Syst. Biol. 0(0):1–17, 2022

https://doi.org/10.1093/sysbio/syac017
Advance Access publication March 4, 2022

Origin and Diversification of the Saguaro Cactus (*Carnegiea gigantea*): A Within-Species Phylogenomic Analysis

MICHAEL J. SANDERSON[1,*], ALBERTO BÚRQUEZ[2], DARIO COPETTI[3], MICHELLE M. MCMAHON[4], YICHAO ZENG[1], AND MARTIN F. WOJCIECHOWSKI[5]

[1]*Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA;* [2]*Instituto de Ecología, Unidad Hermosillo, Universidad Nacional Autónoma de México, Hermosillo, Sonora, Mexico;* [3]*Arizona Genomics Institute, School of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA;* [4]*School of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA and* [5]*School of Life Sciences, Arizona State University, Tempe, AZ 85287, USA*

**Correspondence to be sent to: Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA; E-mail: sanderm@email.arizona.edu.*

Received 25 March 2020; reviews returned 18 February 2022; accepted 25 February 2022
Associate Editor: Claudia Solis-Lemus

Abstract.—Reconstructing accurate historical relationships within a species poses numerous challenges, not least in many plant groups in which gene flow is high enough to extend well beyond species boundaries. Nonetheless, the extent of tree-like history within a species is an empirical question on which it is now possible to bring large amounts of genome sequence to bear. We assess phylogenetic structure across the geographic range of the saguaro cactus, an emblematic member of Cactaceae, a clade known for extensive hybridization and porous species boundaries. Using 200 Gb of whole genome resequencing data from 20 individuals sampled from 10 localities, we assembled two data sets comprising 150,000 biallelic single nucleotide polymorphisms (SNPs) from protein coding sequences. From these, we inferred within-species trees and evaluated their significance and robustness using five qualitatively different inference methods. Despite the low sequence diversity, large census population sizes, and presence of wide-ranging pollen and seed dispersal agents, phylogenetic trees were well resolved and highly consistent across both data sets and all methods. We inferred that the most likely root, based on marginal likelihood comparisons, is to the east and south of the region of highest genetic diversity, which lies along the coast of the Gulf of California in Sonora, Mexico. Together with striking decreases in marginal likelihood found to the north, this supports hypotheses that saguaro's current range reflects postglacial expansion from the refugia in the south of its range. We conclude with observations about practical and theoretical issues raised by phylogenomic data sets within species, in which SNP-based methods must be used rather than gene tree methods that are widely used when sequence divergence is higher. These include computational scalability, inference of gene flow, and proper assessment of statistical support in the presence of linkage effects. [Phylogenomics; phylogeography; rooting; Sonoran Desert.]

Some of the first quantitative approaches to phylogenetic inference focused on inferring population trees—trees within species in which each leaf represents a population of individuals (Edwards and Cavalli-Sforza 1964; Thompson 1975). With the increasing availability of genomic data in recent years, many other approaches have been developed for inferring within-species trees (Bryant et al. 2012; Pickrell and Pritchard 2012; De Maio et al. 2015; Hey et al. 2018; Zhu and Nakhleh 2018). However, though phylogenetic trees of alleles (or individual nucleotide sites) in a genome are usually well-defined within species (Kingman 1982; Hein et al. 2005), trees within and between populations in sexually reproducing species may not be. Much depends on factors such as migration, gene flow, demography and genomic architecture—and the kind and extent of genome sequence data that can be used (Cutter 2013; Leaché et al. 2014; Hey et al. 2018; Long and Kubatko 2018; Shi and Yang 2018; Thawornwattana et al. 2018; Bravo et al. 2019; Li et al. 2019; Mason et al. 2020; Olave and Meyer 2020). Integrative approaches have been proposed to tease apart these factors, ranging from those with strong emphasis on modeling continuous gene flow, demography, and population genetics in a relatively small number of populations (Gutenkunst et al. 2009; Jouganous et al. 2017; Hey et al. 2018; Jones 2019), to those that emphasize inferring the topology of a population network with discrete gene flow events (Pickrell and Pritchard 2012; Zhu and Nakhleh 2018; Flouri et al. 2020). Much of the latter work extends the extensive work on phylogenetic methods for species tree inference under the multispecies coalescent (MSC) model (Degnan and Rosenberg 2009; Xu and Yang 2016; Liu et al. 2019).

How tree-like a species' history is between its populations might affect the strength of downstream inferences based on that history, including those concerning phylogeography or the location of the root. If a phylogenetic history is better modeled as a rooted network rather than tree (Huson et al. 2010; Pickrell and Pritchard 2012; Zhu and Nakhleh 2018), this network has potentially many more "parameters" which must be inferred from the same amount of data. Though the data might support this with a higher likelihood score (Solís-Lemus and Ané 2016; Wen et al. 2018; Blair and Ané 2020), the bias–variance tradeoff (Posada and Buckley 2004; Burnham and Anderson 2010; Wen et al. 2018) could have ripple effects on analyses that depend on these results. Consequently, a reasonable first step in empirical studies is to assess the robustness of the simpler tree-like model.

