

Phylogenomic analyses reveal a deep history of hybridization and polyploidy in the Neotropical genus *Lachemilla* (Rosaceae)

Diego F. Morales-Briones^{1,2,3,5} , Aaron Liston⁴ and David C. Tank^{1,2,3}

¹Department of Biological Sciences, University of Idaho, 875 Perimeter Drive MS 3051, Moscow, ID 83844-3051, USA; ²Institute for Bioinformatics and Evolutionary Studies, University of Idaho, 875 Perimeter Drive MS 3051, Moscow, ID 83844-3051, USA; ³Stillinger Herbarium, University of Idaho, 875 Perimeter Drive MS 3051, Moscow, ID 83844-3051, USA;

⁴Department of Botany and Plant Pathology, Oregon State University, 2082 Cordley Hall, Corvallis, OR 97331, USA; ⁵Present address: Department of Plant and Microbial Biology, College of Biological Sciences, University of Minnesota, 140 Gortner Laboratory, 1479 Gortner Avenue, Saint Paul, MN 55108, USA

Summary

Author for correspondence:

Diego F. Morales-Briones

Tel: + 1 208 596 5006

Email: dfmoralesb@gmail.com

Received: 30 November 2017

Accepted: 9 February 2018

New Phytologist (2018) **218**: 1668–1684

doi: 10.1111/nph.15099

Key words: cytonuclear discordance, gene flow, hybridization, introgression, *Lachemilla* (Rosaceae), phylogenetic networks, polyploidy, species trees.

- Hybridization, incomplete lineage sorting, and phylogenetic error produce similar incongruence patterns, representing a great challenge for phylogenetic reconstruction. Here, we use sequence capture data and multiple species tree and species network approaches to resolve the backbone phylogeny of the Neotropical genus *Lachemilla*, while distinguishing among sources of incongruence.
- We used 396 nuclear loci and nearly complete plastome sequences from 27 species to clarify the relationships among the major groups of *Lachemilla*, and explored multiple sources of conflict between gene trees and species trees inferred with a plurality of approaches.
- All phylogenetic methods recovered the four major groups previously proposed for *Lachemilla*, but species tree methods recovered different topologies for relationships between these four clades. Species network analyses revealed that one major clade, Orbiculate, is likely of ancient hybrid origin, representing one of the main sources of incongruence among the species trees. Additionally, we found evidence for a potential whole genome duplication event shared by *Lachemilla* and allied genera.
- *Lachemilla* shows clear evidence of ancient and recent hybridization throughout the evolutionary history of the group. Also, we show the necessity to use phylogenetic network approaches that can simultaneously accommodate incomplete lineage sorting and gene flow when studying groups that show patterns of reticulation.

Introduction

Hybridization is now recognized as a fundamental process in the evolution of animals, plants, and fungi (Giraud *et al.*, 2008; Schwenk *et al.*, 2008; Soltis & Soltis, 2009; Payseur & Rieseberg, 2016), but it seems to be particularly common in plants, where hybrid speciation, especially through polyploidy, is a well-established mechanism (Linder & Rieseberg, 2004; Mallet, 2007; Whitney *et al.*, 2010). Many plant species might be of direct hybrid origin or descended from a hybrid species in the recent past (Soltis & Soltis, 1995), and estimates reveal that 40–70% of all plant species are polyploids (Otto & Whitton, 2000), suggesting that hybridization may indeed be a common mechanism for spurring adaptive radiations in plants.

Reticulate processes often lead to incongruence between nuclear and plastid phylogenies and discordant phylogenetic histories between independent nuclear loci and/or alleles (Rieseberg & Soltis, 1991; Doyle, 1992; Wendel & Doyle, 1998; Linder & Rieseberg, 2004). Cytonuclear discordance has been widely detected in plants, and continues to be a good first approximation for the detection of reticulate evolution (e.g. Sang *et al.*,

1995; Soltis & Kuzoff, 1995; Fehrer *et al.*, 2007; Lundberg *et al.*, 2009; Pirie *et al.*, 2009; de Kuppler *et al.*, 2015; Scheunert & Heubl, 2017). That said, incongruence may also be the product of several other processes, the most frequent being phylogenetic error and incomplete lineage sorting (ILS) (Pamilo & Nei, 1988; Rieseberg & Soltis, 1991; Doyle, 1992; Maddison, 1997; Wendel & Doyle, 1998). Therefore, to establish hybridization as the main source of discordance, several approaches have been used to identify and/or quantify phylogenetic error (e.g. Reid *et al.*, 2012; Buddenhagen *et al.*, 2016; Arcila *et al.*, 2017), and to distinguish ILS from hybridization (e.g. Buckley *et al.*, 2006; Maureira-Butler *et al.*, 2008; Joly *et al.*, 2009; Konowalik *et al.*, 2015; Meyer *et al.*, 2017).

Several species tree methods that model ILS using multilocus sequence data have been implemented and are now widely used (reviewed in Edwards *et al.*, 2016; Mirarab *et al.*, 2016; Xu & Yang, 2016). Additionally, the evaluation and comparison in performance of species tree methods and traditional implementations, specifically concatenation, have received great attention (e.g. Warnow, 2015; Edwards *et al.*, 2016; Springer & Gatesy, 2016). Differences between inferred trees from species tree and

concatenation methods have been explained by the presence of ILS (Warnow, 2015) or gene tree estimation error (Springer & Gatesy, 2016). However, recent studies (Solís-Lemus *et al.*, 2016; Long & Kubatko, 2018) revealed that species tree and concatenation methods can also be inconsistent in the presence of gene flow. Moreover, Solís-Lemus *et al.* (2016) also showed that phylogenetic network methods perform better at finding the species tree when gene flow is present, indicating that approaches that can accommodate ILS and gene flow simultaneously should be applied when studying groups that show patterns of reticulation.

Recently, methods to estimate phylogenetic species networks from sequence data that incorporate gene-tree uncertainty and discordance due to ILS and gene flow have been developed (e.g. Yu *et al.*, 2014; Yu & Nakhleh, 2015; Solís-Lemus & Ané, 2016; Wen *et al.*, 2016a; Wen & Nakhleh, 2017; Zhang *et al.*, 2018; Zhu *et al.*, 2018). Although, these methods are still computationally intensive and limited to a small number of species and reticulation events (Hejase & Liu, 2016), their usage to detect patterns of reticulation is rapidly increasing (e.g. Wen *et al.*, 2016b; Copetti *et al.*, 2017; Crawl *et al.*, 2017; Meyer *et al.*, 2017).

The genus *Lachemilla* (Focke) Rydb. is a group of about 60 species that includes perennial rosette-forming herbs, stoloniferous herbs, trailing herbs, procumbent herbs, subshrubs, and dwarf shrubs (Romoleroux, 1996, 2004; Gaviria, 1997; Morales-Briones *et al.*, 2018a). *Lachemilla* is distributed between 2200 and 5000 m throughout the high mountains of the western American tropics from northern Mexico to northern Argentina and Chile (Gaviria, 1997; Romoleroux, 2004), and is especially common and diverse in the high-elevation ecosystems of the northern Andes, where the clade has undergone a rapid ecological radiation associated with the most recent Andean orogeny (Morales-Briones *et al.*, 2018a).

Previous phylogenetic analyses based on the internal transcribed spacer of the nuclear ribosomal DNA cistron and the chloroplast intergenic spacer *trnL-F* have identified clades within *Lachemilla* that correspond in part to traditional, morphologically defined sections (Gehrke *et al.*, 2008; Morales-Briones *et al.*, 2018a). Furthermore, Morales-Briones *et al.* (2018a) identified four well-supported lineages within *Lachemilla*. The Tripartite clade comprises ascending and procumbent herbs with tripartite leaves that often appear to have five divisions due to the bifid lateral segments of some species (Fig. 1a). The Verticillate clade includes subshrubs with erect or decumbent stems and reduced leaves that fuse with the stipules to form verticillate sheaths (Fig. 1b). The Orbiculate clade encompasses species with a stoloniferous habit and palmately lobed leaves (Fig. 1c). Finally, the Pinnate clade includes species with repent or decumbent stems and pinnate or bipinnatifid basal leaves (Fig. 1d). These clades are in part congruent with previous morphological classifications of the group (Perry, 1929; Rothmaler, 1937), but the relationships among them remain largely unresolved.

Lachemilla also shows widespread signs of hybridization and polyploidy. Recently, Morales-Briones *et al.* (2018a) used multiple sources of evidence, including patterns of cytonuclear discordance, detection of outliers, and phylogenetic network reconstruction (from multilabeled trees) to establish evidence of at least 24 potential hybrid species involving all four major lineages of *Lachemilla*. Moreover, several of those hybrid species have been identified as putative allopolyploids (Morales-Briones *et al.*, 2018a). Chromosome numbers in *Lachemilla* range from diploid (e.g. *Lachemilla mandoniana*: $2n=16$) to dodecaploid (e.g. *Lachemilla jaramilloi*: $2n=96$), with several species showing multiple ploidy levels (Morales-Briones *et al.*, 2018a).

Genomic data provide an excellent opportunity to detect hybridization (Twyford & Ennos, 2012; Payseur & Rieseberg,



Fig. 1 Representative species of the four major clades of *Lachemilla*: (a) Tripartite, *L. aphanoides*; (b) Verticillate, *L. nivalis*; (c) Orbiculate, *L. pectinata*; and (d) Pinnate, *L. pinnata*. Line drawings illustrate representative leaf morphologies of each major clade; illustrations modified from Romoleroux (1996).

2016); however, in groups where hybridization is widespread across the clade, phylogenetic data from multiple independent single- or low-copy nuclear genes is required (Pamilo & Nei, 1988; Sang & Zhang, 1999). Targeted sequence capture and extensions of these methods allow for the sequencing of hundreds of low-copy nuclear loci and high-copy genomic targets, like the chloroplast and/or mitochondrial genomes (Cronn *et al.*, 2012; Lemmon & Lemmon, 2013; Mandel *et al.*, 2014; Weitemier *et al.*, 2014; Folk *et al.*, 2015), and have been used in multiple groups of plants to resolve phylogenetic relationships (e.g. Stephens *et al.*, 2015; Heyduk *et al.*, 2016; Sass *et al.*, 2016; Moore *et al.*, 2017) and investigate patterns of hybridization (e.g. Grover *et al.*, 2015; Crowl *et al.*, 2017; Folk *et al.*, 2017; García *et al.*, 2017; Kamneva *et al.*, 2017; Mitchell *et al.*, 2017). In this paper, we use a phylogenomic dataset of 396 nuclear loci and complete plastomes assembled via targeted sequence capture to (1) estimate the phylogeny of *Lachemilla* with a focus on relationships among the major clades, (2) reexamine the source of incongruence between the plastid and nuclear phylogenies using genome-scale data, and (3) investigate the sources of discordance among gene trees and species trees. Using a plurality of phylogenetic approaches, we find clear evidence of both ancient and recent gene flow in the group, and demonstrate the necessity of simultaneously accommodating both ILS and gene flow when studying groups that show patterns of hybridization.

Materials and Methods

Taxon sampling

We sampled 29 individuals from 27 species of *Lachemilla* (Table 1), representing *c.* 50% of the total described diversity of the group, and most of the morphological variation within the four major clades of *Lachemilla* (Morales-Briones *et al.*, 2018a). Additionally, two species of *Alchemilla*, representing the Eurasian and African clades, one species of *Aphanes*, and one species of *Fragaria* were included as outgroups. Complete voucher information is listed in Supporting Information Table S1.

