Disentangling Sources of Gene Tree Discordance in Phylotranscriptomic Datasets: A Case Study from Amaranthaceae s.l.

Diego F. Morales-Briones[1*], Gudrun Kadereit[2], Delphine T. Tefarikis[2], Michael J. Moore[3], Stephen A. Smith[4], Samuel F. Brockington[5], Alfonso Timoneda[5], Won C. Yim[6], John C. Cushman[6], Ya Yang[1*]

[1] Department of Plant and Microbial Biology, University of Minnesota-Twin Cities, 1445 Gortner Avenue, St. Paul, MN 55108, USA

[2] Institut für Molekulare Physiologie, Johannes Gutenberg-Universität Mainz, D-55099, Mainz, Germany

[3] Department of Biology, Oberlin College, Science Center K111, 119 Woodland Street, Oberlin, OH 44074-1097, USA

[4] Department of Ecology & Evolutionary Biology, University of Michigan, 830 North University Avenue, Ann Arbor, MI 48109-1048, USA

[5] Department of Plant Sciences, University of Cambridge, Tennis Court Road, Cambridge, CB2 3EA, United Kingdom

[6] Department of Biochemistry and Molecular Biology, University of Nevada, Reno, NV, 89577, USA

[*] Correspondence to be sent to: Diego F. Morales-Briones and Ya Yang. Department of Plant and Microbial Biology, University of Minnesota, 1445 Gortner Avenue, St. Paul, MN 55108, USA, Telephone: +1 612-625-6292 (YY) Email: dfmoralesb@gmail.com; yangya@umn.edu

Abstract.— Phylogenomic datasets have become common and fundamental to understanding the phylogenetic relationships of recalcitrant groups across the Tree of Life. At the same time, working with large genomic or transcriptomic datasets requires special attention to the processes that generate gene tree discordance, such as data processing and orthology inference, incomplete lineage sorting, hybridization, model violation, and uninformative gene trees. Methods to estimate species trees from phylogenomic datasets while accounting for all sources of conflict are not available, but a combination of multiple approaches can be a powerful tool to tease apart alternative sources of conflict. Here using a phylotranscriptomic analysis in combination with reference genomes, we explore sources of gene tree discordance in the backbone phylogeny of the plant family Amaranthaceae s.l. The dataset was analyzed using multiple phylogenetic approaches, including coalescent-based species trees and network inference, gene tree discordance analyses, site pattern test of introgression, topology test, synteny analyses, and simulations. We found that a combination of processes might have acted, simultaneously and/or cumulatively, to generate the high levels of gene tree discordance in the backbone of Amaranthaceae s.l. Furthermore, other analytical shortcomings like uninformative genes as well as misspecification of the model of molecular evolution seem to contribute to tree discordance signal in this family. Despite the comprehensive phylogenomic dataset and detailed analyses presented here, no single source can confidently be pointed out to account for the strong signal of gene tree discordance, suggesting that the backbone of Amaranthaceae s.l. might be a product of an ancient and rapid lineage diversification, and remains —and probably will remain— unresolved even with genome-scale data. Our work highlights the need to test for multiple sources of conflict in phylogenomic analyses and provide a set of recommendations moving forward in disentangling ancient and rapid diversification.

Keywords: Amaranthaceae; gene tree discordance; hybridization; incomplete lineage sorting; phylogenomics; transcriptomics; species tree; species network.

The detection of gene tree discordance is ubiquitous in the phylogenomic era. As large phylogenomic datasets are becoming more common (e.g. Jarvis et al. 2014; Misof et al. 2014; Wickett et al. 2014; Hughes et al. 2018; Walker et al. 2018; Laumer et al. 2019; Varga et al. 2019), exploring gene tree heterogeneity in such datasets (e.g. Salichos et al. 2014; Smith et al. 2015; Huang et al. 2016; Arcila et al. 2017; Pease et al. 2018, is essential for inferring phylogenetic relationships while accommodating and understanding the underlying processes that produce gene tree conflict.

Discordance among gene trees can be the product of multiple sources. These include errors and noise in data assembly and filtering, hidden paralogy, incomplete lineage sorting (ILS), gene duplication/loss (Pamilo and Nei 1988; Doyle 1992; Maddison 1997; Galtier and Daubin 2008), random noise from uninformative genes, as well as misspecified model parameters of molecular evolution such as substitutional saturation, codon usage bias, or compositional heterogeneity (Foster 2004; Cooper 2014; Cox et al. 2014; Li et al. 2014; Liu et al. 2014). Among these potential sources of gene tree discordance, ILS is the most studied in the systematics literature (Edwards 2009), and a number of phylogenetic inference methods have been developed that accommodate ILS as the source of discordance (reviewed in Edwards et al. 2016; Mirarab et al. 2016; Xu and Yang 2016). More recently, methods that account for additional processes such as hybridization or introgression have gained attention. These include methods that estimate phylogenetic networks while accounting for ILS and hybridization simultaneously (Yu et al. 2014; Yu and Nakhleh 2015; Solís-Lemus and Ané 2016; Wen et al. 2016b; Wen and Nakhleh 2018; Zhang et al. 2018a; Zhu et al. 2018; Zhu et al. 2019), and methods that detect introgression based on site patterns or phylogenetic invariants (Green et al.

2010; Durand et al. 2011; Patterson et al. 2012; Eaton and Ree 2013; Pease and Hahn 2015; Elworth et al. 2018; Glémin et al. 2019; Kubatko and Chifman 2019).

The above sources of gene tree discordance can act alone, but most often multiple sources may contribute to gene tree heterogeneity (Holder et al. 2001; Buckley et al. 2006; Maureira-Butler et al. 2008; Joly et al. 2009; Meyer et al. 2017; Knowles et al. 2018; Glémin et al. 2019). However, at present no method can estimate species trees from phylogenomic data while modeling multiple sources of conflict and molecular substitution simultaneously. To overcome these limitations, the use of multiple phylogenetic tools and data partitioning schemes in phylogenomic datasets have become a common practice in order to disentangle sources of gene tree heterogeneity and resolve recalcitrant relationships at deep and shallow nodes of the Tree of Life (e.g. Duchêne et al. 2018; Prasanna et al. 2019; Alda et al. 2019; Roycroft et al. 2019; Widhelm et al. 2019).

Here we explore these issues in the plant family Amaranthaceae s.l., including the previously segregated family Chenopodiaceae (Hernández-Ledesma et al. 2015; The Angiosperm Phylogeny Group et al. 2016). With c. 2050 to 2500 species in 181 genera and a worldwide distribution (Hernández-Ledesma et al. 2015), Amaranthaceae s.l. are iconic for the repeated evolution of complex traits representing adaptations to extreme environments such as C_4 photosynthesis in hot and often dry environments (e.g. Kadereit et al. 2012; Bena et al. 2017), various modes of extreme salt tolerance (e.g. Flowers and Colmer 2015; Piirainen et al. 2017) that in several species are coupled with heavy metal tolerance (Moray et al. 2016), and very fast seed germination and production of multiple diaspore types on one individual (Kadereit et al. 2017). Amaranthaceae s.l. contains a number of crops, some of them with a long cultivation history, such as the pseudocereals quinoa and amaranth (Jarvis et al. 2017), and some that have been taken under cultivation more recently, such as sugar beet (Dohm et al. 2014), spinach, glassworts, and *Salsola soda*. Many species of the family are important fodder plants in arid regions and several are currently being investigated for their soil ameliorating and desalinating effects. Reference genomes are available for *Beta vulgaris* (sugar beet, subfamily Betoideae; Dohm et al. 2014), *Chenopodium quinoa* (quinoa, Chenopodioideae; Jarvis et al. 2017), *Spinacia oleracea* (spinach; Chenopodioideae; Xu et al. 2017) and *Amaranthus hypochondriacus* (amaranth; Amaranthoideae; Lightfoot et al. 2017), representing three of the 13 currently recognized subfamilies (sensu Kadereit et al. 2003; Kadereit et al. 2017).

Within the core Caryophyllales the previously recognized families Amaranthaceae s.s. and Chenopodiaceae have always been regarded as closely related and their separate family status has been subjected to phylogenetic and taxonomic debate repeatedly (see Kadereit et al. 2003; Masson and Kadereit 2013; Hernández-Ledesma et al. 2015; Walker et al. 2018; Fig. 1). Their common ancestry was first concluded from a number of shared morphological, anatomical and phytochemical synapomorphies and later substantiated by molecular phylogenetic studies with the Achatocarpaceae as sister group (see Kadereit et al. 2003 and references therein). Amaranthaceae s.s. has a predominant tropical and subtropical distribution with the highest diversity found in the Neotropics, eastern and southern Africa and Australia (Müller and Borsch 2005), while the previously segregated family Chenopodiaceae predominantly occurs in temperate regions and semi-arid or arid environments of subtropical regions (Kadereit et al. 2003). The key problem has always been the species-poor and heterogeneous subfamilies Polycnemoideae and Betoideae, which do not fit comfortably morphologically in either the Chenopodiaceae or Amaranthaceae s.s. (cf. Table 5 in Kadereit et al. 2003). Polycnemoideae are similar in ecology and distribution to Chenopodiaceae but share important floral traits such as

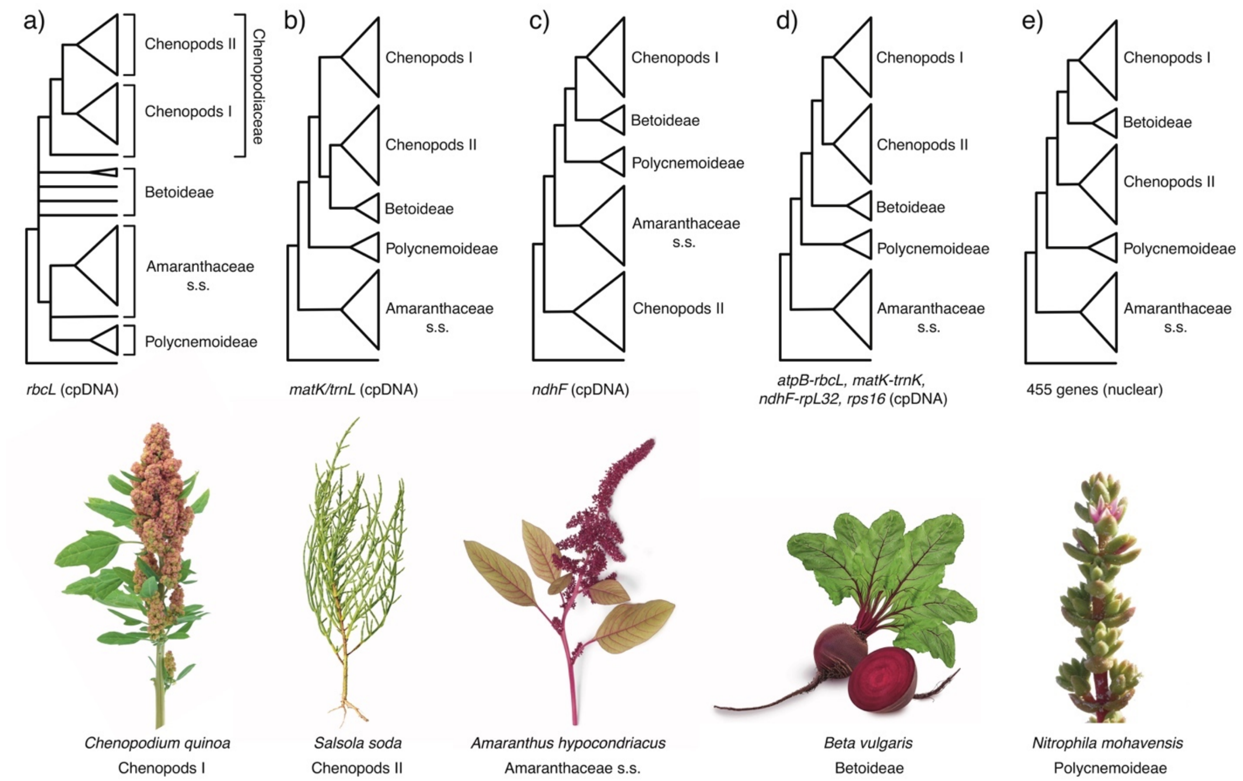


FIGURE 1. Phylogenetic hypothesis of Amaranthaceae s.l. from previous studies. a) Kadereit et al. (2003) using the chloroplast (cpDNA) *rbcL* coding region. b) Müller and Borsch (2005); using the chloroplast *matK* coding region and partial *trnL* intron. c) Hohmann et al. (2006) using the chloroplast *ndhF* coding region. d) Kadereit et al. (2017) using the chloroplast *atpB-rbcL* spacer, *matK* with *trnL* intron, *ndhF-rpL32* spacer, and *rps16* intron e) Walker et al. (2018) using 455 nuclear genes from transcriptome data. Major clades of Amaranthaceae s.l. named following the results of this study. Image credits: *Amaranthus hypochondriacus* by Picture Partners, *Beta vulgaris* by Olha Huchek, *Chenopodium quinoa* by Diana Mower, *Nitrophila mohavensis* by James M. André, and *Salsola soda* by Homeydesign.

petaloid tepals, filament tubes and 2-locular anthers with Amaranthaceae s.s. Morphologically, Betoideae fit into either of the two traditionally circumscribed families but have a unique fruit type—a capsule that opens with a circumscissile lid (Kadereit et al. 2006). Both Betoideae and

Polycnemoideae show strongly disjunct distribution patterns, occurring each with only a few species on three different continents. Furthermore, the genera of both subfamilies display a number of morphologically dissociating features. Both intercontinental disjunctions of species-poor genera and unique morphological traits led to the hypothesis that Betoideae and Polycnemoideae might be relicts of, or from hybridization among early-branching lineages in Amaranthaceae s.l. (Hohmann et al. 2006; Masson and Kadereit 2013).

Previous molecular phylogenetic analyses struggled to resolve the relationships among Betoideae, Polycnemoideae and the rest of the Amaranthaceae s.l. (Kadereit et al. 2003; Müller and Borsch 2005; Kadereit et al. 2012; Masson and Kadereit 2013; Walker et al. 2018). The first phylogenomic study of Amaranthaceae s.l. by Walker et al. (2018) revealed that gene tree discordance mainly occurred at deeper nodes of the phylogeny involving Betoideae. Polycnemoideae was resolved as sister to Chenopodiaceae in Walker et al. (2018), albeit with low (17%) gene tree concordance, which contradicted previous analyses based on chloroplast data (Masson and Kadereit 2013). However, only a single species of Betoideae (the cultivated beet and its wild relative) was sampled in Walker et al. (2018). In addition, sources of conflicting signals among species trees remained unexplored.

In this study, we leverage 71 publicly available transcriptomes, 17 newly sequenced transcriptomes, and 4 reference genomes that span all 13 subfamilies of Amaranthaceae s.l. and include increased taxon sampling in Betoideae. Consistent with previous analyses, we identified high levels of gene tree discordance in the backbone phylogeny of Amaranthaceae s.l. Using a combination of phylogenetic approaches, we explored multiple sources that can explain such conflict. We tested for 1) ancient hybridization, focusing on the hypothesis of the hybrid origin of Polycnemoideae and Betoideae, between Amaranthaceae s.s. and Chenopodioideae, 2) discordance produced by misspecifications of model of molecular evolution, and 3) discordance due to ILS as a result of short internal branches in the backbone phylogeny of Amaranthaceae s.l. In addition, we comprehensively updated the phylotranscriptomic pipeline of Yang and Smith (2014) with additional features of filtering isoforms and spurious tips. Our results showed that both species network and site pattern methods that model gene flow while accounting for ILS detected signals of multiple hybridization events in Amaranthaceae s.l. However, when these hybridization events were analyzed individually, most of the gene tree discordance could be explained by uninformative gene trees. In addition, the high level of gene tree discordance in Amaranthaceae s.l. could also be explained by three consecutive short branches that produce anomalous gene trees. Combined, our results showed that multiple processes might have contributed to the gene tree discordance in Amaranthaceae s.l., and that we might not be able to distinguish among these processes even with genomic-scale sampling and synteny information. Finally, we make recommendations on strategies for disentangling multiple sources of gene tree discordance in phylogenomic datasets.

Materials and Methods

An overview of all dataset and phylogenetic analyses can be found in Figure S1. Scripts for raw data processing, assembly, translation, and homology and orthology search can be found at https://bitbucket.org/yanglab/phylogenomic_dataset_construction/ as part of an updated ‘phylogenomic dataset construction’ pipeline (Yang and Smith 2014).

Taxon sampling, transcriptome sequencing

We sampled 92 species (88 transcriptomes and four genomes) representing all 13 currently recognized subfamilies and 16 out of 17 tribes of Amaranthaceae s.l. (sensu [Kadereit et al. 2003; Kadereit et al. 2017]). In addition, 13 outgroups across the Caryophyllales were included (ten transcriptomes and three genomes; Table S1). We generated 17 new transcriptomes for this study (Table S2). For *Tidestromia oblongifolia*, tissue collection, RNA isolation, library preparation was carried out using the KAPA Stranded mRNA-Seq Kits (KAPA Biosystems, Wilmington, Massachusetts, USA). The library was multiplexed with 10 other samples from a different project on an Illumina HiSeq2500 platform with V4 chemistry at the University of Michigan Sequencing Core (Yang et al. 2017). For the remaining 16 samples total RNA was isolated from c. 70-125 mg leaf tissue collected in liquid nitrogen using the RNeasy Plant Mini Kit (Qiagen) following the manufacturer's protocol (June 2012). A DNase digestion step was included with the RNase-Free DNase Set (Qiagen). Quality and quantity of RNA were checked on the NanoDrop (Thermo Fisher Scientific) and the 2100 Bioanalyzer (Agilent Technologies). Library preparation was carried out using the TruSeq® Stranded Total RNA Library Prep Plant with RiboZero probes (96 Samples. Illumina, #20020611). Indexed libraries were normalized, pooled and size selected to 320bp +/- 5% using the Pippin Prep HT instrument to generate libraries with mean inserts of 200 bp, and sequenced on the Illumina HiSeq2500 platform with V4 chemistry at the University of Minnesota Genomics Center. Reads from all 17 libraries were paired-end 125 bp.