Some of the factors that can add both to the complexity of actual within-species histories and impede their accurate reconstruction are of particular concern in plants. A priori the saguaro cactus (*Carnegiea gigantea* (Engelman.) Britton & Rose) would appear to raise the

spectre of a number of these challenges: it belongs to the Cactaceae, a clade notorious for hybridization and indistinct species boundaries; it is broadly distributed over a range extending north to south over 1000 km, with many populations having large census sizes (O'Brien and Swann 2021); and its pollen and seeds are dispersed by wide-ranging volant bats and birds, some known to travel tens of km a day (Goldshtein et al. 2020). All these factors limit local isolation and differentiation and might be expected to foster nontree-like histories.

In addition, low sequence divergence within or between species can limit the resolving power of phylogenomic methods to infer complex histories. Whether phylogenomic methods use inferred gene trees as input (Maddison 1997; Liu et al. 2019) or gene alignments directly (Rannala and Yang 2017; Hey et al. 2018; Flouri et al. 2020), accurate species tree reconstruction in the context of low variation requires increasing numbers of genes (Dasarathy et al. 2015). Of the thousands of gene trees constructed by Copetti et al. (2017) from loci in saguaro and four other cacti, a sizable fraction were still poorly supported because of low sequence divergence even after tens of millions of years of divergence.

The first goal of this article is to use whole genome resequencing data to test whether the low sequence divergence observed between saguaro and its relatives is matched by low sequence diversity within saguaro. Our data comprise widely scattered single nucleotide polymorphisms (SNPs) across a fairly large genome assembly (1 Gbp). We will see that these are relatively rare: roughly five SNPs per gene (150,000 sites, 28,293 genes; see Results section), which is likely below what would be needed to reconstruct a large number of well-supported gene trees (Yang 1998). The second goal therefore is to exploit phylogenomic methods aimed explicitly at such dispersed SNPs ("SNP methods") to assess the extent of tree-like history within saguaro. Finally, the third goal, assuming a tree can be inferred reliably, is to infer the root of the tree from the within-species sequence data, which bears on long-standing hypotheses about climatic impacts on saguaro's geographic range (Nason et al. 2002; Bustamante et al. 2016). Like other recent phylogenomic studies of plants (Pease et al. 2016; Stein et al. 2018; Wang et al. 2020; Brandrud et al. 2020; Choi et al. 2021; Meleshko et al. 2021), an overarching goal is to assess how well the quantity of genome sequence data can overcome some of the many potential obstacles to phylogenetic inference mentioned above.

For the second goal of the article, to assess the strength and form of treelike signal across the species range, we used a diverse suite of five SNP methods. Three of these, PoMo (De Maio et al. 2015; Schrempf et al. 2016), SVDquartets (Chifman and Kubatko 2014), and SNAPP (Bryant et al. 2012), have been developed in the context of the mathematical assumptions of the multi*species* coalescent model, but are frequently used within species to infer population trees. Another method, TreeMix (Pickrell and Pritchard 2012) uses a model of continuous gene frequency change with migration episodes. Finally, we show how to "recycle" an older method, polymorphism parsimony (PP: Felsenstein, 1979) that provides a computationally efficient solution to tree building by minimizing deep coalescences (MDC), a discrete parsimony analog of the MSC (Maddison 1997; Ma et al. 2001; Maddison and Knowles 2006; Than and Nakhleh 2009; Bansal et al. 2010; Zhang 2011).

Finally, for the third goal of the article, to assess the information content of the genome data for inferring the root of the phylogenetic tree, we used two of these methods, SNAPP and PP, both of which permit inferences without outgroup data. Outgroup rooting of within-species trees in saguaro is problematic. Genome assemblies available for potential outgroup cacti related to saguaro are based on lower coverage sequencing and are much more fragmented (Copetti et al. 2017) until a distant relative, *Hylocereus undatus*, is reached at a divergence of perhaps 20 million years (Zheng et al. 2021).

Although the problem of rooting trees from ingroup sequence data alone is well known (Huelsenbeck et al. 2002; Naser-Khdour et al. 2021), many theoretical and computational questions remain, especially in the context of dispersed SNP data. Felsenstein (1981) showed that the likelihood score for a model and gene tree was independent of the root location of that gene tree—and therefore not identifiable from the alignment alone—if the model was (i) mathematically "reversible" and (ii) there were no constraints on the edge lengths of the tree, such as those imposed by a molecular clock assumption. Much work has focused on clock or relaxed-clock edge constraints (Drummond et al. 2006; Tria et al. 2017), and recent studies have shown that nonreversible models can provide reasonable evidence about the root (Naser-Khdour et al. 2021; Bettisworth and Stamatakis 2021; Cherlin et al. 2018). The root can also be inferred from the interplay between gene trees and species trees, such as with gene duplication, and via the MSC model, even if gene trees are unrooted Allman et al. (2011). Parsimony methods can also be used for rooting when there is an asymmetric step matrix (Huelsenbeck et al. 2002), and parsimony-like criteria such as minimizing the number of deep coalescences in species tree inference can also provide information about the root (Yu et al., 2011, though see also Alanzi and Degnan, 2017). Many of these theoretical results should carry over to SNP methods, but formal proofs are lacking. We use the present analysis to evaluate some of these issues empirically, alongside drawing inferences specific to saguaro phylogeny.

