



## RESEARCH ARTICLE

# Coalescent-based species delimitation in North American pinyon pines using low-copy nuclear genes and plastomes

José-Rubén Montes<sup>1</sup>  | Pablo Peláez<sup>2</sup> | Alejandra Moreno-Letelier<sup>3</sup> | David S. Gernandt<sup>4</sup> 

<sup>1</sup>Posgrado en Ciencias Biológicas, Instituto de Biología, Universidad Nacional Autónoma de México, 04510, Ciudad de México, Mexico

<sup>2</sup>Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, 62210, Cuernavaca, Morelos, Mexico

<sup>3</sup>Jardín Botánico, Instituto de Biología, Universidad Nacional Autónoma de México, 04510, Ciudad de México, Mexico

<sup>4</sup>Departamento de Botánica, Instituto de Biología, Universidad Nacional Autónoma de México, 04510, Ciudad de México, Mexico

## Correspondence

David S. Gernandt, Departamento de Botánica, Instituto de Biología, Universidad Nacional Autónoma de México, 04510, Ciudad de México, Mexico.  
 Email: [dgernandt@ib.unam.mx](mailto:dgernandt@ib.unam.mx)

**Abstract**

**Premise:** Accurate species delimitation is essential for evolutionary biology, conservation, and biodiversity management. We studied species delimitation in North American pinyon pines, *Pinus* subsection *Cembroides*, a natural group with high levels of incomplete lineage sorting.

**Methods:** We used coalescent-based methods and multivariate analyses of low-copy number nuclear genes and nearly complete high-copy number plastomes generated with the Hyb-Seq method. The three coalescent-based species delimitation methods evaluated were the Generalized Mixed Yule Coalescent (GMYC), Poisson Tree Process (PTP), and Trinomial Distribution of Triplets (Tr2). We also measured admixture in populations with possible introgression.

**Results:** Our results show inconsistencies among GMYC, PTP, and Tr2. The single-locus based GMYC analysis of plastid DNA recovered a higher number of species (up to 24 entities, including singleton lineages and clusters) than PTP and the multi-locus coalescent approach. The PTP analysis identified 10 species whereas Tr2 recovered 13, which agreed closely with taxonomic treatments.

**Conclusions:** We found that PTP and GMYC identified species with low levels of ILS and high morphological divergence (*P. maximartinezii*, *P. pinceana*, and *P. rzedowskii*). However, GMYC method oversplit species by identification of more divergent samples as singletons. Moreover, both PTP and GMYC were incapable of identifying some species that are readily identified morphologically. We suggest that the divergence times between lineages within North American pinyon pines are so disparate that GMYC results are unreliable. Results of the Tr2 method coincided well with previous delimitations based on morphology, DNA, geography, and secondary chemistry.

**KEYWORDS**

coalescent theory, conifers, GMYC, Hyb-Seq, incomplete lineage sorting, Pinaceae, pines, PTP, Tr2

Species are fundamental units of study in several areas of the biological sciences. The accurate delimitation of species boundaries is essential for evolutionary biology, conservation biology, and biodiversity management (Sites and Crandall, 1997; Sites and Marshall, 2003; Yang and Rannala, 2010). Species delimitation is related to species concepts, which have been long debated by biologists because there are discrepancies among species definitions (de Queiroz,

2007). Definitions are independent of methodological aspects of lineage delimitation, but criteria for inferring species boundaries have been used both for conceptualization and species delimitation. Nonetheless, species concepts exhibit an underlying unity that provides the basis for a unified concept of species as separately evolving lineages, which allows us to address the problem of delimitation more directly (de Queiroz, 2007).

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2022 The Authors. *American Journal of Botany* published by Wiley Periodicals LLC on behalf of Botanical Society of America.

Interest in integrative species delimitation has grown thanks to the increasing availability of molecular data, new approaches, and methods (Schlick-Steiner et al., 2010; Fujita et al., 2012; Zhang et al., 2013) that have been developed to test species boundaries. Many coalescent-based methods have been developed, including single-locus methods such as the Generalized Mixed Yule Coalescent (GMYC; Pons et al., 2006) and Poisson Tree Processes (PTP; Zhang et al., 2013; Kapli et al., 2017). Multi-locus species delimitation methods include Brownian processes (Brownie; O'Meara et al., 2006; O'Meara, 2010) and Species Tree Estimation under Maximum Likelihood (spedeSTEM; Ence and Carstens, 2011). Methods have also been developed for biallelic genetic markers such as Single Nucleotide Polymorphisms and Amplified Fragment Length Polymorphisms Phylogenies (SNAPP; Bryant et al., 2012), Bayesian Phylogenetics and Phylogeography (BP&P; Yang, 2015), Division of Individuals into Species using Sequences and Epsilon-Collapsed Trees (DISSECT; Jones et al., 2015), and the Trinomial Distribution method (Tr2; Fujisawa et al., 2016). Coalescent-based methods provide an attractive alternative for studying the evolutionary processes that contribute to speciation, inferring the relationships among species, and delimiting independent evolutionary lineages objectively in the presence of gene-tree conflict (Fujita et al., 2012; Smith et al., 2015, 2020; Luo et al., 2018).

In pines and other conifers, there are few studies using objective methods for the delimitation of species. *Pinus* L. classification and species delimitation initially used a small number of morphological characters. Over time additional evidence was incorporated from anatomy, reproductive biology, biochemistry, and molecular markers (Price et al., 1998; Gernandt et al., 2001, 2003, 2005; Syring et al., 2005; Parks et al., 2012; Willyard et al., 2021). Nevertheless, each source of evidence used undergoes processes that can introduce error, such as plasticity of certain morphological characters in response to the environment (e.g., number of needles per fascicle and needle length and width), transfer of genetic information among genomic compartments, low interspecific variability (e.g., plastid DNA markers), and uniparental inheritance of plastid DNA, which is susceptible to "plastid capture" (Mirov, 1967; Rieseberg and Soltis, 1991; Liston et al., 1999; Gernandt et al., 2003; Kan et al., 2007; Mort et al., 2007; Poulos and Berlyn, 2007; Cole et al., 2008; Tsutsui et al., 2009; Turna and Güney, 2009; Nobis et al., 2012; Cole et al., 2013). Lineage delimitation of trees is also difficult because of complex evolutionary histories caused by incomplete lineage sorting (ILS) and reticulation resulting from hybridization and introgression (Rosenberg, 2003; Hernández-León et al., 2013; Zhang et al., 2014).

Trees are commonly characterized by large population sizes, longevity, slower mutation and speciation rates, and longer generation times (Petit and Hampe, 2006). Studies of species boundaries have been carried out in *Populus* L. (Wang et al., 2011), *Cycas* L. (Feng et al., 2016), and *Pinus* (Moreno-Letelier et al., 2013; Zhang et al., 2014; López-Reyes et al., 2015; Willyard et al., 2017). Species delimitation has been done using

clustering algorithms (Pritchard et al., 2000) and grouping individuals in populations but without evaluating the evolutionary divergence of clusters. An exception was an attempt to delimit species of North American hard pines using plastid DNA and a single-locus based coalescent-based method (Hernández-León et al., 2013).

*Pinus* subsection *Cembroides* Engelm. is a clade of North American pinyon pines with a fossil record that extends to the Late Oligocene (Wolfe and Schorn, 1990). The pinyon pines are restricted to arid or semi-arid environments extending from the southwestern United States to south central Mexico (Critchfield and Little, 1966). They comprise approximately 15 pine taxa of exceptional ecological importance (Lanner, 1981; Farjon and Styles, 1997). The North American pinyon pines are small to medium-sized trees or shrubs with 1 to 5 secondary leaves with deciduous fascicle sheaths, ovulate cones with a short peduncle, and seeds that are functionally wingless in all species except *P. rzedowskii* Madrigal & M. Caball. (Madrigal and Caballero, 1969; Malusa, 1992). The International Union for Conservation of Nature (IUCN, 2019) lists five pinyon pine taxa as vulnerable or endangered whereas the Mexican government lists nine as protected (SEMARNAT, 2010).

The circumscription of recognized species in *Pinus* subsect. *Cembroides* differs in recent works (Gernandt et al., 2005; Farjon and Filer, 2013; Gernandt and Pérez de la Rosa, 2014; Montes et al., 2019) and the taxonomic ranges used for pinyon pines has varied widely, due in part to placing emphasis on different molecular and structural characters (Malusa, 1992; Farjon and Styles, 1997; Gernandt et al., 2005; Farjon and Filer, 2013). Disagreements in the classification of some taxa include whether or not to elevate *P. cembroides* Zucc. subsp. *orizabensis* D.K. Bailey and *P. cembroides* subsp. *lagunae* (Robert-Passini) D.K. Bailey (here treated as *P. lagunae* (Robert-Passini) Passini), to the rank of species (Gernandt et al., 2005; Farjon and Filer, 2013; Montes et al., 2019). Similarly, phylogenetic analyses clearly support separating *P. discolor* D.K. Bailey & F.G. Hawksw. and *P. johannis* M.F. Robert from *P. cembroides* (Gernandt et al., 2003, 2005; Parks et al., 2012; Montes et al., 2019) although they have been treated as infraspecific taxa of *P. cembroides* (Farjon and Filer, 2013).

*Pinus californiarum* D.K. Bailey, *P. fallax* (Little) Businský, and *P. monophylla* Torr. & Frém. all have solitary needles (Malusa, 1992). *Pinus californiarum* has been treated as an independent lineage, as a synonym of *P. monophylla*, or as a variety, *P. monophylla* var. *californiarum* (D.K. Bailey) Silba (Silba, 1990; Farjon and Styles, 1997; Price et al., 1998). *Pinus californiarum* was recovered as sister to *P. monophylla* and *P. quadrifolia* Parl. ex Sudw. in a previous phylogenetic study of low-copy nuclear DNA, suggesting that it could be considered as a valid species rather than as an infraspecific taxon of *P. monophylla* (Montes et al., 2019). *Pinus fallax* and *P. californiarum* may also be valid species rather than infraspecific taxa of *P. monophylla* or *P. edulis* Engelm. (Montes et al., 2019). *Pinus fallax* was originally described

as *P. edulis* var. *fallax* Little (1968) but has been treated as *P. californiarum* subsp. *fallax* (Little) D.K. Bailey, or *P. monophylla* subsp. *fallax* (Little) Silba (Farjon and Styles, 1997; Cole et al., 2008; Farjon and Filer, 2013). *Pinus fallax* occurs in environmental conditions with moderate summer rainfall (Malusa, 1992; Cole et al., 2008) unlike *P. monophylla* and *P. californiarum*, which inhabit places with dry summers (Cole et al., 2008).

*Pinus* subsect. *Cembroides* offers an opportunity to study the boundaries among species that have evolved few morphological differences (Price et al., 1998; Gernandt et al., 2008) and species with clear morphological divergences and presumably relatively deep divergences. In this study, our aims were to: (1) infer the species boundaries in *Pinus* subsect. *Cembroides* using coalescent-based methods; (2) compare species delimitation hypotheses of single and multi-locus coalescent methods; (3) reexamine the taxonomic validity of some taxa in the light of multi-locus data; and (4) study admixture in three subgroups: (a) *Pinus cembroides* subsp. *cembroides*, *P. cembroides* subsp. *orizabensis*, and *P. lagunae*; (b) *P. johannis* and *P. discolor*; and (c) *P. californiarum*, *P. fallax*, and *P. monophylla*.

## MATERIALS AND METHODS

### Sampling

We included 3–8 individuals per species from material deposited in the Herbario Nacional de México (MEXU) and from field collections (Appendix S1). Ninety-three individuals were sampled, including 80 corresponding to subsect. *Cembroides*, four to subsect. *Balfourianae* Engelm., and three to subsect. *Nelsoniae* Burgh. These three subsections comprise section *Parrya* Mayr of subgenus *Strobis* Lemmon (Gernandt et al., 2005). From sect. *Quinquefoliae* Duhamel we included two individuals of subsect. *Gerardianae* Loudon, one of subsect. *Krempfianae* Little and Critchfield, and one of subsect. *Strobis* Loudon.

We extracted DNA from haploid seed megagametophyte of 10 individuals using a Wizard genomic DNA purification kit (Promega, Madison, Wisconsin, USA) and from diploid leaf tissue of 83 individuals using the CTAB method (Doyle and Doyle, 1987).

### Illumina library preparation, probe design, and Hyb-Seq sequencing

We used 500 ng of total DNA per sample to prepare the genomic libraries. DNA fragments of ca. 250 bp were size-selected with a bioruptor sonicator and the length distribution of fragments was evaluated by automated electrophoresis using a 2100 Bioanalyzer System (Agilent, Santa Clara, California, USA). Barcode adapters were ligated for Illumina sequencing using a NEBNext library prep kit for three samples (*P. lagunae* BS3-BS5) and a

TruSeq library prep kit for all other samples (Illumina, San Diego, California, USA).

Pools of 24 samples were enriched for nuclear targets with MYbaits version 2.3.1 biotinylated RNA baits (Arbor Biosciences, Ann Arbor, Michigan, USA) following the manufacturer's protocol. Probes were designed for 1045 putative low-copy nuclear genes from *Pinus taeda* L. (Willyard et al., 2007; Neves et al., 2013; Gernandt et al., 2018; see Appendix S2 for more details). The samples were spread across eight sequencing runs (Appendix S3) that also included Pinaceae samples for other studies. Different mixtures of enriched and unenriched libraries were used for successive runs, according to recovery of plastid sequences in prior runs. We combined samples into three different multiplex sets (Appendix S3) to sequence on a single lane each of an Illumina Hi-Seq. 2500 or Hi-Seq. 4000 using the 100, 125, or 150 bp modules with paired reads (Appendix S3).

### Processing of Hyb-Seq data

Illumina reads were demultiplexed based on their barcode and processed with Trimmomatic version 0.32 (Bolger et al., 2014) using the parameters for paired end reads suggested by the authors (Appendix S2). Trimmed reads were assembled with the HybPiper version 1.2 pipeline (Johnson et al., 2016). This pipeline first performs read sorting with the BWA method (Li and Durbin, 2009) using the nuclear gene sequences from the probe design step as references and then assembles each gene using SPAdes version 3.10.1 (Bankevich et al., 2012). Assembled gene files (without introns) were imported into Geneious version R11.0.5 (Kearse et al., 2012) and aligned with MAFFT version 7.0 (Katoh et al., 2002). The genes were filtered under five ad hoc exclusion criteria: (1) missing sequence for one or more samples (341); (2) pairwise identity less than 93% (168); (3) fewer than 50% of identical sites (67); (4) genes detected as possible paralogs (164); and (5) genes with anomalously high substitution rates based on profiling phylogenetic informativeness (22) (Appendix S2). The first three criteria were applied in Geneious and the fourth with the HybPiper script `paralog_investigator.py` (Johnson et al., 2016). The paralog script identified contigs with lengths  $\geq 85\%$  of the reference sequence, indicating multiple long-length matches.