Transcriptome data processing and assembly

We processed raw reads for all 98 transcriptome datasets (except *Bienertia sinuspersici*) used in this study (88 ingroups + 10 outgroups; Table S1). Sequencing errors in raw reads were corrected with Rcorrector (Song and Florea 2015) and reads flagged as uncorrectable were removed. Sequencing adapters and low-quality bases were removed with Trimmomatic v0.36 (SLIDINGWINDOW:4:5 LEADING:5 TRAILING:5 MINLEN:25; Bolger et al. 2014). Additionally, reads were filtered for chloroplast and mitochondrial reads with Bowtie2 v 2.3.2 (Langmead and Salzberg 2012) using publicly available Caryophyllales organelle genomes from the Organelle Genome Resources database (RefSeq; [Pruitt et al. 2007]; last accessed on October 17, 2018) as references. Read quality was assessed with FastQC v 0.11.7 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Finally, overrepresented sequences detected with FastQC were discarded. *De novo* assembly was carried out with Trinity v 2.5.1 (Haas et al. 2013) with default settings, but without in silico normalization. Assembly quality was assessed with Transrate v 1.0.3 (Smith-Unna et al. 2016). Low quality and poorly supported transcripts were removed using individual cut-off values for three contig score components of Transrate: 1) proportion of nucleotides in a contig that agrees in identity with the aligned read, s(Cnuc) ≤ 0.25; 2) proportion of nucleotides in a contig that have one or more mapped reads, s(Ccov) ≤0.25; and 3) proportion of reads that map to the contig in correct orientation, s(Cord) ≤ 0.5. Furthermore, chimeric transcripts (*trans*-self and *trans*-multi-gene) were removed following the approach described in Yang and Smith (2013) using *Beta vulgaris* as the reference proteome, and percentage similarity and length cutoffs of 30 and 100, respectively. In order to remove isoforms and assembly artifacts, filtered reads were remapped to filtered transcripts with Salmon v 0.9.1 (Patro et al. 2017) and putative genes were clustered with

Corset v 1.07 (Davidson and Oshlack 2014) using default settings, except that we used a minimal of five reads as threshold to remove transcripts with low coverage (-m 5). Only the longest transcript of each putative gene inferred by Corset was retained. Our previous benchmark study have shown that Corset followed by selecting the longest transcript for each putative gene performed well in reducing isoforms and assembly artifacts, especially in polyploid species (Chen et al. 2019). Filtered transcripts were translated with TransDecoder v 5.0.2 (Haas et al. 2013) with default settings and the proteome of *Beta vulgaris* and *Arabidopsis thaliana* to identify open reading frames. Finally, translated amino acid sequences were further reduced with CD-HIT v 4.7 (-c 0.99; [Fu et al. 2012]) to remove near-identical amino acid sequences.

Homology and orthology inference

Initial homology inference was carried out following Yang and Smith (2014) with some modification. First, an all-by-all BLASTN search was performed on coding sequences (CDS) using an *E* value cutoff of 10 and max_target_seqs set to 100. Raw BLAST output was filtered with a hit fraction of 0.4. Then putative homologs groups were clustered using MCL v 14-137 (van Dongen 2000) with a minimal minus log-transformed *E* value cutoff of 5 and an inflation value of 1.4. Finally, only clusters with a minimum of 25 taxa were retained. Individual clusters were aligned using MAFFT v 7.307 (Katoh and Standley 2013) with settings '–genafpair –maxiterate 1000'. Aligned columns with more than 90% missing data were removed using Phyx (Brown et al. 2017). Homolog trees were built using RAxML v 8.2.11 (Stamatakis 2014) with a GTR-CAT model and clade support assessed with 200 rapid bootstrap (BS) replicates. Spurious or outlier long tips were detected and removed with TreeShrink v 1.0.0 (Mai and Mirarab 2018). Monophyletic and paraphyletic tips that belonged to the same taxon were removed keeping the tip with the highest number of characters in the trimmed alignment. After visual inspection of ca. 50 homolog trees, internal branches longer than 0.25 were likely representing deep paralogs. These branches were cut apart, keeping resulting subclades with a minimum of 25 taxa. Homolog tree inference, tip masking, outlier removal, and deep paralog cutting was carried out for a second time using the same settings to obtain final homologs. Orthology inference was carried out following the 'monophyletic outgroup' approach from Yang and Smith (2014), keeping only ortholog groups with at least 25 ingroup taxa. The 'monophyletic outgroup' approach filters for clusters that have outgroup taxa being monophyletic and single-copy, and therefore filters for single- and low-copy genes. It then roots the gene tree by the outgroups, traverses the rooted tree from root to tip, and removes the side with less taxa when gene duplication is detected at any given node.

Chloroplast assembly

Although DNase treatment is carried out to remove genomic DNA, due to its high copy number, chloroplast sequences are often carried over in RNA-seq libraries. In addition, as young leaf tissue was used for RNA-seq, RNA from chloroplast genes are expected to be represented, especially in libraries prepared using a RiboZero approach. To investigate phylogenetic signal from plastome sequences, *de novo* assemblies were carried out with the Fast-Plast v.1.2.6 pipeline (https://github.com/mrmckain/Fast-Plast) using the organelle reads from the filtering step.lNo complete or single-contig plastomes were obtained. Filtered contigs produced by Spades v 3.9.0 (Bankevich et al. 2012) were mapped to the closest available reference plastome (with an Inverted Repeat removed; Table S3) and manually edited in Geneious v.11.1.5 (Kearse et al. 2012) to produce final oriented contigs.

Assessment of recombination

Coalescent species tree methods assume that there is no recombination within loci and free recombination between loci. To determine the presence of recombination in our dataset, we used the Φ (pairwise homoplasy index) test for recombination, as implemented in PhiPack (Bruen et al. 2006). We tested recombination on the final set of ortholog alignments (with a minimum of 25 taxa) with the default sliding window size of 100 bp.

Nuclear phylogenetic analysis

We used concatenation and coalescent-based methods to reconstruct the phylogeny of Amaranthaceae s.l. Sequences from final orthologs were aligned with MAFFT, columns were trimmed with Phyx requiring a minimal occupancy of 30%, and alignments with at least 1,000 characters and 99 out of 105 taxa were retained. We first estimated a maximum likelihood (ML) tree of the concatenated matrix with RAxML using a partition-by-gene scheme with GTR-CAT model for each partition and clade support assessed with 200 rapid bootstrap (BS) replicates. To estimate a coalescent-based species tree, first we inferred individual ML gene trees using RAxML with a GTR-CAT model and 200 BS replicates to assess clade support. Individual gene trees were then used to estimate a species tree with ASTRAL-III v5.6.3 (Zhang et al. 2018b) using local posterior probabilities (LPP; Sayyari and Mirarab 2016) to assess clade support.

Detecting and visualizing nuclear gene tree discordance

To explore discordance among gene trees, we first calculated the internode certainty all (ICA) value to quantify the degree of conflict on each node of a target tree (i.e. species tree) given individual gene trees (Salichos et al. 2014). In addition, we calculated the number of conflicting and concordant bipartitions on each node of the species trees. We calculated both the ICA scores and the number of conflicting/concordant bipartitions with Phyparts (Smith et al. 2015), mapping against the estimated ASTRAL species trees, using individual gene trees with BS support of at least 50% for the corresponding node. Additionally, in order to distinguish strong conflict from weakly supported branches, we evaluated tree conflict and branch support with Quartet Sampling (QS; Pease et al. 2018) using 100 replicates. Quartet Sampling subsamples quartets from the input tree and alignment and assess the confidence, consistency, and informativeness of each internal branch by the relative frequency of the three possible quartet topologies (Pease et al. 2018)

Furthermore, in order to visualize conflict, we built a cloudogram using DensiTree v2.2.6 (Bouckaert and Heled 2014). We filtered the final ortholog alignments to include only 41 species (38 ingroup and 3 outgroups) in order to include as many orthologs as possible while representing all main clades of Amaranthaceae s.l. (see results). Individual gene trees were inferred as previously described. Trees were time-calibrated with TreePL v1.0 (Smith and O'Meara 2012) by fixing the crown age of Amaranthaceae s.l. to 66–72.1 based on a pollen record of *Polyporina cribraria* from the late Cretaceous (Maastrichtian; Srivastava 1969), and the root for the reduced 41-species dataset (most common recent ancestor of Achatocarpaceae and Aizoaceae) was set to 95 Ma based on the time-calibrated plastome phylogeny of Caryophyllales from Yao et al. (2019).

Chloroplast phylogenetic analysis

Assembled contigs (excluding one inverted repeat region) were aligned with MAFFT with the setting '--auto'. Two samples (*Dysphania schraderiana* and *Spinacia turkestanica*) were removed due to low sequence occupancy. Using the annotations of the reference genomes (Table S3), the coding regions of 78 genes were extracted and each gene alignment was visually inspected in Geneious to check for potential misassemblies. From each gene alignment taxa with short sequences (i.e. < 50% of the aligned length) were removed and realigned with MAFFT. The genes *rpl32* and *ycf2* were excluded from downstream analyses due to low taxon occupancy (Table S4). For each individual gene we performed extended model selection (Kalyaanamoorthy et al. 2017) followed by ML gene tree inference and 1,000 ultrafast bootstrap replicates for branch support (Hoang and Chernomor 2018) in IQ-Tree v.1.6.1 (Nguyen et al. 2015). For the concatenated matrix we searched for the best partition scheme (Lanfear et al. 2012) followed by ML gene tree inference and 1,000 ultrafast bootstrap replicates for branch support in IQ-Tree. Additionally, we evaluated branch support with QS using 1,000 replicates and gene tree discordance with PhyParts in the ML and species tree. Finally, to identify the origin of the chloroplast reads (i.e. genomic or RNA), we predicted RNA editing from CDS alignments using PREP (Mower 2009) with the alignment mode (PREP-aln), and a cutoff value of 0.8.

Species network analysis using a reduced 11-taxon dataset

We inferred species networks that model ILS and gene flow using a maximum pseudo-likelihood approach (Yu and Nakhleh 2015). Species network searches were carried out with PhyloNet v.3.6.9 (Than et al. 2008) with the command 'InferNetwork_MPL' and using the individual gene trees as input. Due to computational restrictions, and given our main focus was to search for potential reticulating events among major clades of Amaranthaceae s.l., we reduced our taxon sampling to one outgroup and ten ingroup taxa including two representative species from each of the five well-supported major lineages in Amaranthaceae s.l. (see results). We filtered the final 105-taxon ortholog alignments to include genes that have all 11 taxa [referred herein as 11-taxon(net) dataset]. After realignment and trimming we kept genes with a minimum of 1,000 aligned base pairs and individual ML gene trees were inferred with RAxML with a GTR-GAMMA model and 200 bootstrap replicates. We carried out 10 independent network searches allowing for up to five hybridization events for each search. To estimate the optimum number of hybridizations, first we optimized the branch lengths and inheritance probabilities and computed the likelihood of the best scored network from each of the five maximum hybridization events searches. Network likelihoods were estimated given the individual gene trees, as implemented in Yu et al. (2012), using the command ‘CalGTProb’ in PhyloNet. Then, we performed model selection using the bias-corrected Akaike information criterion (AICc; Sugiura 1978), and the Bayesian information criterion (BIC; Schwarz 1978). The number of parameters was set to the number of branch lengths being estimated plus the number of hybridization probabilities being estimated. The number of gene trees used to estimate the likelihood was used to correct for finite sample size. To compare network models to bifurcating trees, we also estimated ML and coalescent-based species trees as well as a chloroplast tree with the same taxon sampling used in the network searches. Tree inferences were carried out as previously described for the ML, coalescent-based, and chloroplast trees, respectively.

Hypothesis testing and detecting introgression using four-taxon datasets

Given the signal of multiple clades potentially involved in hybridization events detected by PhyloNet (see results), we next conducted quartet analyses to explore a single event at a time. First, we further reduced the 11-taxon(net) dataset to six taxa that included one outgroup genome (*Mesembryanthemum crystallinum*) and one ingroup from each of the five major ingroup clades: *Amaranthus hypochondriacus* (genome)*, Beta vulgaris* (genome)*, Chenopodium quinoa* (genome)*, Caroxylon vermiculatum* (transcriptome)*,* and *Polycnemum majus* (transcriptome) to represent Amaranthaceae s.s., Betoideae, 'Chenopods I', 'Chenopods II' and Polycnemoideae, respectively. We carried out a total of ten quartet analyses using all ten four-taxon combinations that included three out of five ingroup species and one outgroup. We filtered the final set of 105-taxon ortholog alignments for genes with all four taxa for each combination and inferred individual gene trees as described before. For each quartet we carried out the following analyses. We first estimated a species tree with ASTRAL and explored gene tree conflict with PhyParts. We then explored individual gene tree resolution by calculating the Tree Certainty (TC) score (Salichos et al. 2014) in RAxML using the majority rule consensus tree across the 200 bootstrap replicates. Next, we explored potential correlation between TC score and alignment length, GC content and alignment gap proportion using a linear regression model in R v.3.6.1 (R Core Team 2019). Finally, we tested for the fit of gene trees to the three possible rooted quartet topologies for each gene using the approximately unbiased (AU) tests (Shimodaira 2002). We carried out ten constraint searches for each of three topologies in RAxML with the GTR-GAMMA model, then calculated site-wise log-likelihood scores for the three constraint topologies in RAxML using the GTR-GAMMA and carried out the AU test using Consel v.1.20 (Shimodaira and Hasegawa 2001). In order to detect possible introgression among species of each quartet, first we estimated a species network with PhyloNet using a full maximum likelihood approach (Yu et al. 2014) with 100 independent searches while optimizing the likelihood of the branch lengths and inheritance probabilities for every proposed species network. Furthermore, we also carried out the ABBA/BABA test to detect introgression (Green et al. 2010); Durand et al. 2011; Patterson et al. 2012) in each quartet. We calculated the *D*-statistic and associated *z* score for the null hypothesis of no introgression ($D = 0$) following each quartet ASTRAL species tree for taxon order assignment using 100 jackknife replicates and a block size of 10,000 bp with evobiR v1.2 (Blackmon and Adams) in R.

Additionally, to visualize any genomic patterns of the phylogenetic history of *Beta vulgaris* regarding its relationship with Amaranthaceae s.s. and Chenopodiaceae, we first identified syntenic regions between the genomes of *Beta vulgaris* and the outgroup *Mesembryanthemum crystallinum* using the SynNet pipeline (https://github.com/zhaotao1987/SynNet-Pipeline; Zhao and Schranz 2019). We used DIAMOND v.0.9.24.125 (Buchfink et al. 2015) to perform all-by-all inter- and intra-pairwise protein searches with default parameters, and MCScanX (Wang et al. 2012) for pairwise synteny block detection with default parameters, except match score (-k) that was set to five. Then, we plot the nine chromosomes of *Beta vulgaris* by assigning each of the 8,258 orthologs of the quartet composed of *Mesembryanthemum crystallinum* (outgroup), *Amaranthus hypochondriacus, Beta vulgaris,* and *Chenopodium quinoa* (BC1A) to synteny blocks and to one of the three possible quartet topologies based on best likelihood score.

Assessment of substitutional saturation, codon usage bias, compositional heterogeneity, and model of sequence evolution misspecification

We refiltered the final 105-taxon ortholog alignments to again include genes that have the same 11 taxa (referred herein as 11-taxon(tree) dataset used for the species network analyses. We realigned individual genes using MACSE v.2.03 (Ranwez et al. 2018) to account for codon structure and frameshifts. Codons with frameshifts were replaced with gaps, and ambiguous alignment sites were removed using GBLOCKS v0.9b (Castresana 2000) while accounting for codon alignment (-t=c -b1=6 -b2=6 -b3=2 -b4=2 -b5=h). After realignment and removal of ambiguous sites, we kept genes with a minimum of 300 aligned base pairs. To detect potential saturation, we plotted the uncorrected genetic distances against the inferred distances as described in Philippe and Forterre (1999). The level of saturation was determined by the slope of the linear regression between the two distances where a shallow slope (i.e < 1) indicates saturation. We estimated the level of saturation by concatenating all genes and dividing the first and second codon positions from the third codon positions. We calculated uncorrected, and inferred distances with the TN93 substitution model using APE v5.3 (Paradis and Schliep 2019) in R. To determine the effect of saturation in the phylogenetic inferences we estimated individual gene trees using three partition schemes. We inferred ML trees with an unpartitioned alignment, a partition by first and second codon positions, and the third codon positions, and by removing all third codon positions. All tree searches were carried out in RAxML with a GTR+GAMMA model and 200 bootstrap replicates. A species tree for each of the three data schemes was estimated with ASTRAL and gene tree discordance was examined with PhyParts.

Codon usage bias was evaluated using a correspondence analysis of the Relative Synonymous Codon Usage (RSCU), which is defined as the number of times a particular codon is observed relative to the number of times that the codon would be observed in the absence of any codon usage bias (Sharp and Li 1986). RSCU for each codon in the 11-taxon concatenated alignment was estimated with CodonW v.1.4.4 (Peden 1999). Correspondence analysis was carried out using FactoMineR v1.4.1(Lê et al. 2008) in R. To determine the effect of codon usage bias in the phylogenetic inferences we estimated individual gene trees using codon-degenerated alignments. Alignments were recoded to eliminate signals associated with synonymous substitutions by degenerating the first and third codon positions using ambiguity coding using DEGEN v1.4 (Regier et al. 2010; Zwick et al. 2012). Gene tree inference and discordance analyses were carried out on the same three data schemes as previously described.

To examine the presence of among-lineage compositional heterogeneity, individual genes were evaluated using the compositional homogeneity test that uses a null distribution from simulations as proposed by Foster (2004). We performed the compositional homogeneity test by optimizing individual gene trees with a GTR-GAMMA model and 1,000 simulations in P4 (Foster 2004). To assess if compositional heterogeneity had an effect in species tree inference and gene tree discordance, gene trees that showed the signal of compositional heterogeneity were removed from saturation and codon usage analyses and the species tree and discordance analyses were rerun.

To explore the effect of sequence evolution model misspecification, we reanalyzed the datasets from the saturation and codon usage analyses using inferred gene trees that accounted for model selection. We performed extended model selection followed by ML gene tree inference and 1,000 ultrafast bootstrap replicates for branch support in IQ-Tree. Species tree inference, conflict analysis and removal of genes with compositional heterogeneity were carried out as previously described.

Finally, we also used amino acid alignments from MACSE to account for substitutional saturation. Amino acid positions with frameshifts were replaced with gaps, and ambiguous alignment sites were removed with Phyx requiring a minimal occupancy of 30%. We inferred individual gene trees with IQ-tree to account for a model of sequence evolution and carried out species tree inference, conflict analysis, and removal of genes with compositional heterogeneity as described for the nucleotide alignments.

Polytomy test

To explore if the gene tree discordance among the main clades of Amaranthaceae s.l. could be explained by polytomies instead of bifurcating nodes, we carried out the polytomy test by Sayyari and Mirarab (2018) as implemented in ASTRAL. This test uses quartet frequencies to assess whether a branch should be replaced with a polytomy while accounting for ILS. We performed the polytomy test using the gene trees inferred from the saturation and codon usage analyses [11-taxon(tree) dataset]. Because this test can be sensitive to gene tree error (Syyari and Mirarab 2018), we ran the analyses using the original gene trees and also using gene trees where branches with less than 75% of bootstrap support were collapsed.

Coalescent simulations

To investigate if gene tree discordance can be explained by ILS alone, we carried out coalescent simulations similar to Cloutier et al. (2019) An ultrametric species tree with branch lengths in mutational units (μT) was estimated by constraining an ML tree search of the 11-taxon(net) concatenated alignment (from individual MAFFT gene alignment) to the ASTRAL species tree topology with a GTR+GAMMA model while enforcing a strict molecular clock in PAUP v4.0a (build 165; Swofford 2002). The mutational branch lengths from the constrained tree and branch lengths in coalescent units (τ= T/4N_e) from the ASTRAL species trees were used to estimate the population size parameter theta (Θ= μT/τ; Degnan and Rosenberg 2009) for internal branches. Terminal branches were set with a population size parameter theta of one. We used the R package Phybase v. 1.4 (Liu and Yu 2010) which uses the formula from Rannala and Yang (2003) to simulate 10,000 gene trees using the constraint tree and the estimated theta values. Then the tree-to-tree distances using the Robinson and Foulds (1981) metric was calculated between the species tree and each gene tree and compared with the distribution of tree-to-tree distances between the species tree and the simulated gene tree. Tree-to-tree distances were calculated using the R package Phangorn v2.5.3 (Schliep 2011). We ran simulations in seven species trees and associated gene tree distribution to represent the trees and gene tree distributions from the saturation, codon usage and model selection analyses that accounted for branch length variation in the species trees and individual gene tree inference. Following Maureira-Butler et al. (2008), if the tree-to-tree distances between the species trees and gene trees were larger than 95% of the distribution of tree-to-tree distances of the species trees and the simulated gene trees then ILS alone is considered unlikely to explain the gene tree heterogeneity.