DNA extraction, hybrid enrichment, and sequencing

Total genomic DNA was isolated from fresh, silica-dried, or herbarium material using the DNeasy Plant Mini Kit (Qiagen, Valencia, CA, USA) following the manufacturer's protocol, or with a modified 2 × CTAB method (Doyle & Doyle, 1987). We used baits designed for *Fragaria* that target 257 putatively single-copy orthologous genes (1419 exons) identified in Rosaceae via a comparison of the apple (*Malus*), peach (*Prunus*), and strawberry (*Fragaria*) genomes (Kamneva *et al.*, 2017). Genomic DNAs were sheared by nebulization at 30 psi for 70 s, yielding an average shear size of 500 bp as measured by a Bioanalyzer High-Sensitivity Chip (Agilent Technologies Inc., Santa Clara, CA, USA). Libraries were constructed using the Illumina TruSeq library preparation kit (Illumina Inc. San Diego, CA, USA) and NEXTflex DNA barcodes (Bioo Scientific, Austin, TX, USA), standardized at 2 nM, and pooled in 16-plexes before hybrid

Table 1 Species of *Lachemilla* sampled in the current study (for complete voucher information see Supporting Information Table S1)

Taxon	Distribution	Major clade ^a	Hybrid ^b
<i>L. andina</i>	Colombia–Bolivia	Tripartite	No
<i>L. aphanooides</i>	Mexico–Bolivia	Tripartite	No
<i>L. barbata</i>	Peru–Bolivia	Pinnate	Yes
<i>L. diplophylla</i>	Ecuador–Argentina/Chile	Pinnate	No
<i>L. erodiifolia</i>	Colombia–Bolivia	Pinnate	No
<i>L. fulvescens</i>	Venezuela–Peru	—	Yes
<i>L. galioides</i>	Venezuela–Peru	Verticillate	No
<i>L. hirta</i>	Venezuela–Ecuador	Tripartite	Yes
<i>L. hispidula</i>	Colombia–Bolivia	Verticillate	No
<i>L. holosericea</i>	Colombia–Peru	—	Yes
<i>L. jamesonii</i>	Ecuador	Tripartite	No
<i>L. jaramilloi</i>	Colombia–Ecuador	Pinnate	Yes
<i>L. mandoniana</i>	Costa Rica–Bolivia	Pinnate	No
<i>L. nivalis</i>	Venezuela–Ecuador	Verticillate	No
<i>L. orbiculata</i>	Venezuela–Peru	Orbiculate	No
<i>L. pectinata</i>	Mexico–Bolivia	Orbiculate	No
<i>L. pinnata</i>	Mexico–Argentina	Pinnate	No
<i>L. polylepis</i>	Costa Rica–Colombia	Verticillate	Yes
<i>L. procumbens</i>	Mexico–Costa Rica	Tripartite	Yes
<i>L. pseudovenusta</i>	Peru	—	Yes
<i>L. rupestris</i>	Ecuador–Peru	—	Yes
<i>L. sprucei</i>	Colombia–Ecuador	Verticillate	No
<i>L. talamanquensis</i>	Costa Rica	—	Yes
<i>L. tanacetifolia</i>	Venezuela–Bolivia	Pinnate	No
<i>L. uniflora</i>	Colombia–Ecuador	—	Yes
<i>L. verticillata</i>	Costa Rica–Colombia	Verticillate	Yes
<i>L. vulcanica</i>	Mexico–Bolivia	Tripartite	No

^aSpecies without a designated major clade represent inter-clade hybrids/allopolyplids of two or more clades, as identified by Morales-Briones *et al.* (2018a).

^bSpecies previously identified as intra-clade hybrids/allopolyplids by Morales-Briones *et al.* (2018a).

enrichment. Library concentrations were determined using the KAPA qPCR kit (KK4835) (Kapa Biosystems, Woburn, MA, USA) on an ABI StepOnePlus Real-Time PCR System (Life Technologies, Grand Island, NY, USA). Solution-based hybridization with MYbaits biotinylated RNA baits (MYcroarray, Ann Arbor, MI, USA) and enrichment followed Weitemier *et al.* (2014). The target-enriched libraries were then sequenced on an Illumina HiSeq 2000 with 150 bp paired-end reads at the Genomics Core Facility at the University of Oregon.

Read processing and assembly

To remove sequencing adaptors and low-quality bases (Phred scores < 20), demultiplexed reads were cleaned with SEQCLEAN v.1.8.10 (<https://github.com/ibest/seqclean>) using default settings. Assemblies of nuclear loci were carried out with HYBPIPER v.1.1 (Johnson *et al.*, 2016) using *Fragaria vesca* L. exon sequences as references. Assembly of complete genes was attempted, but we identified several alignments that included chimeric gene sequences, likely the product of identifying multiple paralogs from many genes (see Exon assembly subsection); therefore, we assembled exons individually. Exons with an expected size ≥ 300 bp were assembled (400 exons from 225

genes), and paralog assessment was run for all samples and exons. Plastome assembly was carried out using ALIGNREADS v.2.5.2 (Straub *et al.*, 2011) in an iterative process. First, all samples were assembled using the *F. vesca* chloroplast genome as a reference (Genbank accession no. JF345175). Then, for all samples, a consensus sequence of this first assembly was used as reference for a second round of assembly for each sample. The resulting plastome assemblies were annotated using *F. vesca* and *Dasiphora fruticosa* (L.) Rydb. (Genbank accession KC507758) chloroplast genomes as references in GENEIOUS v.7.1.9 (Kearse *et al.*, 2012).

Nuclear data processing

HYBPIPER assemblies of nuclear loci resulted in multiple copies for most loci (Table S2). To choose the appropriate gene copy for downstream analyses, we aligned each exon with MAFFT v.7.037b (Katoh & Standley, 2013) using the automatic alignment strategy, and inferred gene trees for each alignment using FASTTREE2 (Price *et al.*, 2010) with the ‘-slow’, ‘-gtr’, and ‘-gamma’ options. Using the resulting gene trees and alignments, we used a tree-based approach implemented in PHYLOTREEPRUNER (Kocot *et al.*, 2013) to screen for evidence of paralogy. The maximal inclusive subtree with a minimum support of 0.75 and at least 25 taxa, where each taxon was represented by no more than one sequence, was selected. In cases where multiple sequences from the same taxon formed a clade, we retained the longest sequence. Detected paralogs are likely the product of a whole genome duplication event predating the diversification of *Lachemilla* (see the Discussion section), and as a result, 351 exons have multiple subtrees that met the pruning criteria. Because the output of PHYLOTREEPRUNER is a single alignment, we ran the paralog search for each exon after excluding sequences that were pruned in the first search. When multiple paralog alignments were obtained for an exon, we kept the alignment with the most number of taxa. In some cases, multiple subtree alignments had the same number of taxa, and therefore, we kept all alignments and treated each exon copy as an independent locus for downstream analyses. Occasionally, multiple exon copies that did not form single clades were recovered from some species (likely representing allopolyploid derived alleles; Morales-Briones *et al.*, 2018a), preventing PHYLOTREEPRUNER from pruning paralogs. In these cases, we randomly retained only one copy for those samples, and reran the paralog pruning search. Finally, individual loci were realigned, and ambiguously aligned positions were removed with GBLOCKS v.0.91b using default parameters (Castresana, 2000; Talavera & Castresana, 2007).

Phylogenetic analyses

We used concatenation and coalescent-based methods to reconstruct the phylogeny of *Lachemilla*. We performed the nuclear phylogenetic analyses on four datasets: the COMPLETE dataset that includes all sampled species, the HYBRID-REDUCED dataset, which excludes the hybrid species identified previously by Morales-Briones *et al.* (2018a), the ORBICULATE-REDUCED dataset that excludes the hybrid species and the

Orbiculate clade (identified as a hybrid clade; see the Results section), and the NO-RECOMBINATION dataset that excludes the hybrid species and loci that show evidence of recombination (see Assessment of recombination subsection). For each dataset, we first estimated phylogenetic relationships using the concatenated matrix using RAxML v.8.0.3 (Stamatakis, 2014) with a partition-by-locus scheme selected using PARTITIONFINDER v.2.1.1 (Lanfear *et al.*, 2017). All partitions used a GTR + G model, 100 searches for the best tree were performed, and clade support was assessed with 1000 bootstrap (BS) replicates. To estimate coalescent-based species trees, we used three different approaches. First, we used two summary statistic methods: ASTRAL-II (Mirarab & Warnow, 2015) and MP-EST (Liu *et al.*, 2010). Individual locus gene trees were estimated using RAxML with a GTR + G model, 10 searches for the best tree, and 100 BS replicates to assess clade support. Individual gene trees and BS replicates were used to estimate species trees in ASTRAL-II and MP-EST with 100 BS replicates. A third method, SVDquartets (Chifman & Kubatko, 2014; implemented in PAUP v.4.0a152, Swofford, 2002), which utilizes the full data to estimate the species trees, was used on the concatenated matrix with 100 BS replicates to assess clade support.

For the chloroplast phylogenetic analyses, complete plastome sequences (excluding one inverted repeat region) were aligned using the automatic alignment strategy in MAFFT. We used RAxML with a partition by coding and noncoding regions strategy, as selected using PARTITIONFINDER. All partitions used a GTR + G model, 100 searches for the best tree were performed, and clade support was assessed with 1000 BS replicates.

Concordance analyses

To explore discordance between gene tree and species tree estimates, we first calculated the internode certainty all (ICA), a measure that quantifies the degree of conflict on each node of a target tree (i.e. species tree estimates) given individual gene trees (Salichos *et al.*, 2014). In addition, we identified the number of conflicting and concordant bipartitions on the species trees. ICA values close to 1 indicate strong concordance in the bipartition of interest, while ICA values close to 0 indicate equal support for one or more conflicting bipartitions. Negative ICA values indicate that the bipartitions of interest conflict with one or more bipartitions that have a higher frequency, and ICA values close to -1 indicate the absence of concordance for the bipartition of interest (Salichos *et al.*, 2014). We calculated ICA and the number of conflicting/concordant bipartitions with PHYPARTS (Smith *et al.*, 2015), using the estimated species trees as the map tree and the individual gene trees with a BS support cutoff of 50%. We also summarized phylogenetic conflict across the genome using a Bayesian concordance analysis with BUCKY v.1.4.4 (Ané *et al.*, 2007; Larget *et al.*, 2010). First, we estimated posterior distributions of individual gene trees with MRBAYES v.3.2.6 (Ronquist *et al.*, 2012). Analyses consisted of two independent runs with four Markov chain Monte Carlo chains of 30 million generations each, sampling every 30 000th generation using a GTR + G model. Convergence of parameter estimates resulting from the

two independent Markov chain Monte Carlo runs was assessed using TRACER 1.6 (Rambaut *et al.*, 2014). Only loci that had reached convergence by 30 million generations, and which had complete taxon sampling (excluding *Aphanes australis*, to increase the final number of loci), were used for the Bayesian concordance analysis. BUCKY was run using the posterior distribution of gene trees after discarding 10% as burn-in, and multiple values of the *a priori* discordance parameter ($\alpha = 2, 20, 200, 2000$), to test for the impact of this parameter.

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 calculated the test for recombination, Φ (or pairwise homoplasy index; Bruen *et al.*, 2006), using PHIPACK (Bruen *et al.*, 2006) with the default sliding window size of 100 bp. Loci that showed a signal of recombination were removed (NO-RECOMBINATION dataset), and the concatenated and species tree inferences were rerun, as described earlier.

Assessment of hybridization

We used coalescent simulations similar to Folk *et al.* (2017) to test whether ILS alone could explain plastid and nuclear incongruence in the COMPLETE dataset. We simulated 10 000 plastid species trees under the coalescent with DENDROPY v.4.1.0 (Sukumaran & Holder, 2010) using the MP-EST species trees as a guide tree with branch lengths scaled by four to account for organellar inheritance. Clade frequency of the simulated plastid genes was obtained by summarizing them on the chloroplast tree (RAxML tree). In a scenario of ILS alone, we expect to find clades from the empirical plastid tree to be present in the simulated gene trees and have a high frequency. By contrast, in a scenario of hybridization, we expect clades to be unique to the empirical plastid tree and be absent or at low frequency in the simulated gene trees (García *et al.*, 2017).

Gene genealogy interrogation analysis

To distinguish incompatible signals regarding the relationship of the four major clades of *Lachemilla*, we used gene genealogy interrogation (GGI; Arcila *et al.*, 2017), a recently described method that discerns between estimation error and actual biological conflict explaining gene tree discordance. GGI identifies the best-supported hypothesis for each locus by enforcing monophyly of the clades of interest and performing constrained maximum likelihood searches for each hypothesis. Then, constrained gene trees are ranked based on their probabilities estimated using the approximately unbiased topology test (Shimodaira, 2002). We performed the GGI analyses using the HYBRID-REDUCED and NON-RECOMBINATION datasets, and tested the four possible topologies obtained from the nuclear species trees, chloroplast tree, and concordance analysis (see Results section).

Species network analysis

We inferred species networks that model ILS and gene flow using a maximum pseudo-likelihood approach (Yu & Nakhleh, 2015). Species network searches were carried out with PHYLONET v.3.6.1 (Than *et al.*, 2008) with the command 'InferNetwork_MPL' and using the individual gene trees from the HYBRID-REDUCED dataset as input. Networks searches were performed using only nodes in the gene trees that have BS support of at least 75%, allowing for up to three hybridization events and optimizing the branch lengths and inheritance probabilities of the returned species networks under the full likelihood. To estimate the best number of hybridizations and test whether the species network fits our gene trees better than a strictly bifurcating species tree, we computed the likelihood scores of the four tree topologies used in the GGI analyses, given the individual gene trees, as implemented in Yu *et al.* (2012), using the command 'CalGTPProb' in PHYLONET (again using individual gene trees and only nodes with BS support of at least 75%). Finally, we performed model selection using three information criteria – the Akaike information criterion (Akaike, 1973), the bias-corrected Akaike information criterion (Sugiura, 1978), and the Bayesian information criterion (Schwarz, 1978), – where number of parameters equals the number of branch lengths being estimated, plus the number of hybridization probabilities being estimated, and number of gene trees used to estimate the likelihood, to correct for finite sample size. We chose as the best model the one with the lowest information criterion value. The major tree (also referred to as the backbone tree), which is obtained by removing the reticulation branches with smaller inheritance probabilities from the networks (Solís-Lemus & Ané, 2016; Zhu *et al.*, 2016), was obtained from the best supported network using the function 'majorTree' in the phylogenetic network package PHYLONETWORKS (Solís-Lemus *et al.*, 2017).