Materials and Methods

Sampling and Genome Sequence Data

We surveyed genomic variation across the geographic range of saguaro cactus (*Carnegiea gigantea* (Engelman.) Britton & Rose) using short read Illumina genome sequence data (Table 1). We sampled individuals from each of 10 widely separated localities across the range

TABLE 1. Sequencing read statistics for samples

Sample Name	Population	Reads ($\times 10^6$)	Total Bases ($\times 10^9$)	Coverage ($\times$)	Mean Length (base pairs)	Mean Quality Score
Sample_C60	Caborca	87.18	11.64	8.3	133	33
Sample_C61	Caborca	92.95	12.28	8.8	132	33
Sample_D17	El Dipo	65.93	6.26	4.5	94	37
Sample_D73	El Dipo	74.01	7.05	5.0	95	37
Sample_G50	Guásimas	83.68	11.19	8.0	133	33
Sample_G81	Guásimas	81.44	10.90	7.8	133	33
Sample_K26	Kino Bay	87.85	11.78	8.4	134	33
Sample_K4	Kino Bay	56.04	7.44	5.3	132	33
Sample_J34	La Joyita	59.52	5.67	4.0	95	37
Sample_J45	La Joyita	65.96	6.29	4.5	95	37
Sample_M69	Masiaca	71.93	6.84	4.9	95	37
Sample_M85	Masiaca	43.39	4.12	2.9	94	37
SGP-OR3	Orégano	89.82	11.69	8.3	130	34
SGP-OR6	Orégano	76.94	9.99	7.1	129	34
SGP-SM1	San Marcial	28.90	3.78	2.7	130	34
SGP-SM7	San Marcial	22.98	3.03	2.2	131	34
SGP5_L004	Tucson	256.20	24.49	17.5	95	38
Sample_CU	Tucson	76.08	7.24	5.2	95	37
Sample_119	Wickenburg	110.12	10.48	7.5	95	37
Sample_129	Wickenburg	73.01	6.96	5.0	95	37

of saguaro. The minimum distance between localities ranged from 64 km (San Marcial-Orégano) to 286 km (Tucson-Wickenburg), and the north–south maximum extent between localities was nearly 900 km. We sampled tissue from at least eight individuals at each locality separated by less than 150 m between them and selected two from each for whole genome resequencing. Operationally, each locality was considered to be a local "population" in the sense required for within-species tree inference methods that pool SNPs from individuals. We therefore use the term "population" synonymously with "locality" to describe those phylogenetic analyses. This design aimed to maximize sequencing coverage and thereby quality of SNP calling, as well as geographic coverage, but was inadequate to test hypotheses *within* these putative populations.

We used FASTQC v.0.11.8 (Andrews 2018) to examine quality statistics for all read sets and Trimmomatic 0.38 (Bolger et al. 2014) to trim low quality ends. Options for the latter were "ILLUMINACLIP:TruSeq3-PE:2:30:10 HEADCROP:3 LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:36," except "CROP:147" was appended for low quality 3′ ends found in the Orégano and San Marcial samples. Trimmed reads were mapped to the saguaro SGP5 v.1.3 assembly ("SGP5" henceforth; Copetti et al. 2017; GenBank accession GCA_002740515.1) using the Bowtie2 v. 2.3.4.3 (Langmead and Salzberg 2012) short read aligner in paired end mode, discarding unaligned, mixed, and discordant pairs. To remove experimentally produced duplicate reads, deduplication was implemented with SAMtools v. 1.9. (Li et al. 2009; Danecek et al. 2021) with the command sequence (sort, fixmate, sort, markup-r) described in their documentation, and resulting BAM files were indexed with SAMtools. Statistics for resulting BAM files were obtained using the SAMtools "stats" command, and coverage was computed based on a SGP5 genome size of 1.403 Gbp from the C-value database (Bennett and Leitch 2012). All reads have been deposited in the NCBI SRA archive under BioProject accession number PRJNA767819.

Two subsets of sites in gene regions were extracted from BAM files based on the SGP5 annotation (Copetti et al. 2017). The 0-fold and 4-fold degenerate sites in codons of all protein coding genes ("0-fold" and "4-fold" data sets) were identifiable with high confidence, because their annotations stemmed from both evidence-based and gene modeling methods. They were extracted using a custom PERL script adapted from "identify_4D_sites.pl" (Sackton 2014). All custom scripts are available with the Supplementary material available on Dryad at https://doi.org/10.5061/dryad.wdbrv15q4.