Test of anomaly zone

Anomaly zone occurs where a set of short internal branches in the species tree produces gene trees that differ from the species tree more frequently than those that are concordant [a(x); as defined in equation 4 of Degnan and Rosenberg (2006)]. To explore if gene tree discordance observed in Amaranthaceae s.l. is a product of the anomaly zone, we estimated the boundaries of the anomaly zone [a(x); as defined in equation 4 of Degnan and Rosenberg (2006)] for the internal nodes of the species tree. Here, x is the branch length (coalescent units) in the species tree that has a descendant internal branch. If the length of the descendant internal branch (y) is smaller than a(x), then the internode pair is in the anomaly zone and is likely to produce anomaly gene trees (AGTs). We carried out the calculation of a(x) following Linkem et al. (2016) in the same 11-taxon(tree) ASTRAL species trees used for coalescent simulations to account for branch length variation. Additionally, to establish the frequency of gene trees that were concordant with the estimated species trees, we quantified the frequency of all 105 possible rooted gene trees (when clades of Amaranthaceae sl. are monophyletic). We calculated tree-to-tree distances between the 105 possible topologies and all 5,936 gene trees and counted how many times a topology had a distance of zero among the set of gene trees.

RESULTS

Transcriptome sequencing, assembly, translation, and quality control

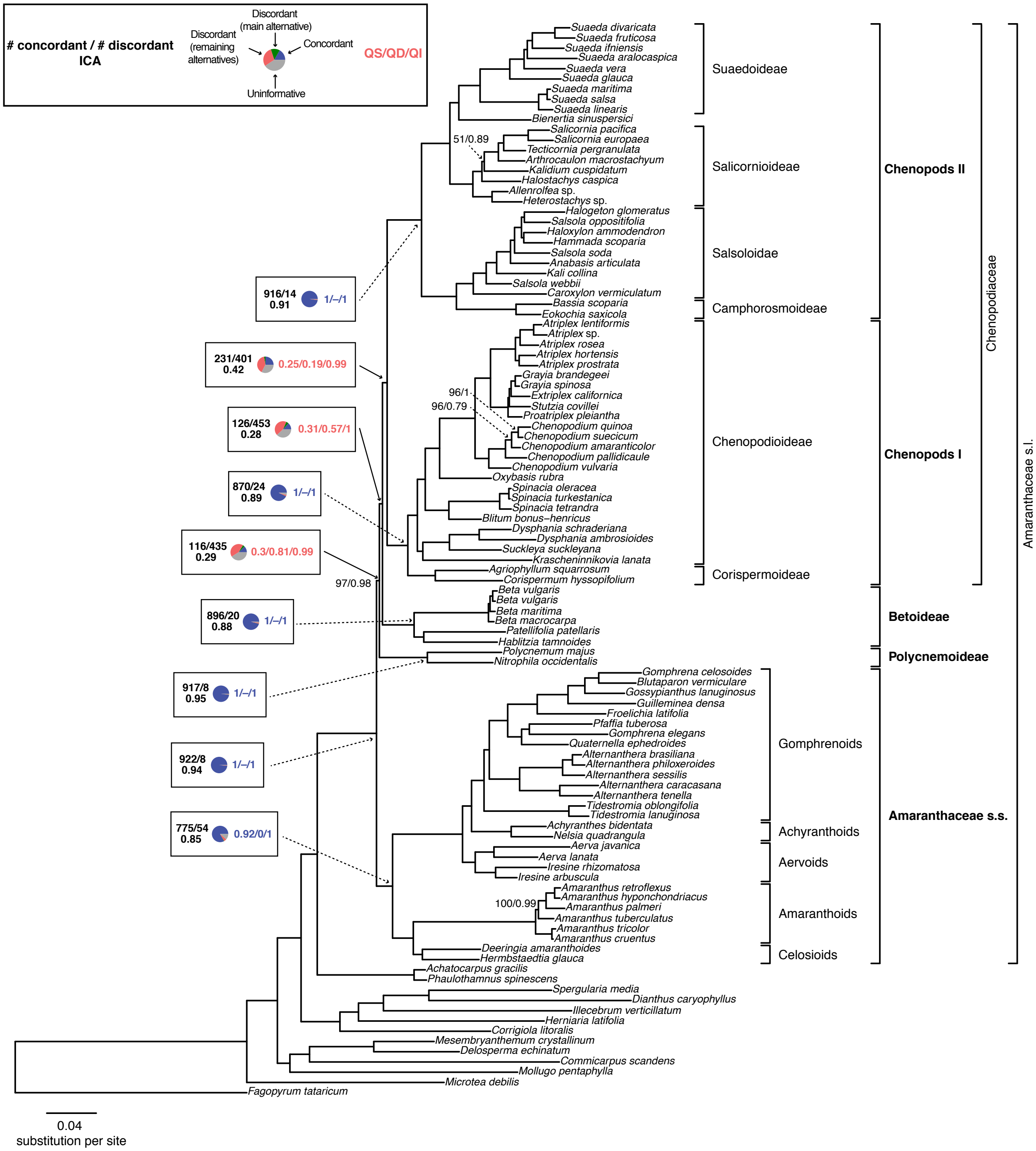
We generated 17 new transcriptomes of Amaranthaceae s.l. for this study. Raw reads are available from the NCBI Sequence Read Archive (BioProject: XXXX; Table S2). The number of raw read pairs ranged from 17 to 27 million. For the 16 samples processed using RiboZero organelle reads accounted for 15% to 52% of read pairs (Table S2). For *Tidestromia oblongifolia* that poly-A enrichment was carried out in library prep with ~5% of raw reads were from organelle (Table S2). Number of final CDS (after quality control and redundancy reduction) used for all-by-all homology search can be found in Table S5. The final number of orthologs from the ‘monophyletic outgroup’ approach was 13,024 with a mean of 9,813 orthologs per species (Table S6).

Assessment of recombination

The test for recombination, Φ, identified 82 out of the 13,024 genes from the final set of orthologs (with a minimum of 25 taxa) with a strong signal of recombination ($p \leq 0.05$; Table S7). Alignments that showed signal of recombination were removed from all subsequent phylogenetic analyses.

Analysis of the nuclear dataset of Amaranthaceae s.l.

The final set of nuclear orthologous genes included 936 genes with at least 99 out of 105 taxa and 1,000 bp in aligned length after removal of low occupancy columns (the 105-taxon dataset, Fig. S1). The concatenated matrix consisted of 1,712,054 columns with a gene and character occupancy of 96% and 82%, respectively. The species tree from ASTRAL and the concatenated ML tree from RAxML recovered the exact same topology with most clades with the highest support [i.e. bootstrap percentage (BS) = 100, local posterior probabilities (LPP) = 1; Fig. 2; Figs S2–S3]. Our phylogenetic analyses recovered Chenopodiaceae as monophyletic with the subfamilies and relationships among them similar to Kadereit et al. (2017). Betoideae was placed as sister of Chenopodiaceae, while Polycnemoideae was placed as sister (BS = 97, LPP = 0.98) to the clade composed of Chenopodiaceae and Betoideae. Finally, we recovered Amaranthaceae s.s. with an overall topology concordant to Kadereit et al. (2017), with the exception of *Iresine* that is placed among the Aervoids (Fig. 2; Figs S2–S3).



566

567 **FIGURE 2.** Maximum likelihood phylogeny of Amaranthaceae s.l. inferred from RAxML 568 analysis of the concatenated 936-nuclear gene supermatrix. All nodes have full support 569 (Bootstrap = 100/Local posterior probability = 100) unless noted next to nodes. Boxes contain 570 gene tree conflict and Quartet Sampling (QS) scores for major clades (see Figs S2–S3 for all 571 node scores). In each box, numbers on the upper left indicate the number of gene trees 572 concordant/conflicting with that node in the species tree, and the number on the lower left

indicate the Internode Certainty All (ICA) score. Pie charts present the proportion of gene trees that support that clade (blue), the proportion that support the main alternative bifurcation (green), the proportion that support the remaining alternatives (red), and the proportion (conflict or support) that have < 50% bootstrap support (gray). Number on the right of the pie chart indicates QS scores: Quartet concordance/Quartet differential/Quartet informativeness. QS scores in blue indicate support for individual major clades of Amaranthaceae s.l., while red scores indicate strong support for alternative relationships among them. Branch lengths are in number of substitutions per site (scale bar on the bottom).

The conflict analyses confirmed the monophyly of Amaranthaceae s.l. with most gene trees being concordant (922; ICA= 0.94) and full QS support (1/–/1; i.e. all sampled quartets supported that branch), but also recovered significant discordance in the backbone of the family (Fig. 2; Figs S2–S3). The monophyly of Chenopodiaceae s.s. was supported only by 231 out of 632 informative gene trees (ICA = 0.42) and the QS score (0.25/0.19/0.99) suggested weak quartet support with a skewed frequency for an alternative placement of two well-defined clades within Chenopodiaceae s.s., herein referred to as ‘Chenopods I’ and ‘Chenopods II’ (Fig. 2; Figs S2–S3). ‘Chenopods I’ and ‘Chenopods II’ were each supported by the majority of gene trees, 870 (ICA = 0.89) and 916 (ICA = 0.91), respectively and full QS support. The placement of Betoideae and Polycnemoideae as successive sisters of Chenopodiaceae also showed significant conflict (Fig. 2; Figs S2–S3). The placement of Betoideae was supported only by 126 out of 579 informative gene trees (ICA = 0.28) and the QS score (0.31/0.57/1) also showed low support with the presence of supported alternative placements close to the same frequency. Similarly, the placement of Polycnemoideae was supported by only 116 out of 511 informative gene trees (ICA = 0.29) and low QS support (0.3/0.81/0.99) with alternative topologies close to equal frequencies. The monophyly of Amaranthaceae s.s. was highly supported by 755 gene trees (ICA

=0.85) and the QS score (0.92/0/1) also indicated high quartet support and no support for a single alternative topology.

Congruent with the overall low support in the backbone of Amaranthaceae s.l. from BS, LPP, ICA, QS, and PhyParts, the cloudogram of 41 species using 1,242 gene trees also showed significant conflict in the backbone of Amaranthaceae s.l. where no clear pattern can be identified regarding the relationships of the five main clades of Amaranthaceae s.l. (Fig. 3). In summary, analysis of nuclear genes recovered five well-supported clades in Amaranthaceae s.l.: Amaranthaceae s.s., Betoideae, 'Chenopods I', 'Chenopods II', and Polycnemoideae. However, relationships among these five clades showed a high level of conflict among genes (ICA scores and gene counts [pie charts]) and among subsampled quartets (QS scores), despite having high support from both BS and LPP scores.

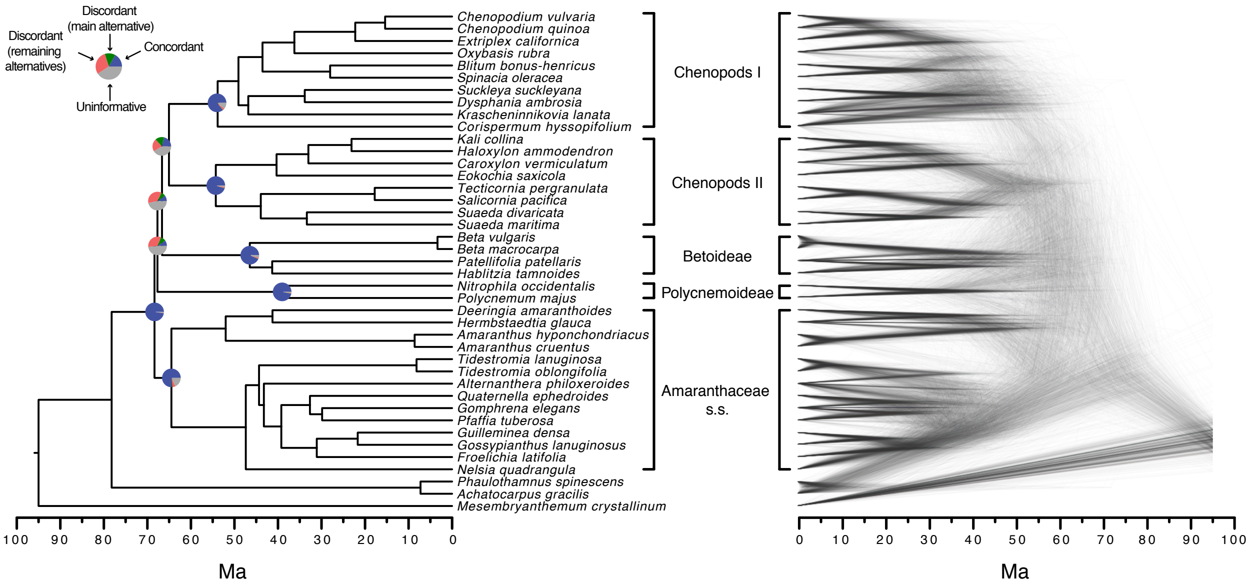


FIGURE 3. ASTRAL species tree (left) and cloudogram (right) inferred from 1,242 nuclear genes for the 41-taxon dataset of Amaranthaceae s.l. Pie charts on nodes present the proportion of gene trees that support that clade (blue), the proportion that support the main alternative bifurcation

(green), the proportion that support the remaining alternatives (red), and the proportion (conflict or support) that have < 50% bootstrap support (gray).

Chloroplast phylogenetic analysis of Amaranthaceae s.l.

The final alignment from 76 genes included 103 taxa and 55,517 bp in aligned length. The ML tree recovered the same five main clades within Amaranthaceae s.l. with the highest support (BS = 100; Fig. 4; Figs S4–S5). Within each main clade, relationships were fully congruent with (Kadereit et al. 2017) and mostly congruent with our nuclear analyses. However, the relationship among the five main clades differed from the nuclear tree. Here, Betoideae was retrieved as sister (BS = 100) of 'Chenopods I', while Amaranthaceae s.s. and Polycnemoideae were also recovered as sister clades (BS =100). Furthermore, the clade formed by Betoideae and 'Chenopods I', and Amaranthaceae s.s. and Polycnemoideae were recovered as sister groups (BS = 73), leaving ‘Chenopods II’ as sister to the former two. Conflict analysis confirmed the monophyly of Amaranthaceae s.l. with 51 out of 76 gene trees supporting this clade (ICA = 0.29) and full QS support (1/–/1). On the other hand, and similar to the nuclear phylogeny, conflict and QS analyses showed significant discordance in the backbone of the family (Fig. 4; Figs S4–S5). The sister relationship of Betoideae and 'Chenopods I' was supported by only 20 gene trees (ICA = 0.06), but it had a strong support from QS (0.84/0.88/0/94). The relationship between Amaranthaceae s.s. and Polycnemoideae was supported only by 15 gene trees (ICA = 0.07), while QS showed weak support (0.41/0.21.0.78) with signals of a supported secondary evolutionary history. The clade uniting Betoideae, 'Chenopods I', Amaranthaceae s.s., and Polycnemoideae was supported by only four-gene trees, with counter-support from both QS (-0.29/0.42/0.75) and ICA (-0.03), suggesting that most gene trees and sampled quartets supported alternative topologies. RNA editing prediction analysis revealed editing sites only on CDS sequences of reference plastome genomes (Table S3), suggesting that cpDNA reads in RNA-seq libraries come from RNA rather than DNA contamination from incomplete DNase digestion during sample processing.

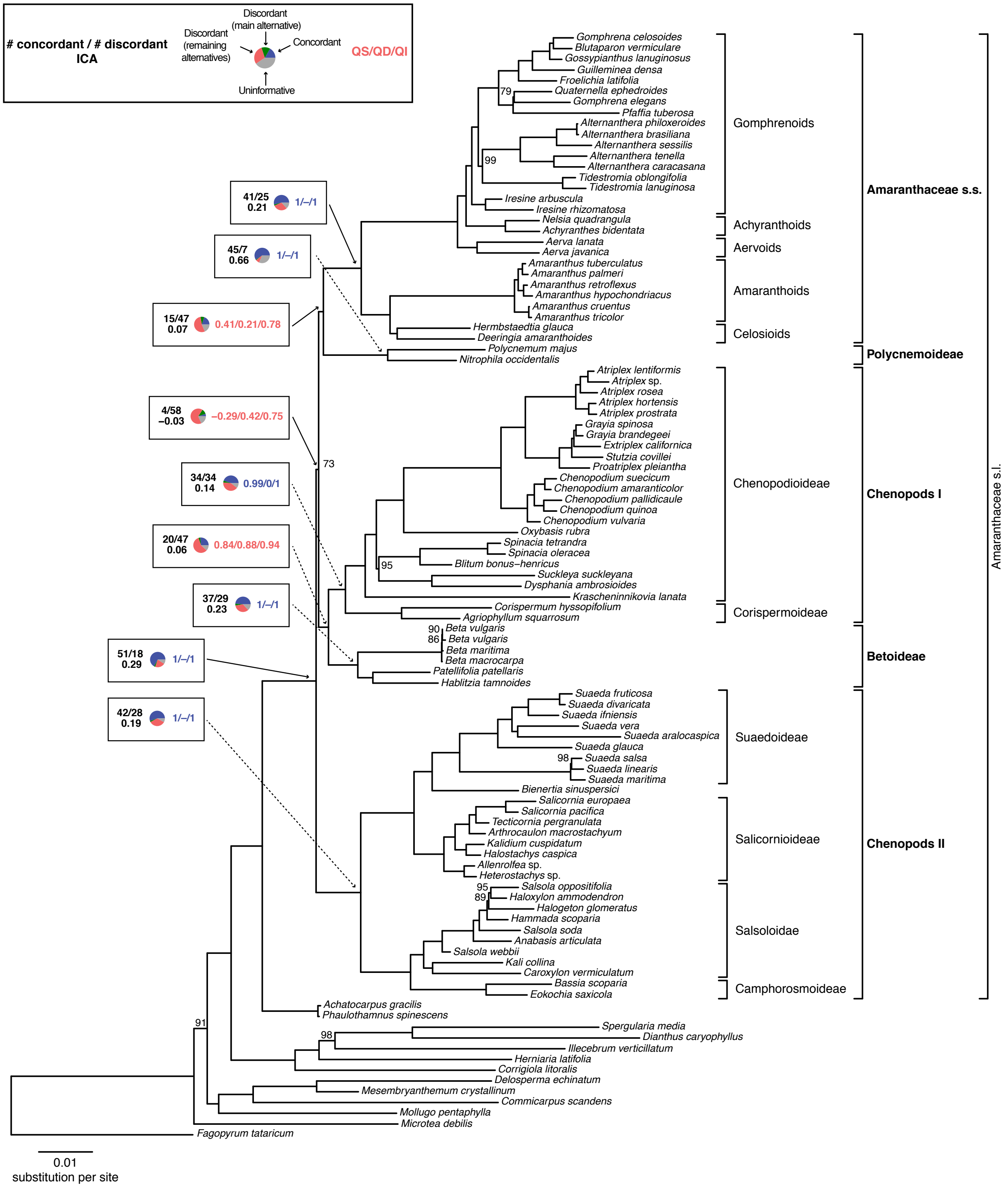


FIGURE 4. Maximum likelihood phylogeny of Amaranthaceae s.l. inferred from IQ-tree analysis of concatenated 76-chloroplast gene supermatrix. All nodes have full support (Bootstrap = 100/Local posterior probability = 100) unless noted next to nodes. Boxes contain gene tree conflict and Quartet Sampling (QS) scores for major clades (see Figs S2–S3 for all node scores). In each box, numbers on the upper left indicate the number of gene trees concordant/conflicting with that node in the species tree, and the number on the lower left indicate the Internode Certainty All (ICA) score. Pie charts present the proportion of gene trees that support that clade (blue), the proportion that support the main alternative bifurcation (green), the proportion that support the remaining alternatives (red), and the proportion (conflict or support) that have < 50% bootstrap support (gray). Numbers on the right of the pie chart indicate QS scores: Quartet concordance/Quartet differential/Quartet informativeness. QS scores in blue indicate support for individual major clades of Amaranthaceae s.l., while red scores indicate strong support for alternative relationships among them. Branch lengths are in number of substitutions per site (scale bar on the bottom).