Data accessibility

Raw Illumina data from sequence capture is available at the Sequence Read Archive (SRA) under accession SRP132080 (see Table S1 for individual sample SRA accession numbers). DNA alignments, phylogenetic trees and results from all analyses and datasets can be found in the Dryad data repository doi:10.5061/dryad.vj2s888 (Morales-Briones *et al.*, 2018b).

Results

Exon assembly

The assembly resulted in sequences of up to 392 exons (≥ 300 bp) per species (Table S2). HYBPIPER identified paralogous copies for up to 284 exons per species (Table S2). We found up to six paralogs per exon in *Lachemilla*, and identified that these represent complete gene copies (i.e. all or most species have all copies; Table S2; Fig. S11). After paralog pruning and removal of exons with poor coverage across samples (≤ 24 samples), we kept 333 exons from 196 different genes. Additionally, 63 of those exons

showed the presence of two (60) and three (three) paralogs copies that met the pruning requirements, giving us a total of 396 loci. The resulting concatenated matrix had an aligned length of 265 028 bp with 21 004 parsimony-informative sites, a minimum locus size of 277 bp, and a maximum locus size of 5739 bp. The chloroplast matrix (with one inverted repeat excluded) had an aligned length of 118 846 bp with 3733 parsimony-informative sites.

Nuclear phylogenetic analyses: COMPLETE dataset

All analyses recovered the four main clades of *Lachemilla* proposed by Morales-Briones *et al.* (2018a), but relationships among and within the four clades varied in each analysis (Table 2). The concatenated analysis supports ‘Topology 1’, where the Verticillate and Tripartite clades are monophyletic and sister to the clade formed by the Orbiculate and Pinnate clades (Fig. 2). ASTRAL-II, SVDquartets, and MP-EST analyses all recovered ‘Topology 2,’ where the Verticillate and Tripartite clades form a clade with the Orbiculate and Pinnate clades, representing successive sister groups (Fig. 3). With the exception of the concatenated analyses, most of the major clades and relationships within and among them were well supported ($BS \geq 75\%$). However, the concordance analyses and ICA scores revealed that most gene trees are actually in conflict with the species trees (Fig. 3). All Bayesian concordance analyses with different *a priori* discordance parameter resulted in identical results. BUCKY recovered ‘Topology 3’, where the Verticillate and Tripartite clades are again monophyletic, with the Pinnate clade and then Orbiculate clade as successive sister groups (Fig. 4a). The Bayesian concordance factors were low for most clades, suggesting a high degree of conflict among the gene trees (Fig. 4a; for simplified versions of all topologies, refer to Fig. 6).

Chloroplast phylogenetic analyses and evidence of hybridization: COMPLETE dataset

Phylogenetic analysis of the plastome dataset also recovered the four major clades in *Lachemilla*, but resulted in the fourth distinct topology with respect to relationships among these four lineages (‘Topology 4’), where the Verticillate clade is sister to a clade formed by the Orbiculate and Tripartite clades, with the Pinnate clade sister to all of them (Figs 2, 4b). Most of the

relationships are supported with $BS = 100$, but the level of discordance between the nuclear and chloroplast trees is high (Figs 2, 4b), with multiple species (e.g. *Lachemilla uniflora*, *Lachemilla verticillata*, *Lachemilla fulvescens*) located in different clades, and with different relationships among the four groups. Coalescent simulations under the organellar model did not produce gene trees that resembled the observed chloroplast tree. When the simulated plastid gene trees were summarized on the observed chloroplast tree, most clade frequencies were 0%, especially for the clades involving previously detected hybrid species and for the clades formed by the four major groups of *Lachemilla* (Fig. 4b). This clearly suggests that ILS alone cannot explain the high level of cytonuclear discordance observed in *Lachemilla*.

Nuclear phylogenetic analyses: HYBRID-REDUCED dataset

With previously identified hybrid species (Morales-Briones *et al.*, 2018a) removed, the concatenated and ASTRAL-II analyses both recovered ‘Topology 1’, but BS support for the sister group relationship of the Pinnate and Orbiculate clades was low (63% and 21% respectively; Figs 5a, S1). Although relationships within each major clade were identical in these analyses, the ASTRAL-II analysis recovered low support for relationships within the Verticillate clade (Fig. 5a). SVDquartets and MP-EST analyses both recovered ‘Topology 2’ with high BS for all clades (Figs 5b, S2). Relationships within the major clades were constant in both analyses, but the position of *Lachemilla sprucei* in the Verticillate clade and *Lachemilla tanacetifolia* in the Pinnate clade varied with respect to the concatenated and ASTRAL-II topologies (Fig. 5). Concordance analyses and ICA scores continue to reveal a high level of incongruence between individual gene trees and species tree estimations, even after the removal of the previously identified hybrids. As with the ASTRAL-II and concatenated analyses, BUCKY analyses of this dataset recovered ‘Topology 1’, and concordance factors remain low for most clades (Fig. S3).

Recombination analyses

The test for recombination, Φ , identified 131 loci with a strong signal of recombination for the HYBRID-REDUCED dataset ($P < 0.05$; Table S3). Concatenated, MP-EST, SVDquartets and BUCKY phylogenetic analyses of the NO-RECOMBINATION dataset (after removal of recombinant loci) recovered identical

Table 2 Datasets used in this study, indicating the number of taxa, number of loci, and topology (following Fig. 6) recovered for *Lachemilla* in each analysis

Dataset	Number of taxa	Number of loci/Topology recovered					
		Concatenation (RAXML)	ASTRAL-II	MP-EST	SVDquartets	BUCKY	Chloroplast (RAXML)
COMPLETE	29	396/1	396/2	396/2	396/2	208/3	1/4
HYBRID-REDUCED	15	396/1	396/1	396/2	396/2	219/1	1/4
NO-RECOMBINATION	15	265/1	265/2	265/2	265/4	160/1	NA/NA
ORBICULATE-REDUCED	13	396/NA	396/NA	396/NA	396/NA	221/NA	1/NA

NA, not applicable.

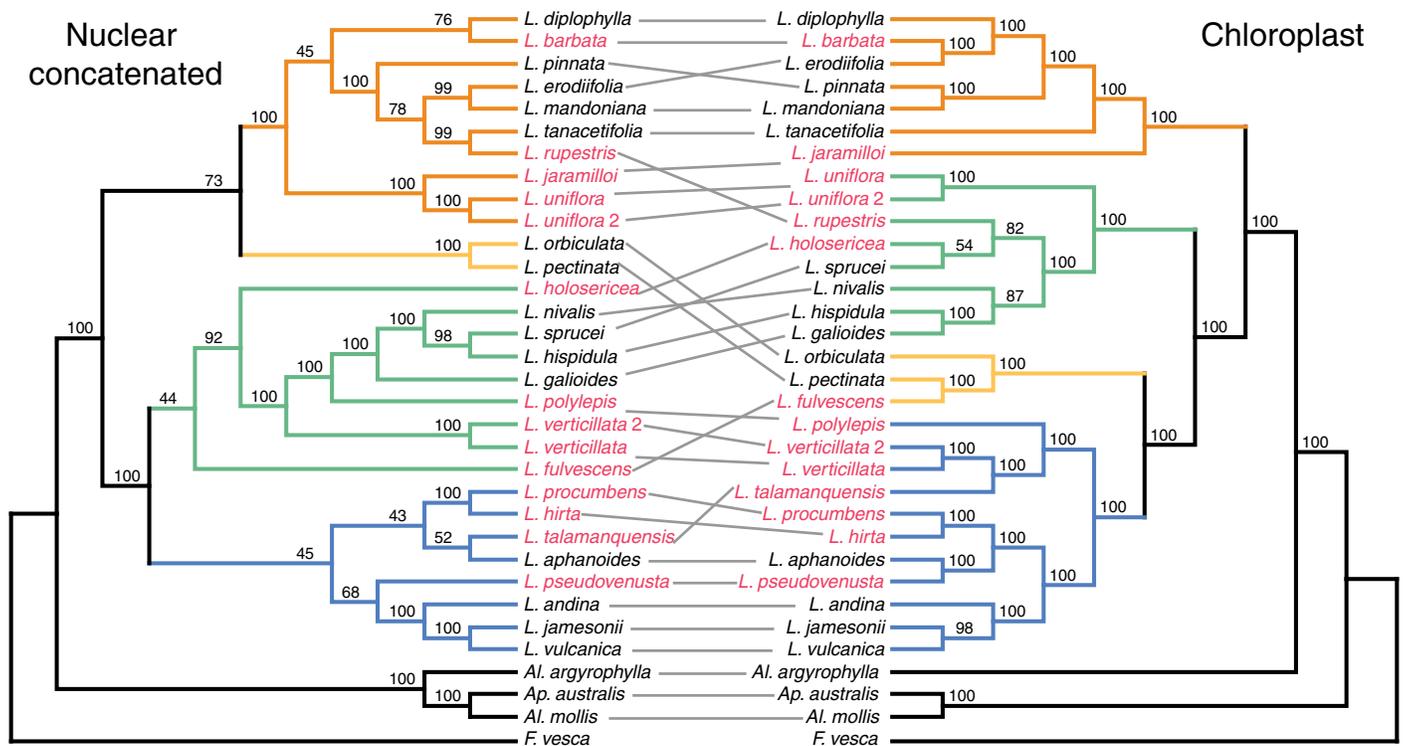


Fig. 2 Tanglegram of the nuclear concatenated (left) and chloroplast (right) phylogenies of the COMPLETE dataset. Gray lines connect taxa between the phylogenies. Maximum likelihood bootstrap support values are shown above branches. Branches are colored by major clades within *Lachemilla*: orange, Pinnate; yellow, Orbiculate; green, Verticillate; blue, Tripartite. Taxa previously identified as hybrids by Morales-Briones *et al.* (2018a) are highlighted in red.

topologies to the analyses of the HYBRID-REDUCED dataset with all loci included. As in the HYBRID-REDUCED dataset, ‘Topology 1’ was inferred for all analyses, with the exception of the ASTRAL-II analysis, where ‘Topology 2’ was recovered (Fig. S4).

Gene genealogy interrogation and network analysis: HYBRID-REDUCED dataset

The GGI analysis indicated the largest support for ‘Topology 3’, with 73 gene trees supporting this topology; however, this analysis also shows that the majority of gene trees do not provide significant support for any of the four alternative topologies ($P < 0.05$; Fig. 6; Table S4). GGI analysis of the NO-RECOMBINATION dataset showed similar results (Fig. S5; Table S5).

Species network analyses recovered topologies with up to three hybridization events. All networks recovered the four major clades of *Lachemilla* (Fig. S6), with the Orbiculate clade always identified as a reticulate node. All three information criteria indicated that the species networks with hybridization events involving the Orbiculate clade provided a better fit for our data than any of the four strictly bifurcating hypotheses (Table 3; Fig. 6). The network with two hybridization events (Fig. 7a) had the best support for the three information criteria. With this best species network, the first reticulation event involves *Lachemilla aphanoides* and rest of the Tripartite clade. The inheritance

probabilities show that the ancestral lineage of the clade formed by *Lachemilla andina*, *Lachemilla jamesonii*, and *Lachemilla vulcanica* has a genomic contribution of 38.8% from *L. aphanoides*. The second reticulation event also reveals ancestral gene flow in the Orbiculate clade. Inferred inheritance probabilities for this event indicate that the largest genomic contribution to the Orbiculate clade (86.3%) comes from an ancestral lineage of the Tripartite clade, and only a small portion (13.7%) comes from an ancestral or unsampled lineage within the Pinnate clade. The major tree obtained from the best supported network shows ‘Topology 4’.

Phylogenetic analyses: ORBICULATE-REDUCED dataset

After the removal of the Orbiculate clade, all phylogenetic analyses recovered the same well-supported topology with the Verticillate and Tripartite clades sister to each other, and the Pinnate clade sister to that clade (Figs 7b, S7–S10). Despite this consistent result, the levels of gene tree discordance with this topology were still high, especially with respect to relationships within the Verticillate and Pinnate clades.