We estimated the informativeness of characters with the PhyDesign web application (Townsend, 2007; López-Giráldez and Townsend, 2011). The input files for PhyDesign were the concatenated alignment with 229 gene partitions (those remaining after filtering) and the ultrametric tree. Maximum likelihood trees were inferred in RAxML version 8.2.10 (Stamatakis, 2014) under the general time reversible model with the gamma parameter (GTR + G) and 1000 bootstrap searches. The trees were ultrametricized with clock-based likelihood in PAUP\* version 4.0a150 (Swofford, 2002) using the HKY85 substitution model and “Thorne” parameterization for

clock optimization. Genes with unusually high substitution rates, resulting in recent illusory spikes in the informativeness plot were eliminated (Appendix S4).

We used the modified protocol by Aguirre-Dugua and Gernandt (2017) to assemble plastomes. We removed duplicated reads in Geneious and trimmed low-quality bases with Trimmomatic. De novo assembly was then performed on the sequences that mapped to the reference using SPAdes. The resulting scaffolds were imported into Geneious, eliminating those that were <500 bp in length. We mapped the scaffolds to the consensus sequence that was previously generated in the mapping step and extracted a new consensus sequence. Finally, we remapped the reads to the consensus sequence to produce a final consensus. From the assembled plastomes, we chose those with the highest coverage for phylogenetic analyses (Appendix S5).

The plastome sequences were aligned with MAFFT in Geneious. Poorly aligned or divergent regions (with elevated numbers of differences, insertions, and deletions) were deleted using the Gblocks webserver version 0.91b (Castresana, 2000) with the following options: (1) smaller final blocks; (2) gap position within the final blocks; and (3) less strict flanking positions.

## Phylogenetic analysis

We performed a maximum likelihood analysis of the 207 nuclear genes and 93 terminals in RAxML (Stamatakis, 2014). We performed heuristic searches with the `-#autoMREoption`, which automatically determines the sufficient number of bootstrap replicates (bs) and a GTR + G model to calculate the heterogeneity of rates for each of the multiple alignments. We also concatenated the 207 gene alignments and performed a maximum likelihood analysis in RAxML with 1000 heuristic searches and the GTR + G model. The plastome sequences were analyzed in RAxML with 1000 heuristic searches on the alignment partitioned into coding and noncoding regions and applying the GTR + G model to each.

We also performed Bayesian inference using MrBayes version 3.2.7 (Huelsenbeck and Ronquist, 2001) on the plastid DNA alignment with partition blocks for coding and noncoding regions. The nucleotide substitution model was chosen using the Akaike Information Criterion test (AIC) in jModelTest version 2.1.10 (Miller et al., 2010; Darriba et al., 2012). The analysis was conducted using the GTR model allowing both invariant sites and rate heterogeneity (I + G). The analysis was run using three heated chains and one cold chain, and a heating of 0.2. Two independent runs of 40,000,000 generations were performed with sampling every 1000 generations, discarding 0.25 as a burn-in fraction. We used Tracer version 1.7.1. (Rambaut et al., 2018) to corroborate chain convergence. Tree topologies were summarized in a 50% majority rule tree and the consensus tree was imported into FigTree version 1.4.0 for further editing (Rambaut, 2012).

## Identification of single nucleotide polymorphisms

To provide an alternative method for assessing variable sites for the species delimitation analyses, SNP calling was performed through the alignment of the Hyb-Seq data to the *Pinus taeda* genome. SNPs were initially called from all 93 samples. Quality of demultiplexed sequence reads was assessed with FastQC version 0.11.7 and MultiQC version 1.5 (Andrews, 2010; Ewels et al., 2016). Sequence quality trimming and adapter removal were performed using Trimmomatic with default parameters (Bolger et al., 2014). Paired cleaned reads were mapped against the *P. taeda* genome version 2.01 using BWA-MEM (Li, 2013 [Preprint]; Neale et al., 2014; Wegrzyn et al., 2014). BAM files were sorted and uniquely mapped reads were extracted with the sort and view routines of SAMtools version 0.1.19 (Li et al., 2009). The Picard tool MarkDuplicates version 2.5.0 was used to discard duplicate reads (website: <http://broadinstitute.github.io/picard>). SAMtools mpileup was used to call variants with the following parameters: (1) biallelic variants only; (2) no-BAQ; (3) minimum mapping quality of 20; and (4) minimum base quality of 25 (Li, 2011). A first step filter was applied with VCFtools version 0.1.13 to keep variants genotyped in 50% of the samples, having a minimum quality score of 25, a minor allele count less than three, and a minimum mean depth of three reads (Danecek et al., 2011). In a second filtering step, SNPs were removed if genotypes were not present across all samples (100%), minimum mean depth was below 10, and minimum quality score was below 30. With respect to the SNPs called for the *P. cembroides* complex (17 samples), the *P. johannis-discolor* complex (11 samples), and the *P. californiarum-fallax-monophylla* complex (27 samples), samples were extracted separately after the first step filter with VCFtools and then subjected to the second filter. The variant calling format (VCF) file containing all 93 samples was converted into a tab-delimited text file, and subsequently SNPs were concatenated into a FASTA file including heterozygous sites. A multiple sequence alignment was generated with MAFFT.

## Lineage tree estimation

Lineage tree inference was performed with the coalescent method ASTRAL-III version 5.7.3 (Mirarab and Warnow, 2015; Zhang et al., 2018). We used 207 nuclear gene trees with 93 terminals estimated previously with RAxML as input for ASTRAL-III. Branch lengths, coalescent units, and local posterior probabilities (lpp) were estimated with the lineage tree. We also explored gene conflict using gene-concordance factors (gCFs) in IQ-TREE-2 version 2.1.2 (Minh et al., 2020).

Tree inference based on SNPs was performed with the coalescent method SVDquartets (Chifman and Kubatko, 2014) in PAUP\* (Swofford, 2002). SVDquartets performs



well even when there is variation in effective population sizes, presence of ILS, and gene flow (Long and Kubatko, 2018, 2019). It infers unrooted trees from quartets based on multi-locus and unlinked SNP data (Chifman and Kubatko, 2014). For this method we used a NEXUS input file that included 26,180 SNPs and 93 terminals. All possible quartets were analyzed and branch support was estimated with 1000 bootstrap replicates.

## Coalescent-based species delimitation

Species boundaries were inferred under the coalescent framework with both low-copy nuclear genes and plastid DNA. For plastomes, we employed the Generalized Mixed Yule Coalescent (GMYC) using single and multiple thresholds models (sGMYC; Pons et al., 2006; mGMYC; Monaghan et al., 2009), and the Poisson Tree Processes method (PTP; Zhang et al., 2013; Kapli et al., 2017). The rooted triplets method (Tr2; Fujisawa et al., 2016) was used to perform species delimitation with the nuclear genes.

The GMYC was designed to test species delimitation with single-locus data. The method distinguishes between Yule (branching events inter-specific) and coalescence processes (branching events intraspecific) based on the difference in branching rates across all species in the phylogeny (Pons et al., 2006). A likelihood ratio test (LRT) is used to assess the timing of branching events from a null model (same species) and an alternative model (different species). The likelihood score from a chi-square test allows detecting significant changes between branching events inter-specific and branching events intraspecific (Pons et al., 2006). We used the ultrametric plastid tree with branch lengths estimated using maximum likelihood as input and performed the analyses in RStudio version 1.2.1335 (RStudio team, 2019; website: <http://www.rstudio.com>) using the “*splits*” package.

PTP uses two functions (speciation/coalescence) for modeling the transition point (node) between inter-specific and intraspecific branching events in a phylogeny (Zhang et al., 2013; Kapli et al., 2017). PTP uses an exponential distribution that represents the number of accumulated substitutions ( $k$ ) for speciation events ( $n$ ) (Kapli et al., 2017) and assumes that speciation and coalescence rates are different. PTP implements Markovian chains to assess delimitation support on a phylogenetic tree (Kapli et al., 2017). We used the (non-ultrametric) tree estimated from the plastid DNA alignment with MrBayes as input for PTP version 0.51. The bPTP analysis was run for 500,000 generations (as recommended by the authors), with a thinning of 100, and discarding 0.1 generations as a burn-in fraction.

Species delimitation was also performed with Tr2 in Python version 2.7 (Fujisawa et al., 2016). Tr2 uses Bayesian model comparison and reduces the likelihood calculations because phylogenies are decomposed into rooted triplet topologies. The method explores the best delimitation model from a guide tree and multi-locus data. Posterior

probability is used to find the best delimitation model from a set of possible hypotheses. The best model is the one with a posterior probability close to 0 (Fujisawa et al., 2016). Tr2 estimates a null model (without a priori assignment of individuals to species) and allows alternate assignment of individuals to species to test among hypotheses. The 207 maximum likelihood gene trees obtained in RAxML were the input for Tr2. We compared the likelihood scores of five alternative hypotheses based on taxonomic classifications (Appendix S6).

## Identifying genetic clustering from SNPs

We identified genetic clusters from three subsets of SNPs using a principal components analysis (PCA) and discriminant analysis of principal components (DAPC). Analyses were performed separately for subsets of taxa from three species complexes. The first subset corresponded to *P. discolor* and *P. johannis* ( $K=2$ ), the second to *P. cembroides* subsp. *cembroides*, *P. cembroides* subsp. *orizabensis*, and *P. lagunae* ( $K=2-3$ ), and the third to *P. californiarum*, *P. fallax*, *P. monophylla*, *P. edulis*, and *P. quadrifolia* ( $K=4-5$ ); this clade was called “one-needle pines + *P. edulis* + *P. quadrifolia*”. Both PCA and DAPC analyses were carried out following the code by Grünwald et al. (2016) in RStudio with the “*ape*” version 5.6 (Paradis and Schliep, 2019) and “*poppr*” version 2.8.5 packages (Kamvar et al., 2014). The first two components were used for plotting using *ggplot2* package version 3.2.2 (Villanueva et al., 2016). The DAPC was performed to maximize the discrimination between groups with the same parameters as in the PCA (see Grünwald et al., 2016). We assigned the samples a priori to each species and evaluated the species assignments based on the results of the PCA. We illustrated the probability of population membership and assigned probability of populations membership.

## Admixture analysis

To estimate the genetic admixture proportions within the three sample complexes the VCF files were converted to ordinary PLINK files using PLINK version 1.9 (Purcell et al., 2007). The optimal  $K$  value (evaluated from 2-10) and the admixture proportions (Q-values) up to  $K=6$  were obtained with ADMIXTURE version 1.3.0 (Alexander et al., 2009). The Clumpak program version 1.1 was used for the clustering and plotting of Q-matrices (Kopelman et al., 2015).

## RESULTS

### Pre-processing and processing of data

For nuclear DNA, paired reads from 99 individuals were assembled to 996 reference genes in HybPiper. The average

number of reads mapped to the references was 5,064,027. The number of genes recovered was 969 with a mean coverage of 249.3 $\times$ . We retained 207 genes after filtering (see Materials and Methods).

For plastid DNA, we assembled 82 plastomes, 69 corresponding to subsect. *Cembroides* and 13 to close relatives. The mean number of reads per sample was 8,985,593 and the mean number of reads that mapped to the plastid reference was 86,426. The plastomes had a mean length of 117,610 bp and a mean coverage of 73.5 $\times$ , ranging from 8.5 to 1925 $\times$ . From the 82 assembled plastomes, we chose 59 with a coverage of 20 $\times$  or higher and with a mean coverage of 102.2 $\times$  (Appendix S5). The three samples with the highest coverage (*P. aristata* Engelm. AZ2; 1926.2 $\times$ , *P. bungeana* Zucc. ex Engelm. CN; 1098.7 $\times$ , and *P. cembroides* subsp. *orizabensis* PL; 437.2 $\times$ ) were sequenced with version 1 of the probes, which included two plastome regions. This caused a spike in total coverage at two places in the plastome.

## Phylogenetic analyses

We assembled a concatenated nuclear alignment with 140,845 bp and 11,281 informative sites. The best ML tree with partitions by gene (Appendix S7) agreed in topology with the best tree without partitions (Appendix S8), except for minor differences in relationships among different samples of the same species and bootstrap values. *Pinus* subsection *Cembroides* was monophyletic, and samples for six species were recovered as monophyletic lineages (Appendix S7).

The plastid DNA alignment was 117,210 bp after removal of ambiguously aligned sites. It included 3378 informative characters and 1884 variable but parsimony uninformative characters. The Bayesian inference analysis of the plastid DNA alignment recovered the monophyly of subsects. *Cembroides*, *Balfourianae*, and *Gerardianae*. In subsect. *Cembroides*, *P. cembroides* subsp. *cembroides*, *P. maximartinezii*, *P. monophylla*, *P. quadrifolia*, *P. remota* (Little) D.K. Bailey & F.G. Hawksw., and *P. rzedowskii* were recovered as monophyletic lineages. Of the taxa recovered as monophyletic, only *P. remota* had an incongruent phylogenetic position between the nuclear and plastid trees (Appendices S7 and S9).

## Lineage tree estimation

The SVDquartets and ASTRAL analyses recovered *Pinus* subsection *Cembroides* as monophyletic and a greater number of taxa as monophyletic lineages than were recovered with analyses of plastid DNA and concatenated nuclear genes. In both coalescent trees, *P. discolor*, *P. culminicola* Andresen & Beaman, *P. johannis*, *P. lagunae*, *P. maximartinezii*, *P. monophylla*, *P. pinceana* Gordon, *P. quadrifolia*, *P. remota*, and *P. rzedowskii* were recovered as exclusive lineages, whereas *P. cembroides* subsp. *cembroides*, *P. cembroides* subsp.

*orizabensis*, *P. californiarum*, *P. edulis*, and *P. fallax* were nonmonophyletic (Figures 1 and 2). The majority of monophyletic lineages recovered in both SVDquartets and ASTRAL received high support (>80% bs and 0.8 lpp, respectively) except for *P. discolor* and *P. johannis* in the ASTRAL tree (Figure 2). The interrelationships in the trees based on SNPs and low-copy nuclear genes were different in small-cone species but identical in big-cone species (Figures 1 and 2). The level of ILS in the ASTRAL species tree was very high (0.5), indicating that only 50% of the total (331,985,270) quartets estimated from gene trees agree with the lineage tree. The local posterior probabilities for a third of branches were >0.8 and coalescent branch lengths in the ASTRAL tree were short for most relationships in the small-cone clade with an exception in *P. culminicola* (Figure 2). The percent of gene discordance showed high levels of conflict in all branches within subsect. *Cembroides*. In some branches, no gene tree agreed (Figure 2).