Genotypes were inferred from the 20 BAM files with BCFtools (Li et al., 2009; Danecek et al., 2021: options: mpileup -q 20 -Q 20 -a "FORMAT/DP"), followed by BCFtools (options: call -m) to create files with all reads. The FORMAT option allows downstream filtering by individual and site depth. Each of these files was then hard filtered by BCFtools for genotype call quality and to remove indels (options: -filter –SnpGap 3 -i "%QUAL>20 && TYPE != 'indel'"). Resulting VCF files were normalized with BCFtools (options: norm -d all) and checked again for duplicates with a custom PERL script. Further filtering by coverage depth and missing genotype calls was effected with our VCFParse.pl script. In general, a genotype call was required to have coverage of between 2–100$\times$ to be considered nonmissing. Analyses in the pooled sample of 20 were required to have 13 of 20 genotype calls present at a site. The individual population samples were each required to have genotypes present for both individuals to keep a site. These were obtained by using BCFtools (option: view -S) to subsample the VCF file for the entire data set.

Genetic diversity.—Genetic diversity estimates were obtained from genotype calls derived from the BCFtools pipeline described above. Both Tajima's and Watterson's nucleotide diversity estimators (π_T and π_W, respectively; Tajima, 1983; Watterson, 1975) were obtained with our VCFParse.pl script. Watterson's estimator was corrected for missing genotypes using the formula in Ferretti et al. (2012). Missing individuals at a site were omitted in the π_T calculation.

The spatial distribution of genomic diversity was interpolated using point kriging with no drifts with a linear variogram and no nugget effect as recommended for small, noisy samples. The grid was generated using Surfer (Golden Software, Golden, CO, USA) overlaid on a shaded relief base map derived from GMTED2010 (Danielson and Gesch 2011).

Phylogenomic Analysis

Tree terminology.—In various contexts, we consider three kinds of phylogenetic trees depending on what the leaves represent. For any data set of N diploid individuals in M populations, these may be (i) an *allele tree* with $2N$ leaves; (ii) a *population tree* with M leaves; and (iii) a *genotype tree* or *individual tree* having N leaves, one per individual organism. The latter commonly arises implicitly when only a single individual is sampled per population or species but might theoretically be relevant even with broader sampling if there is phylogenetic structure within defined "populations." (In the case of a diploid, the individual then is a surrogate for a "population" of two haploid alleles.)

Tree inference methods.—We used five methods to infer trees. These differ in whether they infer trees of populations or individuals, their computational overhead, the underlying evolutionary models used, their inference framework, whether they infer a rooted tree, and whether they model migration/gene flow.

The first three methods return unrooted trees and/or networks.

(i) *PoMo*. PoMo (De Maio et al. 2015; Schrempf et al. 2016; Schrempf et al. 2019) uses a reversible stochastic model of changes in nucleotide allele counts within and between species or populations caused by drift and mutation to infer a species/population tree. For example, in a population of $N=4$ alleles, evolution from all A in the population to all G involves both a mutation and a stepwise process of genetic drift traversing through each polymorphic state of $n=3,2,1$ A alleles, until all are lost. We used the implementation of PoMo in IQTREE 2 (Minh et al. 2020). Genotypes within each population were combined to generate required allele counts following the PoMo input format. Model options were set to "-m JC+P+N5." Because there is no currently implemented correction for ascertainment bias, the developers recommend either fixing the heterozygosity at its observed value or adding constant sites to the data set (https://groups.google.com/g/iqtree/c/xVxurAjvoIY/m/QayzWsZ_AwAJ). We tried both but settled on the latter and added additional constant sites in which the base frequencies were set equal and bases randomly assigned to allele. The number of added sites was chosen so that the Watterson estimate of diversity, π_W inferred by PoMo was the same as the pooled value observed across populations (0.0025 for 4-fold sites and 0.0013 for 0-fold sites: see Results section). This was obtained quite closely by adding 400,000 constant sites for 4-fold data sets and 800,000 for 0-fold sites. Because the model of allele count change is reversible the trees returned are unrooted. Sites are assumed to be unlinked (De Maio et al., 2015, p. 1029).

(ii) *SVDquartets* (Chifman and Kubatko 2014) relies on the mathematical result that an unrooted species tree quartet is identifiable from site pattern frequency distributions of SNPs sampled across a genome, given the MSC model and a reversible model of substitution (Chifman and Kubatko 2015). Quartets of taxa are ultimately combined into an unrooted species tree. Suppose some quartet of taxa, i,j,k,l has observed nucleotide site pattern frequencies, $\hat{p}_{ijkl}$. These form a 4D array with $4\times4\times4\times4=256$ elements. There are three possible nontrivial bipartitions, $ij|kl, ik|jl, il|jk$. For each of these, it is possible to rewrite or "flatten" the original 4D array into a 16×16 2D array, and the rank of this array, namely, the degree to which it contains nondegenerate information, can be used to compute an SVD score that captures how well the data fit these alternative splits under the MSC. We used the implementation in PAUP* (Swofford, 2002), assigning individual genotypes to 10 populations acting as "species" for the MSC. To force SVDquartets to treat individuals as genotypes, each individual was recoded as two sequences, setting heterozygotes to two alternate states in the two sequences. It was not necessary to randomly phase heterozygotes because the method's calculations assume coalescent independence of sites (Chifman and Kubatko, 2014, p. 3317).