Discussion

Our results show clear evidence of cytonuclear discordance and extensive conflict between individual gene trees and species trees in *Lachemilla*. Moreover, we established that these conflicts are

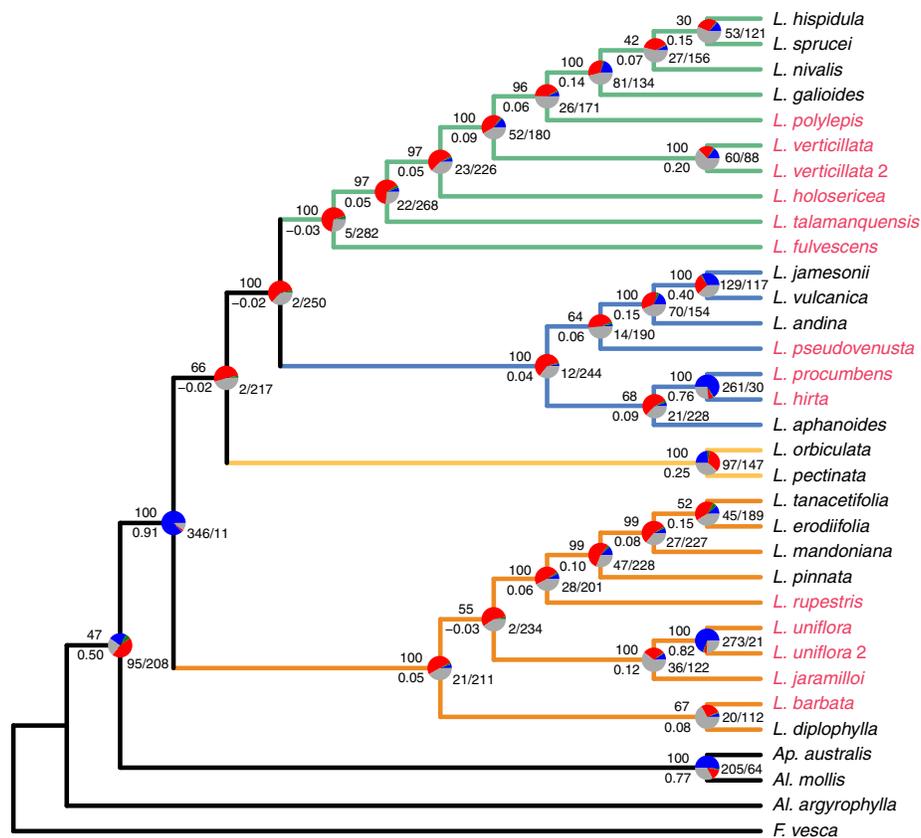


Fig. 3 Species trees of the COMPLETE dataset inferred with ASTRAL-II. Maximum likelihood bootstrap support values and internode certainty all scores are shown above and below main branches respectively. Pie charts next to the nodes present the proportion of gene trees that support that clade (blue), the proportion that support the main alternative for that clade (green), the proportion that support the remaining alternatives (red), and the proportion (conflict or support) that have < 50% bootstrap support (gray). Numbers next to pie charts indicate the number of gene trees concordant/conflicting with that node in the species tree. Branches are colored by major clades within *Lachemilla*: orange, Pinnate; yellow, Orbiculate; green, Verticillate; blue, Tripartite. Taxa previously identified as hybrids by Morales-Briones *et al.* (2018a) are highlighted in red.

the product of both ancient and recent hybridization throughout the evolutionary history of the group. We also established that the conflict between different species tree estimations is not a product of phylogenetic error, but rather the presence of ancestral gene flow. Specifically, using a phylogenetic network approach that can accommodate ILS and hybridization simultaneously, we determined that the Orbiculate clade, one of the four major lineages of *Lachemilla*, may be of ancient hybrid origin. Furthermore, we found evidence for a whole genome duplication event shared by *Lachemilla* and allied genera. These findings are discussed in detail below.

Cytonuclear discordance and evidence of hybridization

Evidence of extensive hybridization has been previously detected in *Lachemilla*, with at least 24 species identified as hybrids (Morales-Briones *et al.*, 2018a). The extensive analyses performed here revealed a similar pattern of cytonuclear discordance, where the hybrid species (Table 1) were recovered in different positions between the nuclear and chloroplast phylogenies, with some of these species (e.g. *L. fulvescens* and *Lachemilla talamanquensis*) having placements with very low support (Fig. 1). Additionally, the Bayesian concordance analysis and ICA scores revealed a large

amount of conflict between individual gene trees and the species tree estimates. Although these patterns may also be attributable to other processes, like ILS and phylogenetic error, our coalescent simulations showed that the observed cytonuclear discordance cannot be explained by ILS alone; furthermore, this is emerging as a common pattern in plant systems (e.g. Maureira-Butler *et al.*, 2008; Blanco-Pastor *et al.*, 2012; Reginato & Michelangeli, 2016; Folk *et al.*, 2017; García *et al.*, 2017; Vargas *et al.*, 2017).

Although removal of identified hybrid lineages reduces conflicting signals across gene trees, ICA values and concordance factors indicate that discordant signals are still persistent for some clades, suggesting that ILS and/or unidentified hybrid lineages continue to obscure our understanding of relationships in *Lachemilla*. For example, species like *Lachemilla diplophylla*, *L. sprucei*, and *L. tanacetifolia*, which have not previously been identified as hybrid taxa, show conflicting positions between species tree estimates and the chloroplast tree, suggesting that these species may also be of hybrid origin.

Additional work identifying parental lineages of putative hybrid species using allelic information from single-copy nuclear genes – for example, statistical phasing of alleles from sequence capture data and/or isolating individual alleles via molecular cloning and/or bioinformatically from high-throughput

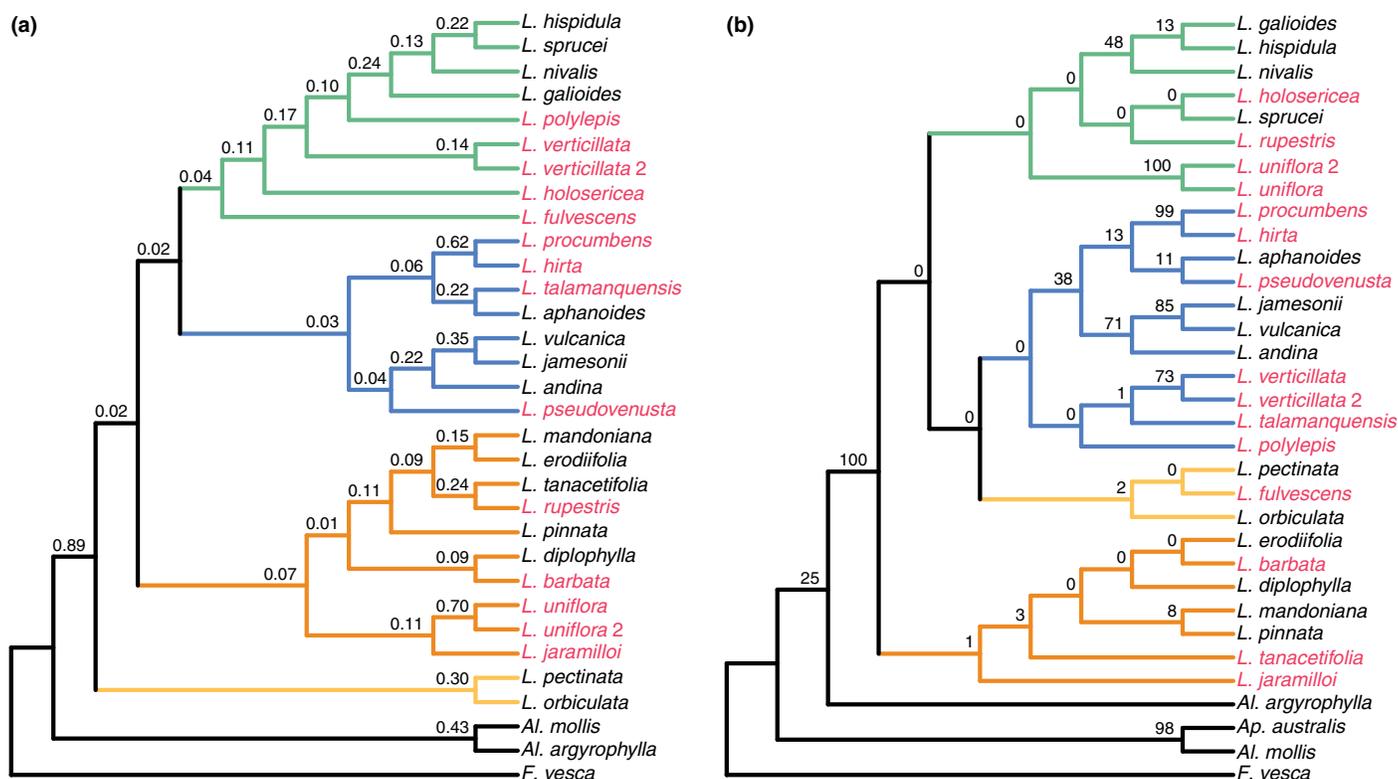


Fig. 4 (a) Concordance tree of the COMPLETE dataset; numbers above branches represent concordance factors. (b) Chloroplast phylogeny of the COMPLETE dataset; numbers above branches represent clade frequencies of the simulated gene trees. Branches are colored by major clades within *Lachemilla*: orange, Pinnate; yellow, Orbiculate; green, Verticillate; blue, Tripartite. Taxa previously identified as hybrids by Morales-Briones *et al.* (2018a) are highlighted in red.

amplicon datasets (e.g. Pyron *et al.*, 2016; Uribe-Convers *et al.*, 2016; Motazed *et al.*, 2017; Rothfels *et al.*, 2017; Blischak *et al.*, 2018) – remains to be done in *Lachemilla*. Kamneva *et al.* (2017) implemented a pipeline to assemble single-copy nuclear gene haplotypes from sequence capture data, but the presence of multiple gene copies in *Lachemilla* and relatives (Fig. S11, Table S2) makes this task a nontrivial problem that deserves further exploration.

Discordance among individual gene trees and species trees

Our analysis of concordance also reveals that a significant number of bipartitions on individual gene trees are not well supported, implying low phylogenetic information in the sampled loci. However, low support values can also be the product of the inclusion of hybrid lineages, and the removal of these taxa from our analyses does result in a general improvement of support measures (although a significant amount of weakly supported bipartitions is still recovered; Fig. 4). Our species tree analyses produced well-supported and congruent trees after the removal of hybrid taxa, suggesting that the low phylogenetic signal in the individual gene trees is not necessarily negatively affecting species tree estimation, as has been seen in other studies that use capture data (e.g. Blom *et al.*, 2017; Mitchell *et al.*, 2017).

Although the four main well-supported clades of *Lachemilla* have been previously recognized, relationships among these clades have remained largely unresolved (Morales-Briones *et al.*,

2018a). Our phylogenetic analyses recover the same four major lineages; however, depending on the dataset used and the phylogenetic approaches employed, these relationships vary considerably. Phylogenetic analyses of the COMPLETE dataset recovered four distinct topologies, and even after removal of previously identified hybrid species three of those topologies were consistently recovered (Table 2; Fig. 6). The major difference between these hypotheses is with respect to the placement of the Orbiculate clade that, with the exception of chloroplast tree, is associated with low concordance and support values, suggesting that the Orbiculate clade might be involved in a hybridization event.

Although recombination was detected for >30% of the loci, our analyses with these loci removed (NO-RECOMBINATION dataset) were largely the same as with them included. Some studies (e.g. Gatesy & Springer, 2013; Springer & Gatesy, 2016) argue that recombination might affect coalescent-based phylogenetic analyses, but simulation studies have shown that methods for species tree inference may be largely robust to intra-locus recombination (Lanier & Knowles, 2012; Wang & Liu, 2016), and a recent empirical study showed that, despite a large amount of recombinant loci (~42%), ASTRAL-II still recovered the same topology with these loci included or excluded from species tree analyses (Folk *et al.*, 2017). With respect to our ASTRAL-II analyses, the only difference with and without recombinant loci is again in the placement of the Orbiculate clade.