## Coalescent-based species delimitation

From the ultrametric tree inferred with 59 plastid genomes, we performed delimitation analyses with both single and multiple threshold GMYC models. Both analyses were congruent in all independent coalescent groups identified, except in the “D”, “E”, and “I” groups (Figure 3). The number of species (clusters and singletons) described below considers only subsect. *Cembroides*. Single threshold GMYC identified 21 species ( $P = 0.05$ ). The single threshold coalescent model exhibited significantly better fit over the null model ( $\log L_{\text{GMYC}} = -529.7634$ ,  $\log L_{\text{null}} = -524.9513$ ,  $\text{LRt} = 9.624205$ ,  $P = 0.008130748^{**}$ ) (Appendix S10). Multiple threshold GMYC recovered four speciation-coalescence transition events (Appendix S11) and identified 24 species ( $P = 0.05$ ). The multiple threshold model exhibited significantly better fit than the null model ( $\log L_{\text{GMYC}} = -530.1472$ ,  $\log L_{\text{null}} = -524.9513$ ,  $\text{LRt} = 10.39178$ ,  $P = 0.005539288^{**}$ ). A comparison between single and multiple GMYC models revealed that they were not significantly different from each other ( $X^2 = 0.1918$ ,  $P_{0.05} = 0.661437$ ). From 21 species identified by the single threshold GMYC model, only 5 taxa from subsect. *Cembroides* have been treated as separate species based on morphological and molecular evidence (Gernandt et al., 2005; Montes et al., 2019). These taxa (*P. monophylla*, *P. maximartinezii*, *P. quadrifolia*, *P. remota*, and *P. rzedowskii*) also were exclusive lineages in the Bayesian inference tree (Appendix S9) whereas with multiple threshold GMYC only 3 taxa identified in the subsect. *Cembroides* clade correspond to recognized species (*P. monophylla*, *P. maximartinezii*, and *P. remota*). Moreover, multiple threshold GMYC separated individuals of *P. quadrifolia* and *P. rzedowskii* into multiple species. Others were treated as a single entity with both single and multiple threshold models, including the *P. cembroides* subsp. *cembroides* + *P. cembroides* subsp. *orizabensis* + *P. lagunae* clade and the *P. californiarum* + *P. fallax* + *P. edulis* clade.

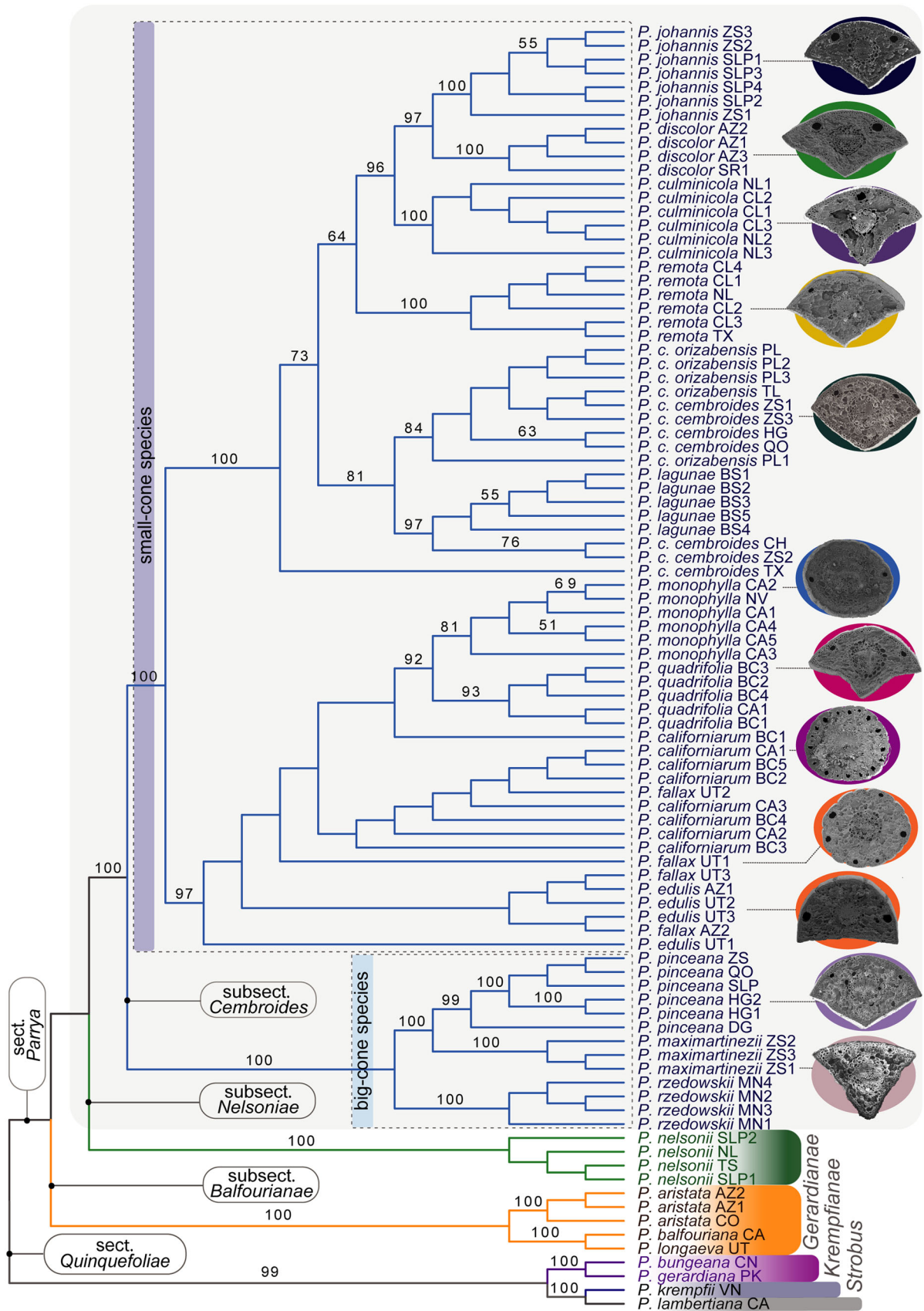
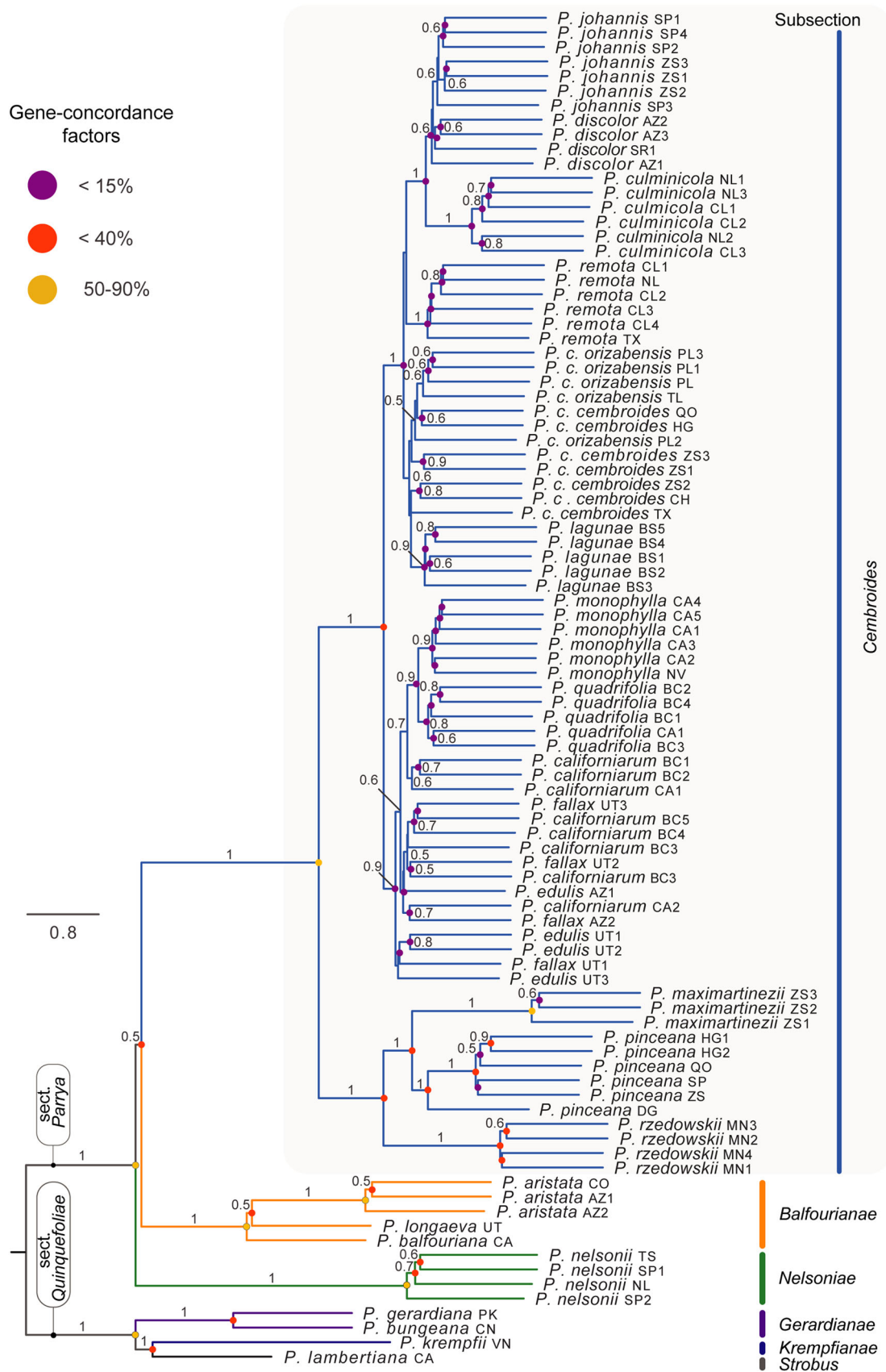


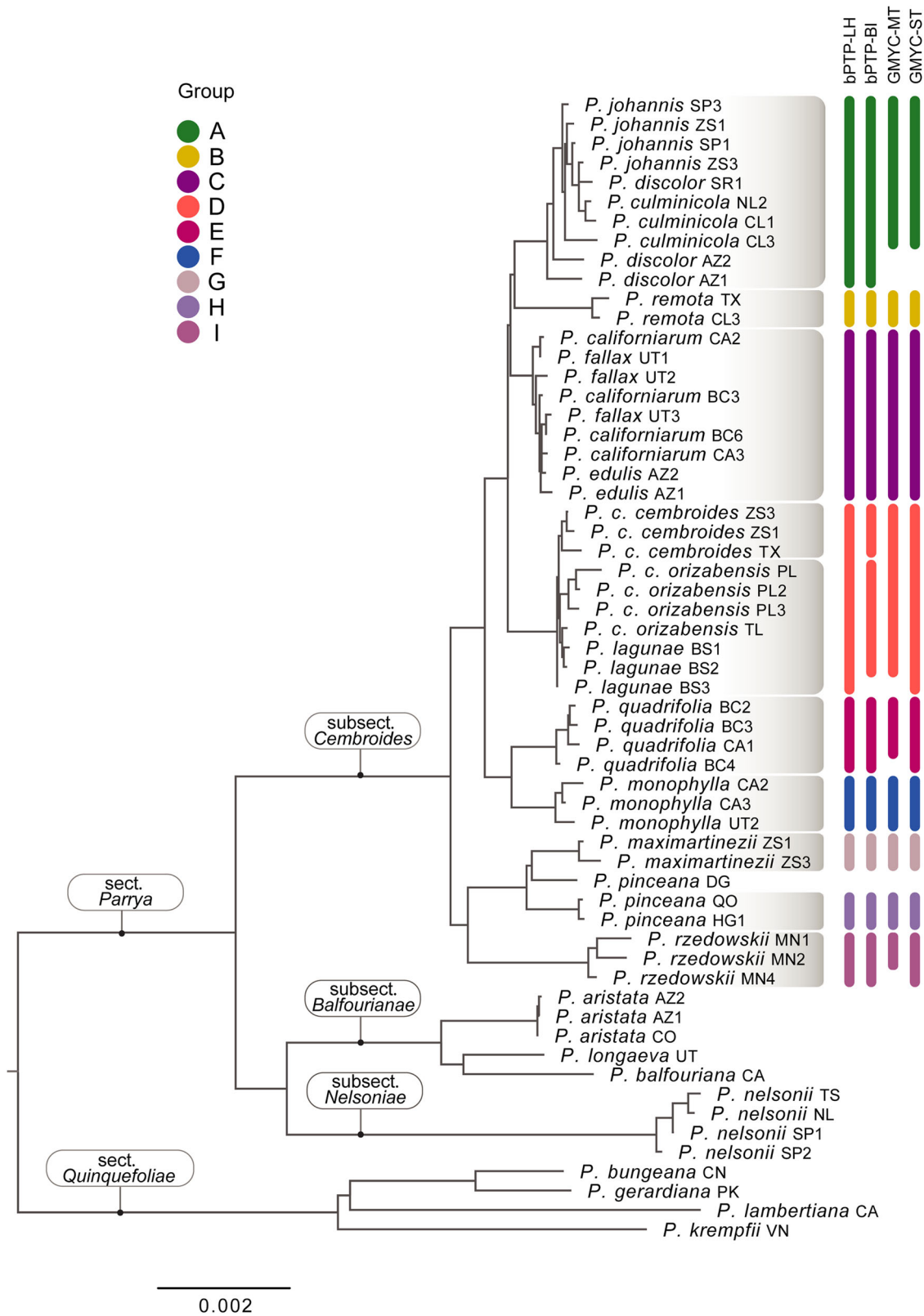
FIGURE 1 SVDquartets lineage tree based on SNPs. The small-cone and big-cone clades of *Pinus* subsection *Cembroides* are indicated. Transverse sections of needles by SEM show the needle shape, number of resin canals, and other internal structures. Bootstrap values >50% are shown on branches.





**FIGURE 2** ASTRAL lineage tree based on low-copy nuclear genes. Results were inferred based on 207 trees inferred with RAxML. Branch lengths represent coalescent units. Local posterior probability values  $\geq 0.5$  are shown on branches. The gene conflict using gene concordance factors is shown on nodes with colored circles.