(iii) *TreeMix* (Pickrell and Pritchard 2012). TreeMix v. 1.13 uses a Gaussian model of genetic drift of gene frequencies within and between populations to compute a composite likelihood of the unrooted population tree. Then, it sequentially adds m "migration" edges in a greedy fashion to improve the score, where m is set by the user. These are meant to reflect the effect of both common ancestry and gene flow on the distribution of gene frequencies. Calculations of the variance–covariance matrix of gene frequencies are combined in contiguous blocks of size "-k" option to account for sitewise autocorrelation (value chosen to exceed the thin distance, see below). Because of our small population samples, we set the "-noss" option, which removes a correction useful in larger samples, as recommended by the manual

The next two methods return rooted trees.

(iv) *SNAPP* (Bryant et al. 2012) uses full computations of the likelihood function under a reversible model of nucleotide substitution combined with the MSC, across a set of coalescent independent sites (Bryant et al., 2012, p. 1919). It is implemented via Bayesian MCMC methods

with a plugin to the BEAST v. 2.6.4 package (Bouckaert et al. 2019). BEAUTi was used to format XML files for BEAST. Integer re-coded (i.e., 0-1-2) genotype data for individuals was used as input and the two individuals from the each population were grouped for the purpose of the MSC calculations. Sites were specified as only polymorphic to correct for ascertainment bias.

For tree search, to improve mixing, we ran Metropolis-coupled MCMC using the "coupledMCMC" package add-on for BEAST2 (Müller and Bouckaert 2020); largely following protocols described in the BEAST documentation. This consisted of mixtures of three hot and one cold chain run for 100,000 generations. Each of the two SNP data sets was run five times to allow combination of samples to improve ESS (effective sample size) values after a burnin period of 25% for each. Convergence of the tree topology was assessed using visualization tools in the R package "rwty" (Warren et al. 2017), which examines traces of bipartition frequencies, and the "approximate ESS" statistic proposed for the tree topology (Lanfear et al. 2016). The prior on the population size parameter for each edge, θ, was set to a gamma distribution with mean of 0.25 (parameters $\alpha = 1; \beta = 4$), reflecting prior information based on empirical estimates from our data (restricted to SNPs only). Likelihood calculations assumed exclusion of constant sites (i.e., were adjusted for ascertainment bias – though see Stange et al., 2018 for caveats). The weight parameter of the NodeSwapper operator was increased by 5× to 2.5. Forward and backward mutation rates were set equal, and therefore to ensure a stationary state frequency of 50% for the two alleles, we swapped allele state definitions at sites as necessary to achieve this. Since genotypes are called relative to the (arbitrary) saguaro genome sequence, this is appropriate to avoid any bias in inference.

The posterior distribution of trees was summarized with a maximum clade credibility (MCC) tree built with TreeAnnotator.

We tested whether SNAPP's inferred root location was sensitive to the algorithm's starting tree by repeating runs with one thinned 4-fold data set using each of 10 distinct rerootings of the UPGMA tree described above as a starting tree. Other parameters were kept the same. To summarize agreement across rerootings, we used a majority rule consensus tree across MCMC samples and rerootings.

To compare the statistical support for different rootings, we estimated the marginal likelihoods (Oaks et al. 2019) of each possible rerooting of the MCC tree found for the single 4-fold data set analyzed. These runs used simple MCMC (75,000 generations; 25% burnin) from the fixed tree (i.e., the NodeSwapper operator was turned off) rather than Metropolis-coupled MCMC. Estimates of marginal likelihoods can be biased, and several corrections have been proposed in the literature (Lartillot and Philippe 2006; Xie et al. 2011; Oaks et al. 2019). For each of the 17 possible rooted binary trees for 10 populations, we computed the harmonic mean and two other estimators based on path sampling: the trapezoidal approximation (Lartillot and Philippe 2006) and the stepping stone estimator (Xie et al. 2011). The path sampling approach was implemented using the PathSampling options in the BEAST ModelSelection module, which entails $K = 8$ separate runs of entire chains as described above for each rerooted tree. In all, this entailed 8×17 MCMC runs. Sensitivity of results to the rate prior described above was assessed by rerunning with SNAPP's default uniform prior and also with a much narrower gamma prior with the same mean but 1/10th the standard deviation as above (i.e., $\alpha = 100; \beta = 400$).

(v) *Polymorphism parsimony* was first described in the context of morphological or cytological traits having two alternate states, in which some individuals can exhibit both—that is, be polymorphic (Farris 1978; Felsenstein 1979). This character type clearly fits biallelic SNP genotypes, but it has not been used for such data before. PP is a discrete parsimony analog to model-based methods that use the MSC model for likelihood calculations. We explore this more formally in the Supplementary material available on Dryad, where we prove that PP is equivalent to species tree inference approaches that use the criterion of MDC (see Theorem 1 of the Supplementary material available on Dryad), under the assumption of an infinite sites model (Kimura 1969). PP evaluates trees by their the *polymorphism parsimony score*, which is the minimal number of edges of the tree that are polymorphic at both endpoints (Fig. S1 of the Supplementary material available on Dryad), summed across sites (implicitly assuming independence: Felsenstein, 1979, p. 50).