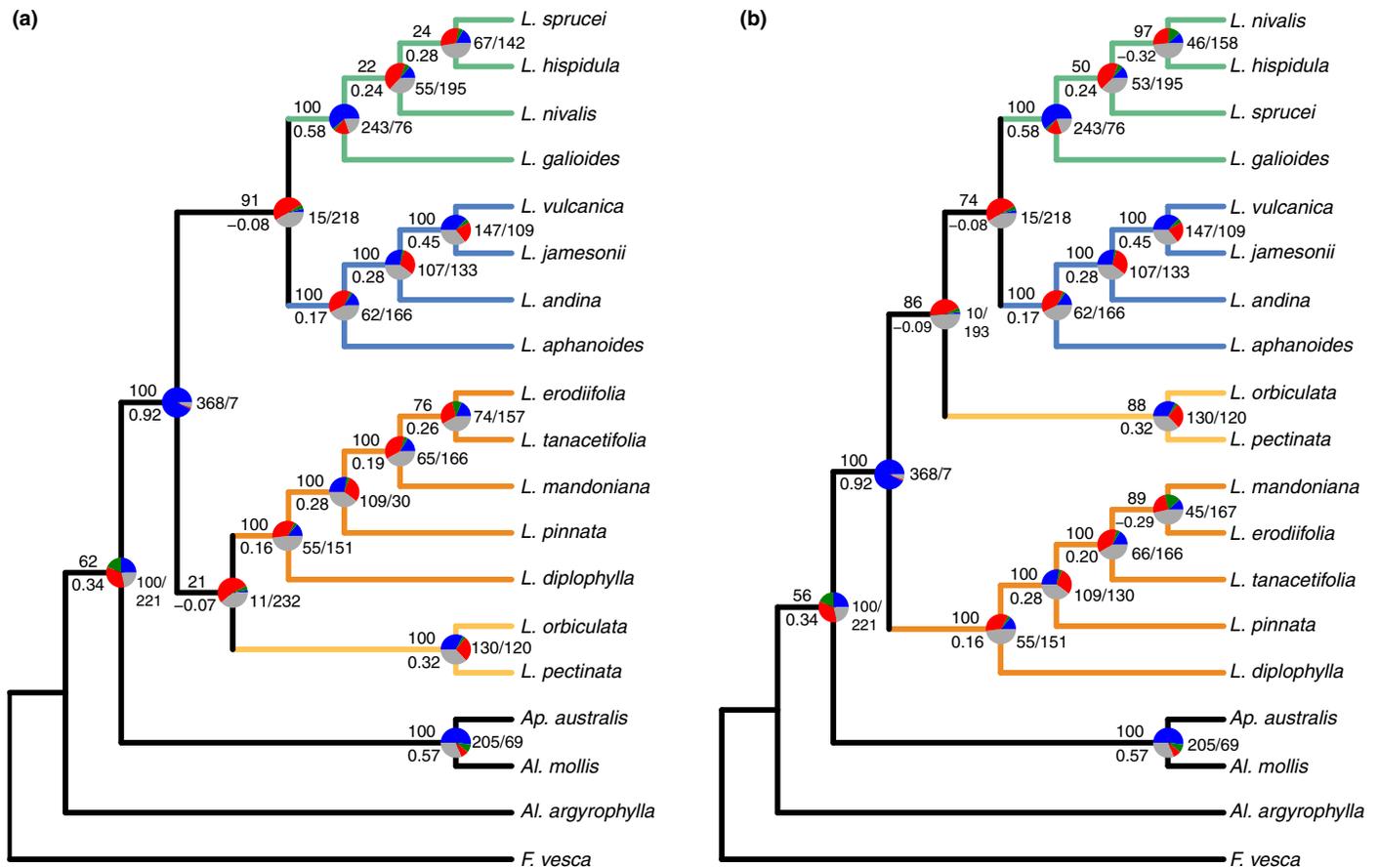


Fig. 5 Species tree topologies recovered for the HYBRID-REDUCED dataset. (a) Species trees inferred with ASTRAL-II. (b) Species trees inferred with SVDquartets. Maximum likelihood bootstrap support values and internode certainty all scores are shown above and below branches respectively. Pie charts next to the nodes represent the proportion of gene trees that support that clade (blue), the proportion that support the main alternative for that clade (green), the proportion that support the remaining alternatives (red), and the proportion (conflict or support) that have < 50% bootstrap support (gray). Numbers next to pie charts indicate the number of gene trees concordant/conflicting with that node in the species tree.

Phylogenetic networks model gene flow between populations, and this gene flow can be in the form of hybridization, introgression, or horizontal gene transfer. Although these processes are biologically different, phylogenetic networks model these reticulation processes in the same way, and do not distinguish between them (Solís-Lemus & Ané, 2016). Based on inheritance probabilities, Solís-Lemus *et al.* (2017) suggest that a small contribution (~0.10) from a parental population to a reticulate node may suggest introgression, as seen in species of the North American columnar cacti (Copetti *et al.*, 2017). On the other hand, inheritance probabilities close to 0.50 may suggest that the reticulate node is the product of hybrid speciation between the parental populations. Crowl *et al.* (2017), based on near-equal inheritance probabilities and genome size estimation, showed the hybrid (allopolyploid) origin of an octoploid lineage of *Campula erinus* L. (Campanulaceae). Our results show that the parental contributions to the reticulation events detected in *Lachemilla* (Fig. 7a) are unequal. Within the Tripartite clade, the inheritance contributions (0.388 and 0.612) support a hybridization event between *L. aphanoides* and the ancestral lineage of the rest of species of the Tripartite clade. The second reticulation event reveals that there has been extensive gene flow between the Orbiculate clade and

the Tripartite and the Pinnate clades. Given extensive history of hybridization and allopolyploidy in *Lachemilla* (Morales-Briones *et al.*, 2018a), we argue that the Orbiculate clade may be of hybrid origin between ancestral lineages of the Tripartite and the Pinnate clades. However, given the small inheritance contribution from the Pinnate clade (0.137), it is also plausible that ancestral gene flow from the Pinnate clade to the Orbiculate clade (or an ancestral lineage of this clade) could also produce this result (Fig. 7). It is also important to keep in mind, as noted by Solís-Lemus *et al.* (2017), that inheritance probabilities can be altered by many biological factors, and additional biological information is necessary for a robust interpretation of these values.

The varying placements of the Orbiculate clade when analyzing different datasets and/or using different approaches to estimate the species tree seems to be primarily the product of the inconsistency of species tree estimation in the presence of gene flow (Solís-Lemus *et al.*, 2016; Long & Kubatko, 2018). Our network analysis of the HYBRID-REDUCED dataset using PHYLONET revealed that all models involving reticulation events fit our data better than any model with strict bifurcating trees (Table 3). When the major tree (Fig. 7b), which displays the major vertical inheritance pattern in the data (Solís-Lemus *et al.*,

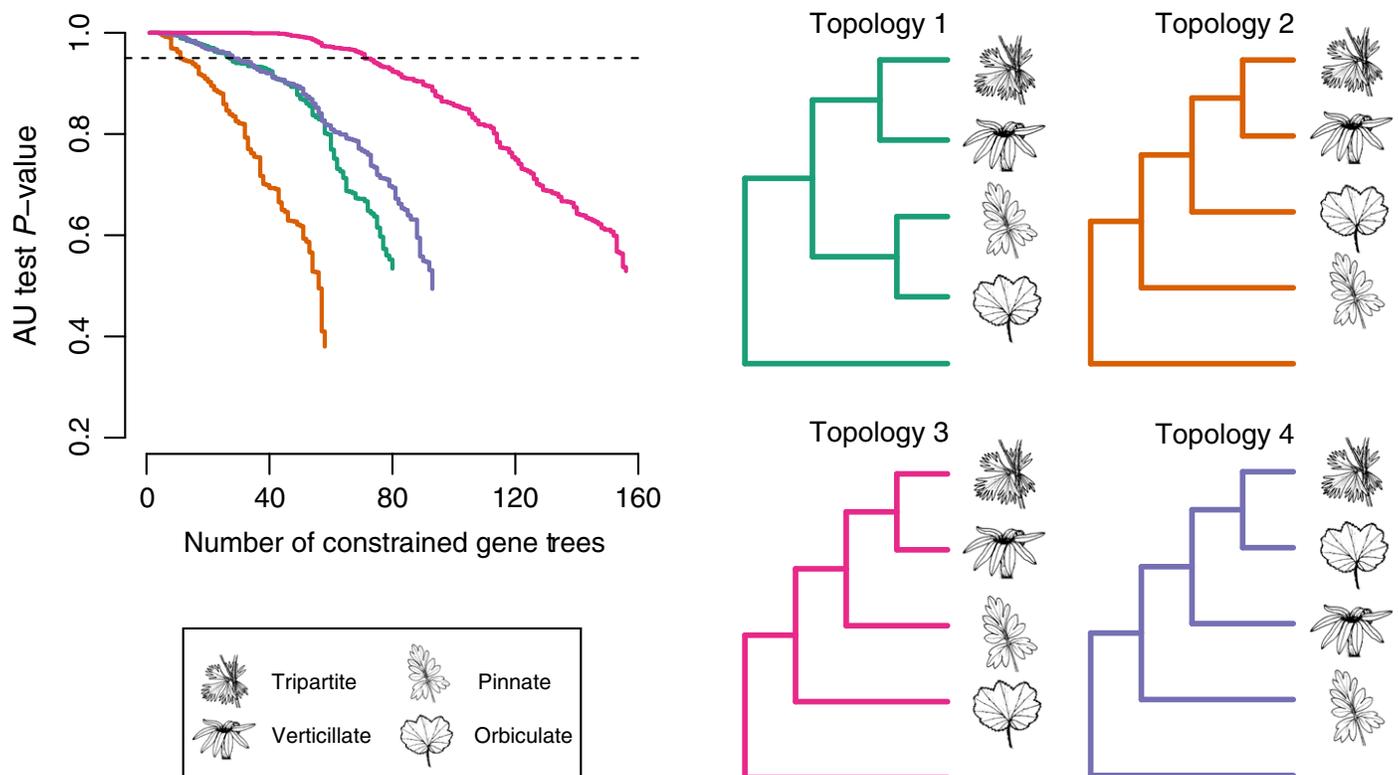


Fig. 6 Gene genealogy interrogation results testing the four topologies inferred for the four major clades of *Lachemilla*. Embedded plot represents the cumulative number of genes supporting each topology with highest probability, and their P -values from the approximately unbiased (AU) tests. Values above the dashed line indicate topologies that are significantly better than the alternatives ($P \leq 0.05$). Line drawings illustrate representative leaf morphologies of each major clade; illustrations modified from Romoleroux (1996).

Table 3 Model selection between the different species trees and species networks recovered for *Lachemilla*

Topology	$\text{Log}_e L$	Parameters	Loci	Number of hybridizations	Information criterion		
					AIC	AICc	BIC
Tree topology 1	-6147.752	35	222	NA	12365.504	12379.052	12484.598
Tree topology 2	-6156.017	35	222	NA	12382.035	12395.583	12501.128
Tree topology 3	-6148.437	35	222	NA	12366.874	12380.423	12485.968
Tree topology 4	-6262.415	35	222	NA	12594.831	12608.379	12713.924
Network 1	-6083.621	36	222	1	12239.243	12253.643	12361.738
Network 2	-6072.542	37	222	2	12219.084	12234.266	12344.983
Network 3	-6092.135	39	222	3	12262.269	12279.412	12394.974

The model with the lowest information criterion was selected as the best one (highlighted in bold). Topological hypotheses follow Fig. 6.

2016), is extracted from our best supported network, we can see that this tree displays 'Topology 4', indicating that the majority of the genome is congruent with the chloroplast tree, where the Orbiculate and Tripartite clades are sisters (Fig. 2). This is in direct conflict with our model selection results that show 'Topology 4' is the worst model (Table 3), indicating that designating a strictly bifurcating tree to *Lachemilla* might not be adequate. Furthermore, Zhu *et al.* (2016) found that, in the presence of deep coalescence, the most likely gene tree is not necessarily one of the backbone (major) trees inside the network. These empirical results corroborate simulation studies that have shown that phylogenetic species network methods that simultaneously model

discordance due to ILS and hybridization should be the preferred approach for investigating phylogenetic relationships in groups where gene flow is prominent (Solís-Lemus *et al.*, 2016).