**FIGURE 3** Single-locus, coalescent-based species delimitation. Results presented are based on the plastome tree resulting from Bayesian inference (BI). Vertical bars corresponding to each lineage are potential species. Colored vertical bars represent delimited clusters between both bPTP and GMYC methods. The first vertical bars correspond to bPTP with maximum likelihood estimation, the second to bPTP with Bayesian inference, the third to GMYC with multiple thresholds, and the final vertical bars indicate clusters recovered by GMYC with a single threshold.

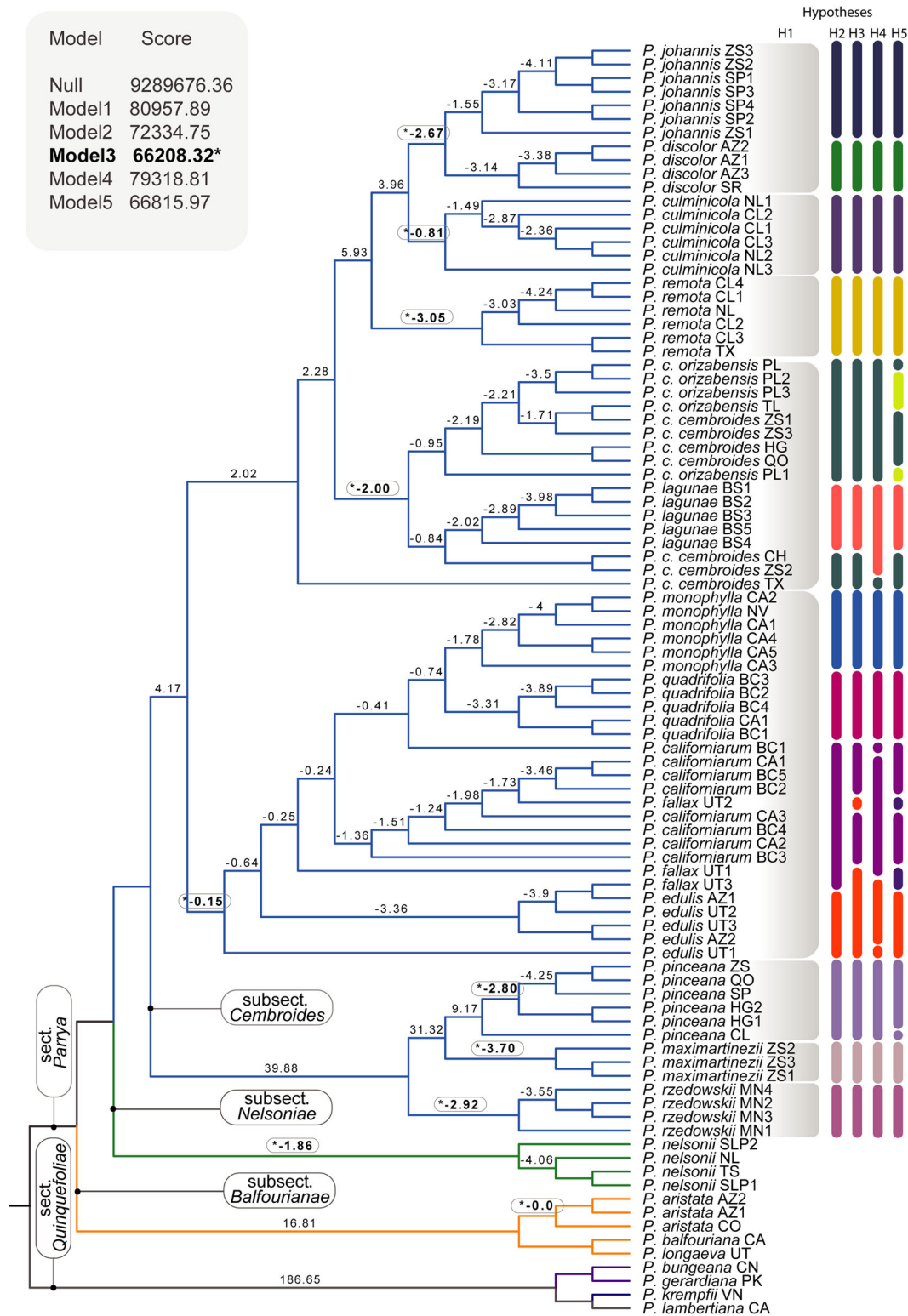
Poisson tree processes estimated 9 to 10 species within subsect. *Cembroides* (Figure 3). Bayesian and maximum likelihood solutions agreed in all independent coalescent groups identified with the exception of the monophyletic *P. cembroides* group (Figure 3). The number of independent coalescent groups estimated with a moderately to well-supported partition in both analyses represents 50% of all taxa (Acceptance<sub>rate</sub> = 0.53). The likelihood plot indicated convergence of Markovian chains (Appendix S12). The Bayesian solution of PTP estimated more well-supported species than the maximum likelihood solution (Appendices S12 and S13). The bPTP-BI analysis estimated 10 species within subsect. *Cembroides*, and the bPTP-ML analysis estimated 9 (Figure 3). Of 10 species estimated by bPTP-BI, only *P. cembroides* subsp. *cembroides*, *P. monophylla*, *P. maximartinezii*, *P. quadrifolia*, *P. remota*, and *P. rzedowskii* were monophyletic lineages in the Bayesian inference tree (Appendix S9). As with GMYC, bPTP solutions identified *P. californiarum*, *P. fallax*, and *P. edulis* as a single species (Figure 3) and all individuals from the *P. discolor* + *P. johannis* + *P. culminicola* clade as a single species. Based on the bPTP-BI solution, *P. cembroides* is divided into two taxa: *P. cembroides* subsp. *cembroides* and *P. lagunae* + *P. cembroides* subsp. *orizabensis*. However, the bPTP-BI solution did not identify all individuals of *P. lagunae* as part of the same species (*P. lagunae* BS3). This result was not supported by the bPTP-ML solution (Figure 3).

Based on the guide tree obtained with SVDquartets, five different hypotheses were tested with Tr2. The alternative model (Ha = 66208.32) was better than the null model (Ho = 9289676.36). Based on the null hypothesis, Tr2 estimated 8 species in subsect. *Cembroides* (Figure 4) corresponding to *P. culminicola*, *P. maximartinezii*, *P. pinceana*, *P. remota*, *P. rzedowskii*, the *P. discolor* + *P. johannis* clade, *P. cembroides* (together with *P. lagunae* and *P. cembroides* subsp. *orizabensis*), and the one-needle pines + *P. edulis* + *P. quadrifolia*. The best delimitation model was H<sub>3</sub>, which identified thirteen species. Ten belonged to the small-cone pinyons clade: *Pinus californiarum*, *P. culminicola*, *P. cembroides* subsp. *cembroides* + *P. cembroides* subsp. *orizabensis*, *P. discolor*, *P. edulis* + *P. fallax*, *P. johannis*, *P. lagunae*, *P. monophylla*, *P. quadrifolia*, and *P. remota*, and three to the large-cone pinyons: *P. maximartinezii*, *P. pinceana*, and *P. rzedowskii*. *Pinus fallax* and *P. cembroides* subsp. *orizabensis* were not identified as independent evolutionary lineages or species. Particularly, *P. fallax* was not recovered as a monophyletic lineage because two samples (UT3, AZ2) grouped with *P. edulis* (Figure 4). These individuals of *P. fallax* had one and two needles on the same tree. Another two samples (UT1, UT2) grouped with *P. californiarum* (Figure 4). All individuals of this latter cluster had predominantly solitary needles but both samples of *P. fallax* also had one and two needles on the same tree.

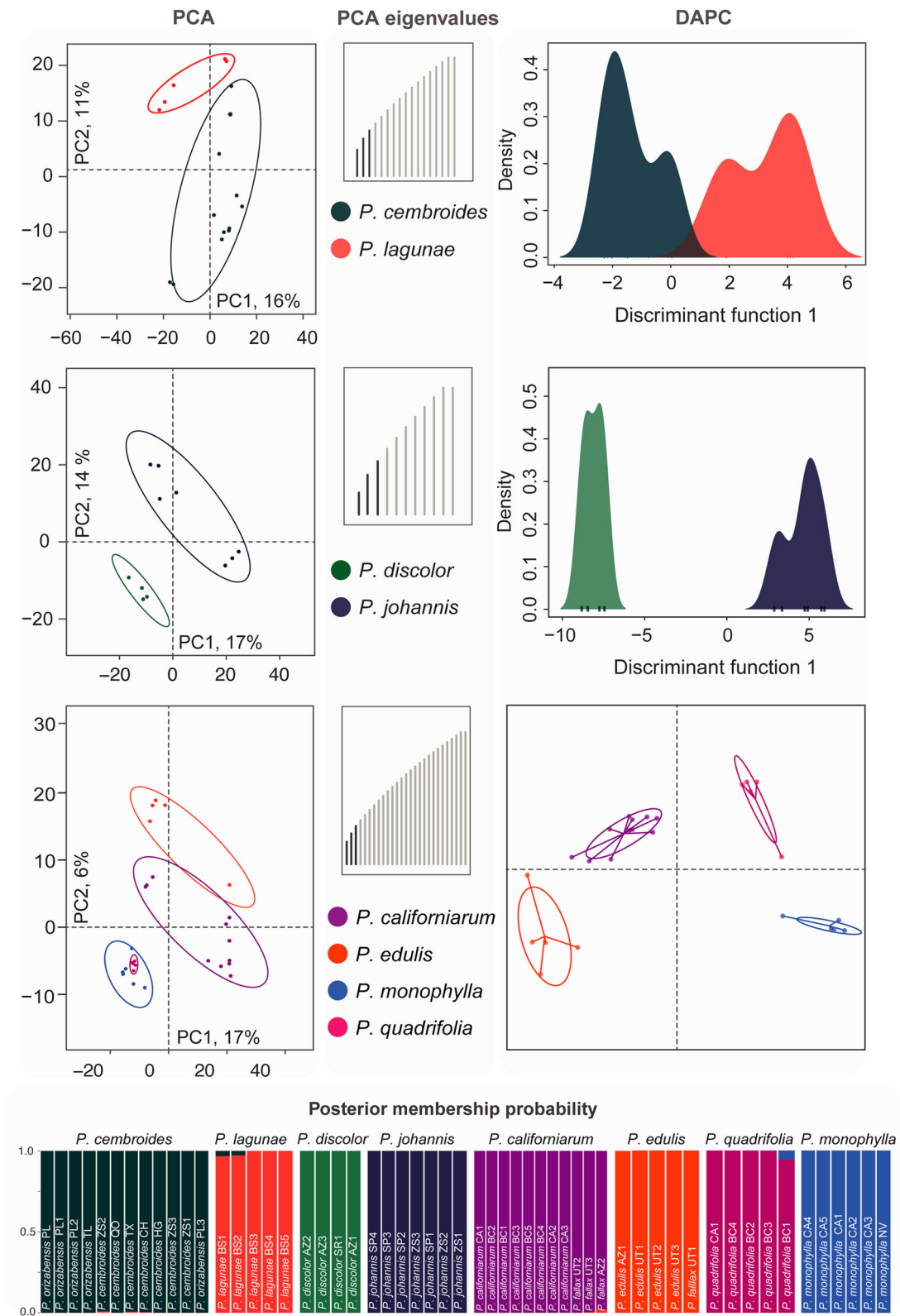
## Identifying genetic clustering from SNP subsets

The first subset of SNPs for the *P. cembroides* complex had 22,500 variants. The first three principal components (PCs)

represent 33.1% of the total variation, 16%, 11%, and 6.1%, respectively (Figure 5). The  $K=2$  analysis identified and separated two different clusters, one comprising all individuals of *P. cembroides* subsp. *cembroides* and *P. cembroides* subsp. *orizabensis* and the second comprising individuals of *P. lagunae*. The PC1 had low and positive values, mainly for *P. cembroides* subsp. *cembroides* and *P. cembroides* subsp. *orizabensis*. The PC2 differentiated *P. lagunae* from the other taxa. The second set of SNPs for *P. johannis* and *P. discolor* had 29,601 variants. Similarly, the first three principal components represent 42% of the total variation, 17%, 14%, and 11%, respectively (Figure 5). The  $K=2$  analysis differentiated two clusters, one comprising all individuals of *P. johannis* and the other all individuals of *P. discolor*. The PC1 had negative values for *P. discolor*. The PC2 had mainly positive values and differentiated *P. johannis* from *P. discolor*. The last set of SNPs for solitary needle species, *P. edulis* and *P. quadrifolia* had 30,666 variants. The first three principal components represent 28.8% of the total variation, 17%, 6%, and 5.8%, respectively (Figure 5). The  $K=4$  analysis differentiated four clusters, one composed of individuals of *P. monophylla*, a second of individuals of *P. quadrifolia*, which is nested in *P. monophylla*, a third of individuals of *P. edulis* and only one sample of *P. fallax* (UT1) from Utah, and the last cluster is composed of individuals of *P. californiarum* and the remaining *P. fallax* samples from Utah and Arizona (UT2, UT3, AZ2). The PC1 had positive values mainly for *P. californiarum* and *P. fallax* UT2, UT3, and AZ2, whereas *P. monophylla* and *P. quadrifolia* had negative values. In the PC2 both *P. monophylla* and *P. quadrifolia* were differentiated from the other species. The analysis also separated *P. edulis* and *P. fallax* UT1 from a cluster of all other species. We performed three discriminant analyses of principal components (DAPC) retaining the same principal components of the PCA (3 PCs). The DAPC results coincide with the PCA (Figure 5). Eight clusters were differentiated: (I) *Pinus cembroides* subsp. *cembroides* and *P. cembroides* subsp. *orizabensis* distributed from southwestern USA and north to south-central Mexico; (II) *P. lagunae*, endemic to Baja California Sur, Mexico; (III) *P. discolor* distributed in southwestern USA (Arizona and New Mexico), northwestern Mexico and southern San Luis Potosí; (IV) *P. johannis* distributed in the Sierra Madre Oriental; (V) *P. californiarum*, *P. fallax* UT2, UT3, and AZ2 are distributed in southwestern USA and Baja California, Mexico; (VI) *P. edulis* and *P. fallax* (UT1) distributed in southwestern USA; (VII) *P. monophylla* distributed in southwestern USA; and (VIII) *P. quadrifolia* distributed in southwestern USA (California) and Baja California, Mexico. All individuals of this last cluster, including the only sample of *P. fallax*, share two needles per fascicle, and. All individuals of this last cluster have predominantly solitary needles but two samples of *P. fallax* had both solitary and two-needle fascicles. The posterior membership probability calculated for all clusters was high (>95%), indicating a correct assignment of the individuals to species (Figure 5).