Species network analysis of Amaranthaceae s.l.

Due to the computational limit of species network analyses, we reduced our full 105-taxon dataset to ten ingroup taxa plus one outgroup taxon. In this reduced dataset two taxa were used to represent the diversity for each of the five well-supported ingroup clades within Amaranthaceae s.l. The reduced 11-taxon(net) dataset included 4,138 orthologous gene alignments with no missing taxon and a minimum of 1,000 bp (aligned length after removal of low occupancy columns). The 11-taxon(net) ASTRAL species tree was congruent with the 105-taxon tree, while both the nuclear and chloroplast ML trees from concatenated supermatrices both had different topologies than their corresponding 105-taxon trees (Fig. 5). PhyloNet identified up to five hybridization events among the clades of Amaranthaceae s.l. (Fig. 5), with the best model having five hybridization events involving all five clades (Table 1). ‘Chenopods II’ was involved in hybridization events in all networks with one to five hybridization events. Model selection indicated that any species network was a better model than the bifurcating nuclear or chloroplast trees (Table 1).

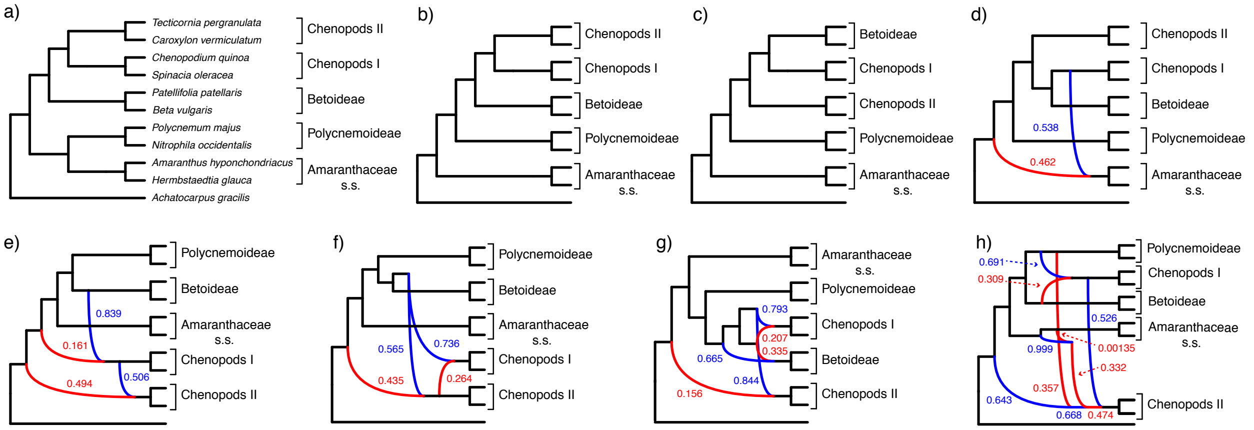


FIGURE 5. Species trees and species networks of the reduced 11-taxon(net) dataset of Amaranthaceae s.l. a) Maximum likelihood phylogeny inferred from RAxML analysis of the concatenated 4,138-nuclear gene supermatrix. b) Species tree inferred with ASTRAL using 4,138 nuclear genes. c) Maximum likelihood tree inferred from IQ-tree analysis of the concatenated 76-chloroplast gene supermatrix. d–h). Best species network inferred from PhyloNet pseudolikelihood analyses with 1 to 5 maximum number of hybridizations. Red and blue indicates the minor and major edges, respectively, of hybrid nodes. Number next to the branches indicates inheritance probabilities for each hybrid node.

TABLE 1. Model selection between maximum number of hybridizations in species networks searches.

Topology	Maximum number of hybridizations allowed	Number of inferred hybridizations	ln(L)	Parameters	Number of loci	AICc	ΔAICc	BIC	ΔBIC
RAxML ML tree	NA	NA	-24486.33124	19	4138	49048.84703	20589.66354	49130.89387	20546.62287
ASTRAL species tree	NA	NA	-23448.39741	19	4138	46972.97939	18513.79589	47055.02622	18470.75522
Chloroplast ML tree	NA	NA	-24568.33287	19	4138	49212.8503	20753.66681	49294.89713	20710.62614
Network 1	1	1	-21177.79113	21	4138	42439.80675	13980.62326	42530.46958	13946.19859
Network 2	2	2	-17275.62523	23	4138	34643.51881	6184.335324	34742.79372	6158.522728
Network 3	3	2	-16741.99114	23	4138	33576.25064	5117.067147	33675.52555	5091.254551
Network 4	4	3	-15415.80012	25	4138	30931.91638	2472.73289	31039.79943	2455.528435
Network 5	**5**	**5**	**-14171.37996**	**29**	**4138**	**28459.18349**	**0**	**28584.27099**	**0**

Four-taxon analyses

To test for hybridization events one at a time, we further reduced the 11-taxon(net) dataset to 10 four-taxon combinations that each included one outgroup and one representative each from three out of the five major ingroup clades. Between 7,756 and 8,793 genes were used for each quartet analysis (Table 2) and each quartet topology can be found in Figure 6. Only five out of the ten bifurcating quartet species trees (H0 and more frequent gene tree) were compatible with the nuclear species tree inferred from the complete 105-taxon dataset. The other five quartets compatible with the complete-taxon species tree corresponded to the second most frequent quartet gene trees, except for the quartet of Betoideae, 'Chenopods II' and Polycnemoideae (PBC2, which correspond to the least frequent gene tree).

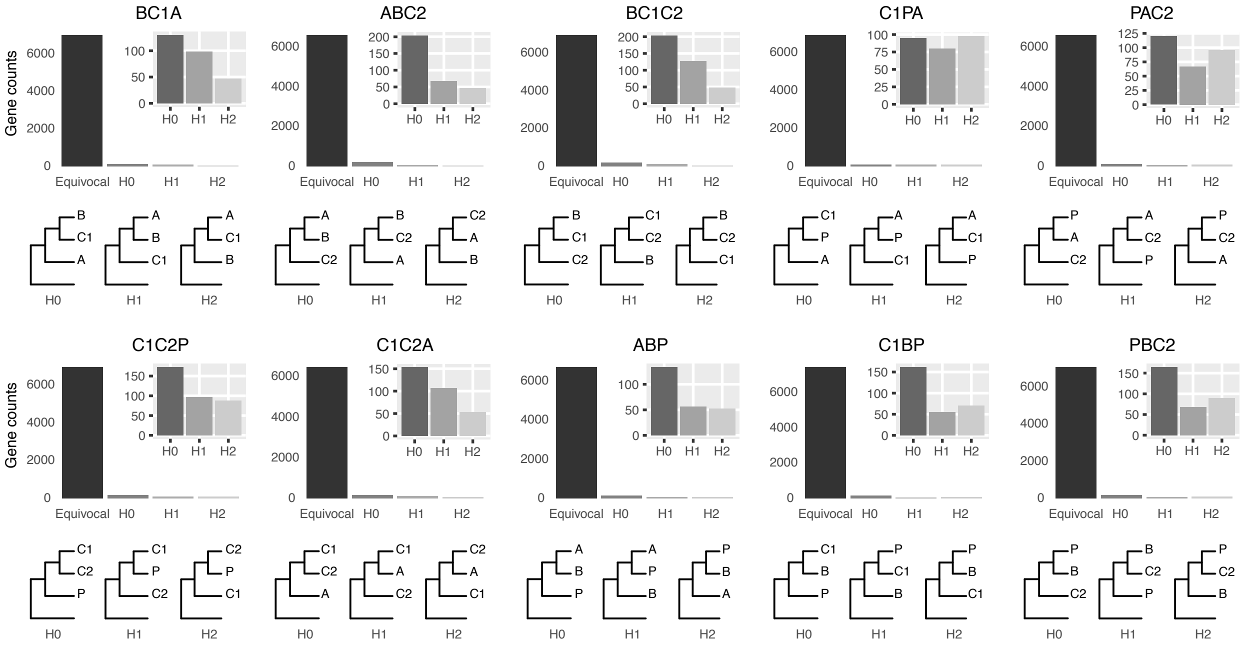


FIGURE 6. Gene counts from Approximate-Unbiased (AU) topology test of the 10 quartets from the five main clades of Amaranthaceae s.l. AU tests were carried out between the three possible topologies of each quartet. H0 represents the ASTRAL species tree of each quartet. Equivocal

indicates gene trees that fail to reject all three alternative topologies for a quartet with p ≤ 0.05. Gene counts for each of the three alternative topologies represent gene trees supporting unequivocally one topology by rejecting the other two alternatives with p ≤ 0.05. Insets represent gene count only for unequivocally topology support. Each quartet is named following the species tree topology, where the first two species are sister to each other (all topologies can be found in Figure S1). A = Amaranthaceae s.s. (represented by *Amarantus hypocondriacus*), B = Betoideae (*Beta vulgaris*), C1 = Chenopods I (*Chenopodium quinoa*), C2 = Chenopods II (*Caroxylum vermiculatum*), P = Polycnemoideae (*Polycnemum majus*). All quartets are rooted with *Mesembryanthemum crystallinum.*

Similar to the 105-taxon and the 11-taxon(net) datasets, the conflict analyses recovered significant conflict among all three possible rooted quartet topologies in all ten quartets. In each of the ten quartets, the ASTRAL species tree topology (H0) was the most frequent among individual gene trees (raw counts) but only with 35%–41% of occurrences while the other two topologies varied between similar or slightly skewed frequencies (Fig. S6a; Table S8). Gene counts based on the raw likelihood scores from the constraint analyses showed similar patterns (Fig. S6b; Table S8). Furthermore, when gene counts were filtered by significant likelihood support (i.e. **Δ**AICc ≥ 2), the number of trees supporting each of the three possible topologies dropped between 34% and 45%, but the species tree remained to be the most frequent topology for all quartets (Fig. S6b; Table S8). The AU topology tests failed to reject ($p \leq 0.05$) approximately 85% of the gene trees for any of the three possible quartet topologies and rejected all but a single topology in only 3%–4.5% of cases. Among the unequivocally selected gene trees, the frequencies among the three alternative topologies were similar to ones based on raw likelihood scores and overall the species tree was the most common topology for each quartet

(Fig 6; Table S8). Furthermore, the topology test clearly showed that most genes were uninformative for resolving the relationships among the major groups of Amaranthaceae s.l.

Across all ten quartets we found that most genes had very low TC scores (for any single node the maximum TC value is 1; Supplemental Fig. S7), showing that individual gene trees had also large conflict among bootstrap replicates, which is also a signal of uninformative genes and is concordant with the AU topology test results. Additionally, the linear models did not show any significant correlation between TC scores and alignment length, GC content or alignment gapless (Table S9), suggesting that filtering genes by any of these criteria are unlikely to increase the information content of the dataset.

Species network analyses followed by model selection using each of the four-taxon datasets showed that in seven out of the ten total quartets, the network with one hybridization event was a better model than any bifurcating tree topology. However, each of the best three networks from PhyloNet had very close likelihood scores and no significant **Δ**AICc among them. For the remaining three quartets the most common bifurcating tree (H0; C1PA, C1BP, PBC2) was the best model (Table 2; Figs 6, S6, S8).

TABLE 2. Model selection between quartet tree topologies and species networks. Trees correspond to each of the three possible quartet topologies where H0 is the ASTRAL quartet species tree. Networks correspond to the best three networks for searches with one hybridization event allowed.

Quartet[a]	**Topology**[b]	**ln(*L*)**	**Parameters**	**Number of loci**	**AICc**	**ΔAICc**	**BIC**	**ΔBIC**
BC1A	H0	-9014.809786	5	8258	18049.62684	24.73436754	18074.71426	14.70279692
	H1	-9072.456373	5	8258	18164.92002	140.0275408	18190.00743	129.9959702
	H2	-9073.888783	5	8258	18167.78484	142.8923611	18192.87225	132.8607905
	Net 1	**-8998.43945**	**7**	**8258**	**18024.89248**	**0**	**18060.01146**	**0**
	Net 2	-8998.439526	7	8258	18024.89263	0.000151947	18060.01162	0.000151947
	Net 3	-8998.441478	7	8258	18024.89653	0.004056302	18060.01552	0.004056302
ABC2	H0	-8516.854413	5	7811	17053.71651	12.87079823	17078.52527	2.950887757
	H1	-8581.563051	5	7811	17183.13379	142.2880731	17207.94254	132.3681626
	H2	-8582.670875	5	7811	17185.34944	144.5037223	17210.15819	134.5838118
	Net 1	**-8506.415681**	**7**	**7811**	**17040.84572**	**0**	**17075.57438**	**0**
	Net 2	-8506.415769	7	7811	17040.84589	0.000176519	17075.57456	0.000176519
	Net 3	-8506.42071	7	7811	17040.85577	0.010057548	17075.58444	0.010057548
BC1C2	H0	-9140.191425	5	8385	18300.39001	156.347016	18325.55385	146.2848258
	H1	-9201.981045	5	8385	18423.96925	279.9262567	18449.13309	269.8640665
	H2	-9214.405292	5	8385	18448.81775	304.7747517	18473.98158	294.7125615

	Net 1	**-9058.014812**	**7**	**8385**	**18144.04299**	**0**	**18179.26902**	**0**
	Net 2	-9058.019338	7	8385	18144.05205	0.009052497	18179.27807	0.009052497
	Net 3	-9058.024046	7	8385	18144.06146	0.018468011	18179.28749	0.018468011
C1PA	**H0**	**-8932.927759**	**5**	**8134**	**17885.8629**	**0**	**17910.87456**	**0**
	H1	-8936.145955	5	8134	17892.29929	6.436391285	17917.31095	6.436391285
	H2	-8936.481125	5	8134	17892.96963	7.106730999	17917.98129	7.106730999
	Net 1	-8932.077808	7	8134	17892.1694	6.306498884	17927.18227	16.30771403
	Net 2	-8932.078011	7	8134	17892.16981	6.306905172	17927.18268	16.30812032
	Net 3	-8932.078714	7	8134	17892.17121	6.308310587	17927.18408	16.30952573
PAC2	H0	-8530.661274	5	7784	17081.33026	40.10000797	17106.12168	30.18704595
	H1	-8552.9448	5	7784	17125.89731	84.66706025	17150.68873	74.75409823
	H2	-8548.291438	5	7784	17116.59059	75.36033576	17141.382	65.44737374
	Net 1	**-8506.607925**	**7**	**7784**	**17041.23025**	**0**	**17075.93463**	**0**
	Net 2	-8506.609795	7	7784	17041.23399	0.00373969	17075.93837	0.00373969
	Net 3	-8506.618966	7	7784	17041.25233	0.02208072	17075.95671	0.02208072
C1C2P	H0	-9119.250871	5	8341	18258.50894	12.50997925	18283.64643	2.458344441
	H1	-9163.685997	5	8341	18347.37919	101.38023	18372.51669	91.32859519
	H2	-9164.83263	5	8341	18349.67246	103.6734974	18374.80995	93.62186263
	Net 1	**-9108.992761**	**7**	**8341**	**18245.99896**	**0**	**18281.18809**	**0**
	Net 2	-9108.994383	7	8341	18246.00221	0.003244509	18281.19133	0.003244509

	Net 3	-9108.994843	7	8341	18246.00313	0.0041636	18281.19225	0.0041636
C1C2A	H0	-8447.623029	5	7756	16915.2538	63.6063012	16940.02717	53.70057058
	H1	-8520.509174	5	7756	17061.02609	209.378593	17085.79946	199.4728624
	H2	-8522.764578	5	7756	17065.5369	213.889401	17090.31027	203.9836704
	Net 1	**-8411.816521**	**7**	**7756**	**16851.6475**	**0**	**16886.3266**	**0**
	Net 2	-8411.819912	7	7756	16851.65428	0.006781956	16886.33338	0.006781956
	Net 3	-8411.820308	7	7756	16851.65507	0.007573446	16886.33417	0.007573446
ABP	H0	-9008.115816	5	8206	18036.23895	3.307596079	18061.29474	-6.711300872
	H1	-9015.941176	5	8206	18051.88967	18.95831519	18076.94546	8.939418238
	H2	-9014.738462	5	8206	18049.48424	16.55288764	18074.54003	6.533990688
	Net 1	**-9002.458846**	**7**	**8206**	**18032.93135**	**0**	**18068.00604**	**0**
	Net 2	-9002.460142	7	8206	18032.93395	0.002592568	18068.00863	0.002592568
	Net 3	-9002.464397	7	8206	18032.94246	0.011102577	18068.01714	0.011102577
C1BP	**H0**	**-9557.910518**	**5**	**8793**	**19135.82787**	**0**	**19161.22959**	**0**
	H1	-9661.475396	5	8793	19342.95762	207.1297559	19368.35935	207.1297559
	H2	-9661.009687	5	8793	19342.0262	206.1983365	19367.42793	206.1983365
	Net 1	-9556.24034	7	8793	19140.49343	4.665563813	19176.05266	14.82306554
	Net 2	-9556.243036	7	8793	19140.49882	4.670955519	19176.05805	14.82845724
	Net 3	-9556.246261	7	8793	19140.50527	4.677405326	19176.0645	14.83490705
PBC2	**H0**	**-9158.309463**	**5**	**8379**	**18336.62609**	**0**	**18361.78635**	**0**

H1	-9206.127177	5	8379	18432.26152	95.63542753	18457.42177	95.63542753
H2	-9205.933131	5	8379	18431.87343	95.24733612	18457.03368	95.24733612
Net 1	-9158.016519	7	8379	18344.04642	7.42032489	18379.26742	17.48107897
Net 2	-9158.017286	7	8379	18344.04795	7.421858749	18379.26896	17.48261282
Net 3	-9158.017377	7	8379	18344.04813	7.422042036	18379.26914	17.48279611

[a] Each quartet is named following the species tree topology, where the first two are sister. A = Amaranthaceae. s.s. (*Amaranthus hypochondriacus*), B = Betoideae (*Beta vulgaris*), C1 = Chenopods I (*Chenopodium quinoa*), C2 = Chenopods II (*Caroxylum vermiculatum*), P = Polycnemoideae (*Polycnemum majus*).