Because of the large amount of conflict between gene trees, we also used GGI (Arcila *et al.*, 2017) to assess the potential for gene tree estimation error as the reason for the pattern of incongruence among species tree topologies. Although this method can be useful for distinguishing between estimation error and actual biological conflict in explaining gene tree discordance, as pointed out by Arcila *et al.* (2017), additional analyses are necessary to correctly interpret the signal of gene tree discordance when other processes like ILS or hybridization might also contribute to the

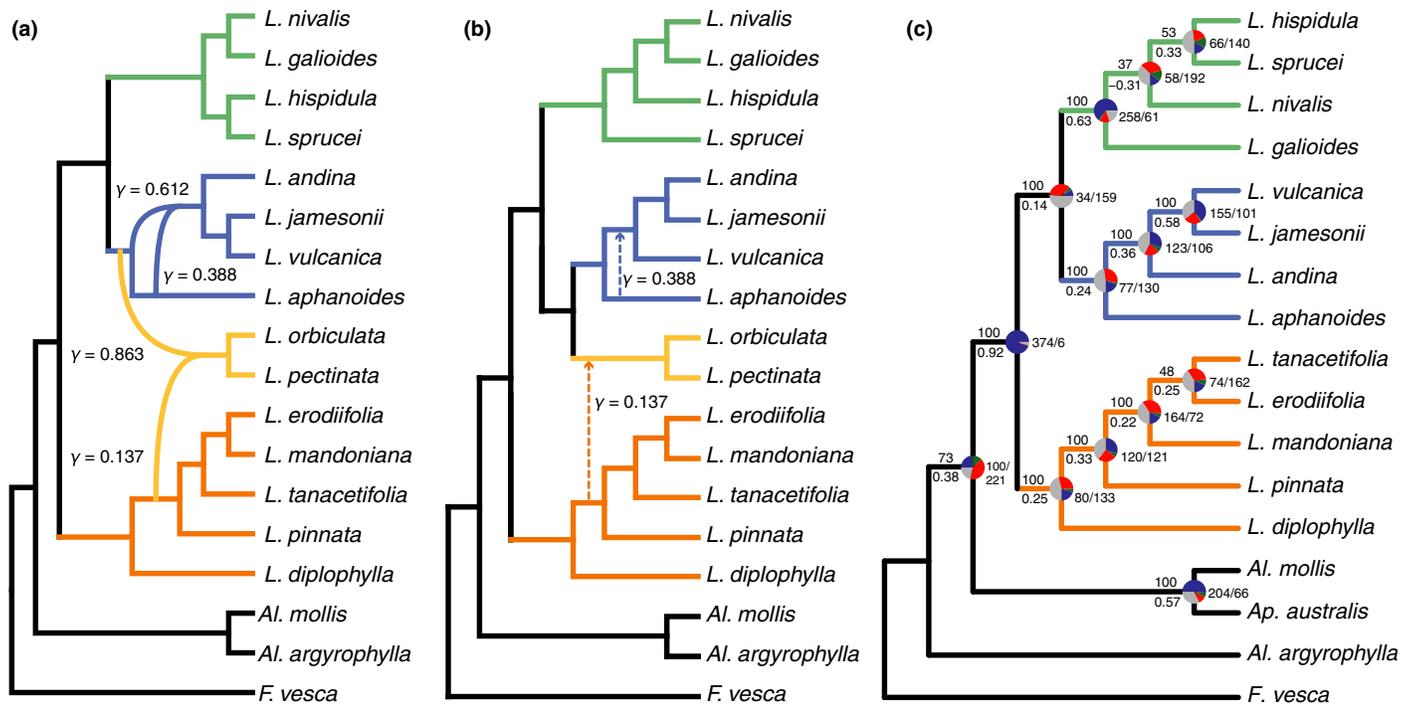


Fig. 7 (a) Best supported species network of the HYBRID-REDUCED dataset inferred with PHYLONET. Numbers next to the hybrid branches indicate inheritance probabilities. (b) Major tree obtained from the best supported species network. Dotted lines represent minor hybrid edges (edges with an inheritance contribution < 0.50). (c) Species tree of the ORBICULATE-REDUCED dataset inferred with ASTRAL-II. Maximum likelihood bootstrap support values and internode certainty all scores are shown above and below branches respectively. Pie charts next to the nodes represent the proportion of gene trees that support that clade (blue), the proportion that support the main alternative for that clade (green), the proportion that support the remaining alternatives (red), and the proportion (conflict or support) that have < 50% bootstrap support (gray). Numbers next to pie charts indicate the number of gene trees concordant/conflicting with that node in the species tree.

observed conflict. In our case, GGI selects ‘Topology 3’ as the hypothesis with the highest support from individual gene trees (Fig. 6), but it is likely that this topology was chosen over the alternative hypotheses, because by placing the Orbiculate clade sister to the rest of *Lachemilla*, it removes the source of conflict between the other three clades. This interpretation is corroborated by the convergence on the same topology by all phylogenetic methods using the dataset with Orbiculate clade removed (Fig. 7c).

Relationships among major clades of *Lachemilla* and systematic implications

Based mainly on foliar characters, Perry (1929) divided *Lachemilla* into six groups, and recent phylogenetic analyses recover four main clades of *Lachemilla* that have a partial correspondence with four of Perry’s groups (Morales-Briones *et al.*, 2018a). This partial correspondence is the product of the inclusion of a number of species, now recognized to be of hybrid origin from taxa in distinct groups that have incongruent positions in molecular phylogenies. The other two groups (both monotypic – *Lachemilla polylepis* and *L. diplophylla*) were found to be distinctive members of two of the major clades, where *L. polylepis* belongs to the Verticillate clade and *L. diplophylla* to the Pinnate clade, although in both cases these species have different overall morphologies when compared with the rest of the

clade. Although, these major clades were identified with strong support by Morales-Briones *et al.* (2018a), relationships between them remained unresolved, probably due to the limited amount of DNA sequence data used, as well as the hybrid origin of the Orbiculate clade identified here.

Our analyses strongly support the sister group relationship of the Verticillate and Tripartite clades (Fig. 7c). The Verticillate clade, mainly characterized by the highly modified leaf blades that fuse with the stipules to simulate a whorl of simple, elongate leaves, was considered by Perry (1929) as transition from the Tripartite clade, which has tripartite leaves that often appear to have five divisions due to the bifid lateral segments of some species and usually bifid, leaf-like stipules. Gaviria (1997) also recognized this leaf transition, although it is worth noting that some of the species used to identify this transition correspond to hybrid species between the two groups (Morales-Briones *et al.*, 2018a).

The Tripartite clade as defined by Perry (1929; series Aphanoides) was subdivided into six subgroups, where four of them are actually composed of only hybrid species between this group and the other three major clades, while the other two correspond to the Tripartite clade (Morales-Briones *et al.*, 2018a). Here, we identified an additional hybridization event between *L. aphanoides* and an extinct or unsampled lineage that led to a clade of three species. Interestingly, *L. aphanoides* belongs to one of the two Tripartite clades that is characterized by glomerulate inflorescences, while the other three species belong to the second

Tripartite clade that is characterized by loose inflorescences and pubescence in the inner part of the hypanthium; several hybridization events between species of these two clades within the Tripartite clade were also identified by Morales-Briones *et al.* (2018a), and Notov & Kusnetzova (2004) found the distinction of these two groups rather ambiguous, likely due to the promiscuity of members of this clade with respect to interspecific hybridization.

Perry (1929) interpreted the Orbiculate clade (series Orbiculateae), characterized by species with a stoloniferous habit and palmately lobed leaves, as most closely related to series Aphanoides, again probably due to the presence of numerous hybrid species between the Tripartite and Orbiculate clades. Here, we find evidence for the hybrid origin of the Orbiculate clade, with genomic contributions from taxa of the Pinnate and Tripartite clades (Fig. 7a). Multiple regional treatments (e.g. Rothmaler, 1935; Gaviria, 1997) have proposed infrageneric groups within *Lachemilla* that do not reflect phylogenetic relationships, and often several hybrid species and/or species belonging to the Orbiculate clade are used as transitional states for these groups. Thus, it is significant that our analyses have clarified the role that hybridization has played in the morphological complexity of *Lachemilla*, and especially in future taxonomic treatments of the clade.

Multiple gene copies and evidence of whole genome duplication

All loci targeted in this study appear to be single-copy genes in *Fragaria* and across Rosaceae (Kamneva *et al.*, 2017); however, our results show that > 70% of these loci have multiple copies in *Lachemilla*, *Alchemilla*, and *Aphanes* (Table S2; Fig. S11). A similar pattern of multicopy genes recovered from exon capture data has been reported in *Artocarpus* (Moraceae; Johnson *et al.*, 2016), which is known to have undergone at least one whole genome duplication (Gardner *et al.*, 2016). This suggests that the pattern detected in *Lachemilla*, *Alchemilla*, and *Aphanes* might also be the result of an ancient whole genome duplication that predates the diversification of the clade. While, there is not apparent doubling in chromosome number to support this whole genome duplication (as in *Artocarpus*; Gardner *et al.*, 2016), *Lachemilla*, *Alchemilla*, and *Aphanes* are the only members of subtribe Fragariinae that have a haploid chromosome number of eight instead of seven (Lundberg *et al.*, 2009). Although, this change could potentially be explained by dysploidy alone, there is evidence in other groups of Rosaceae that have undergone dysploidy following a whole genome duplication event (e.g. Evans & Campbell, 2002). Although, we do not have definitive evidence for this in *Lachemilla* and relative genera, it remains a plausible hypothesis, and more detailed studies of chromosome evolution in the clade are warranted. Moreover, in a recent transcriptome-based phylogenomic analysis of Rosaceae, Xiang *et al.* (2017) identified multiple whole genome duplication events across the family, and > 33% of genes used in their analyses showed evidence of duplication in the two species of *Alchemilla* sampled in their study. The precise phylogenetic position of this putative duplication

remains unresolved until additional members of subtribe Fragariinae (including *Alchemilla* and *Aphanes*) are sampled, and statistical methods to detect whole genome duplications are applied (e.g. Jiao *et al.*, 2011; Rabier *et al.*, 2014; Huang *et al.*, 2016; Tiley *et al.*, 2016).

Conclusions

Gene flow, in the form of hybridization and introgression, is a common pattern, and has played a fundamental role in the evolution of animals and plants (Soltis & Soltis, 2009; Mallet *et al.*, 2016; Payseur & Rieseberg, 2016). However, when investigating the evolutionary history of species, typically, strictly bifurcating species tree methods that account only for ILS are applied, and the potential impact of gene flow is not taken in account during the inference process. Moreover, mounting evidence that species tree methods are inconsistent in the presence of gene flow (Solís-Lemus *et al.*, 2016; Long & Kubatko, 2018) demonstrates the need to incorporate methods that account for ILS and gene flow simultaneously in phylogenetic studies. Here, we present a clear example of the utility of these methods to clarify the evolutionary history of *Lachemilla*. Our results provide strong evidence that both ancient and recent hybridization events have shaped the evolutionary history of this group. Reticulation, in addition to ILS, has resulted in extensive gene tree discordance, and has obscured phylogenetic inference in this group. Furthermore, discordance among species tree estimations in *Lachemilla*, due to gene flow, demonstrates the need for phylogenetic network approaches when studying groups that show patterns of reticulation. The recent explosion of new methods to estimate phylogenetic species networks (e.g. Yu *et al.*, 2014; Yu & Nakhleh, 2015; Solís-Lemus & Ané, 2016; Wen *et al.*, 2016a; Wen & Nakhleh, 2017; Zhang *et al.*, 2018; Zhu *et al.*, 2018) will facilitate more comprehensive studies of reticulation in groups like *Lachemilla*. Moreover, with the emergence of approaches for performing phylogenetic comparative methods on networks (Jhwueng & O'Meara, 2015; Bastide *et al.*, 2017), we hope the results presented here will help us to investigate broad questions regarding trait evolution, biogeography, and diversification dynamics in *Lachemilla*, as well as an evolutionarily informed classification system that reflects the complex (reticulate) history of the group.

Acknowledgements

We thank K. Romoleroux for access to DNA samples and illustrations of *Lachemilla*, K. Wieteimer for assistance with the sequence capture protocol, M. Johnson for assistance with HYBPIPER, and Ya Yang and three anonymous reviewers for comments on earlier versions of the manuscript. This work was funded in part by a Secretaría de Educación Superior, Ciencia, Tecnología e Innovación del Ecuador (SENESCYT) doctoral scholarship to D.F.M.-B., Graduate Student Research Grants from the Botanical Society of America, American Society of Plant Taxonomists, International Association of Plant Taxonomists, and the University of Idaho Stillinger Herbarium Expedition Funds to D.F.M.-B., and a National Science Foundation

Doctoral Dissertation Improvement Grant to D.C.T. for D.F.M-B. (DEB-1502049). Access to genomic and computational resources was granted through the University of Idaho Institute for Bioinformatics and Evolutionary Studies (IBEST) supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health (P30 GM103324).

Author contributions

D.F.M-B., A.L., and D.C.T. conceived the study. D.F.M-B. performed the research and analyzed the data. D.F.M-B. wrote the manuscript with main contributions from D.C.T. and comments from A.L.