**FIGURE 4** Multi-locus coalescent-based species delimitation scenarios. Results presented are based on the SNPs tree resulting from SVDquartets. Vertical bars correspond to each lineage recovered as a potential species. Colored vertical bars represent different hypotheses tested (H#). The gray boxes correspond to the null hypothesis. The best model of delimitation is hypothesis three. The numbers on the branches indicate average differences of posterior probability scores. “\*” indicates the best delimitation according to the null model. Positive values = between-species branches, negative values = within species. Values without symbols do not have enough samples to split.



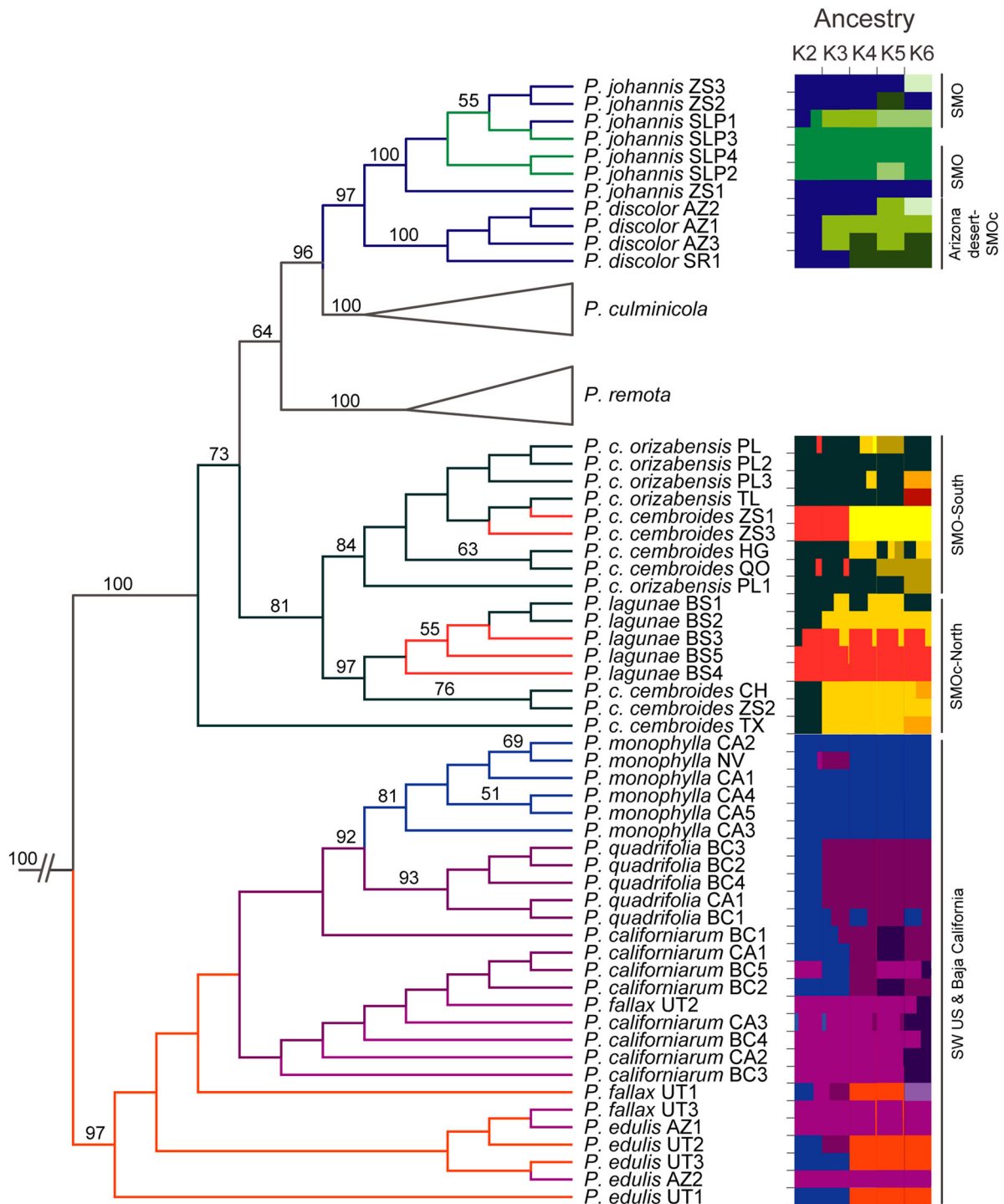
**FIGURE 5** PCA and DAPC plots based on SNPs. Plots show PC1 and PC2 for all species and specimens. Each species used in PCA and DAPC graphs is represented by a different color. Specimens and groups of individuals of the same species in DAPC are enclosed by ellipses that included 95% of the data for each group. Bar graphs depict the percentage variance of significant PCs (eigenvalues). The colored bar-graphs represent the posterior membership probability of specimens to species.



### Admixture analysis

In total, 29,506 SNPs were obtained for the *P. johannis* and *P. discolor* group, 22,454 for the *P. cembroides* group, and 30,609 for the one-needle pines, *P. edulis*, and *P. quadrifolia* group.

The average mean depth of the SNPs for the three sets was 200x. For the three groups analyzed, the optimal clustering was for two subpopulations ( $K=2$ ); however, a cluster range of 2 to 6 was evaluated to explore the dynamics of the classifications (Figure 6). Overall, the two



**FIGURE 6** Genetic admixture proportions in the three subgroups of small-cone pinyons. Analyses were performed for K values ranging from 2 to 6 with three different matrices containing 22,500 variants for the *P. cembroides* subgroup, 29,601 variants for *P. johannis* and *P. discolor*, and 30,666 variants for single-needle species, *P. edulis*, and *P. quadrifolia*. Different colors represent different clusters. The combination of different colors in a bar indicates the degree of admixture. Samples in the admixture model are in the same topological order as in the SVDquartets tree. Bootstrap values >50% are shown above the branches. SMO = Sierra Madre Oriental and SMOc = Sierra Madre Occidental.

subpopulations partitioned by the optimal clustering in each of the three groups did not match exactly with the clade organization in the SNP tree. In the *P. johannis* and *P. discolor* group, optimal clustering inferred two subpopulation structures for individuals of *P. johannis*, sharing one of these structures with *P. discolor*. Only one individual (*P. johannis* SLP1) in this group showed signs of admixture. For most of the subsequent higher  $K$  values, the subpopulation composed of only individuals of *P. johannis* was kept, and at least one of the other subpopulations was composed of individuals of the two species. In the *P. cembroides* group, one subpopulation ( $K=2$ ) included individuals of *P. cembroides* subsp. *cembroides* and *P. lagunae*, whereas the other subpopulation had individuals of all three taxa. Three individuals showed signs of admixture, including one of *P. cembroides* subsp. *orizabensis*. Three of five individuals of *P. lagunae* that do not have the same structure (BS3, BS4, and BS5) group with individuals of *P. cembroides* from Zacatecas, Mexico (ZS1 and ZS3) (SMOc-South; Figure 6). The *P. lagunae* BS3 individual was admixed with the main subpopulation from the north. Individuals predominantly from the south (Sierra Madre Oriental-South), forming one clade, were partitioned in more subpopulations than the individuals from the north, and thus had more admixed individuals in the successive  $K$  values. The subpopulations of the one-needled pines, *P. edulis*, and *P. quadrifolia* best matched the order of the clades in the tree (Figure 6) across most of the  $K$  values evaluated. The admixture run of  $K=2$  in this group resulted in a subpopulation composed of all the individuals of *P. monophylla* and *P. quadrifolia*, three of *P. californiarum*, one *P. fallax*, and three *P. edulis*. The other subpopulation comprised individuals of *P. californiarum*, *P. fallax*, and *P. edulis*, suggesting higher genetic variation and admixture among the individuals of these species than those of *P. monophylla* and *P. quadrifolia*. For  $K=3$ , a new subpopulation containing mainly individuals of *P. quadrifolia* was observed. For the remaining values of  $K$ , the individuals of *P. monophylla* comprised one reliable subpopulation. Only one individual from *P. quadrifolia* showed strong admixture proportions with this subpopulation. Also, for different numbers of ancestral clusters, individuals of *P. californiarum* tended to be assigned to two different subpopulations denoting structure within the species. For  $K>3$ , certain individuals of *P. fallax* and *P. edulis* were grouped in the same subpopulation as some individuals of *P. californiarum*; however, certain individuals of these two species also constituted an independent subpopulation.

## DISCUSSION

### Plastid and Nuclear DNA data

Our analyses of plastomes and low copy nuclear genes allowed us to re-evaluate previous studies of *Pinus* subsection *Cembroides* based on much smaller data sets (Gernandt et al., 2001, 2005). We analyzed plastomes for a comprehensive taxonomic sampling from multiple individuals per species

within subsect. *Cembroides*. The alignment length (117,210 bp) was shorter, and the number of informative sites (11,281 bp) was fewer in our plastome alignment of *Pinus* subgen. *Strobus* than the genus-wide alignment by Parks et al. (2012), which had a total length of 141,265 aligned sites and 15,151 informative characters but only one individual per species. The plastid DNA relationships among pinyon pines, including those supported by high bootstrap values, changed somewhat between the studies. In contrast to Parks et al. (2012), *P. monophylla* and *P. quadrifolia* formed a well-supported clade (1.0 posterior probability [pp]). Likewise, *P. maximartinezii*, *P. pinceana*, and *P. rzedowskii* formed another well-supported clade (1.0 pp).

For nuclear DNA, we used fewer genes for phylogenetic estimates compared to a previous study (Montes et al., 2019) because we included an additional criterion to eliminate genes with exceptionally high substitution rates. Nonetheless, we added 36 more individuals than Montes et al. (2019). The individuals cover a greater part of the geographical distribution of *Pinus* subsect. *Cembroides*, including several individuals per species from edges of their ranges. For instance, we included eight individuals of *P. californiarum*, three from California, and five from Baja California (including one population from La Asamblea, the southern limit of the species). We also added four more individuals of *P. pinceana* from different populations in Mexico (Durango, Hidalgo, Querétaro, and Zacatecas), and expanded sampling of *P. cembroides* subsp. *cembroides* to include individuals from three more populations (Querétaro, Zacatecas, Mexico and Texas, USA). The phylogenetic position of taxa with expanded sampling is consistent with our previous study (Montes et al., 2019) although there are some differences in poorly supported relationships of closely related taxa.

### Coalescent-based species delimitation

Species delimitations were inconsistent among GMYC, PTP, and Tr2 methods in *Pinus* subsection *Cembroides*. Single-locus based coalescent analyses accurately delimited the species of the big-cone pinyons, *P. maximartinezii*, *P. pinceana*, and *P. rzedowskii*. These three Mexican endemics are easily recognized based on their morphologically divergent needles, wood anatomy, cones, and seeds (Malusa, 1992). Also, both GMYC and bPTP methods identified *P. monophylla*, *P. quadrifolia*, and *P. remota* as species. These taxa are consistently recognized in morphological treatments (Price et al., 1998) and have been recovered as distinct lineages in molecular phylogenetic studies (Gernandt et al., 2003; Montes et al., 2019). Our results support treating *P. quadrifolia* from southwestern USA and Baja California, Mexico as separate from the single-needle pinyon pines, *P. monophylla* from southwestern USA and *P. californiarum* distributed in California and Baja California. Lanner (1974) proposed that *P. quadrifolia* originates from recent hybridization

between *P. californiarum* and a five-needled species that he named *P. juarezensis* Lanner (which we treat as a synonym of *P. quadrifolia*). The evidence suggests that *P. quadrifolia* is not a hybrid (Montes et al., 2019), although *P. quadrifolia* show signs of admixture with some individuals of *P. californiarum* and *P. monophylla* (Buck et al., 2020).

GMYC analyses of plastid DNA recovered more lineages (up to 24; Figure 3) than bPTP and Tr2. Nonetheless, GMYC lumped together greater numbers of putative species in the small-cone clade, mainly in the “A” and “C” groups. Our results failed to separate *P. culminicola*, *P. discolor*, and *P. johannis* (Group A). *Pinus culminicola* has consistently been treated as a valid species, whereas *P. discolor* and *P. johannis* have been treated as separate species (Perry, 1991, Price et al., 1998) or as a single variety of *P. cembroides* (Farjon and Styles, 1997). The three taxa have been recovered as a clade in phylogenetic studies (Malusa, 1992; Gernandt et al., 2003; Ortiz-Medrano et al., 2016; Montes et al., 2019).

The single-locus methods GMYC and bPTP did not separate *P. californiarum*, *P. edulis*, and *P. fallax* (Group C). Montes et al. (2019) recognized *P. fallax* as a valid species based on its phylogenetic position in an analysis of nuclear genes but the results from the three species delimitation methods presented do not support that conclusion here (the nuclear results are discussed below). Henceforth, we refer to this taxon as *P. edulis* var. *fallax*, as originally proposed by Little (1968). *Pinus californiarum* and *P. edulis* have been recognized based on morphology, phylogeny, geographic distribution, and distinctive precipitation regimens (Bailey, 1987; Cole et al., 2008; Montes et al., 2019). However, the species of clade “C” cluster together and were identified as only one species by bPTP and GMYC. This result corroborates the study by LaHood (1995), who found that populations of *P. edulis*, *P. edulis* var. *fallax*, and *P. californiarum* share plastid DNA due to introgression.

Our results reflect both the nature of the molecular data and the limitations of both bPTP and GMYC methods. Hernández-León et al. (2013) suggested that including longer sequences of plastid DNA could improve species estimates with GMYC in another pine clade, *Pinus* sect. *Trifoliae* DuRoi, but despite obtaining relatively high resolution among terminals using nearly complete plastomes (~117,000 bp) in this study (Appendix S9), bPTP identified fewer species than expected whereas GMYC identified more species.

Our results suggest that even though sampling was increased and longer plastid DNA sequences were used, both GMYC and bPTP are not accurate for delimiting pine species. Many pine and other tree species have large effective population sizes (Petit and Hampe, 2006), and the discriminatory power of GMYC and bPTP in recognizing species with plastome data could be associated with different biological phenomena, such as plastid capture, gene flow, and population size (Luo et al., 2018).

