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology*, *56*(4), 564–577. https://doi.org/10.1080/10635150701472164

Tan, G., Muffato, M., Ledergerber, C., Herrero, J., Goldman, N., Gil, M., & Dessimoz, C. (2015). Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. *Systematic Biology*, *64*(5), 778–791. https://doi.org/10.1093/sysbio/syv033

Thomas, A. E., Igea, J., Meudt, H. M., Albach, D. C., Lee, W. G., & Tanentzap, A. J. (2021). Using target sequence capture to improve the phylogenetic resolution of a rapid radiation in New Zealand Veronica. *American Journal of Botany*, *108*(7), 1289–1306. https://doi.org/10.1002/ajb2.1678

Vargas, O. M., Ortiz, E. M., & Simpson, B. B. (2017). Conflicting phylogenomic signals reveal a pattern of reticulate evolution in a recent high-Andean diversification (Asteraceae: Astereae: Diplostephium). *New Phytologist*, *214*(4), 1736–1750. https://doi.org/10.1111/nph.14530

Whitfield, J. B., & Lockhart, P. J. (2007). *Deciphering ancient rapid radiations*. *22*(5). https://doi.org/10.1016/j.tree.2007.01.012

Wilkinson, M. (1996). Majority-rule reduced consensus trees and their use in bootstrapping. *Molecular Biology and Evolution*, *13*(3), 437–444. https://doi.org/10.1093/oxfordjournals.molbev.a025604

Xi, Z., Liu, L., Rest, J. S., & Davis, C. C. (2014). Coalescent versus Concatenation Methods and the Placement of Amborella as Sister to Water Lilies. *Systematic Biology*, *63*(6), 919–932. https://doi.org/10.1093/sysbio/syu055

Yan, Z., Smith, M. L., Du, P., Hahn, M. W., & Nakhleh, L. (2022). Species tree inference methods intended to deal with incomplete lineage sorting are robust to the presence of paralogs. *Systematic Biology*, *71*(2), 367–381. https://doi.org/10.1093/sysbio/syab056

Zhang, C., Rabiee, M., Sayyari, E., & Mirarab, S. (2018). ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, *19*(6), 15–30. https://doi.org/10.1186/s12859-018-2129-y

Zhang, C., Zhao, Y., Braun, E. L., & Mirarab, S. (2021). TAPER: Pinpointing errors in multiple sequence alignments despite varying rates of evolution. *Methods in Ecology and Evolution*, *12*(11), 2145–2158. https://doi.org/10.1111/2041-210X.13696

Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., Farooq, H., Herdean, A., Ariza, M., & Scharn, R. (2019). CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution*, *10*(5), 744–751. https://doi.org/10.1111/2041-210X.13152

Zwickl, D. J., Stein, J. C., Wing, R. A., Ware, D., & Sanderson, M. J. (2014). Disentangling methodological and biological sources of gene tree discordance on Oryza (Poaceae) chromosome 3. *Systematic Biology*, *63*(5), 645–659. https://doi.org/10.1093/sysbio/syu027