[b] All quartet tree topologies can be found in Figure 6 and quartet network topologies in Figure S8.

The ABBA/BABA test results showed a significant signal of introgression within each of the ten quartets (Table 3). The possible introgression was detected between six out of the ten possible pairs of taxa. Potential introgression between Betoideae and Amaranthaceae s.s., 'Chenopods I' or 'Chenopods II', and between 'Chenopods I' and Polycnemoideae was not detected.

TABLE 3. ABBA/BABA test results of Amaranthaceae s.l. five main groups quartets.

Quartet (H0)[a]	Number of loci	Sites in alignment	ABBA	BABA	Raw D-statistic	Z-score	P-value	Introgression direction
BC1A[b]	8258	12778649	287226	254617	0.06018164	41.1085	≤ 0.001	A⇔C1
ABC2	7811	12105324	252772	376755	-0.1969463	124.4161	≤ 0.001	A⇔C2
BC1C2	8385	13192317	306570	258349	0.08535914	54.59751	≤ 0.001	C1⇔C2
C1PA[b]	8134	12635201	342350	286813	0.08827124	64.62297	≤ 0.001	A⇔P
PAC2	7784	12049734	344726	405627	-0.08116313	42.88069	≤ 0.001	C2⇔P
C1C2P[b]	8341	13127397	445384	276652	0.2336892	136.0151	≤ 0.001	C2⇔P
C1C2A[b]	7756	12114778	396219	292561	0.1504951	101.3243	≤ 0.001	A⇔C2
ABP	8206	12622625	276319	312060	-0.06074486	36.64264	≤ 0.001	A⇔P
C1BP[b]	8793	13712853	273286	261620	0.02180944	18.08364	≤ 0.001	B⇔P
PBC2	8379	13074019	217549	415616	-0.3128205	196.8972	≤ 0.001	C2⇔P

[a]Each quartet is named following the species tree topology, where the first two are sister. A = Amaranthaceae. s.s. (*Amaranthus hypochondriacus*), B = Betoideae (*Beta vulgaris*), C1 = Chenopods I (*Chenopodium quinoa*), C2 = Chenopods II (*Caroxylum vermiculatum*), P = Polycnemoideae (*Polycnemum majus*). H0 topologies can be found in Figure 6

[b]Quartet compatible with the complete 105-taxon species trees

The synteny analysis between the diploid ingroup reference genome *Beta vulgaris* and the diploid outgroup reference genome *Mesembryanthemum crystallinum* recovered 22,179 (out of 52,357) collinear genes in 516 syntenic blocks. With the collinear ortholog pair information, we found that of the 8,258 orthologs of the BC1A quartet 6,941 contained orthologous genes within 383 syntenic blocks. The distribution of the BC1A quartet topologies along the chromosomes of *Beta vulgaris* did not reveal any spatial clustering along the chromosomes (Fig. S9).

Assessment of substitutional saturation, codon usage bias, compositional heterogeneity, and sequence evolution model misspecification

We assembled a second 11-taxon(tree) dataset that included 5,936 genes and a minimum of 300 bp (aligned length after removal of low occupancy columns) and no missing taxon. The saturation plots of uncorrected and predicted genetic distances showed that the first and second codon position are unsaturated (y = 0.8841002x), while the slope of the third codon positions (y = 0.5710071x) showed a clear signal of saturation (Fig. S10). The correspondence analyses of RSCU show that some codons are more frequently used in different species, but overall the codon usage seems to be randomly dispersed among all species and not clustered by clade (Fig. S11). This suggests that the phylogenetic signal is unlikely to be driven by differences in codon usage bias among clades. Furthermore, 549 (~9%) genes showed signal of compositional heterogeneity ($p < 0.05$) (Table S10). The topology and support (LPP = 1.0) for all branches was the same for the ASTRAL species trees obtained from the different data schemes while accounting for saturation, codon usage, compositional heterogeneity, and model of sequence evolution, and was also congruent with the ASTRAL species tree and concatenated ML from the full-taxon analyses (Fig. 7). In general, the proportion of gene trees supporting each bipartition remained the same in every analysis and showed high levels of conflict among the main clades of Amaranthaceae s.l. (Fig 7). Gene trees inferred accounting for selection of model of sequence evolution had higher bootstrap support resulting in higher proportion of both concordant and discordant trees (Fig 7b, 7d, 7e), but the proportion among them is the same as in the gene trees that used a single model of sequence evolution (Fig 7a–7c).

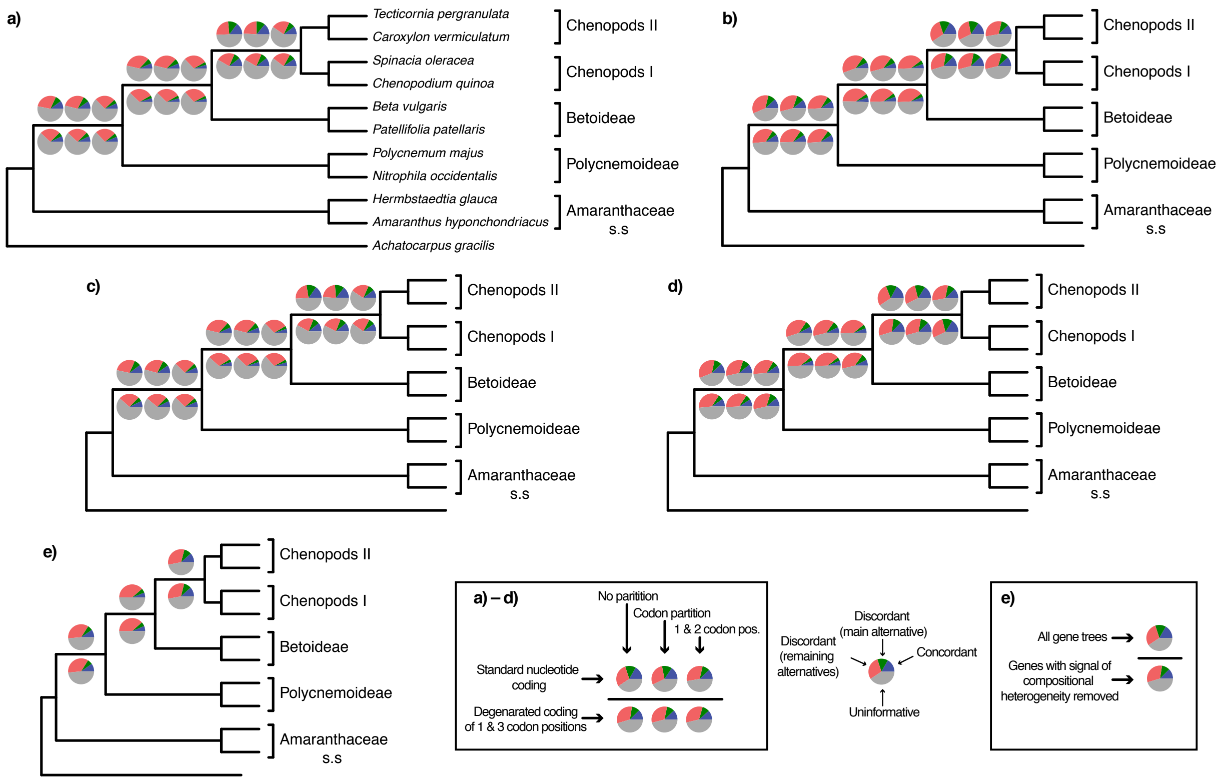


FIGURE 7. ASTRAL species trees from the 11-taxon(net) dataset estimated from gene trees inferred using multiple data schemes. a) Gene trees inferred with RAxML with a GTR-GAMMA model. b) Gene trees inferred with IQ-tree allowing for automatic model selection of sequence evolution. c) Gene trees inferred with RAxML with a GTR-GAMMA model and removal of genes that had signal of compositional heterogeneity. d) Gene trees inferred with IQ-tree allowing for automatic model selection of sequence evolution and removal of genes that had

signal of compositional heterogeneity. a–d) Gene trees were inferred with no partition, codon partition (first and second codon, and third codon) and, only first and second codon positions (third codon position removed and no partition). Gene trees were inferred using codon alignments with standard nucleotide coding, and alignments with degenerated coding of the first and third codon positions. e) All gene trees and gene trees after removal of genes that had signal of compositional heterogeneity, inferred with IQ-tree using amino acid sequences allowing for automatic model selection of sequence evolution. Pie charts on nodes present the proportion of gene trees that support that clade (blue), the proportion that support the main alternative bifurcation (green), the proportion that support the remaining alternatives (red), and the proportion (conflict or support) that have < 50% bootstrap support (gray).

Polytomy test

The ASTRAL polytomy test resulted in the same bifurcating species tree for the 11-taxon(tree) dataset and rejected the null hypothesis that any branch is a polytomy ($p < 0.01$ in all cases). These results were identical when using gene trees with collapsed branches.

Coalescent simulations

The distribution of tree-to-tree distances of the empirical and simulated gene trees to the species tree largely overlapped in all seven partition schemes tested (Fig. 8), suggesting that ILS alone was able to largely account for the gene tree heterogeneity seen in the 11-taxon(tree) dataset. The least overlap between empirical vs. simulated gene trees was observed in the dataset that only included the first and second codon positions in CDS, and in the amino acid dataset. This can be attributed to higher gene tree inference error due to removal of informative sites from the third codon position.

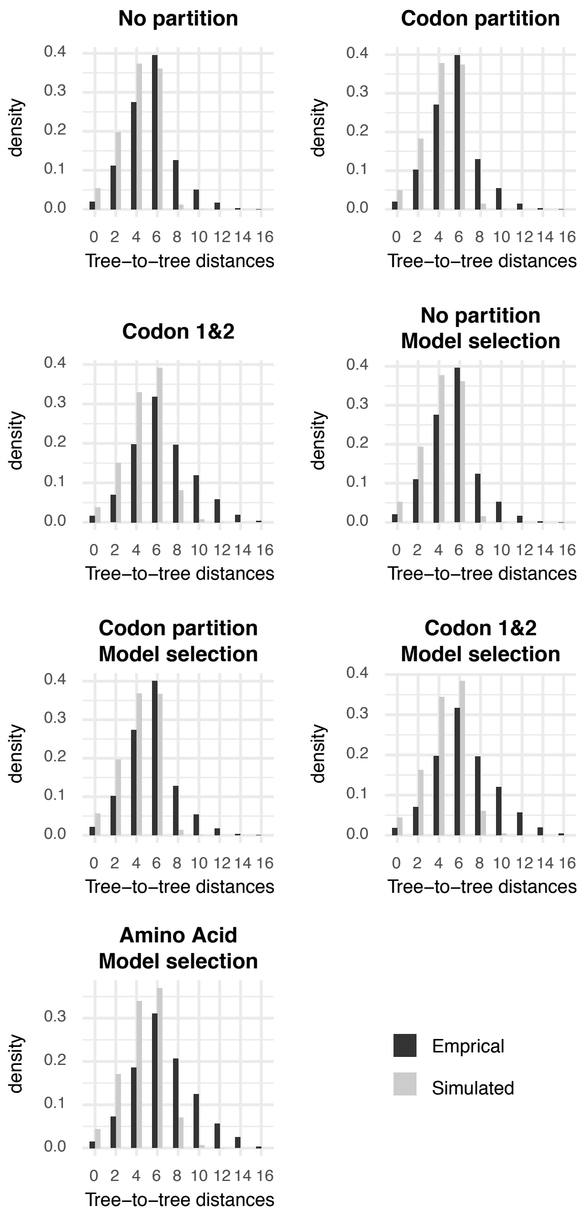


FIGURE 8. Distribution of tree-to-tree distances from empirical gene trees and species tree versus coalescent simulation. Simulations were carried out using the ASTRAL species trees from the 11-taxon(tree) dataset estimated from gene trees inferred using seven data schemes. Species trees used for the coalescent simulation can be seen in Figure 9.

Test of anomaly zone

The anomaly zone limit calculations using species trees from the 11-taxon(tree) dataset revealed that two pairs of internodes in the Amaranthaceae s.l. species tree fell into the anomaly zone. These internodes are characterized by having very short branches relative to the rest of the tree. The branch lengths among species trees from the seven different data schemes varied among the trees, but the same internodes were identified under the anomaly zone in all cases. The first pair of internodes is located between the clade comprised of all Amaranthaceae s.l. and the clade that includes Chenopods I, Chenopods II, and Betoideae. The second pair of internodes is located between the clade that includes Chenopods I, Chenopods II, and Betoideae and the clade composed of Chenopods I and Chenopods II (Table 4, Fig. 9).

TABLE 4. Anomaly zone limit calculations in 11-taxon species trees. Bold rows show pair of internodes in the anomaly zone when y < a(x).

Species tree[a]	Clade (x)[b]	Clade (y)[b]	x	y	a(x)
No partition	(C1, C2)	(C1)	0.1467	2.722	0.1799
	(C1, C2)	(C2)	0.1467	2.1102	0.1799
	((C1, C2), B)	**(C1, C2)**	**0.1045**	**0.1467**	**0.3084**
	((C1, C2), B)	(B)	0.1045	2.6081	0.3084
	(((C1, C2), B), P)	**((C1, C2), B)**	**0.0846**	**0.1045**	**0.4003**
	(((C1, C2), B), P)	(P)	0.0846	3.5424	0.4003
Codon partition	(C1, C2)	(C1)	0.1433	2.6766	0.1882
	(C1, C2)	(C2)	0.1433	2.0655	0.1882
	((C1, C2), B)	**(C1, C2)**	**0.0931**	**0.1433**	**0.3572**
	((C1, C2), B)	(B)	0.0931	2.5862	0.3572
	(((C1, C2), B), P)	**((C1, C2), B)**	**0.0734**	**0.0931**	**0.4669**
	(((C1, C2), B), P)	(P)	0.0734	3.5366	0.4669
Codon 1&2	(C1, C2)	(C1)	0.118	1.3092	0.26
	(C1, C2)	(C2)	0.118	1.7645	0.26
	((C1, C2), B)	**(C1, C2)**	**0.1009**	**0.118**	**0.3231**
	((C1, C2), B)	(B)	0.1009	1.6229	0.3231
	(((C1, C2), B), P)	**((C1, C2), B)**	**0.0673**	**0.1009**	**0.5102**

	(((C1, C2), B), P)	(P)	0.0673	2.3625	0.5102
No partition - Model selection	(C1, C2)	(C1)	0.1439	2.1199	0.1865
	(C1, C2)	(C2)	0.1439	2.7013	0.1865
	((C1, C2), B)	**(C1, C2)**	**0.1079**	**0.1439**	**0.2951**
	((C1, C2), B)	(B)	0.1079	2.5846	0.2951
	(((C1, C2), B), P)	**((C1, C2), B)**	**0.0803**	**0.1079**	**0.4242**
	(((C1, C2), B), P)	(P)	0.0803	3.58	0.4242
Codon partition - Model selection	(C1, C2)	(C1)	0.1427	2.6398	0.1896
	(C1, C2)	(C2)	0.1427	2.0911	0.1896
	((C1, C2), B)	**(C1, C2)**	**0.0945**	**0.1427**	**0.351**
	((C1, C2), B)	(B)	0.0945	2.6252	0.351
	(((C1, C2), B), P)	**((C1, C2), B)**	**0.0721**	**0.0945**	**0.4758**
	(((C1, C2), B), P)	(P)	0.0721	3.5978	0.4758
Codon 1&2 - Model selection	(C1, C2)	(C1)	0.1232	1.9043	0.2432
	(C1, C2)	(C2)	0.1232	1.415	0.2432
	((C1, C2), B)	**(C1, C2)**	**0.1024**	**0.1232**	**0.317**
	((C1, C2), B)	(B)	0.1024	1.7399	0.317
	(((C1, C2), B), P)	**((C1, C2), B)**	**0.07**	**0.1024**	**0.4906**
	(((C1, C2), B), P)	(P)	0.07	2.5666	0.4906
Amino Acid - Model selection	(C1, C2)	(C1)	0.115	1.887	0.269
	(C1, C2)	(C2)	0.115	1.317	0.269
	((C1, C2), B)	**(C1, C2)**	**0.122**	**0.115**	**0.247**
	((C1, C2), B)	(B)	0.122	1.744	0.247
	(((C1, C2), B), P)	**((C1, C2), B)**	**0.077**	**0.122**	**0.444**
	(((C1, C2), B), P)	(P)	0.077	2.471	0.444

[a] Species tree topologies can be found in Figure 7.

[b] B = Betoideae (*Beta vulgaris*), C1 = Chenopods I (*Chenopodium quinoa*), C2 = Chenopods II (*Caroxylum vermiculatum*), P = Polycnemoideae (*Polycnemum majus*).

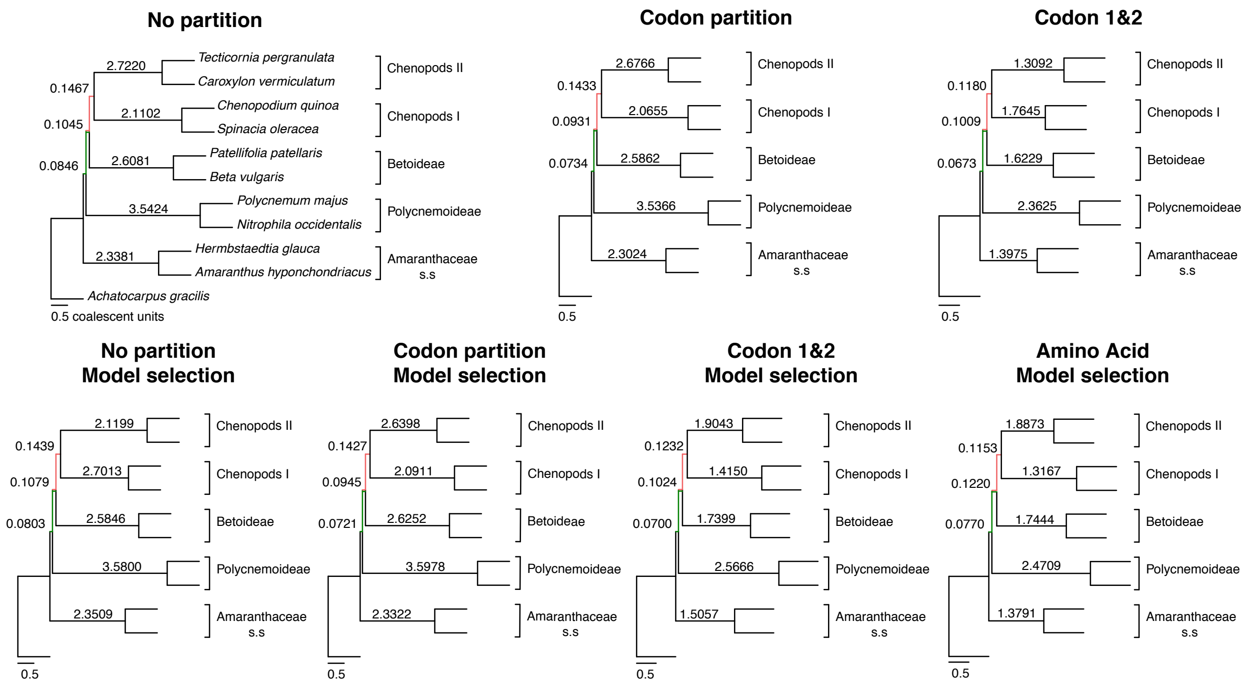


FIGURE 9. ASTRAL species trees from the 11-taxon(tree) dataset estimated from individual gene trees inferred with seven data schemes. Number next or above branches represent branch length in coalescent units. Colored branches represent pairs of internodes that fall in the anomaly zone (see Table 4 for anomaly zone limits).