Multi-locus analysis with the Tr2 method recovered 13 species (*P. californiarum*, *P. cembroides*, *P. culminicola*, *P. discolor*, *P. edulis*, *P. johannis*, *P. lagunae*, *P. maximartinezii*,

*P. monophylla*, *P. pinceana*, *P. quadrifolia*, *P. remota*, and *P. rzedowskii*). Moreover, our delimitation results with Tr2 and multivariate analyses agreed with recent phylogenetic analyses (Montes et al., 2019) but failed to identify *P. edulis* var. *fallax* as a distinct species. Although we included one individual from near the type locality in Gila, Arizona, it grouped with *P. edulis* from Utah and Arizona. Thus, our results support the hypothesis by Little (1968) that single-needle pinyon pines from Arizona are a taxonomic variety of *P. edulis*, which predominantly has two needles per fascicle. *Pinus edulis* and *P. edulis* var. *fallax* are parapatric in distribution (Malusa, 1992; Cole et al., 2008) and some populations co-occur in Arizona and New Mexico. Both taxa occupy areas with similar seasonal precipitation (Cole et al., 2008). Needle numbers seem to be associated with distinct precipitation regimens and seasonality, and particularly in *P. edulis* var. *fallax* the needle numbers could be an adaptation to water deficit in summer (Cole et al., 2008). Nevertheless, two individuals of *P. edulis* var. *fallax* with one and two-needles on the same tree (UT1 and UT2) grouped with *P. californiarum*. Bailey (1987) studied the morphological similarities mainly in needles, resin canals, cones, and seeds of single-needle pinyons from Arizona and concluded that they are a taxonomic variety of *P. californiarum* (predominantly one-needle).

The phylogenetic position of *P. edulis* var. *fallax* individuals could be reflecting the presence of shared ancestral polymorphism or introgression from *P. edulis* var. *fallax* into *P. californiarum*. However, the hypothesis of introgression from *P. edulis* var. *fallax* into *P. californiarum* was not recovered using phylogenetic network analysis (Than et al., 2008) of nuclear genes by Montes et al. (2019). The coalescent, paleoclimatic, ecological, and genetic evidence do not reinforce the species boundaries between *P. edulis* var. *fallax* and *P. californiarum*. Increased sampling of *P. edulis* var. *fallax* populations is required for a more complete perspective of this taxon.

In morphology-based classifications, *P. californiarum* and *P. monophylla* are united by possessing solitary needles (Malusa, 1992) and *P. californiarum* has been considered a taxonomic variety of *P. monophylla* (Silba, 1990). Bailey (1987) segregated *P. californiarum* from *P. monophylla* based on differences in fascicle sheath length, the number of leaf resin canals, the shape of the base of the seed cone, and seed size. Our results support Bailey (1987) in recognizing this taxon as a valid species. Coalescent phylogenetic analyses recovered *P. californiarum* as an independent lineage and sister to *P. monophylla* and *P. quadrifolia* (Figures 7 and 8 in Montes et al., 2019). *Pinus californiarum* occurs in southeastern and central California and northern Baja California (Silba, 1990) and co-occurs with *P. quadrifolia* in both regions. Although *P. californiarum* occurs in sympatry with *P. quadrifolia* at some sites, it is easy to distinguish one taxon from the other because *P. quadrifolia* has approximately four needles per fascicle. *Pinus monophylla* and *P. californiarum* both occur in California (Bailey, 1987) but are parapatric or allopatric in distribution (Critchfield and Little,

1966). The geographic distribution of *P. monophylla* extends to western Utah, northwestern Arizona, southern Idaho, and western Nevada (Critchfield and Little, 1966; Bailey, 1987). Both single-needle pinyon pines occur in regions with high winter precipitation (Cole et al., 2008). Morphological, phylogenetic, paleoclimatic, geographical, ecological, and genetic evidence support the recognition of *P. californiarum* and *P. monophylla* as separate species.

*Pinus quadrifolia* and *P. monophylla* present divergent morphology in the trunk, number of leaves per fascicle, number of rows of stomata on the leaves, and leaf anatomy (Bailey, 1987; Farjon and Styles, 1997). The species are genetically distinct (Montes et al., 2019; Buck et al., 2020) but in our PCA analysis, *P. quadrifolia* is indistinguishable from *P. monophylla*. In contrast, DAPC was capable of separating the two. *Pinus quadrifolia* is sister to *P. monophylla* and likely they share alleles due to shared ancestral polymorphism or introgression (see below).

*Pinus cembroides* is distributed in the southern USA and widespread in Mexico (Critchfield and Little, 1966). This species was segregated into three taxa (Bailey, 1983; Passini, 1987) based on morphology and distribution. All three taxa share characters including a tree growth form with a monopodial and short trunk, short (4–6 mm) and loosely imbricate fascicle sheaths, ovoid to cylindrical vegetative buds, rough shoots with non-decurrent or short decurrent pulvini, and a pinkish megagametophyte (Farjon and Styles, 1997). Our results support the treatment by Passini (1987) of *P. lagunae* as a distinct species from *P. cembroides* but do not support the proposal by Bailey and Hawksworth (1992) to elevate *P. cembroides* subsp. *orizabensis* to specific status. The *P. cembroides* subsp. *cembroides* + *P. cembroides* subsp. *orizabensis* clade is identified as a single species (*P. cembroides*).

The rooted triplets method had discriminatory power to delimit *P. lagunae* and *P. cembroides*. *Pinus lagunae* is endemic to Baja California Sur and is widely separated geographically from the rest of *P. cembroides*, whereas *P. cembroides* subsp. *cembroides* and *P. cembroides* subsp. *orizabensis* are allopatric or parapatric (Bailey, 1983). Based on morphology, *P. lagunae* differs from *P. cembroides* in height, substantially longer leaves (4–7 cm; Passini and Pinel, 1987), and in internal and external leaf cuticular characteristics such as an elliptical to rectangular stomatal apparatus, and circular stomata that usually lack a plug (Whang et al., 2001). *Pinus lagunae* occurs in a subtropical climate on granitic slopes whereas *P. cembroides* inhabits a broad elevational range with vegetation types ranging from semi-desert to montane forest (Bailey, 1983; Passini and Pinel, 1987; Farjon and Styles, 1997). Moreover, climatic changes during glacial or interglacial periods in the Pleistocene may have affected the geographic range and genetic composition of pine species (Moreno-Letelier and Piñero, 2009). We hypothesize that the divergence of *P. cembroides* and *P. lagunae* occurred during the Pleistocene and the current geographic distribution of *P. lagunae* may be the result of climatic fluctuations during

that time, resulting in an expansion of its range towards the south and its subsequent persistence in the southern Baja California refuge. A similar history was reported in the columnar cactus *Pachycereus pringlei* (S. Watson) Britton & Rose distributed in the Baja California Peninsula and the Sonoran Desert where the interaction of climatic fluctuations, historical vicariance, and dispersal can explain its current biogeographic pattern (Gutiérrez-Flores et al., 2016).

The rooted triplets method and multivariate analyses also were congruent in recovering *P. discolor* and *P. johannis* as separate species. Our results support previous studies (Flores-Rentería et al., 2013; Ortiz-Medrano et al., 2016; Montes et al., 2019) in treating *P. discolor* and *P. johannis* as distinct. *Pinus discolor* is not a synonym of *P. johannis* as suggested by Passini (1994). *Pinus johannis* and *P. discolor* are morphologically similar, although some differences have been identified. For instance, the number of cotyledons was reported as 6–11 in *P. johannis* and 8–15 in *P. discolor* (Robert, 1978; Little, 1968), height was reported as 2 to 6 m in *P. johannis* and 5 to 12 m in *P. discolor* (Bailey and Hawksworth, 1983), and growth form was described as a multi-stemmed shrub or tree in *P. johannis* compared to a tree in *P. discolor* (Bailey and Hawksworth, 1979) but these claims have caused confusion and disagreement. The two taxa have yet to evolve clear morphological differences although apparently, they are geographically separated. The distribution of *P. johannis* is restricted to the Sierra Madre Oriental whereas *P. discolor* occurs in the Sky Islands of the southwestern USA, the Sierra Madre Occidental, and the southern Sierra Madre Oriental in San Luis Potosí (Bailey and Hawksworth, 1979). Moreover, *P. johannis* develops on sand-textured lithic rendzina or calcareous soils (Robert, 1978) whereas *P. discolor* occurs on arid slopes and ridges (Farjon and Styles, 1997). Populations of *P. discolor* are adapted to a mild winter climate (Little, 1968), whereas *P. johannis* is adapted to a longer winter period from October to March (Robert, 1978).

*Pinus discolor* and *P. johannis* differ significantly in the concentration of sabinene-related monoterpenes such as sabinene, thujene,  $\gamma$ -terpinene, terpinolene, and *p*-cymene. *Pinus discolor* produces a higher quantity of these monoterpenes compared to *P. johannis*, which has them in trace amounts (Zavarin and Snajberk, 1986). Traditionally, terpenes have been used as a character to differentiate species of pines (Mitić et al., 2017) but the differences are not always significant at the species level because the composition of monoterpenes differs very little among species or there is great variability in composition of monoterpenes within species (Zavarin et al., 1980; Snajberk and Zavarin, 1986; Zavarin and Snajberk, 1987).

## Admixture analysis

Our results provide a picture of possible interbreeding in the three groups of small-cone pinyon pines: (1) *Pinus cembroides* subsp. *cembroides*, *P. cembroides* subsp.



*orizabensis*, and *P. lagunae*; (2) *P. johannis* and *P. discolor*; and (3) *P. californiarum*, *P. edulis* var. *fallax*, and *P. monophylla*. For all values of  $K$  analyzed (2 to 6), two subpopulations ( $K=2$ ) were the optimal clustering in this study (Figure 6). The number of subpopulations partitioned by the optimal clustering did not coincide with the number of lineages in the SNPs tree, particularly in the single-needled pinyons, *P. edulis*, and *P. quadrifolia* subgroup where five previously hypothesized taxa are considered (Figure 6) and the best model of  $K$  in admixture resulted in two subpopulations. Distribution of ancestry fractions indicate that *P. monophylla* (s.s.) shares genetic diversity with individuals of *P. quadrifolia*, *P. californiarum*, *P. edulis*, and *P. edulis* var. *fallax*. Buck et al. (2020) reported from 1868 SNPs that interbreeding does occur between *P. californiarum* and *P. quadrifolia*, as well as between *P. quadrifolia* and *P. monophylla*, and less commonly between *P. monophylla* and *P. californiarum*. Also, Montes et al. (2019) detected gene flow in *P. edulis* from *P. monophylla* using nuclear genes and SNPs. The other subpopulation indicates that individuals of *P. edulis* and *P. edulis* var. *fallax* are introgressed with *P. californiarum* but the direction of gene flow was not determined (Montes et al., 2019). Our results suggest higher genetic variation and admixture among the individuals of *P. edulis*, *P. edulis* var. *fallax*, and *P. californiarum* than those of *P. monophylla* and *P. quadrifolia*. Although interbreeding occurs between *P. monophylla* and *P. quadrifolia* it is less common (Buck et al., 2020). We observed genetic structure in *P. californiarum*, *P. monophylla*, and *P. quadrifolia* for  $K=3$  to  $K=6$ , supporting that they are genetically distinct species. Conversely, we did not observe genetic structure between *P. edulis* and *P. edulis* var. *fallax*.

Distribution of ancestry fractions indicates that *P. lagunae* is introgressed with *P. cembroides*. According to Montes et al. (2019), evidence of reticulation between *P. lagunae* and *P. cembroides* was weak but signatures of admixture could be the result of long-distance pollen dispersal. Both species release their pollen in a short interval of time from May to July (Farjon and Styles, 1997). However, *P. cembroides* and *P. lagunae* are sisters and it is more likely that their shared genetic diversity is due to retention of ancestral polymorphism.

Distribution of ancestry fractions indicates that *P. johannis* is introgressed with *P. discolor* but this was not detected by Montes et al. (2019). *Pinus discolor* and *P. johannis* are closely related and share little genetic diversity that can be caused by introgression or ILS. This sign of admixture occurs only in one individual of *P. johannis* from San Luis Potosí where *Pinus discolor* and *P. johannis* come into close contact, separated by a ca. 120 km between San Miguelito Mountains and Las Charcas. Our results showed genetic structure in these species and support that they are genetically distinct. Nonetheless, more field work is needed in San Luis Potosí to explore the evidence of historical contact between *P. discolor* and *P. johannis*.

## Perspectives for the use of coalescent-based methods of species delimitation

Employing data from target enrichment and genome skimming (Hyb-Seq; Weitemier et al., 2014) permitted us to delimit species using plastome and nuclear DNA sequences in a group of pines. Moreover, we examined the utility of using coalescent-based models to assess the boundaries with single-locus and multi-locus data. Our study revealed the potential of using single and multi-locus methods to estimate species in the presence of ILS and recent divergence (Gernandt et al., 2008; Montes et al., 2019). The multi-locus method Tr2 provided an estimate that better matched our expectations based on morphology, geography, and previous genetic studies.

## AUTHOR CONTRIBUTIONS

J.R.M. performed both field and laboratory work, assembled DNA sequences, performed the phylogenetic, delimitation, and multivariate analyses, and was the primary author for the manuscript. J.R.M., A.M.L., and D.S.G. designed the study and performed fieldwork. P.P. provided SNP data and performed admixture analyses. All authors reviewed and edited the manuscript.



## ACKNOWLEDGMENTS

The authors are grateful to Nelly López for assisting in genomic library construction and María Inés Badillo for a valuable revision of a previous version of the manuscript. We also thank José Delgadillo for providing logistical assistance during collecting trips to Baja California. We thank Jorge Pérez de la Rosa and Abisai García for sharing collections of *Pinus lagunae*, and Laura Figueroa for collecting and providing material of *Pinus pinceana*. We thank Xitlali Aguirre-Dugua, Eng. Mario S. Montes-Montiel, PT. Angélica Castolo P. for their participation in fieldwork. We also thank Dra. Lidia I. Cabrera and the LANABIO of the Instituto de Biología, UNAM. Two anonymous reviewers provided valuable comments on a previous draft of the manuscript. This project was funded by PAPIIT-DGAPA, UNAM Grant IN209816 and CONACyT Grant 221694, and is part of the Ph.D. dissertation of J. R. Montes in the Posgrado de Ciencias Biológicas, Instituto de Biología, UNAM.

## DATA AVAILABILITY STATEMENT

The nuclear, SNP, and plastid DNA sequence alignments analyzed in this study can be found in Appendices S14, S15, and S16, respectively.

## ORCID

José-Rubén Montes  <http://orcid.org/0000-0001-6441-5983>  
David S. Gernandt  <http://orcid.org/0000-0002-3592-994X>

## REFERENCES

- Aguirre-Dugua, X., and D. S. Gernandt. 2017. Complete plastomes of three endemic Mexican pine species (*Pinus* subsection *Australes*). *Mitochondrial DNA part B, Resources* 2: 562–565.