PP was implemented using the "dollop" program in Phylip 3.695 (2013) with settings of "PP," "Ancestral" states set to "?," and 100 replicated random addition sequences using Phylip's "Jumble" option. A bug in the program's output of the total PP score was fixed (Felsenstein, personal communication). Because dollop is not optimized for searching for the best *rooting* per se (Felsenstein, personal communication), the optimal tree reported by dollop was rerooted in all possible ways using a PERL wrapper script around the Newick Utilities program nw_reroot (Junier and Zdobnov 2010), and each rerooted tree was scored with dollop as described above. The best rooted tree was retained. Note that PP can return different ancestral state reconstructions depending on which homozygous genotype is assumed to be ancestral. The Phylip implementation computes the PP score for both choices and selects the smaller of the two. Thus, although the evolutionary model is implicitly rooted, the data are used to select the rooting that is most parsimonious at each site (however, the overall tree can still have different scores for different rootings).

Phylogenetic nonindependence of SNPs and data set construction.—Two "complete" data sets were constructed consisting of the diploid genotypes at biallelic sites from 0-fold and 4-fold degenerate sites in protein coding

genes. Linkage is expected to introduce autocorrelation between nearby sites in the genome (Pollard et al. 2006; Slatkin and Pollack 2006), which must be factored into the five methods we used. We estimated the empirical rate of decay of statistical dependence between sites in our SNP data and then created "thinned" subsampled data sets of putatively independent sites. To do this, we implemented a generalization of Hudson and Kaplan's (1985) "four-gamete test" for a pair of unphased diploid genotypes (Wang 2013). A pair of sites for N individuals is *pairwise genotype compatible* if and only if there exist haplotype "assignments" for all genotypes (i.e., arrangements of allele states from the two sites into haplotypes) consistent with a "perfect phylogeny" (one in which these sites evolve with no homoplasy). This is what is expected under the infinite sites model (Kimura 1969). Wang (2013) used this concept to identify blocks of compatible sites but did not explicitly describe it; for completeness, we show the computation in the Supplementary material available on Dryad. For a sample of 10,000 regularly spaced sites in each of the two complete data sets, we computed the fraction of sites at lag distance of λ sites downstream ($1 \leq \lambda \leq 100$) that were pairwise genotype compatible. This provided an estimate of autocorrelation as a function of coordinate distance. We then inferred a *thinning distance*, k_{thin} from the minimum distance at which the pairwise compatibility fraction decreases to its genome-wide level.

Except for TreeMix, these SNP-based phylogenomic inference methods either assume explicitly that SNP sites are statistically independent ("coalescent independent") in computing their optimality scores, or use this assumption in deriving key mathematical results (e.g., Chifman and Kubatko, 2014). We computed trees separately for all thinned data sets for PoMo, SVDquartets, and PP. For SNAPP, because of computational expense, we analyzed one thinned data set from each of the data partitions. Note that any thinned data set is a sample across the entire gene space of the assembly, so there is little bias toward a particular region of the genome. TreeMix uses the complete data sets but computes on contiguous blocks of sites of a size set by the user. We set this size so that it exceeded the thinning distance determined above.

Estimates of phylogenetic support.—When sites are dependent, estimates of statistical support in phylogenetic trees should account for the dependence structure. Standard bootstrap approaches are likely to be in inaccurate in general (Carlstein 1986; Künsch 1989), and though there is little theory specific to phylogenetics to guide methods to overcome this, general approaches can be adapted. Carlstein (1986) proposed drawing nonoverlapping spaced samples (subseries) of data, where the interval length depends on the strength of dependency. We used this approach for PoMo, SVDquartets, and PP. From a data set, D, with m sites, we construct k_{thin} subsampled data sets, D_i, where sites are sampled at successive intervals spaced k_{thin} sites apart, beginning at positions, 1,2, ..., etc.:

$$D_i = \{s_j : j = i, i + k_{\text{thin}}, i + 2k_{\text{thin}}, \ldots\}, j \leq m, 1 \leq i \leq k_{\text{thin}}.$$

Let $\hat{\Psi}_i(D_i)$ be the inferred tree and $\hat{\Psi}_{MR}(\hat{\Psi}_1, \hat{\Psi}_2, \ldots, \hat{\Psi}_{k_{\text{thin}}})$ the majority rule consensus tree based on all subsamples. Support for clades/splits can then be read from the frequency of clades/splits found on $\hat{\Psi}_{MR}$ as with conventional bootstrap analyses. We used the majority rule consensus algorithm in PAUP* (Swofford 2002) with the LE50 option to score clades/splits consistent with the ML tree but having support lower $\leq 50\%$. Majority rule trees were implemented with unrooted or rooted input trees as appropriate for the method.

TreeMix takes an approach suggested by Künsch (1989), of bootstrapping blocks of contiguous sites, where the block size is chosen so that it exceeds the distance at which autocorrelation decays to background levels. We used a length of 50 sites, chosen to exceed the thin distances described above (TreeMix option "-k 50 -bootstrap").

Finally, for SNAPP, which is a Bayesian method, we estimated posterior probabilities from the MCMC chain, annotated on the MCC tree.

Results

Sequence Data and Genetic Diversity

We obtained a total of 198 Gbp of short read sequence for 20 individuals in 10 localities (Table 1), with individual coverage ranging from 2.2× to 17.5×.