The gene tree counts showed that the species tree was not the most common gene tree topology in four of the seven data schemes analyzed, as expected for the anomaly zone (Fig. S12). When gene trees were inferred with no partition or partitioned by codon, the species tree was the fourth most common gene tree topology (119 out of 5,936 gene trees), while the most common gene tree topologies occurred between 170 and 149 times (Fig. 10). Similar patterns were identified for gene trees inferred while accounting from model of sequence evolution selection (Fig. 10). Interestingly, for the gene tree sets inferred using only the first and second codons, and amino acids, the species tree was the most common topology.

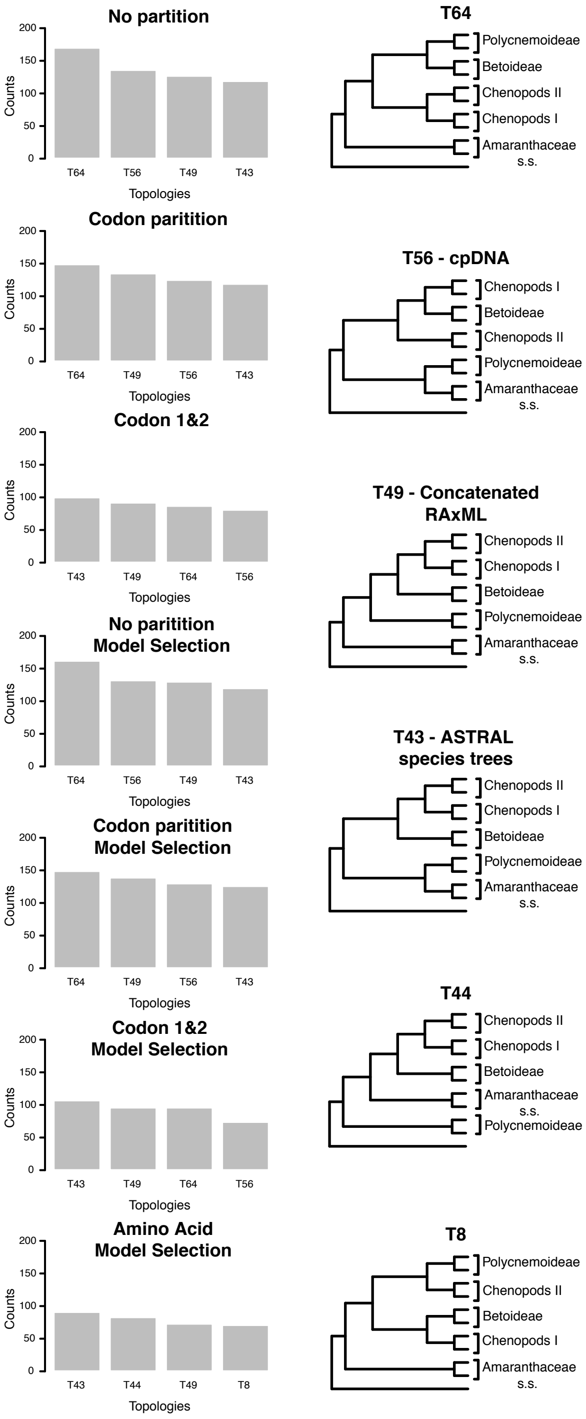


FIGURE 10. Gene tree counts (left) of the four most common topologies (right) of 11-taxon(tree) dataset inferred with seven data schemes. Gene trees that do not support the monophyly of any of the five major clades were ignored.

DISCUSSION

Using a phylotranscriptomic dataset in combination with reference genomes representing major clades, we have shown the prevalence of gene tree discordance in the backbone phylogeny of Amaranthaceae s.l. Interestingly, we found that this discordance is also present within the chloroplast dataset. Despite the strong signal of gene tree discordance, we were able to identify five well-supported major clades within Amaranthaceae s.l. that are congruent with morphology and previous taxonomic treatments of the group. Using multiple phylogenetic tools and simulations we comprehensively tested for processes that might have contributed to the gene tree discordance in Amaranthaceae s.l. Phylogenetic network analyses and ABBA-BABA tests both supported multiple reticulation events among the five major clades in Amaranthaceae s.l. At the same time, the patterns of gene tree discordance among these clades can also largely be explained by uninformative gene trees and ILS. We found evidence that three consecutive short internal branches produce anomalous trees contributing to the discordance. Molecular evolution model misspecification (i.e. substitutional saturation, codon usage bias, or compositional heterogeneity) was less likely to account for the gene tree discordance. Taken together, no single source can confidently be pointed out to account for the strong signal of gene tree discordance, suggesting that the discordance results primarily from ancient and rapid lineage diversification. Furthermore, the backbone of Amaranthaceae s.l. and remains —and probably will remain— unresolved even with genome-scale data. Our work highlights the need to test for multiple sources of conflict in phylogenomic analyses and provide a set of recommendations moving forward in resolving ancient and rapid diversification.

Five well-supported major clades in Amaranthaceae s.l.

Both our nuclear and chloroplast datasets strongly supported five major clades within Amaranthaceae s.l.: Amaranthaceae s.s, ‘Chenopods I’, ‘Chenopods II’, Betoideae, and Polycnemoideae (Figs. 2 & 4). We recovered Amaranthaceae s.s., Betoideae, and Polycnemoideae as monophyletic, which is consistent with morphology and the most recent molecular analyses of these lineages (Hohmann et al. 2006; Masson and Kadereit 2013; Di Vincenzo et al. 2018). In the case of Chenopodiaceae s.s., the nuclear analyses (Fig. 2) suggested the monophyly of this previously segregated family, but gene tree discordance analyses revealed high levels of conflict among two well-defined clades (Fig. 2), ‘Chenopods I’ and ‘Chenopods II’. Moreover, the chloroplast analyses did not support the monophyly of Chenopodiaceae s.s. While we also find evidence of gene tree discordance in the backbone cpDNA phylogeny (see below), a sister relationship between ‘Chenopods I’ and Betoideae had strong QS support (0.84/0.88/0.94; Fig. 4). Weak support and/or conflicting topologies along the backbone on the Amaranthaceae s.l. characterize all previous molecular studies of the lineage (Fig. 1), even with hundreds of loci (Walker et al. 2018). On the other hand, all studies support the five major clades found in our analysis.

For the sake of taxonomic stability, we therefore suggest retaining Amaranthaceae s.l. sensu APG IV (The Angiosperm Phylogeny Group et al. 2016), which includes the previously recognized Chenopodiaceae. Here we recognize five subfamilies within Amaranthaceae s.l.: Amaranthoideae representing Amaranthaceae s.s. (incl. Gomphrenoideae Schinz), Betoideae Ulbr., Chenopodioideae represented as ‘Chenopods I’ here (incl. Corispermoideae Ulbr.), Polycnemoideae Ulbr., and Salicornioideae Ulbr. represented by ‘Chenopods II’ (incl. Salsoloideae Ulbr., Suaedoideae Ulbr. and Camphorosmoideae A.J. Scott). The stem ages of these five subfamilies date back to the early Tertiary (Paleocene, Fig. 3) which agrees with dates based on chloroplast markers (Kadereit et al. 2012; Di Vincenzo et al. 2018; Yao et al. 2019). Due to the gene tree discordance along the backbone, the geographic origin of Amaranthaceae s.l. remains ambiguous.

Gene tree discordance detected among chloroplast genes

Our concatenation-based chloroplast phylogeny (Fig. 4) retrieved the same five major clades of Amaranthaceae s.l. as in the nuclear phylogeny, but the relationships among the major clades are incongruent with the nuclear phylogeny (Fig. 2). Cytonuclear discordance is a well-known process in plants and it has been traditionally attributed to reticulate evolution (Rieseberg and Soltis 1991; Sang et al. 1995; Soltis and Kuzoff 1995). Such discordance continues to be treated as evidence in support of hybridization in more recent phylogenomic studies that assume the chloroplast to be a single, linked locus (e.g. Folk et al. 2017; Vargas et al. 2017; Morales-Briones et al. 2018b; Lee-Yaw et al. 2019). However, cytonuclear discordance can also be attributed to other processes like ILS (Doyle 1992; Ballard and Whitlock 2004). Recent work shows that chloroplast protein-coding genes may not necessarily act as a single locus, and high levels of tree conflict has been detected (Gonçalves et al. 2019; Walker et al. 2019).

In Amaranthaceae s.l., previous studies based on chloroplast protein-coding genes or introns (Kadereit et al. 2003; Müller and Borsch 2005; Hohmann et al. 2006; Kadereit et al. 2017) resulted in different relationships among the five main clades and none in agreement with our 76-gene phylogeny. Our conflict and QS analyses of the chloroplast dataset (Fig. 4; Figs S4–S5) revealed strong signals of gene tree discordance among the five major clades of Amaranthaceae s.l. The strong conflicting signal in the chloroplast genome may be attributed to heteroplasmy and difference in individual gene phylogenetic information (Walker et al. 2019), although the exact sources of conflict are yet to be clarified (Gonçalves et al. 2019). Unlike the results found by Walker et al. (2019), nodes showing conflicting signals in individual gene trees in our dataset were mostly highly supported (i.e. BS ≥ 70, Fig S4), suggesting that low phylogenetic information is not the source of conflict in our chloroplast dataset.

Our results support previous studies showing RNA-seq data can be a reliable source for plastome assembly (Smith 2013; Osuna-Mascaró et al. 2018). While the approach has been used for deep-scale phylogenomic reconstruction in green plants (Gitzendanner et al. 2018), at present extracting plastome data is not part of routine phylotranscriptomic pipelines. RNA-seq libraries can contain some genomic DNA due to incomplete digestion during RNA purification (Smith 2013) and given the AT-rich nature of plastomes, this allows plastome DNA to survive the poly-A selection during mRNA enrichment (Schliesky et al. 2012). However, our results showed that Amaranthaceae s.l. cpDNA assemblies came from RNA rather than DNA contamination regardless of library preparation strategies. Similarly, Osuna-Mascaró et al. (2018) also found highly similar plastome assemblies (i.e. general genome structure, and gene number and composition) from RNA-seq and genomic libraries, supports the idea that plastome genomes are fully transcribed in photosynthetic eukaryotes (Shi et al. 2016). Here we implemented additional steps to the Yang and Smith (2014) pipeline to filter chloroplast and mitochondrial reads prior to *de novo* transcriptome assembly, which allowed us to assemble plastome sequences from RNA-seq libraries, build a plastome phylogeny, and compare it to gene trees constructed from nuclear genes. Furthermore, the backbone topology of our cpDNA tree built mainly from RNA-seq data (97 out of 105 samples) was consistent with a recent complete plastome phylogeny of Caryophyllales (Yao et al. 2019), showing the potential value of using cpDNA from RNA-seq data. Nonetheless, RNA editing might be problematic when combining samples from RNA and DNA, especially when trying to resolve phylogenetic relationships among closely related species.

Hybridization

Rapid advances have been made in recent years in developing methods to infer species networks in the presence of ILS (reviewed in Elworth et al. 2019). These methods have been increasingly used in phylogenetic studies (e.g. Marcussen et al. 2014; Wen et al. 2016a; Copetti et al. 2017); Morales-Briones et al. 2018a; Crowl et al. 2019). To date, however, species network inference is still computationally intensive and limited to a small number of species and a few hybridization events (Hejase and Liu 2016; but see Hejase et al. 2018 and Zhu et al. 2019). Furthermore, studies evaluating the performance of different phylogenetic network inference approaches are scarce and restricted to simple hybridization scenarios. (Kamneva and Rosenberg 2017) showed that likelihood methods like Yu et al. (2014) are often robust to ILS and gene tree error when symmetric hybridization (equal genetic contribution of both parents) events are considered, and while it usually does not overestimate hybridization events, it fails to detect skewed hybridization (unequal genetic contribution of both parents) events in the presence of significant ILS. Methods developed to scale to larger numbers of species and hybridizations like the ones using pseudo-likelihood approximations (i.e. Solís-Lemus and Ané 2016; Yu and Nakhleh 2015) are yet to be evaluated independently, but in the case of Yu and Nakhleh (2015), a method based on rooted triples, it has been shown that this method cannot distinguish the correct network when other networks can produce the same set of triples (Yu and Nakhleh 2015). The result of our 11-taxon(net) phylogenetic analysis using a pseudo-likelihood approach detected up to five hybridization events involving all five major clades of Amaranthaceae s.l. (Fig. 5). Model selection, after calculating the full likelihood of the obtained networks, chose the 5-reticulation species as the best model. Also, any species network had a better score than a bifurcating tree (Table 1). However, a further look of these hybridization events by breaking the 11-taxon dataset into ten quartets showed that full likelihood networks searches with up to one hybridization event are indistinguishable from each other (Table 2), resembling a random gene tree distribution. This pattern can probably be explained by the high levels of gene tree discordance and lack of phylogenetic signal in the inferred quartet gene trees (Fig. 6), suggesting that the 11-taxon(net) network searches can potentially overestimate reticulation events due to high levels of gene tree error or ILS.

Using the *D*-Statistic (Green et al. 2010; Durand et al. 2011) we also found signals of introgression in seven possible directions among the five main groups of Amaranthaceae s.l. (Table 3). The inferred introgression events agreed with at least one of the reticulation scenarios from the phylogenetic network analysis. However, the *D*-Statistic did not detect any introgression that involves Betoideae, which was detected in the phylogenetic network analysis with either four or five reticulations events. The *D*-Statistic has been shown to be robust to a wide range of divergence times, but it is sensitive to relative population size (Zheng and Janke 2018), which agrees with the notion that large effective population sizes and short branches increase the chances of ILS (Pamilo and Nei 1988) and in turn can dilute the signal for the *D*-Statistic (Zheng and Janke 2018). Recently, Elworth et al. (2018) found that multiple or ‘hidden’ reticulations can cause the signal of the *D*-statistic to be lost or distorted. Furthermore, when multiple reticulations are present, the traditional approach of subsetting datasets into quartets can be problematic as it largely underestimates *D* values (Elworth et al. 2018). Given short internal branches in the backbone of Amaranthaceae s.l. and the phylogenetic network results showing multiple hybridizations, it is plausible that our *D*-statistic may be affected by these issues. Our analysis highlights the uncertainty of relying *D*-statistic as the only test for detecting reticulation events, especially in cases of ancient and rapid diversification.

ILS and the Anomaly Zone

Incomplete Lineage Sorting, or ILS, is ubiquitous in multi-locus phylogenetic datasets. In its most severe cases ILS produces the ‘anomaly zone’, defined as a set of short internal branches in the species tree that produce anomalous gene trees (AGTs) that are more likely than the gene tree that matches the species tree (Degnan and Rosenberg 2006). Rosenberg (2013) expanded the definition of the anomaly zone to require that a species tree contain two consecutive internal branches in an ancestor–descendant relationship in order to produce AGTs. To date, only a few examples of an empirical anomaly zone have been reported (Linkem et al. 2016; Cloutier et al. 2019). Furthermore, Huang and Knowles (2009) have pointed out that the gene tree discordance produced from the anomaly zone can also be produced by uninformative gene trees and that for species trees with short branches the most probable gene tree topology is a polytomy rather than an AGT. Our results show that the species tree of Amaranthaceae s.l. have three consecutive short internal branches that lay within the limits of the anomaly zone (i.e. $y < a(x)$; Fig. 9; Table 4). While this is clear evidence that gene tree discordance in Amaranthaceae s.l may be product of AGTs, it is important to point out that our quartet analysis showed that most quartet gene trees were equivocal (94–96%; Fig. 6), and were therefore uninformative gene trees. Nonetheless, the ASTRAL polytomy test rejected a polytomy along the backbone of Amaranthaceae s.l. in any of the gene tree sets used. While we did not test for polytomies in individual gene trees, our

ASTRAL polytomy test was also carried using gene trees with branches collapsed if they had <75% bootstrap support, and obtained the same species tree with polytomy being rejected. Furthermore, we found that for most of the partition schemes tested, the species tree is not the most frequent gene tree (Fig. 10). The distribution of gene tree frequency in combination with short internal branches in the species tree supports the presence of an anomaly zone in Amaranthaceae s.l.

Considerations in distinguishing sources of gene tree discordance

With the frequent generation of phylogenomic datasets, the need to explore and disentangle gene tree discordance has become a fundamental step to understand the phylogenetic relationships of recalcitrant groups across the Tree of Life. Recently, development of tools to identify and visualize gene tree discordance has received great attention (e.g. Salichos et al. 2014; Smith et al. 2015; Huang et al. 2016; Pease et al. 2018). New tools have facilitated the detection of conflict, which has led to the development of downstream phylogenetic analyses to attempt to characterize it. Although exploring sources of conflicting signal in phylogenomic data is now common, this is typically focused on data filtering approaches and its effect on concatenation-based vs. coalescent-based tree inference methods (e.g. Alda et al. 2019; Mclean et al. 2019; Roycroft et al. 2019). Methods to estimate species trees from phylogenomic dataset while accounting for multiple sources of conflict and molecular substitution simultaneously are not available, but by combining transcriptomes and genomes, we were able to create a rich and dense dataset to start to tease apart alternative hypotheses concerning the sources of conflict in the backbone phylogeny of Amaranthaceae s.l. Nonetheless, we could not attribute the strong gene tree discordance signal to a single main source. Instead, we found that gene tree heterogeneity observed in Amaranthaceae s.l. is likely to be explained by a combination of processes, including ILS, hybridization, uninformative genes, and molecular evolution model misspecification, that might have acted simultaneously and/or cumulatively.

Our results highlight the need to test for multiple sources of conflict in phylogenomic analyses, especially when trying to resolve phylogenetic relationships in groups with a long history of phylogenetic conflict. We consider that special attention should be put in data processing, orthology inference, as well as the informativeness of individual gene trees. Furthermore, we need to be aware of the strengths and limitations of different phylogenetic methods and be cautious against relying on a single analysis, for example in the usage of phylogenetics species networks over coalescent-based species trees (also see Blair and Ané 2019). While the backbone phylogeny of Amaranthaceae s.l. remains difficult to resolve despite employing genome-scale data, a question emerges whether this is an atypical case, or as we leverage more phylogenomic datasets and explore gene tree discordance in more detail, we could find similar patterns in other groups, especially in those that are products of ancient and rapid lineage diversification (Widhelm et al. 2019). Ultimately, such endeavor will be instrumental in our fundamental understanding of the biology of the organisms.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/.[NNNN]

Acknowledgments

The authors thank H. Freitag, J.M. Bena and the Millennium Seed Bank for providing seeds; Ursula Martiné for assisting with RNA extraction; Alexandra Crum for help revising the manuscript; X anonymous reviewers for providing helpful comments; the Minnesota Supercomputing Institute (MSI) at the University of Minnesota for providing access to computational resources. This work was supported by the University of Minnesota, University of Michigan, and the US National Science Foundation (DEB 1354048).