- Alexander, D. H., J. Novembre, and K. Lange. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* 19: 1655–1664.
- Andrews, S. 2010. FastQC: A quality control tool for high throughput sequence data. Website: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>
- Bailey, D. K. 1983. A new allopatric segregate form and a new combination in *Pinus cembroides* Zucc. at its southern limits. *Phytologia* 54: 89–100.
- Bailey, D. K. 1987. A study of *Pinus* subsection *Cembroides*. I: The single-needle pinyons of the Californias and the Great Basin. *Notes from the Royal Botanic Garden Edinburgh* 44: 275–310.
- Bailey, D. K., and F. G. Hawksworth. 1979. Pinyons of the Chihuahuan Desert region. *Phytologia* 44: 129–133.
- Bailey, D. K., and F. G. Hawksworth. 1983. Pinaceae of the Chihuahuan Desert region. *Phytologia* 53: 226–234.
- Bailey, D. K., and F. G. Hawksworth. 1992. Change in status of *Pinus cembroides* subsp. *orizabensis* (Pinaceae) from Central Mexico. *Novon* 2: 306–307.
- Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Leslin, et al. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* 19: 455–477.
- Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114–2120.
- Bryant, D., R. Bouckaert, J. Felsenstein, N. A. Rosenberg, and A. RoyChoudhury. 2012. Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution* 29: 1917–1932.
- Buck, R., S. Hyasat, A. Hossfeld, and L. Flores-Rentería. 2020. Patterns of hybridization and cryptic introgression among one- and four-needled pinyon pines. *Annals of Botany* 126: 401–411.
- 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., and L. Kubatko. 2014. Quartet inference from SNP data under the coalescent model. *Bioinformatics* 30: 3317–3324.
- Cole, K. L., J. F. Fisher, S. T. Arundel, J. Cannella, and S. Swift. 2008. Geographical and climatic limits of needle types of one- and two-needled pinyon pines. *Journal of Biogeography* 35: 257–269.
- Cole, K. L., J. F. Fisher, K. Ironside, J. I. Mead, and P. Koehler. 2013. The biogeographic histories of *Pinus edulis* and *Pinus monophylla* over the last 50,000 years. *Quaternary International* 310: 96e110.
- Critchfield, W. B., and E. L. Little. 1966. Geographic distribution of the pines of the world, Miscellaneous Publication 991. U.S. Department of Agriculture, Washington, District of Columbia, USA.
- Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, et al. 2011. The variant call format and VCFtools. *Bioinformatics* 27: 2156–2158.
- Darriba, D., G. L. Taboada, R. Doallo, and D. Posada. 2012. jModelTest 2: More models, new heuristics and parallel computing. *Nature Methods* 9: 772.
- de Queiroz, K. 2007. Species concepts and species delimitation. *Systematic Biology* 56: 879–886.
- Doyle, J. J., and J. L. Doyle. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* 19: 11–15.
- Ence, D. D., and B. C. Carstens. 2011. SpedeSTEM: A rapid and accurate method for species delimitation. *Molecular Ecology Resources* 11: 473–480.
- Ewels, P., M. Magnusson, S. Lundin, and M. Käller. 2016. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32: 3047–3048.
- Farjon, A., and D. Filer. 2013. An atlas of the world's conifers: An analysis of their distribution, biogeography, diversity and conservation status. Brill Press, Leiden, Netherlands.
- Farjon, A., and B. Styles. 1997. *Pinus* (Pinaceae). Flora neotropica monograph 75. New York Botanical Garden, Bronx, New York, USA.
- Feng, X., J. Liu, and X. Gong. 2016. Species delimitation of the *Cycas segmentifida* complex (Cycadaceae) resolved by phylogenetic and distance analyses of molecular data. *Frontiers in Plant Science*. 7: 134.
- Flores-Rentería, L., A. Wegier, D. Ortega Del Vecchyo, A. Ortiz-Medrano, D. Piñero, A. V. Whipple, F. Molina-Freaner, et al. 2013. Genetic, morphological, geographical and ecological approaches reveal phylogenetic relationships in complex groups, an example of recently diverged pinyon pine species (subsection *Cembroides*). *Molecular Phylogenetics and Evolution* 69: 940–949.
- Fujisawa, T., A. Aswad, and T. G. Barraclough. 2016. A rapid and scalable method for multilocus species delimitation using Bayesian model comparison and rooted triplets. *Systematic Biology* 65: 759–771.
- Fujita, M. K., A. D. Leaché, F. T. Burbrink, J. A. McGuire, and C. Moritz. 2012. Coalescent-based species delimitation in an integrative taxonomy. *Trends in Ecology & Evolution* 27: 480–488.
- Gernandt, D. S., X. Aguirre-Dugua, A. Vázquez-Lobo, A. Willyard, A. Moreno Letelier, J. A. Pérez de la Rosa, D. Piñero, and A. Liston. 2018. Multi-locus phylogenetics, lineage sorting, and reticulation in *Pinus* subsection *Australes*. *American Journal of Botany* 105: 711–725.
- Gernandt, D. S., G. Geada López, S. Ortiz García, and A. Liston. 2005. Phylogeny and classification of *Pinus*. *Taxon* 54: 29–42.
- Gernandt, D. S., A. Liston, and D. Piñero. 2001. Variation in the nrDNA ITS of *Pinus* subsection *Cembroides*: Implications for molecular systematic studies of pine species complexes. *Molecular Phylogenetics and Evolution* 21: 449–467.
- Gernandt, D. S., A. Liston, and D. Piñero. 2003. Phylogenetics of *Pinus* subsections *Cembroides* and *Nelsoniae* inferred from cpDNA sequences. *Systematic Botany* 28: 657–673.
- Gernandt, D. S., S. A. Magallón, G. Geada López, O. Zerón Flores, A. Willyard, and A. Liston. 2008. Use of simultaneous analyses to guide fossil-based calibrations of Pinaceae phylogeny. *International Journal of Plant Sciences* 169: 1086–1099.
- Gernandt, D. S., and J. A. Pérez de la Rosa. 2014. Biodiversity of Pinophyta (conifers) in Mexico. *Revista Mexicana de Biodiversidad* 85: 126–133.
- Grünwald, N. J., Z. N. Kamvar, and S. E. Everhart. 2016. Population genetics in R. Website: [https://grunwaldlab.github.io/Population\\_Genetics\\_in\\_R/gbs\\_analysis.html](https://grunwaldlab.github.io/Population_Genetics_in_R/gbs_analysis.html)
- Gutiérrez-Flores, C., F. J. García-De León, J. L. León-de la Luz, and J. H. Cota-Sánchez. 2016. Microsatellite genetic diversity and mating systems in the columnar cactus *Pachycereus pringlei* (Cactaceae). *Perspectives in Plant Ecology, Evolution and Systematics* 22: 1–10.
- Hernández-León, S., D. S. Gernandt, J. A. Pérez de la Rosa, and L. Jardón-Barbolla. 2013. Phylogenetic relationships and species delimitation in *Pinus* section *Trifoliae* inferred from plastid DNA. *PLoS One* 8: e70501.
- Huelsenbeck, J. P., and F. Ronquist. 2001. MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* 17: 754–755.
- IUCN. 2019. The IUCN red list of threatened species, version 2019-3. International Union for Conservation of Nature, Gland, Switzerland. Website: <http://www.iucnredlist.org>
- Johnson, M. G., E. M. Gardner, Y. Liu, R. Medina, B. Goffinet, A. J. Shaw, N. J. C. Zerega, et al. 2016. HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in Plant Sciences* 4: 1600016.
- Jones, G., Z. Aydin, and B. Oxelman. 2015. DISSECT: An assignment-free Bayesian discovery method for species delimitation under the multispecies coalescent. *Bioinformatics* 31: 991–998.
- Kamvar, Z. N., J. F. Tabima, and N. J. Grünwald. 2014. Poppr: An R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *Peer J* 2: e281.
- Kan, X.-Z., S.-S. Wang, X. Ding, and X.-Q. Wang. 2007. Structural evolution of nrDNA ITS in Pinaceae and its phylogenetic implications. *Molecular Phylogenetics and Evolution* 44: 765–777.
- Kapli, P., S. Lutteropp, J. Zhang, K. Kobert, P. Pavlidis, A. Stamatakis, and T. Flouri. 2017. Multi-rate Poisson tree processes for single-locus species delimitation under maximum likelihood and Markov chain Monte Carlo. *Bioinformatics* 33: 1630–1638.

- Katoh, K., K. Misawa, K.-I. Kuma, and T. Miyata. 2002. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* 30: 3059–3066.
- Kearse, M., R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, et al. 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28: 1647–1649.
- Kopelman, N. M., J. Mayzel, M. Jakobsson, N. A. Rosenberg, and I. Mayrose. 2015. Clumpak: A program for identifying clustering modes and packaging population structure inferences across *K*. *Molecular Ecology Resources* 15: 1179–1191.
- LaHood, E. 1995. A chloroplast DNA phylogeny of nine taxa in *Pinus Cembroides* subsection. MSc. thesis, Northern Arizona University, Flagstaff, Arizona, USA.
- Lanner, R. M. 1974. A new pine from Baja California and the hybrid origin of *Pinus quadrifolia*. *Southwestern Naturalist* 19: 75–95.
- Lanner, R. M. 1981. The piñon pine: A natural and cultural history. University of Nevada Press, Reno, Nevada, USA.
- Li, H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27: 2987–2993.
- Li, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Arxiv 1303.3997. Website: <https://arxiv.org/abs/1303.3997>
- Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754–1760.
- Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, et al. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25: 2078–2079.
- Liston, A., W. A. Robinson, D. Piñero, and E. R. Alvarez-Buylla. 1999. Phylogenetics of *Pinus* (Pinaceae) based on nuclear ribosomal DNA internal transcribed spacer region sequences. *Molecular Phylogenetics and Evolution* 11: 95–109.
- Little, E. L. 1968. Two new pinyon varieties from Arizona. *Phytologia* 17: 329–342.
- Long, C., and L. Kubatko. 2018. The effect of the gene flow on coalescent-based species-tree inference. *Systematic Biology* 67: 770–785.
- Long, C., and L. Kubatko. 2019. Identifiability and reconstructibility of species phylogenies under a modified coalescent. *Bulletin of Mathematical Biology* 81: 408–430.
- López-Giráldez, F., and J. P. Townsend. 2011. PhyDesign: An online application for profiling phylogenetic informativeness. *BMC Evolutionary Biology* 11: 1–4.
- López-Reyes, A., J. A. Pérez de la Rosa, E. Ortiz, and D. S. Gernandt. 2015. Morphological, molecular, and ecological divergence in *Pinus douglasiana* and *P. maximinoi*. *Systematic Botany* 40: 658–670.
- Luo, A., C. Ling, S. Y. W. Ho, and C.-D. Zhu. 2018. Comparison of methods for molecular species delimitation across a range of speciation scenarios. *Systematic Biology* 67: 830–846.
- Madrigal, S. X., and D. M. Caballero. 1969. Una nueva especie mexicana de *Pinus*. *Boletín Técnico del Instituto Nacional de Investigaciones Forestales* 26: 1–11.
- Malusa, J. 1992. Phylogeny and biogeography of the pinyon pines (*Pinus* subsect. *Cembroides*). *Systematic Botany* 17: 42–66.
- Miller, M. A., W. Pfeiffer, and T. Schwartz. 2010. Creating the CIPRES science gateway for inference of large phylogenetic trees. Proceedings of the gateway computing environments workshop (GCE) 14 November 2010, vol. 14, 1–8. Institute of Electrical and Electronics Engineers, New Orleans, Louisiana, USA. Website: <https://doi.org/10.1109/GCE.2010.5676129>
- Minh, B.Q., H.A. Schmidt, O. Chernomor, D. Schrempf, M.D. Woodhams, A. von Haeseler, and R. Lanfear. 2020. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution* 37: 1530–1534.
- Mirarab, S., and T. Warnow. 2015. ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* 31: i44–i52.
- Mirov, N. T. 1967. The genus *Pinus*. Ronald Press, New York, New York, USA.
- Mitić, Z. S., B. M. Nikolić, M. S. Ristić, V. V. Tešević, S. R. Bojović, and P. D. Marin. 2017. Terpenes as useful markers in differentiation of natural populations of relict pines *Pinus heldreichii*, *P. nigra*, and *P. peuce*. *Chemistry and Biodiversity* 14: e1700093.
- Monaghan, M. T., R. Wild, M. Elliot, T. Fujisawa, M. Balke, D. J. G. Inward, D. C. Lees, et al. 2009. Accelerated species inventory on Madagascar using coalescent-based models of species delineation. *Systematic Biology* 58: 298–311.
- Montes, J. R., P. Peláez, A. Willyard, A. Moreno-Letelier, D. Piñero, and D. S. Gernandt. 2019. Phylogenetics of *Pinus* subsection *Cembroides* Engelm. (Pinaceae) inferred from low-copy nuclear gene sequences. *Systematic Botany* 44: 501–518.
- Moreno-Letelier, A., A. Ortiz-Medrano, and D. Piñero. 2013. Niche divergence versus neutral processes: Combined environmental and genetic analyses identify contrasting patterns of differentiation in recently diverged pine species. *PLoS One* 8: e78228.
- Moreno-Letelier, A., and D. Piñero. 2009. Phylogeographic structure of *Pinus strobiformis* Engelm. across the Chihuahuan Desert filter-barrier. *Journal of Biogeography* 36: 121–131.
- Mort, M. E., J. K. Archibald, C. P. Randle, N. D. Levens, T. R. O'Leary, K. Topalov, C. M. Wiegand, et al. 2007. Inferring phylogeny at low taxonomic levels: Utility of rapidly evolving cpDNA and nuclear ITS loci. *American Journal of Botany* 94: 173–183.
- Neale, D. B., J. L. Wegrzyn, K. A. Stevens, A. V. Zimin, D. Puiu, M. W. Crepeau, C. Cardeno, et al. 2014. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biology* 15: R59.
- Neves, L. G., J. M. Davis, W. B. Barbazuk, and M. Kirst. 2013. Whole-exome targeted sequencing of the uncharacterized pine genome. *Plant Journal* 75: 146–156.
- Nobis, M. P., C. Traiser, and A. Roth-Nebelsick. 2012. Latitudinal variation in morphological traits of the genus *Pinus* and its relation to environmental and phylogenetic signals. *Plant Ecology and Diversity* 5: 1–11.
- O'Meara, B. C. 2010. New heuristic methods for joint species delimitation and species tree inference. *Systematic Biology* 59: 59–73.
- O'Meara, B., C. Ané, M. J. Sanderson, and P. C. Wainwright. 2006. Testing for different rates of continuous trait evolution using likelihood. *Evolution* 60: 922–933.
- Ortiz-Medrano, A., D. P. Scantlebury, A. Vázquez-Lobo, A. Mastretta-Yanes, and D. Piñero. 2016. Morphological and niche divergence of pinyon pines. *Ecology and Evolution* 6: 2886–2896.
- Paradis, E., and K. Schliep. 2019. Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35: 526–528.
- Parks, M., R. Cronn, and A. Liston. 2012. Separating the wheat from the chaff: Mitigating the effects of noise in a plastome phylogenomic data set from *Pinus* L. (Pinaceae). *BMC Evolutionary Biology* 12: 1–17.
- Passini, M.-F. 1987. The endemic pinyon of Lower California *Pinus lagunae* M.-F. Passini. *Phytologia* 63: 337–338.
- Passini, M.-F. 1994. Synonymie entre *Pinus discolor* Bailey & Hawsworth et *Pinus johannis* M.-F. Robert. *Acta Botanica Gallica* 141: 387–388.
- Passini, M.-F., and N. Pinel. 1987. Morphology and phenology of *Pinus lagunae* M.-F. Passini. *Phytologia* 63: 331–336.
- Perry, J. P. 1991. The pines of Mexico and Central America. Timber Press, Portland, Oregon, USA.
- Petit, R. J., and A. Hampe. 2006. Some evolutionary consequences of being a tree. *Annual Review of Ecology, Evolution, and Systematics* 37: 187–214.
- Pons, J., T. G. Barraclough, J. Gomez-Zurita, A. Cardoso, D. P. Duran, S. Hazell, S. Kamoun, et al. 2006. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology* 55: 595–609.
- Poulos, H. M. and G. P. Berlyn. 2007. Variability in needle morphology and water status of *Pinus cembroides* across an elevational gradient in the Davis Mountains of west Texas, USA. *Journal of the Torrey Botanical Society* 134: 281–288.