The BCFtools pipeline inferred from just under 50,000 to nearly 107,000 variants in the 4-fold and 0-fold degenerate site data sets (Table 2). Almost all were biallelic: the fraction of variants having three or four alleles was ≤ 0.0035.

Pooled estimates of neutral 4-fold genetic diversity across the species were 0.0025, with nonsynonymous 0-fold diversity at about half that (Table 3). Tajima's and Watterson's estimators were in broad agreement.

A geographic view of neutral genetic diversity among localities (Fig. 1) indicates the highest diversity is around Kino Bay in Sonora, Mexico (Table 3), but it drops off relatively quickly to the east and south, and more slowly to the north (Fig. S8 of the Supplementary material available on Dryad). Masiaca and Wickenburg, with generally the lowest diversities, are located near the southern and northern distribution margins respectively. Although the interpolated surface is only derived from the genomic data, there is a significant negative linear trend with elevation (Fig. S8 of the Supplementary material available on Dryad), and an association with the vegetation units described by Shreve (1964). The highest genomic diversity is in the Central Gulf Coast, with intermediate values in the Arizona Upland and thornscrub, and the lowest diversity in the transition to the Mojave Desert in the northwestern part of the range.

TABLE 2. Summary statistics for variants and phylogenetic data sets

Data set	0-fold Sites	4-fold Sites
Total effective sites	19,563,072	4,592,714
Number of variants	106,998	47,741
Biallelic sites in complete matrix	106,629	47,511
Tri- or quad-allelic sites	369	230
Fraction of tri- or quad-allelic sites	0.0035	0.0048
Fraction of genotypes homozygous	0.72	0.72
Fraction of genotypes heterozygous	0.19	0.19
Fraction of genotypes missing	0.08	0.09
Fraction of sites with no missing data	0.18	0.17
Sites in each compatibility-thinned matrix	5332	4752

TABLE 3. Nucleotide diversity estimates

		0-fold degenerate sites			4-fold degenerate sites			Ratios[a]	
Population	Sample size	Variants	π_T	π_W	Variants	π_T	π_W	π_T^0/π_T^4	π_W^0/π_W^4
Pooled	20	106,998	0.00124	0.00131	47,741	0.00253	0.00250	0.490	0.524
Caborca	2	40,336	0.00117	0.00115	19,117	0.00237	0.00233	0.494	0.494
El Dipo	2	34,688	0.00108	0.00106	16,645	0.00221	0.00218	0.489	0.486
Guásimas	2	40,603	0.00118	0.00116	19,418	0.00242	0.00238	0.488	0.487
Kino Bay	2	39,744	0.00122	0.00121	18,990	0.00251	0.00248	0.486	0.488
La Joyita	2	33,323	0.00106	0.00105	15,958	0.00218	0.00215	0.486	0.488
Masiaca	2	26,248	0.00097	0.00095	12,736	0.00202	0.00197	0.480	0.482
El Orégano	2	37,817	0.00114	0.00111	18,075	0.00233	0.00227	0.489	0.489
San Marcial	2	11,651	0.00106	0.00103	5457	0.00214	0.00208	0.495	0.495
Tucson	2	36,312	0.00105	0.00104	17,162	0.00212	0.00209	0.495	0.498
Wickenburg	2	32,406	0.00096	0.00093	15,389	0.00193	0.00189	0.497	0.492

Note: π_T and π_W are Tajima (1983) and Watterson (1975) estimators, respectively.
[a]Superscripts denote degeneracy class.

Phylogenetic Data Sets

The two complete biallelic SNP data sets ranged in size from 47,511 to 106,629 sites for 4-fold and 0-fold degenerate sites (Table 2). The fraction of genotypes called as heterozygous ranged from 17% to 20%. The fraction of sites with missing data was 8–9%. Missing data were generally low, but the two San Marcial individuals had 40–54% missing genotypes and one Masiaca individual had 20%.

Analyses of pairwise genotype compatibility in the two complete data sets revealed a strong signal of local statistical dependence along the genome coordinate axis (Fig. 2). The 4-fold data set lost dependence more quickly, within a k_{thin} of 10 sites; the 0-fold data lost dependence after about 20 sites. Note that these distances are in units of sites in the data matrices: two neighboring sites in a phylogenetic data set might be separated by a long coordinate distance along the genomic scaffold because of intervening nonvariable sites.