ORCID

Diego F. Morales-Briones  <http://orcid.org/0000-0003-1535-5739>

References

- Akaike H. 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F, eds. *Second international symposium on information theory*. Budapest, Hungary: Akademiai Kiado, 267–281.
- Ané C, Larget B, Baum DA, Smith SD, Rokas A. 2007. Bayesian estimation of concordance among gene trees. *Molecular Biology and Evolution* 24: 412–426.
- Arcila D, Ortí G, Vari R, Armbruster JW, Stiassny MLJ, Ko KD, Sabaj MH, Lundberg J, Revell LJ, Betancur-R R. 2017. Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. *Nature Ecology and Evolution* 1: 0020.
- Bastide P, Solis-Lemus C, Kriebel R, Sparks KW, Ané C. 2017. Phylogenetic comparative methods on phylogenetic networks with reticulations. *bioRxiv* 194050.
- Blanco-Pastor JL, Vargas P, Pfeil BE. 2012. Coalescent simulations reveal hybridization and incomplete lineage sorting in Mediterranean *Linaria*. *PLoS ONE* 7: e39089.
- Blishchak PD, Latvis M, Morales-Briones DF, Johnson JC, Di Stilio VS, Wolfe AD, Tank DC. 2018. Fluidigm2PURC: automated processing and haplotype inference for double-barcoded PCR amplicons. *bioRxiv* 242677.
- Blom MPK, Bragg JG, Potter S, Moritz C. 2017. Accounting for uncertainty in gene tree estimation: summary-coalescent species tree inference in a challenging radiation of Australian lizards. *Systematic Biology* 66: 352–366.
- Bruen TC, Philippe H, Bryant D. 2006. A simple and robust statistical test for detecting the presence of recombination. *Genetics* 172: 2665–2681.
- Buckley T, Cordeiro M, Marshall D, Simon C. 2006. Differentiating between hypotheses of lineage sorting and introgression in New Zealand alpine cicadas (*Maoricicada* Dugdale). *Systematic Biology* 55: 411–425.
- Buddenhagen C, Lemmon AR, Lemmon EM, Bruhl J, Cappa J, Clement WL, Donoghue M, Edwards EJ, Hipp AL, Kortyna M *et al.* 2016. Anchored phylogenomics of angiosperms I: assessing the robustness of phylogenetic estimates. *bioRxiv* 086298.
- Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution* 17: 540–552.
- Chifman J, Kubatko L. 2014. Quartet inference from SNP data under the coalescent model. *Bioinformatics* 30: 3317–3324.
- Copetti D, Búrquez A, Bustamante E, Charboneau JLM, Childs KL, Eguarte LE, Lee S, Liu TL, McMahon MM, Whiteman NK *et al.* 2017. Extensive gene tree discordance and hemiplasy shaped the genomes of North American columnar cacti. *Proceedings of the National Academy of Sciences, USA* 114: 12003–12008.
- Cronn R, Knaus BJ, Liston A, Maughan PJ, Parks M, Syring JV, Udall J. 2012. Targeted enrichment strategies for next-generation plant biology. *American Journal of Botany* 99: 291–311.
- Crowl C, Myers C, Cellinese N. 2017. Embracing discordance: phylogenomic analyses provide evidence for allopolyploidy leading to cryptic diversity in a Mediterranean *Campanula* (Campanulaceae) clade. *Evolution* 71: 913–922.
- Doyle JJ. 1992. Gene trees and species trees: molecular systematics as one-character taxonomy. *Systematic Botany* 17: 144–163.
- Doyle JJ, Doyle JL. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* 19: 11–15.
- Edwards SV, Xi Z, Janke A, Faircloth BC, McCormack JE, Glenn TC, Zhong B, Wu S, Lemmon EM, Lemmon AR *et al.* 2016. Implementing and testing the multispecies coalescent model: a valuable paradigm for phylogenomics. *Molecular Phylogenetics and Evolution* 94: 447–462.
- Evans RC, Campbell CS. 2002. The origin of the apple subfamily (Maloideae; Rosaceae) is clarified by DNA sequence data from duplicated GBSSI genes. *American Journal of Botany* 89: 1478–1484.
- Fehrer J, Gemeinholzer B, Chrtek J, Bräutigam S. 2007. Incongruent plastid and nuclear DNA phylogenies reveal ancient intergeneric hybridization in *Pilosella* hawkweeds (*Hieracium*, Cichorieae, Asteraceae). *Molecular Phylogenetics and Evolution* 42: 347–361.
- Folk RA, Mandel JR, Freudenstein JV. 2015. Targeted enrichment of intronic sequence markers for recent radiations: a phylogenomic example from *Heuchera* (Saxifragaceae). *Applications in Plant Sciences* 3: 1500039.
- Folk RA, Mandel JR, Freudenstein JV. 2017. Ancestral gene flow and parallel organellar genome capture result in extreme phylogenomic discord in a lineage of angiosperms. *Systematic Biology* 66: 320–337.
- García N, Folk RA, Meerow AW, Chamala S, Gitzendanner MS, de Oliveira RS, Soltis DE, Soltis PS. 2017. Deep reticulation and incomplete lineage sorting obscure the diploid phylogeny of rain-lilies and allies (Amaryllidaceae tribe Hippeastreae). *Molecular Phylogenetics and Evolution* 111: 231–247.
- Gardner EM, Johnson MG, Ragone D, Wickett NJ, Zerega NJC. 2016. Low-coverage, whole-genome sequencing of *Artocarpus camansi* (Moraceae) for phylogenetic marker development and gene discovery. *Applications in Plant Sciences* 4: 1600017.
- Gatesy J, Springer MS. 2013. Concatenation versus coalescence versus “concordance”. *Proceedings of the National Academy of Sciences, USA* 110: E1179.
- Gaviria J. 1997. Sinopsis del género *Lachemilla* (Focke) Rydberg (Rosaceae) para Venezuela. *Plántula* 1: 189–212.
- Gehrke B, Bräuchler C, Romoleroux K, Lundberg M, Heubl G, Eriksson T. 2008. Molecular phylogenetics of *Alchemilla*, *Aphanes* and *Lachemilla* (Rosaceae) inferred from plastid and nuclear intron and spacer DNA sequences, with comments on generic classification. *Molecular Phylogenetics and Evolution* 47: 1030–1044.
- Giraud T, Refrégier G, Le Gac M, de Vienne DM, Hood ME. 2008. Speciation in fungi. *Fungal Genetics and Biology* 45: 791–802.
- Grover CE, Gallagher JP, Jareczek JJ, Page JT, Udall JA, Gore MA, Wendel JF. 2015. Re-evaluating the phylogeny of allopolyploid *Gossypium* L. *Molecular Phylogenetics and Evolution* 92: 45–52.
- Hejase HA, Liu KJ. 2016. A scalability study of phylogenetic network inference methods using empirical datasets and simulations involving a single reticulation. *BMC Bioinformatics* 17: 422.
- Heyduk K, Trapnell DW, Barrett CF, Leebens-Mack L. 2016. Phylogenomic analyses of species relationships in the genus *Sabal* (Arecaceae) using targeted sequence capture. *Biological Journal of the Linnean Society* 117: 106–120.
- Huang CH, Zhang C, Liu M, Hu Y, Gao T, Qi J, Ma H. 2016. Multiple polyploidization events across Asteraceae with two nested events in the early history revealed by nuclear phylogenomics. *Molecular Biology and Evolution* 33: 2820–2835.
- Jhwueng DC, O’Meara B. 2015. Trait evolution on phylogenetic networks. *bioRxiv* 023986.

- Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS *et al.* 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473: 97–100.
- Johnson MG, Gardner EM, Liu Y, Medina R, Goffinet B, Shaw AJ, Zerega NJC, Wickett NJ. 2016. HybPiper: extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in Plant Sciences* 4: 1600016.
- Joly S, McLenachan PA, Lockhart PJ. 2009. A statistical approach for distinguishing hybridization and incomplete lineage sorting. *American Naturalist* 174: E54–E70.
- Kamneva OK, Syring J, Liston A, Rosenberg NA. 2017. Evaluating allopolyploid origins in strawberries (*Fragaria*) using haplotypes generated from target capture sequencing. *BMC Evolutionary Biology* 17: 180.
- Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and stability. *Molecular Biology and Evolution* 30: 772–780.
- Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C *et al.* 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28: 1647–1649.
- Kocot K, Citarella MR, Moroz LL, Halanych KM. 2013. PhyloTreePruner: a phylogenetic tree-based approach for selection of orthologous sequences for phylogenomics. *Evolutionary Bioinformatics* 9: 429–435.
- Konowalik K, Wagner F, Tomasello S, Vogt R, Oberprieler C. 2015. Detecting reticulate relationships among diploid *Leucanthemum* Mill. (Compositae, Anthemideae) taxa using multilocus species tree reconstruction methods and AFLP fingerprinting. *Molecular Phylogenetics and Evolution* 92: 308–328.
- de Kuppler ALM, Fagúndez J, Bellstedt DU, Oliver EGH, León J, Pirie MD. 2015. Testing reticulate versus coalescent origins of *Erica lusitanica* using a species phylogeny of the northern heathers (Ericaceae, Ericaceae). *Molecular Phylogenetics and Evolution* 88: 121–131.
- Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B. 2017. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Molecular Biology and Evolution* 34: 772–773.
- Lanier HC, Knowles LL. 2012. Is recombination a problem for species-tree analyses? *Systematic Biology* 61: 691–701.
- Larget BR, Kotha SK, Dewey CN, Ané C. 2010. BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics* 26: 2910–2911.
- Lemmon EM, Lemmon AR. 2013. High-throughput genomic data in systematics and phylogenetics. *Annual Review of Ecology and Systematics* 44: 99–121.
- Linder CR, Rieseberg LH. 2004. Reconstructing patterns of reticulate evolution in plants. *American Journal of Botany* 91: 1700–1708.
- Liu L, Yu L, Edwards SV. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology* 10: 302.
- Long C, Kubatko L. 2018. The effect of gene flow on coalescent-based species-tree inference. *Systematic Biology* doi: 10.1093/sysbio/syy020.
- Lundberg M, Töpel M, Eriksen B, Nylander JAA, Eriksson T. 2009. Allopolyploidy in Fragariinae (Rosaceae): comparing four DNA sequence regions, with comments on classification. *Molecular Phylogenetics and Evolution* 51: 269–280.
- Maddison WP. 1997. Gene trees in species trees. *Systematic Biology* 46: 523–536.
- Mallet J. 2007. Hybrid speciation. *Nature* 446: 279–283.
- Mallet J, Besansky N, Hahn MW. 2016. How reticulated are species? *BioEssays* 38: 140–149.
- Mandel JR, Dikow RB, Funk VA, Masalia RR, Staton SE, Kozik A, Michelmore RW, Rieseberg LH, Burke JM. 2014. A target enrichment method for gathering phylogenetic information from hundreds of loci: an example from the Compositae. *Applications in Plant Sciences* 2: 1300085.
- Maureira-Butler IJ, Pfeil BE, Muangprom A, Osborn TC, Doyle JJ. 2008. The reticulate history of *Medicago* (Fabaceae). *Systematic Biology* 57: 466–482.
- Meyer BS, Matschiner M, Salzburger W. 2017. Disentangling incomplete lineage sorting and introgression to refine species-tree estimates for Lake Tanganyika cichlid fishes. *Systematic Biology* 66: 531–550.
- Mirarab S, Bayzid MS, Warnow T. 2016. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. *Systematic Biology* 65: 366–380.
- Mirarab S, Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* 31: i44–i52.
- Mitchell N, Lewis PO, Lemmon EM, Lemmon AR, Holsinger KE. 2017. Anchored phylogenomics improves the resolution of evolutionary relationships in the rapid radiation of *Protea* L. *American Journal of Botany* 104: 102–115.
- Moore AJ, De Vos JM, Hancock LP, Goolsby E, Edwards EJ. 2017. Targeted enrichment of large gene families for phylogenetic inference: phylogeny and molecular evolution of photosynthesis genes in the Portulugo clade (Caryophyllales). *Systematic Biology* doi.org/10.1093/sysbio/syx078.
- Morales-Briones DF, Romoleroux K, Kolář F, Tank DC. 2018a. Phylogeny and evolution of the Neotropical radiation of *Lachemilla* (Rosaceae): uncovering a history of reticulate evolution and implications for infrageneric classification. *Systematic Botany*, in press.
- Morales-Briones DF, Liston A, Tank DC. 2018b. Data from: phylogenomic analyses reveal a deep history of hybridization and polyploidy in the Neotropical genus *Lachemilla* (Rosaceae). *Dryad Digital Repository*. doi: 10.5061/dryad.vj2s888.
- Motazed E, Finkers R, Maliepaard C, de Ridder D. 2017. Exploiting next-generation sequencing to solve the haplotyping puzzle in polyploids: a simulation study. *Briefings in Bioinformatics* doi.org/10.1093/bib/bbw126.
- Notov AA, Kusnetzova TV. 2004. Architectural units, axiality and their taxonomic implications in Alchemillinae. *Wulfenia* 11: 85–130.
- Otto SP, Whitton J. 2000. Polyploid incidence and evolution. *Annual Review of Genetics* 34: 401–437.
- Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. *Molecular Biology and Evolution* 5: 568–583.
- Payseur BA, Rieseberg LH. 2016. A genomic perspective on hybridization and speciation. *Molecular Ecology* 25: 2337–2360.
- Perry LM. 1929. A tentative revision of *Alchemilla* § *Lachemilla*. *Contributions from the Gray Herbarium of Harvard University* 84: 1–57.
- Pirie MD, Humphreys AM, Barker NP, Linder HP. 2009. Reticulation, data combination, and inferring evolutionary history: an example from Danthonioideae (Poaceae). *Systematic Biology* 58: 612–628.
- Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5: e9490.
- Pyron RA, Hsieh FW, Lemmon AR, Lemmon EM, Hendry CR. 2016. Integrating phylogenomic and morphological data to assess candidate species-delimitation models in brown and red-bellied snakes (Storeria). *Zoological Journal of the Linnean Society* 4: 937–949.
- Rabier CE, Ta T, Ané C. 2014. Detecting and locating whole genome duplications on a phylogeny: a probabilistic approach. *Molecular Biology and Evolution* 31: 750–762.
- Rambaut A, Suchard MA, Xie D, Drummond AJ. 2014. Tracer v1.6. Available from <http://beast.bio.ed.ac.uk/Tracer>
- Reginato M, Michelangeli FA. 2016. Untangling the phylogeny of *Leandra* s. str. (Melastomataceae, Miconieae). *Molecular Phylogenetics and Evolution* 96: 17–32.
- Reid N, Demboski JR, Sullivan J. 2012. Phylogeny estimation of the radiation of western North American chipmunks (*Tamias*) in the face of introgression using reproductive protein genes. *Systematic Biology* 61: 44–62.
- Rieseberg LH, Soltis DE. 1991. Phylogenetic consequences of cytoplasmic gene flow in plants. *Evolutionary Trends in Plants* 5: 65–84.
- Romoleroux K. 1996. 79. Rosaceae. In: Harling G, Andersson L, eds. *Flora of Ecuador*, vol 56. Göteborg/Stockholm/Quito: University of Gothenburg/Riksmuseum/Pontificia Universidad Católica del Ecuador, 1–152.
- Romoleroux K. 2004. The genus *Lachemilla* (Rosaceae) in the northern Andes of South America. *Lyonia* 7: 21–32.
- Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* 61: 539–542.