- Price, R. A., A. Liston, and S. H. Strauss. 1998. Phylogeny and systematics of *Pinus*. In D. M. Richardson [ed.], *Ecology and biogeography of Pinus*, 49–68. Cambridge University Press, Cambridge, UK.
- Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155: 945–959.
- Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, et al. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* 81: 559–575.
- Rambaut, A. 2012. FigTree. Tree figure drawing tool, version 1.4.0. Website: <https://github.com/rambaut/figtree/releases>
- Rambaut, A., A. J. Drummond, D. Xie, G. Baele, and M. A. Suchard. 2018. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology* 67: 901–904.
- Rieseberg, L. H., and D. E. Soltis. 1991. Phylogenetic consequences of cytoplasmic gene flow in plants. *Evolutionary Trends in Plants* 5: 65–84.
- Robert, M.-F. 1978. Un nouveau pin pignon mexicain: *Pinus johannis*. *Adansonia*, ser. 2, 18: 365–373.
- Rosenberg, N. A. 2003. The shapes of neutral gene genealogies in two species: Probabilities of monophyly, paraphyly and polyphyly in a coalescent model. *Evolution* 57: 1465–1477.
- RStudio team. 2019. RStudio: Integrated development for R. Rstudio Inc., Boston, Massachusetts, USA. Website: <http://www.rstudio.com/>
- Schlick-Steiner, B. C., F. M. Steiner, B. Seifert, C. Stauffer, E. Christian, and R. H. Crozier. 2010. Integrative taxonomy: A multisource approach to exploring biodiversity. *Annual Review of Entomology* 55: 421–438.
- SEMARNAT. 2010. Norma Oficial Mexicana NOM-059-ECOL-2001. Protección ambiental. Especies nativas de México de flora y fauna silvestres. Categorías de riesgo y especificaciones para su inclusión, exclusión o cambio. Lista de especies en riesgo. Diario Oficial de la Federación, 30 de diciembre de 2010. Mexico City, Mexico.
- Silba, J. 1990. A supplement to the international census of the Coniferae, II. *Phytologia* 68: 7–78.
- Sites, J. W., and K. A. Crandall. 1997. Testing species boundaries in biodiversity studies. *Conservation Biology* 11: 1289–1297.
- Sites, J. W., and J. C. Marshall. 2003. Delimiting species: A renaissance issue in systematic biology. *Trends in Ecology and Evolution* 18: 462–470.
- Smith, S. A., M. J. Moore, J. W. Brown, and Y. Yang. 2015. Analysis of phylogenomic datasets reveals conflict, concordance and gene duplications with examples from animals and plants. *BMC Evolutionary Biology* 15: 150.
- Smith S. A., N. Walker-Hale, J. F. Walker, and J. W. Brown. 2020. Phylogenetic conflicts, combinability, and deep phylogenomics in plants. *Systematic Biology* 69: 579–592.
- Snajberk, K., and E. Zavarin. 1986. Monoterpenoid differentiation in relation to the morphology of *Pinus remota*. *Biochemical Systematics and Ecology* 14: 155–163.
- Stamatakis, A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312–1313.
- Swofford, D. L. 2002. PAUP\* Phylogenetic analysis using parsimony (\*and other methods), v. 4.0b10. Sinauer Associates, Sunderland, Massachusetts, USA.
- Syring, J., A. Willyard, R. Cronn, and A. Liston. 2005. Evolutionary relationships among *Pinus* (Pinaceae) subsections inferred from multiple low-copy nuclear loci. *American Journal of Botany* 92: 2086–2100.
- Than, C., D. Ruths, and L. Nakhleh. 2008. PhyloNet: A software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics* 9: 322.
- Townsend, J. P. 2007. Profiling phylogenetic informativeness. *Systematic Biology* 56: 222–231.
- Tsutsui, K., A. Suwa, K. Sawada, T. Kato, T. A. Ohsawa, and Y. Watano. 2009. Incongruence among mitochondrial, chloroplast and nuclear gene trees in *Pinus* subgenus *Strobos* (Pinaceae). *Journal of Plant Research* 122: 509–521.
- Turna, I., and D. Güney. 2009. Altitudinal variation of some morphological characters of Scots pine (*Pinus sylvestris* L.) in Turkey. *African Journal of Biotechnology* 8: 202–208.
- Villanueva, R. A. M., Z. J. Chen, and H. Wickham. 2016. Ggplot2: Elegant graphics for data analysis using the grammar of graphics. Springer-Verlag, New York, New York, USA.
- Wang, J., Y. Wu, G. Ren, Q. Guo, J. Liu, and M. Lascoux. 2011. Genetic differentiation and delimitation between ecologically diverged *Populus euphratica* and *P. pruinosa*. *PLoS One* 6: e26530.
- Wegrzyn, J. L., J. D. Liechty, K. A. Stevens, L-S. Wu, C. A. Loopstra, H. A. Vasquez-Gross, W. M. Dougherty, et al. 2014. Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. *Genetics* 196: 891–909.
- Weitemier, K., S. C. K. Straub, R. C. Cronn, M. Fishbein, R. Schmickl, A. McDonnell, and A. Liston. 2014. Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences* 2: 1400042.
- Whang, S. S., J-H. Pak, R. S. Hill, and K. Kim. 2001. Cuticle micromorphology of leaves of *Pinus* (Pinaceae) from Mexico and Central America. *Botanical Journal of the Linnean Society, Linnean Society of London* 135: 349–373.
- Willyard, A., D. S. Gernandt, B. Cooper, C. Douglas, K. Finch, H. Karemera, E. Lindberg, et al. 2021. Phylogenomics in the hard pines (*Pinus* subsection Ponderosae; Pinaceae) confirms paraphyly in *Pinus ponderosa*, and places *Pinus jeffreyi* with the California big cone pines. *Systematic Botany* 46: 538–561.
- Willyard, A., D. S. Gernandt, K. Potter, V. Hipkins, P. Marquardt, M. F. Mahalovich, S. K. Langer, et al. 2017. *Pinus ponderosa*: A checkered past obscured four species. *American Journal of Botany* 104: 161–181.
- Willyard, A., J. Syring, D. S. Gernandt, A. Liston, and R. Cronn. 2007. Fossil calibration of molecular divergence infers a moderate mutation rate and recent radiations for *Pinus*. *Molecular Biology and Evolution* 24: 90–101.
- Wolfe, J. A., and H. E. Schorn. 1990. Taxonomic revision of the Spermatopsida of the Oligocene Creede flora, Southern Colorado. USGS Bulletin 1923. U.S. Geological Survey, Denver, Colorado, USA.
- Yang, Z. 2015. The BPP program for species tree estimation and species delimitation. *Current Zoology* 61: 854–865.
- Yang, Z., and B. Rannala. 2010. Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences, USA* 107: 9264–9269.
- Zavarin, E., and K. Snajberk. 1986. Monoterpenoid differentiation in relation to the morphology of *Pinus discolor* and *Pinus johannis*. *Biochemical Systematics and Ecology* 14: 1–11.
- Zavarin, E., and K. Snajberk. 1987. Monoterpene differentiation in relation to the morphology of *Pinus culminicola*, *Pinus nelsonii*, *Pinus pinceana* and *Pinus maximartinezii*. *Biochemical Systematics and Ecology* 15: 307–312.
- Zavarin, E., K. Snajberk, and R. Debry. 1980. Terpenoid and morphological variability of *Pinus quadrifolia* and its natural hybridization with *Pinus monophylla* in northern Baja California and adjoining United States. *Biochemical Systematics and Ecology* 8: 225–235.
- Zhang, C., M. Rabiee, E. Sayyari, and S. Mirarab. 2018. ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* 19: 15–30.
- Zhang, D., T. Xia, M. Yan, X. Dai, J. Xu, S. Li, and T. Yin. 2014. Genetic introgression and species boundary of two geographically overlapping pine species revealed by molecular markers. *PLoS One* 9: e101106.
- Zhang, J., P. Kapli, P. Pavlidis, and A. Stamatakis. 2013. A general species delimitation method with applications to phylogenetic placements. *Bioinformatics* 22: 2869–2876.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.



**Appendix S1.** List of individual taxa used in this study. The ID locality is based in ISO 3166-2, RENAPO, MX.

**Appendix S2.** Details for designing target DNA enrichment probes for conifers and processing of Hyb-Seq data.

**Appendix S3.** Conditions and concentrations of the parallel sequencing per sample. Sequencing depth = unenriched:enriched. The ID locality is based in ISO 3166-2, RENAPO, MX.

**Appendix S4.** Informativeness profile. (A) Plot of the genes list identified in PhyDesign with unusually high inferred substitution rates. Colored lines represent nuclear genes. (B) List of nuclear genes with unusually high inferred substitution rates. A total of 22 genes were eliminated from sampling.

**Appendix S5.** List of the individual taxa and statistics for plastome coverage. \*\*Plastomes assembled and used in this study.

**Appendix S6.** Species delimitation hypotheses in Tr2. The hypothesis was tested according to the species reported for subsection *Cembroides* in several studies (Lanner, 1974; Bailey, 1987; Price et al., 1998; Gernandt et al., 2005; Montes et al., 2019). The numbers represent lineages.

**Appendix S7.** Maximum likelihood tree based on low-copy nuclear genes. Results presented are based on the concatenated analysis with partitions. Bootstrap support values >50% are shown on branches. The abbreviations of the states are based on ISO3166-2:MX. Subsections are colored, *Cembroides* by blue branches, *Balfourianae* by orange, *Nelsoniae* by green, *Gerardianae* by purple, *Krempfianae* by navy blue, and *Strobus* by gray.

**Appendix S8.** Maximum likelihood tree based on low-copy nuclear genes. Results presented are based on the concatenated analysis without partitions. Bootstrap support values >50% are shown on branches. The abbreviations of the states are based on ISO3166-2:MX. Subsections are colored, *Cembroides* by blue branches, *Balfourianae* by orange, *Nelsoniae* by green, *Gerardianae* by purple, *Krempfianae* by navy blue, and *Strobus* by gray.

**Appendix S9.** Bayesian inference (BI) tree based on plastome sequences. We included plastomes with coverage >20× with some exceptions reported in this appendix. Posterior probability (PP) values > 0.5 from the BI analysis are shown on branches. The abbreviations of the states are based on ISO3166-2:MX.

**Appendix S10.** Generalized Mixed Yule Coalescent using single thresholds models. Results presented are based on the plastome tree resulting from Bayesian inference (BI). Colored areas corresponding to each cluster are identified as potential species. The blue line represents the transition between coalescence and speciation. The corresponding lineage-through-time plot is given on the upper left. The coalescent model of sGMYP exhibited significantly better fit over the null model ( $\log L_{\text{GMYP}} = -529.7634$ ,  $\log L_{\text{null}} = -524.9513$ ,  $\text{LRt} = 9.624205$ ,  $P = 0.008130748^{**}$ )

**Appendix S11.** Generalized Mixed Yule Coalescent using multiple thresholds models. Results presented are based on the plastome tree resulting from Bayesian inference (BI). Colored areas corresponding to each cluster identified as potential species. The corresponding lineage-through-time plot is given on the upper left. The coalescent model of mGMYP exhibited significantly better fit than the null model ( $\log L_{\text{GMYP}} = -530.1472$ ,  $\log L_{\text{null}} = -524.9513$ ,  $\text{LRt} = 10.39178$ ,  $P = 0.005539288^{**}$ ).

**Appendix S12.** Poisson tree processes using Bayesian inference solution. Results presented are based on the plastome tree resulting from Bayesian inference (BI). Colored areas corresponding to each cluster are identified as potential species. The likelihood plot indicated convergence of Markovian chains on the upper left. Bootstrap values (left) and posterior probability (right) are shown on branches.

**Appendix S13.** Poisson tree processes using maximum likelihood solution. Results presented are based on the plastome tree resulting from Bayesian inference (BI). Colored areas corresponding to each cluster identified as potential species. Bootstrap values are shown on branches.

**Appendix S14.** Concatenated alignment of nuclear genes.

**Appendix S15.** Concatenated alignment of SNPs.

**Appendix S16.** Plastome alignment with partitions, coding, and noncoding blocks.

**How to cite this article:** Montes, J.-R., P. Peláez, A. Moreno-Letelier, and D. S. Gernandt. 2022. Coalescent-based species delimitation in North American pinyon pines using low-copy nuclear genes and plastomes. *American Journal of Botany* 109(5): 706–726. <https://doi.org/10.1002/ajb2.1847>