References

Alda F., Tagliacollo V.A., Bernt M.J., Waltz B.T., Ludt W.B., Faircloth B.C., Alfaro M.E., Albert J.S., Chakrabarty P. 2019. Resolving Deep Nodes in an Ancient Radiation of Neotropical Fishes in the Presence of Conflicting Signals from Incomplete Lineage Sorting. Syst. Biol. 68:573–593.

Arcila D., Ortí G., Vari R., Armbruster J.W., Stiassny M.L.J., Ko K.D., Sabaj M.H., Lundberg J., Revell L.J., Betancur-R. R. 2017. Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. Nat. Ecol. Evol. 1:0020.

Ballard J.W.O., Whitlock M.C. 2004. The incomplete natural history of mitochondria. Mol. Ecol. 13:729–744.

Bankevich A., Nurk S., Antipov D., Gurevich A.A., Dvorkin M., Kulikov A.S., Lesin V.M., Nikolenko S.I., Pham S., Prjibelski A.D., Pyshkin A.V., Sirotkin A.V., Vyahhi N., Tesler G., Alekseyev M.A., Pevzner P.A. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J. Comput. Biol. 19:455–477.

Bena M.J., Acosta J.M., Aagesen L. 2017. Macroclimatic niche limits and the evolution of C4 photosynthesis in Gomphrenoideae (Amaranthaceae). Bot. J. Linn. Soc. 184:283–297.

Blackmon H., Adams R.A. 2015 EvobiR: Tools for comparative analyses and teaching evolutionary biology. doi:10.5281/zenodo.30938

Blair C., Ané C. 2019. Phylogenetic Trees and Networks Can Serve as Powerful and Complementary Approaches for Analysis of Genomic Data. Syst. Biol. syz056, https://doi.org/10.1093/sysbio/syz056

Bolger A.M., Lohse M., Usadel B. 2014. Trimmomatic - a flexible trimmer for Illumina sequence data. Bioinformatics. 30:2112–2120.

Bouckaert R., Heled J. 2014. DensiTree 2: Seeing Trees Through the Forest. BioRxiv. 012401.

Brown J.W., Walker J.F., Smith S.A. 2017. Phyx - phylogenetic tools for unix. Bioinformatics. 33:1886–1888.

Bruen T.C., Philippe H., Bryant D. 2006. A Simple and Robust Statistical Test for Detecting the Presence of Recombination. Genetics. 172:2665–2681.

Buchfink B., Xie C., Huson D.H. 2015. Fast and sensitive protein alignment using DIAMOND. Nat. Methods. 12:59–60.

Buckley T.R., Cordeiro M., Marshall D.C., Simon C. 2006. Differentiating between Hypotheses of Lineage Sorting and Introgression in New Zealand Alpine Cicadas (Maoricicada Dugdale). Syst. Biol. 55:411–425.

Castresana J. 2000. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. Mol. Biol. Evol. 17:540–552.

Chen L.-Y., Morales-Briones D.F., Passow C.N., Yang Y. 2019. Performance of gene expression analyses using de novo assembled transcripts in polyploid species. Bioinformatics. btz620, https://doi.org/10.1093/bioinformatics/btz620

Cloutier A., Sackton T.B., Grayson P., Clamp M., Baker A.J., Edwards S.V. 2019. Whole-Genome Analyses Resolve the Phylogeny of Flightless Birds (Palaeognathae) in the Presence of an Empirical Anomaly Zone. Syst. Biol. syz019, https://doi.org/10.1093/sysbio/syz019

Cooper E.D. 2014. Overly simplistic substitution models obscure green plant phylogeny. Trends Plant Sci. 19:576–582.

Copetti D., Búrquez A., Bustamante E., Charboneau J.L.M., Childs K.L., Eguiarte L.E., Lee S., Liu T.L., McMahon M.M., Whiteman N.K., Wing R.A., Wojciechowski M.F., Sanderson M.J. 2017. Extensive gene tree discordance and hemiplasy shaped the genomes of North American columnar cacti. Proc. Natl. Acad. Sci. 114:12003–12008.

Cox C.J., Li B., Foster P.G., Embley T.M., Civáň P. 2014. Conflicting Phylogenies for Early Land Plants are Caused by Composition Biases among Synonymous Substitutions. Syst. Biol. 63:272–279.

Crowl A.A., Manos P.S., McVay J.D., Lemmon A.R., Lemmon E.M., Hipp A.L. 2019. Uncovering the genomic signature of ancient introgression between white oak lineages (*Quercus*). New Phytol. nph.15842, https://doi.org/10.1111/nph.15842

Davidson N.M., Oshlack A. 2014. Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. Genome Biol. 15:57.

Degnan J.H., Rosenberg N.A. 2006. Discordance of Species Trees with Their Most Likely Gene Trees. PLoS Genet. 2:e68.

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.

Di Vincenzo V., Gruenstaeudl M., Nauheimer L., Wondafrash M., Kamau P., Demissew S., Borsch T. 2018. Evolutionary diversification of the African achyranthoid clade (Amaranthaceae) in the context of sterile flower evolution and epizoochory. Ann. Bot. 122:69–85.

Dohm J.C., Minoche A.E., Holtgräwe D., Capella-Gutiérrez S., Zakrzewski F., Tafer H., Rupp O., Sörensen T.R., Stracke R., Reinhardt R., Goesmann A., Kraft T., Schulz B., Stadler P.F., Schmidt T., Gabaldón T., Lehrach H., Weisshaar B., Himmelbauer H. 2014. The genome of the recently domesticated crop plant sugar beet (Beta vulgaris). Nature. 505:546–549.

van Dongen S.M. 2000. Graph Clustering by Flow Simulation. PhD diss., Ultrecht University.

Doyle J.J. 1992. Gene Trees and Species Trees: Molecular Systematics as One-Character Taxonomy. Syst. Bot. 17:144.

Duchêne D.A., Bragg J.G., Duchêne S., Neaves L.E., Potter S., Moritz C., Johnson R.N., Ho S.Y.W., Eldridge M.D.B. 2018. Analysis of Phylogenomic Tree Space Resolves Relationships Among Marsupial Families. Syst. Biol. 67:400–412.

Durand E.Y., Patterson N., Reich D., Slatkin M. 2011. Testing for Ancient Admixture between Closely Related Populations. Mol. Biol. Evol. 28:2239–2252.

Eaton D.A.R., Ree R.H. 2013. Inferring Phylogeny and Introgression using RADseq Data: An Example from Flowering Plants (Pedicularis: Orobanchaceae). Syst. Biol. 62:689–706.

Edwards S.V. 2009. Is A New and General Theory of Molecular Systematics Emerging? Evolution. 63:1–19.

Edwards S.V., Xi Z., Janke A., Faircloth B.C., McCormack J.E., Glenn T.C., Zhong B., Wu S., Lemmon E.M., Lemmon A.R., Leaché A.D., Liu L., Davis C.C. 2016. Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. Mol. Phylogenet. Evol. 94:447–462.

Elworth R.A.L., Allen C., Benedict T., Dulworth P., Nakhleh L.K. 2018. DGEN: A Test Statistic for Detection of General Introgression Scenarios. WABI.

Elworth R.A.L., Ogilvie H.A., Zhu J., Nakhleh L. 2019. Advances in Computational Methods for Phylogenetic Networks in the Presence of Hybridization. In: Warnow T., editor. Bioinformatics and Phylogenetics: Seminal Contributions of Bernard Moret. Cham: Springer International Publishing. p. 317–360.

Erfan Sayyari, Siavash Mirarab. 2018. Testing for Polytomies in Phylogenetic Species Trees Using Quartet Frequencies. Genes. 9:132.

Flowers T.J., Colmer T.D. 2015. Plant salt tolerance: adaptations in halophytes. Ann. Bot. 115:327–331.

Folk R.A., Mandel J.R., Freudenstein J.V. 2017. Ancestral Gene Flow and Parallel Organellar Genome Capture Result in Extreme Phylogenomic Discord in a Lineage of Angiosperms. Syst. Biol. 66:320-337.

Foster P.G. 2004. Modeling Compositional Heterogeneity. Syst. Biol. 53:485–495.

Fu L., Niu B., Zhu Z., Wu S., Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 28:3150–3152.

Galtier N., Daubin V. 2008. Dealing with incongruence in phylogenomic analyses. Philos. Trans. R. Soc. B Biol. Sci. 363:4023–4029.

Gitzendanner M.A., Soltis P.S., Yi T.-S., Li D.-Z., Soltis D.E. 2018. Plastome Phylogenetics: 30 Years of Inferences Into Plant Evolution. Plastid Genome Evolution. Elsevier. p. 293–313.

Glémin S., Scornavacca C., Dainat J., Burgarella C., Viader V., Ardisson M., Sarah G., Santoni S., David J., Ranwez V. 2019. Pervasive hybridizations in the history of wheat relatives. Sci. Adv. 5:eaav9188.

Gonçalves D.J.P., Simpson B.B., Ortiz E.M., Shimizu G.H., Jansen R.K. 2019. Incongruence between gene trees and species trees and phylogenetic signal variation in plastid genes. Mol. Phylogenet. Evol. 138:219–232.

Green R.E., Krause J., Briggs A.W., Maricic T., Stenzel U., Kircher M., Patterson N., Li H., Zhai W., Fritz M.H.Y., Hansen N.F., Durand E.Y., Malaspinas A.S., Jensen J.D., Marques-Bonet T., Alkan C., Prufer K., Meyer M., Burbano H.A., Good J.M., Schultz R., Aximu-

Petri A., Butthof A., Hober B., Hoffner B., Siegemund M., Weihmann A., Nusbaum C., Lander E.S., Russ C., Novod N., Affourtit J., Egholm M., Verna C., Rudan P., Brajkovic D., Kucan Z., Gusic I., Doronichev V.B., Golovanova L.V., Lalueza-Fox C., de la Rasilla M., Fortea J., Rosas A., Schmitz R.W., Johnson P.L.F., Eichler E.E., Falush D., Birney E., Mullikin J.C., Slatkin M., Nielsen R., Kelso J., Lachmann M., Reich D., Paabo S. 2010. A Draft Sequence of the Neandertal Genome. Science. 328:710–722.

Haas B.J., Papanicolaou A., Yassour M., Grabherr M., Blood P.D., Bowden J., Couger M.B., Eccles D., Li B., Lieber M., MacManes M.D., Ott M., Orvis J., Pochet N., Strozzi F., Weeks N., Westerman R., William T., Dewey C.N., Henschel R., LeDuc R.D., Friedman N., Regev A. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8:1494–1512.

Hejase H.A., Liu K.J. 2016. A scalability study of phylogenetic network inference methods using empirical datasets and simulations involving a single reticulation. BMC Bioinformatics. 17:422.

Hejase H.A., VandePol N., Bonito G.M., Liu K.J. 2018. FastNet: Fast and Accurate Statistical Inference of Phylogenetic Networks Using Large-Scale Genomic Sequence Data. Comp. Genomics.:242–259.

Hernández-Ledesma P., Berendsohn W.G., Borsch T., Mering S.V., Akhani H., Arias S., Castañeda-Noa I., Eggli U., Eriksson R., Flores-Olvera H., Fuentes-Bazán S., Kadereit G., Klak C., Korotkova N., Nyffeler R., Ocampo G., Ochoterena H., Oxelman B., Rabeler R.K., Sanchez A., Schlumpberger B.O., Uotila P. 2015. A taxonomic backbone for the global synthesis of species diversity in the angiosperm order Caryophyllales. Willdenowia. 45:281.

Hoang D.T., Chernomor O. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol. Biol. Evol. 35:518–522.

Hohmann S., Kadereit J.W., Kadereit G. 2006. Understanding Mediterranean-Californian disjunctions: molecular evidence from Chenopodiaceae-Betoideae. TAXON. 55:67–78.

Holder M.T., Anderson J.A., Holloway A.K. 2001. Difficulties in Detecting Hybridization. Syst. Biol. 50:978–982.

Huang H., Knowles L.L. 2009. What Is the Danger of the Anomaly Zone for Empirical Phylogenetics? Syst. Biol. 58:527–536.

Huang W., Zhou G., Marchand M., Ash J.R., Morris D., Van Dooren P., Brown J.M., Gallivan K.A., Wilgenbusch J.C. 2016. TreeScaper: Visualizing and Extracting Phylogenetic Signal from Sets of Trees. Mol. Biol. Evol. 33:3314–3316.

Hughes L.C., Ortí G., Huang Y., Sun Y., Baldwin C.C., Thompson A.W., Arcila D., Betancur-R. R., Li C., Becker L., Bellora N., Zhao X., Li X., Wang M., Fang C., Xie B., Zhou Z., Huang H., Chen S., Venkatesh B., Shi Q. 2018. Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. Proc. Natl. Acad. Sci. 115:6249–6254.

Jarvis D.E., Ho Y.S., Lightfoot D.J., Schmöckel S.M., Li B., Borm T.J.A., Ohyanagi H., Mineta K., Michell C.T., Saber N., Kharbatia N.M., Rupper R.R., Sharp A.R., Dally N., Boughton B.A., Woo Y.H., Gao G., Schijlen E.G.W.M., Guo X., Momin A.A., Negrão S., Al-Babili S., Gehring C., Roessner U., Jung C., Murphy K., Arold S.T., Gojobori T., Linden C.G.V.D., van Loo E.N., Jellen E.N., Maughan P.J., Tester M. 2017. The genome of *Chenopodium quinoa*. Nature. 542:307–312.

Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y.W., Faircloth B.C., Nabholz B., Howard J.T., Suh A., Weber C.C., da Fonseca R.R., Li J., Zhang F., Li H., Zhou L., Narula N., Liu L., Ganapathy G., Boussau B., Bayzid Md.S., Zavidovych V., Subramanian S., Gabaldón T., Capella-Gutiérrez S., Huerta-Cepas J., Rekepalli B., Munch K., Schierup M., Lindow B., Warren W.C., Ray D., Green R.E., Bruford M.W., Zhan X., Dixon A., Li S., Li N., Huang Y., Derryberry E.P., Bertelsen M.F., Sheldon F.H., Brumfield R.T., Mello C.V., Lovell P.V., Wirthlin M., Schneider M.P.C., Prosdocimi F., Samaniego J.A., Velazquez A.M.V., Alfaro-Núñez A., Campos P.F., Petersen B., Sicheritz-Ponten T., Pas A., Bailey T., Scofield P., Bunce M., Lambert D.M., Zhou Q., Perelman P., Driskell A.C., Shapiro B., Xiong Z., Zeng Y., Liu S., Li Z., Liu B., Wu K., Xiao J., Yinqi X., Zheng Q., Zhang Y., Yang H., Wang J., Smeds L., Rheindt F.E., Braun M., Fjeldsa J., Orlando L., Barker F.K., Jønsson K.A., Johnson W., Koepfli K.-P., O'Brien S., Haussler D., Ryder O.A., Rahbek C., Willerslev E., Graves G.R., Glenn T.C., McCormack J., Burt D., Ellegren H., Alström P., Edwards S.V., Stamatakis A., Mindell D.P., Cracraft J., Braun E.L., Warnow T., Jun W., Gilbert M.T.P., Zhang G. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science. 346:1320.

Joly S., McLenachan P.A., Lockhart P.J. 2009. A Statistical Approach for Distinguishing Hybridization and Incomplete Lineage Sorting. Am. Nat. 174:E54–E70.

Kadereit G., Ackerly D., Pirie M.D. 2012. A broader model for C4 photosynthesis evolution in plants inferred from the goosefoot family (Chenopodiaceae s.s.). Proc. R. Soc. B Biol. Sci. 279:3304–3311.

Kadereit G., Borsch T., Weising K., Freitag H. 2003. Phylogeny of Amaranthaceae and Chenopodiaceae and the Evolution of C4 Photosynthesis. Int. J. Plant Sci. 164:959–986.

Kadereit G., Hohmann S., Kadereit J.W. 2006. A synopsis of Chenopodiaceae subfam. Betoideae and notes on the taxonomy of *Beta*. Willdenowia. 36:9–19.

Kadereit G., Newton R.J., Vandelook F. 2017. Evolutionary ecology of fast seed germination—A case study in Amaranthaceae/Chenopodiaceae. Perspect. Plant Ecol. Evol. Syst. 29:1–11.

Kalyaanamoorthy S., Minh B.Q., Wong T.K.F., von Haeseler A., Jermiin L.S. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods. 14:587–589.

Kamneva O.K., Rosenberg N.A. 2017. Simulation-Based Evaluation of Hybridization Network Reconstruction Methods in the Presence of Incomplete Lineage Sorting. Evol. Bioinforma. 13:117693431769193.

Katoh K., Standley D.M. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. 30:772–780.

Kearse M., Moir R., Wilson A., Stones-Havas S., Cheung M., Sturrock S., Buxton S., Cooper A., Markowitz S., Duran C., Thierer T., Ashton B., Meintjes P., Drummond A. 2012.

Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 28:1647–1649.

Knowles L.L., Huang H., Sukumaran J., Smith S.A. 2018. A matter of phylogenetic scale: Distinguishing incomplete lineage sorting from lateral gene transfer as the cause of gene tree discord in recent versus deep diversification histories. Am. J. Bot. 105:376–384.

Kubatko L.S., Chifman J. 2019. An invariants-based method for efficient identification of hybrid species from large-scale genomic data. BMC Evol. Biol. 19:112.

Lanfear R., Calcott B., Ho S.Y.W., Guindon S. 2012. PartitionFinder: Combined Selection of Partitioning Schemes and Substitution Models for Phylogenetic Analyses. Mol. Biol. Evol. 29:1695–1701.

Langmead B., Salzberg S.L. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods. 9:357–359.

Laumer C.E., Fernández R., Lemer S., Combosch D., Kocot K.M., Riesgo A., Andrade S.C.S., Sterrer W., Sørensen M.V., Giribet G. 2019. Revisiting metazoan phylogeny with genomic sampling of all phyla. Proc. R. Soc. B Biol. Sci. 286:20190831.

Lê S., Josse J., Husson F. 2008. FactoMineR : An R Package for Multivariate Analysis. J. Stat. Softw. 25: 1–18.

Lee-Yaw J.A., Grassa C.J., Joly S., Andrew R.L., Rieseberg L.H. 2019. An evaluation of alternative explanations for widespread cytonuclear discordance in annual sunflowers (*Helianthus*). New Phytol. 221:515–526.

Li B., Lopes J.S., Foster P.G., Embley T.M., Cox C.J. 2014. Compositional Biases among Synonymous Substitutions Cause Conflict between Gene and Protein Trees for Plastid Origins. Mol. Biol. Evol. 31:1697–1709.