- Rothfels CJ, Pryer KM, Li FW. 2017. Next-generation polyploid phylogenetics: rapid resolution of hybrid polyploid complexes using PacBio single-molecule sequencing. *New Phytologist* 213: 413–429.
- Rothmaler W. 1935. Alchemillae Columbianae. *Trabajos del Museo Nacional de Ciencias Naturales – Serie Botanica* 31: 1–52.
- Rothmaler W. 1937. Systematische Vorarbeiten zu einer Monographie der Gattung *Alchemilla* (L.) Scop. VII. Aufteilung der Gattung und Nomenklatur. *Repertorium Novarum Specierum Regni Vegetabilis* 42: 164–173.
- Salichos L, Stamatakis A, Rokas A. 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Molecular Biology and Evolution* 31: 1261–1271.
- Sang T, Crawford DJ, Stuessy TF. 1995. Documentation of reticulate evolution in peonies (*Paeonia*) using internal transcribed spacer sequences of nuclear ribosomal DNA: implications for biogeography and concerted evolution. *Proceedings of the National Academy of Sciences, USA* 92: 6813–6817.
- Sang T, Zhang D. 1999. Reconstructing hybrid speciation using sequences of low copy nuclear genes: hybrid origins of five *Paeonia* species based on *Adh* gene phylogenies. *Systematic Botany* 24: 148–163.
- Sass C, Iles WJ, Barrett CF, Smith SY, Specht CD. 2016. Revisiting the Zingiberales: using multiplexed exon capture to resolve ancient and recent phylogenetic splits in a charismatic plant lineage. *PeerJ* 4: e1584.
- Scheuerner A, Heubl G. 2017. Against all odds: reconstructing the evolutionary history of *Scrophularia* (Scrophulariaceae) despite high levels of incongruence and reticulate evolution. *Organisms Diversity and Evolution* 17: 323–349.
- Schwarz G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6: 461–464.
- Schwenk K, Brede N, Streit B. 2008. Introduction. Extent, processes and evolutionary impact of interspecific hybridization in animals. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 363: 2805–2811.
- Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. *Systematic Biology* 51: 492–508.
- Smith SA, Moore MJ, Brown JW, Yang Y. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evolutionary Biology* 15: 150.
- Solis-Lemus C, Ané C. 2016. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genetics* 12: e1005896.
- Solis-Lemus C, Bastide P, Ané C. 2017. PhyloNetworks: a package for phylogenetic networks. *Molecular Biology and Evolution* 34: 3292–3298.
- Solis-Lemus C, Yang M, Ané C. 2016. Inconsistency of species tree methods under gene flow. *Systematic Biology* 65: 843–851.
- Soltis DE, Kuzoff RK. 1995. Discordance between nuclear and chloroplast phylogenies in the *Heuchera* group (Saxifragaceae). *Evolution* 49: 727–742.
- Soltis DE, Soltis PS. 1995. The dynamic nature of polyploid genomes. *Proceedings of the National Academy of Sciences, USA* 92: 8089–8091.
- Soltis PS, Soltis DE. 2009. The role of hybridization in plant speciation. *Annual Review of Plant Biology* 60: 561–588.
- Springer MS, Gatesy J. 2016. The gene tree delusion. *Molecular Phylogenetics and Evolution* 94: 1–33.
- Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312–1313.
- Stephens JD, Rogers WL, Heyduk K, Cruse-Sanders JM, Determann RO, Glenn TC, Malmberg RL. 2015. Resolving phylogenetic relationships of the recently radiated carnivorous plant genus *Sarracenia* using target enrichment. *Molecular Phylogenetics and Evolution* 85: 76–87.
- Straub SC, Fishbein M, Livshultz T, Foster Z, Parks M, Weitemier K, Cronn RC, Liston A. 2011. Building a model: developing genomic resources for common milkweed (*Asclepias syriaca*) with low coverage genome sequencing. *BMC Genomics* 12: 211.
- Sugiura N. 1978. Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics – Theory and Methods* 7: 13–26.
- Sukumaran J, Holder MT. 2010. DendroPy: a Python library for phylogenetic computing. *Bioinformatics* 26: 1569–1571.
- Swofford DL. 2002. PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4. Sunderland, MA, USA: Sinauer Associates.
- Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology* 56: 564–577.
- Than C, Ruths D, Nakhleh L. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics* 9: 322.
- Tiley GP, Ané C, Burleigh JG. 2016. Evaluating and characterizing ancient whole-genome duplications in plants with gene count data. *Genome Biology and Evolution* 8: 1023–1037.
- Twyford AD, Ennos RA. 2012. Next-generation hybridization and introgression. *Heredity* 3: 179–189.
- Uribe-Convers S, Settles ML, Tank DC. 2016. A phylogenomic approach based on PCR target enrichment and high throughput sequencing: resolving the diversity within the South American species of *Bartsia* L. (Orobanchaceae). *PLoS ONE* 11: e0148203.
- Vargas OM, Ortiz EM, Simpson BB. 2017. Conflicting phylogenomic signals reveal a pattern of reticulate evolution in a recent high-Andean diversification (Asteraceae: Astereae: Diplostephium). *New Phytologist* 214: 1736–1750.
- Wang Z, Liu KJ. 2016. A performance study of the impact of recombination on species tree analysis. *BMC Genomics* 17: 785.
- Warnow T. 2015. Concatenation analyses in the presence of incomplete lineage sorting. *PLoS Currents Tree of Life* 7. doi: 10.1371/currents.tol.8d41ac0f13d1abedf4c4a59f5d17b1f7.
- Weitemier K, Straub SC, Cronn RC, Fishbein M, Schmickl R, McDonnell A, Liston A. 2014. Hyb-Seq: combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences* 2: 1400042.
- Wen D, Nakhleh L. 2017. Co-estimating reticulate phylogenies and gene trees from multilocus sequence data. *Systematic Biology*. doi: 10.1093/sysbio/syx085.
- Wen D, Yu Y, Nakhleh L. 2016a. Bayesian inference of reticulate phylogenies under the multispecies network coalescent. *PLoS Genetics* 12: e1006006.
- Wen D, Yu Y, Hahn MW, Nakhleh L. 2016b. Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. *Molecular Ecology* 25: 2361–2372.
- Wendel JF, Doyle JJ. 1998. Phylogenetic incongruence: window into genome history and molecular evolution. In: Soltis DE, Soltis PS, Doyle JJ, eds. *Molecular systematics of plants II*. Boston, MA, USA: Springer, 265–296.
- Whitney KD, Ahern JR, Campbell LG, Albert LP, King MS. 2010. Patterns of hybridization in plants. *Perspectives in Plant Ecology, Evolution and Systematics* 12: 175–182.
- Xiang Y, Huang CH, Hu Y, Wen J, Li S, Yi T, Chen H, Xiang J, Ma H. 2017. Evolution of Rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication. *Molecular Biology and Evolution* 34: 262–281.
- Xu B, Yang Z. 2016. Challenges in species tree estimation under the multispecies coalescent model. *Genetics* 204: 1353–1368.
- Yu Y, Degnan JH, Nakhleh L. 2012. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genetics* 8: e1002660.
- Yu Y, Dong J, Liu KJ, Nakhleh L. 2014. Maximum likelihood inference of reticulate evolutionary histories. *Proceedings of the National Academy of Sciences, USA* 111: 16448–16453.
- Yu Y, Nakhleh L. 2015. A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics* 16: S10.
- Zhang C, Ogilvie HA, Drummond AJ, Stadler T. 2018. Bayesian inference of species networks from multilocus sequence data. *Molecular Biology and Evolution* 35: 504–517.
- Zhu J, Wen D, Yu Y, Meudt HM, Nakhleh L. 2018. Bayesian inference of phylogenetic networks from bi-allelic genetic markers. *PLoS Computational Biology* 14: e1005932.
- Zhu J, Yu Y, Nakhleh L. 2016. In the light of deep coalescence: revisiting trees within networks. *BMC Bioinformatics* 14: 415.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information tab for this article:

Fig. S1 Additional phylogenetic analysis of the HYBRID-REDUCED dataset – RAXML.

Fig. S2 Additional phylogenetic analysis of the REDUCED-HYBIRD dataset – MP-EST.

Fig. S3 Additional phylogenetic analysis of the HYBIRD-REDUCED dataset – BUCKY.

Fig. S4 Phylogenetic analysis of the NO-RECOMBINATION dataset – RAXML.

Fig. S5 Gene genealogy interrogation results of the NO-RECOMBINATION dataset testing the four topologies inferred for the four major clades of *Lachemilla*.

Fig. S6 Best species networks of the HYBRID-REDUCED dataset estimated with PHYLONET.

Fig. S7 Additional phylogenetic analysis of the ORBICULATE-REDUCED dataset – RAXML.

Fig. S8 Additional phylogenetic analysis of the ORBICULATE-REDUCED dataset – MP-EST.

Fig. S9 Additional phylogenetic analysis of the ORBICULATE-REDUCED dataset – SVDquartets.

Fig. S10 Additional phylogenetic analysis of the ORBICULATE-REDUCED dataset – BUCKY.

Fig. S11 Example of approximate-maximum-likelihood phylogenetic trees inferred with FASTTREE2 from gene families in *Lachemilla*.

Table S1 List of species and vouchers used in this study

Table S2 HYBPIPER statistics for assembly exons of 300 bp or greater (400 targets)

Table S3 Results of the test for recombination Φ

Table S4 Results of the Approximately Unbiased (AU) test from GGI for the HYBRID-REDUCED dataset

Table S5 Results of the Approximately Unbiased (AU) test from GGI for the NO-RECOMBINATION dataset

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.



About New Phytologist

- *New Phytologist* is an electronic (online-only) journal owned by the New Phytologist Trust, a **not-for-profit organization** dedicated to the promotion of plant science, facilitating projects from symposia to free access for our Tansley reviews and Tansley insights.
- Regular papers, Letters, Research reviews, Rapid reports and both Modelling/Theory and Methods papers are encouraged. We are committed to rapid processing, from online submission through to publication 'as ready' via *Early View* – our average time to decision is <26 days. There are **no page or colour charges** and a PDF version will be provided for each article.
- The journal is available online at Wiley Online Library. Visit www.newphytologist.com to search the articles and register for table of contents email alerts.
- If you have any questions, do get in touch with Central Office (np-centraloffice@lancaster.ac.uk) or, if it is more convenient, our USA Office (np-usaoffice@lancaster.ac.uk)
- For submission instructions, subscription and all the latest information visit www.newphytologist.com