Lightfoot D.J., Jarvis D.E., Ramaraj T., Lee R., Jellen E.N., Maughan P.J. 2017. Single-molecule sequencing and Hi-C-based proximity-guided assembly of amaranth (*Amaranthus hypochondriacus*) chromosomes provide insights into genome evolution. BMC Biol. 15:74.

Linkem C.W., Minin V.N., Leaché A.D. 2016. Detecting the Anomaly Zone in Species Trees and Evidence for a Misleading Signal in Higher-Level Skink Phylogeny (Squamata: Scincidae). Syst. Biol. 65:465–477.

Liu L., Yu L. 2010. Phybase: an R package for species tree analysis. Bioinformatics. 26:962–963.

Liu Y., Cox C.J., Wang W., Goffinet B. 2014. Mitochondrial Phylogenomics of Early Land Plants: Mitigating the Effects of Saturation, Compositional Heterogeneity, and Codon-Usage Bias. Syst. Biol. 63:862–878.

Maddison W.P. 1997. Gene Trees in Species Trees. Syst. Biol. 46:532–536.

Mai U., Mirarab S. 2018. TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. BMC Genomics. 19:4046.

Marcussen T., Sandve S.R., Heier L., Spannagl M., Pfeifer M., Jakobsen K.S., Wulff B.B.H., Steuernagel B., Mayer K.F.X., Olsen O.-A. 2014. Ancient hybridizations among the ancestral genomes of bread wheat. Science. 345:1250092.

Masson R., Kadereit G. 2013. Phylogeny of Polycnemoideae (Amaranthaceae): Implications for biogeography, character evolution and taxonomy. TAXON. 62:100–111.

Maureira-Butler I.J., Pfeil B.E., Muangprom A., Osborn T.C., Doyle J.J. 2008. The Reticulate History of *Medicago* (Fabaceae). Syst. Biol. 57:466–482.

Mclean B.S., Bell K.C., Allen J.M., Helgen K.M., Cook J.A. 2019. Impacts of Inference Method and Data set Filtering on Phylogenomic Resolution in a Rapid Radiation of Ground Squirrels (Xerinae: Marmotini). Syst. Biol. 68:298–316.

Meyer B.S., Matschiner M., Salzburger W. 2017. Disentangling Incomplete Lineage Sorting and Introgression to Refine Species-Tree Estimates for Lake Tanganyika Cichlid Fishes. Syst. Biol. 66:531–550.

Mirarab S., Bayzid M.S., Warnow T. 2016. Evaluating Summary Methods for Multilocus Species Tree Estimation in the Presence of Incomplete Lineage Sorting. Syst. Biol. 65:366–380.

Misof B., Liu S., Meusemann K., Peters R.S., Donath A., Mayer C., Frandsen P.B., Ware J., Flouri T., Beutel R.G., Niehuis O., Petersen M., Izquierdo-Carrasco F., Wappler T., Rust J., Aberer A.J., Aspock U., Aspock H., Bartel D., Blanke A., Berger S., Bohm A., Buckley T.R., Calcott B., Chen J., Friedrich F., Fukui M., Fujita M., Greve C., Grobe P., Gu S., Huang Y., Jermiin L.S., Kawahara A.Y., Krogmann L., Kubiak M., Lanfear R., Letsch H., Li Y., Li Z., Li J., Lu H., Machida R., Mashimo Y., Kapli P., McKenna D.D., Meng G., Nakagaki Y., Navarrete-Heredia J.L., Ott M., Ou Y., Pass G., Podsiadlowski L., Pohl H., von Reumont B.M., Schutte K., Sekiya K., Shimizu S., Slipinski A., Stamatakis A., Song W., Su X., Szucsich N.U., Tan M., Tan X., Tang M., Tang J., Timelthaler G., Tomizuka S., Trautwein M., Tong X., Uchifune T., Walzl M.G., Wiegmann B.M., Wilbrandt J., Wipfler B., Wong T.K.F., Wu Q., Wu G., Xie Y., Yang S., Yang Q., Yeates D.K., Yoshizawa K., Zhang Q., Zhang R., Zhang W., Zhang Y., Zhao J., Zhou C., Zhou L., Ziesmann T., Zou S., Li Y., Xu X., Zhang Y., Yang H., Wang J., Wang J., Kjer K.M., Zhou X. 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science. 346:763–767.

Morales-Briones D.F., Liston A., Tank D.C. 2018a. Phylogenomic analyses reveal a deep history of hybridization and polyploidy in the Neotropical genus *Lachemilla* (Rosaceae). New Phytol. 218:1668–1684.

Morales-Briones D.F., Romoleroux K., Kolář F., Tank D.C. 2018b. Phylogeny and Evolution of the Neotropical Radiation of *Lachemilla* (Rosaceae): Uncovering a History of Reticulate Evolution and Implications for Infrageneric Classification. Syst. Bot. 43:17–34.

Moray C., Goolsby E.W., Bromham L. 2016. The Phylogenetic Association Between Salt Tolerance and Heavy Metal Hyperaccumulation in Angiosperms. Evol. Biol. 43:119–130.

Mower J.P. 2009. The PREP suite: predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. Nucleic Acids Res. 37:W253–W259.

Müller K., Borsch T. 2005. Phylogenetics of Amaranthaceae Based on *matK/trnK* Sequence Data: Evidence from Parsimony, Likelihood, and Bayesian Analyses. Ann. Mo. Bot. Gard. 92:66–102.

Nguyen L.-T., Schmidt H.A., von Haeseler A., Minh B.Q. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol. Biol. Evol. 32:268–274.

Osuna-Mascaró C., Rubio de Casas R., Perfectti F. 2018. Comparative assessment shows the reliability of chloroplast genome assembly using RNA-seq. Sci. Rep. 8:17404.

Pamilo P., Nei M. 1988. Relationships between Gene Trees and Species Trees. Mol. Biol. Evol. 5:568–583.

Paradis E., Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 35:526–528.

Patro R., Duggal G., Love M.I., Irizarry R.A., Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods. 14:417–419.

Patterson N., Moorjani P., Luo Y., Mallick S., Rohland N., Zhan Y., Genschoreck T., Webster T., Reich D. 2012. Ancient Admixture in Human History. Genetics. 192:1065–1093.

Pease J.B., Brown J.W., Walker J.F., Hinchliff C.E., Smith S.A. 2018. Quartet Sampling distinguishes lack of support from conflicting support in the green plant tree of life. Am. J. Bot. 105:385–403.

Pease J.B., Hahn M.W. 2015. Detection and Polarization of Introgression in a Five-Taxon Phylogeny. Syst. Biol. 64:651–662.

Peden J. 1999. Analysis of Codon Usage. PhD diss., University of Nottingham.

Philippe H., Forterre P. 1999. The Rooting of the Universal Tree of Life Is Not Reliable. J. Mol. Evol. 49:509–523.

Piirainen M., Liebisch O., Kadereit G. 2017. Phylogeny, biogeography, systematics and taxonomy of Salicornioideae (Amaranthaceae/Chenopodiaceae) – A cosmopolitan, highly specialized hygrohalophyte lineage dating back to the Oligocene. Taxon. 66:109–132.

Prasanna A.N., Gerber D., Kijpornyongpan T., Aime M.C., Doyle V.P., Nagy L.G. 2019. Model Choice, Missing Data, and Taxon Sampling Impact Phylogenomic Inference of Deep Basidiomycota Relationships. syz029, https://doi.org/10.1093/sysbio/syz029

Pruitt K.D., Tatusova T., Maglott D.R. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35:D61–D65.

R Core Team. 2019. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Rannala B., Yang Z. 2003. Bayes Estimation of Species Divergence Times and Ancestral Population Sizes Using DNA Sequences From Multiple Loci. Genetics. 166:1645–1656.

Ranwez V., Douzery E.J.P., Cambon C., Chantret N., Delsuc F. 2018. MACSE v2: Toolkit for the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. Mol. Biol. Evol. 35:2582–2584.

Regier J.C., Shultz J.W., Zwick A., Hussey A., Ball B., Wetzer R., Martin J.W., Cunningham C.W. 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature. 463:1079–1083.

Rieseberg L.H., Soltis D.E. 1991. Phylogenetic consequences of cytoplasmic gene flow in plants. Evol. Trends Plants. 5:65–84.

Robinson D.F., Foulds L.R. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.

Rosenberg N.A. 2013. Discordance of Species Trees with Their Most Likely Gene Trees: A Unifying Principle. Mol. Biol. Evol. 30:2709–2713.

Roycroft E.J., Moussalli A., Rowe K.C. 2019. Phylogenomics Uncovers Confidence and Conflict in the Rapid Radiation of Australo-Papuan Rodents. Syst. Biol. syz044, https://doi.org/10.1093/sysbio/syz044

Salichos L., Stamatakis A., Rokas A. 2014. Novel Information Theory-Based Measures for Quantifying Incongruence among Phylogenetic Trees. Mol. Biol. Evol. 31:1261–1271.

Sang T., Crawford D.J., Stuessy T.F. 1995. Documentation of reticulate evolution in peonies (Paeonia) using internal transcribed spacer sequences of nuclear ribosomal DNA: implications for biogeography and concerted evolution. Proc. Natl. Acad. Sci. 92:6813–6817.

Sayyari E., Mirarab S. 2016. Fast Coalescent-Based Computation of Local Branch Support from Quartet Frequencies. Mol. Biol. Evol. 33:1654–1668.

Schliep K.P. 2011. phangorn: phylogenetic analysis in R. Bioinformatics. 27:592–593.

Schliesky S., Gowik U., Weber A.P.M., Bräutigam A. 2012. RNA-Seq Assembly – Are We There Yet? Front. Plant Sci. 3.

Schwarz G. 1978. Estimating the Dimension of a Model. Ann. Stat. 6:461–464.

Sharp P.M., Li W.-H. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. 24:28–38.

Shi C., Wang S., Xia E.-H., Jiang J.-J., Zeng F.-C., Gao L.-Z. 2016. Full transcription of the chloroplast genome in photosynthetic eukaryotes. Sci. Rep. 6:30135.

Shimodaira H. 2002. An Approximately Unbiased Test of Phylogenetic Tree Selection. Syst. Biol. 51:492–508.

Shimodaira H., Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics. 17:1246–1247.

Smith D.R. 2013. RNA-Seq data: a goldmine for organelle research. Brief. Funct. Genomics. 12:454–456.

Smith S.A., Moore M.J., Brown J.W., Yang Y. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. BMC Evol. Biol. 15:745.

Smith S.A., O'Meara B.C. 2012. treePL: divergence time estimation using penalized likelihood for large phylogenies. Bioinformatics. 28:2689–2690.

Smith-Unna R., Boursnell C., Patro R., Hibberd J.M., Kelly S. 2016. TransRate: reference-free quality assessment of de novo transcriptome assemblies. Genome Res. 26:1134–1144.

Solís-Lemus C., Ané C. 2016a. Inferring Phylogenetic Networks with Maximum Pseudolikelihood under Incomplete Lineage Sorting. PLOS Genet. 12:e1005896.

Soltis D.E., Kuzoff R.K. 1995. Discordance between nuclear and chloroplast phylogenies in the Heuchera group (Saxifragaceae). Evolution. 49:727–742.

Song L., Florea L. 2015. Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. GigaScience. 4:48.

Srivastava S.K. 1969. Assorted angiosperm pollen from the Edmonton Formation (Maestrichtian), Alberta, Canada. Can. J. Bot. 47:975–989.

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

Sugiura N. 1978. Further analysts of the data by akaike' s information criterion and the finite corrections. Commun. Stat. - Theory Methods. 7:13–26.

Swofford D. 2002. PAUP*. Phylogenetic analysis using parsimony (*and other methods) version 4. Sunderland MA Sinauer Assoc.

Than C., Ruths D., Nakhleh L. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinformatics. 9:322–16.

The Angiosperm Phylogeny Group, Chase M.W., Christenhusz M.J.M., Fay M.F., Byng J.W., Judd W.S., Soltis D.E., Mabberley D.J., Sennikov A.N., Soltis P.S., Stevens P.F. 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Bot. J. Linn. Soc. 181:1–20.

Varga T., Krizsán K., Földi C., Dima B., Sánchez-García M., Sánchez-Ramírez S., Szöllősi G.J., Szarkándi J.G., Papp V., Albert L., Andreopoulos W., Angelini C., Antonín V., Barry K.W., Bougher N.L., Buchanan P., Buyck B., Bense V., Catcheside P., Chovatia M., Cooper J., Dämon W., Desjardin D., Finy P., Geml J., Haridas S., Hughes K., Justo A., Karasiński D., Kautmanova I., Kiss B., Kocsubé S., Kotiranta H., LaButti K.M., Lechner B.E., Liimatainen K., Lipzen A., Lukács Z., Mihaltcheva S., Morgado L.N., Niskanen T., Noordeloos M.E., Ohm R.A., Ortiz-Santana B., Ovrebo C., Rácz N., Riley R., Savchenko A., Shiryaev A., Soop K., Spirin V., Szebenyi C., Tomšovský M., Tulloss R.E., Uehling J., Grigoriev I.V., Vágvölgyi C., Papp T., Martin F.M., Miettinen O., Hibbett D.S., Nagy L.G. 2019. Megaphylogeny resolves global patterns of mushroom evolution. Nat. Ecol. Evol. 3:668–678.

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 Phytol. 214:1736–1750.

Walker J.F., Walker-Hale N., Vargas O.M., Larson D.A., Stull G.W. 2019. Characterizing gene tree conflict in plastome-inferred phylogenies. PeerJ. 7:e7747.

Walker J.F., Yang Y., Feng T., Timoneda A., Mikenas J., Hutchison V., Edwards C., Wang N., Ahluwalia S., Olivieri J., Walker-Hale N., Majure L.C., Puente R., Kadereit G., Lauterbach M., Eggli U., Flores-Olvera H., Ochoterena H., Brockington S.F., Moore M.J., Smith S.A. 2018. From cacti to carnivores: Improved phylotranscriptomic sampling and hierarchical homology inference provide further insight into the evolution of Caryophyllales. Am. J. Bot. 105:446–462.

Wang Y., Tang H., DeBarry J.D., Tan X., Li J., Wang X., Lee T. -h., Jin H., Marler B., Guo H., Kissinger J.C., Paterson A.H. 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 40:e49–e49.

Wen D., Nakhleh L. 2018. Coestimating Reticulate Phylogenies and Gene Trees from Multilocus Sequence Data. Syst. Biol. 67:439–457.

Wen D., Yu Y., Hahn M.W., Nakhleh L. 2016a. Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. Mol. Ecol. 25:2361–2372.

Wen D., Yu Y., Nakhleh L. 2016b. Bayesian Inference of Reticulate Phylogenies under the Multispecies Network Coalescent. PLOS Genet. 12:e1006006.

Wickett N.J., Mirarab S., Nguyen N., Warnow T., Carpenter E., Matasci N., Ayyampalayam S., Barker M.S., Burleigh J.G., Gitzendanner M.A., Ruhfel B.R., Wafula E., Der J.P., Graham S.W., Mathews S., Melkonian M., Soltis D.E., Soltis P.S., Miles N.W., Rothfels C.J., Pokorny L., Shaw A.J., DeGironimo L., Stevenson D.W., Surek B., Villarreal J.C., Roure B., Philippe H., dePamphilis C.W., Chen T., Deyholos M.K., Baucom R.S., Kutchan T.M., Augustin M.M., Wang J., Zhang Y., Tian Z., Yan Z., Wu X., Sun X., Wong G.K.-S., Leebens-Mack J. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc. Natl. Acad. Sci. 111:E4859–E4868.

Widhelm T.J., Grewe F., Huang J.-P., Mercado-Díaz J.A., Goffinet B., Lücking R., Moncada B., Mason-Gamer R., Lumbsch H.T. 2019. Multiple historical processes obscure phylogenetic relationships in a taxonomically difficult group (Lobariaceae, Ascomycota). Sci. Rep. 9:8968.

Xu B., Yang Z. 2016. Challenges in Species Tree Estimation Under the Multispecies Coalescent Model. Genetics. 204:1353–1368.

Xu C., Jiao C., Sun H., Cai X., Wang X., Ge C., Zheng Y., Liu W., Sun X., Xu Y., Deng J., Zhang Z., Huang S., Dai S., Mou B., Wang Q., Fei Z., Wang Q. 2017. Draft genome of spinach and transcriptome diversity of 120 Spinacia accessions. Nat. Commun. 8:15275.

Yang Y., Moore M.J., Brockington S.F., Timoneda A., Feng T., Marx H.E., Walker J.F., Smith S.A. 2017. An Efficient Field and Laboratory Workflow for Plant Phylotranscriptomic Projects. Appl. Plant Sci. 5:1600128.

Yang Y., Smith S.A. 2013. Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. BMC Genomics. 14:328.

Yang Y., Smith S.A. 2014. Orthology Inference in Nonmodel Organisms Using Transcriptomes and Low-Coverage Genomes: Improving Accuracy and Matrix Occupancy for Phylogenomics. Mol. Biol. Evol. 31:3081–3092.

Yao G., Jin J.-J., Li H.-T., Yang J.-B., Mandala V.S., Croley M., Mostow R., Douglas N.A., Chase M.W., Christenhusz M.J.M., Soltis D.E., Soltis P.S., Smith S.A., Brockington S.F., Moore M.J., Yi T.-S., Li D.-Z. 2019. Plastid phylogenomic insights into the evolution of Caryophyllales. Mol. Phylogenet. Evol. 134:74–86.

Yu Y., Degnan J.H., Nakhleh L. 2012. The Probability of a Gene Tree Topology within a Phylogenetic Network with Applications to Hybridization Detection. PLoS Genet. 8:e1002660–10.

Yu Y., Dong J., Liu K.J., Nakhleh L. 2014. Maximum likelihood inference of reticulate evolutionary histories. Proc. Natl. Acad. Sci. 111:16448–16453.

Yu Y., Nakhleh L. 2015. A maximum pseudo-likelihood approach for phylogenetic networks. BMC Genomics. 16:S10.

Zhang C., Ogilvie H.A., Drummond A.J., Stadler T. 2018a. Bayesian Inference of Species Networks from Multilocus Sequence Data. Mol. Biol. Evol. 35:504–517.

Zhang C., Rabiee M., Sayyari E., Mirarab S. 2018b. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics. 19:523.

Zhao T., Schranz M.E. 2019. Network-based microsynteny analysis identifies major differences and genomic outliers in mammalian and angiosperm genomes. Proc. Natl. Acad. Sci. 116:2165–2174.

Zheng Y., Janke A. 2018. Gene flow analysis method, the D-statistic, is robust in a wide parameter space. BMC Bioinformatics. 19:10.

Zhu J., Liu X., Ogilvie H.A., Nakhleh L.K. 2019. A divide-and-conquer method for scalable phylogenetic network inference from multilocus data. Bioinformatics. 35:i370–i378.

Zhu J., Wen D., Yu Y., Meudt H.M., Nakhleh L. 2018. Bayesian inference of phylogenetic networks from bi-allelic genetic markers. PLOS Comput. Biol. 14:e1005932.

Zwick A., Regier J.C., Zwickl D.J. 2012. Resolving Discrepancy between Nucleotides and Amino Acids in Deep-Level Arthropod Phylogenomics: Differentiating Serine Codons in 21-Amino-Acid Models. PLoS ONE. 7:e47450.