Intraspecific Phylogenies

Trees inferred were quite congruent between data partitions and all five methods. Unrooted topologies from the four population-based inference methods, PoMo (Fig. 3), SVDquartets (Fig. 3), TreeMix (assuming no migration: option "-m 0") (Fig. 3), and SNAPP (Fig. 4) were identical except for slight differences in support values and relationships within a well-supported northern clade of four populations—Caborca, Joyita, Tucson, and Wickenburg. Effects of allowing migration in TreeMix were slight. In the 4-fold data there were a few values of the -m parameter, which controls the number of migration edges allowed, which returned trees that differed by one nearest neighbor interchange from these results. In these, San Marcial and Orégano populations formed one element of a bipartition. The PP (Fig. 5) trees were more of an outlier, but all five methods agreed on a basic pattern of latitudinal differentiation in which more northern populations are phylogenetically distant from southern populations. PP trees had weaker support deeper in the tree relevant to relationships among the southern populations, but the other four methods showed strong agreement. PP trees had individual genomes as leaves and recovered each of the 10 populations of two individuals as unrooted bipartitions, but San Marcial was paraphyletic at the root of the tree. We discuss rooting further below, but at least in an unrooted sense, there are very few notable differences in topological results between methods of inference or data partitions. Moreover, despite the use of three different methods of estimating support for these trees (nonoverlapping subseries, block bootstrap, and Bayesian posteriors), estimates of support were quite consistent in the population trees, with the lowest values generally found within the northern clade.

SNAPP was the most compute-intensive method. We only ran one thinned data set for each data partition, but with five replicates each. Each replicate of

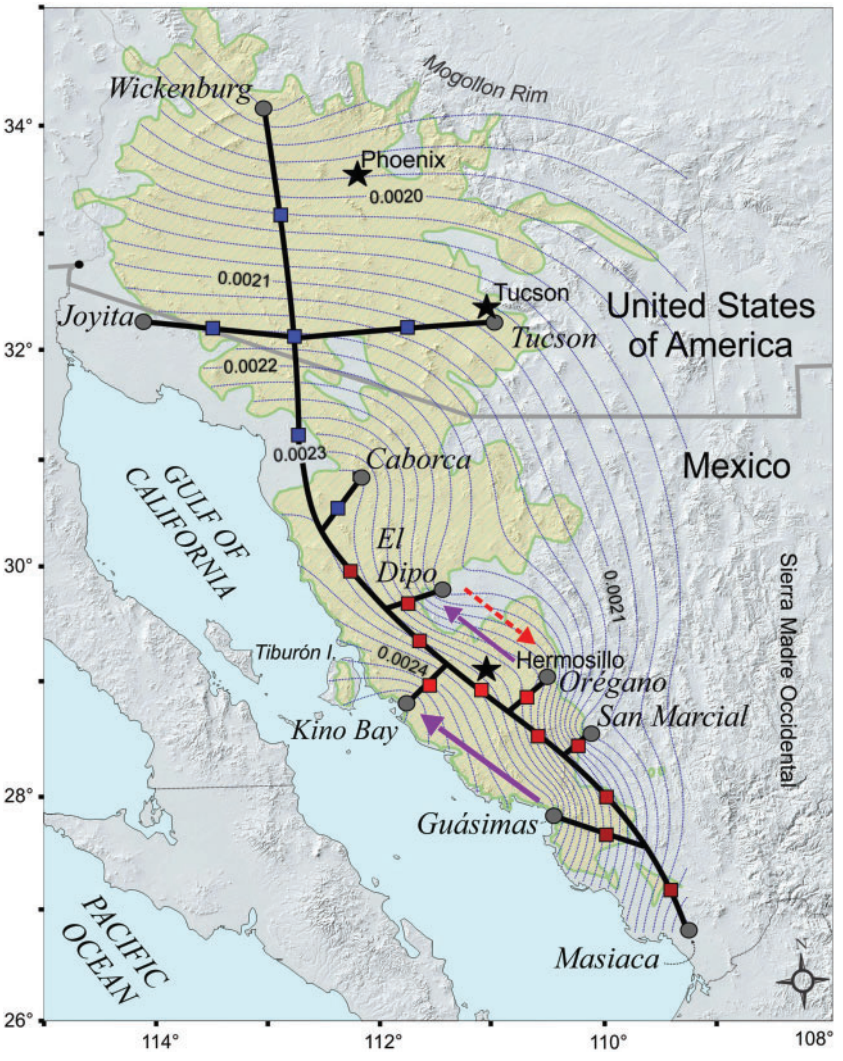

FIGURE 1. Geographic range, genetic diversity, and phylogenetic history within saguaro. Range of saguaro shown in yellow. Gray dots are 10 sample collection localities. Contour plot shows isolines of nucleotide diversity, π_W, in 4-fold degenerate sites. Unrooted phylogenetic tree of saguaro populations (see Fig. 3) is superimposed. Marginal likelihoods of different locations for the root, inferred from SNAPP analyses of 4-fold data set used in Fig. 4b, are shown by colored squares along edges of tree. Red is high marginal likelihood. Blue is low (Table S1 of the Supplementary material available on Dryad). Highest likelihoods are directly east of Kino Bay. Marginal likelihoods shown were estimated with the stepping stone method. The polytomy in the north reflects the uncertainty in relationships among Tucson, Wickenburg, and Joyita populations. Location of squares between neighboring nodes is for visualization purposes only. Purple arrows: substantial migration edges in 0-fold data set inferred by TreeMix; red dashed arrow: substantial migration edge in 4-fold data in some but not all values of maximum number of edges added (Tables S3 and S4 of the Supplementary material available on Dryad).

the Metropolis-coupled MCMC analysis required four chains and 72–96 h of HPC wall time to complete 100,000 generations. Visualizations of the convergence of tree topology in 'rwty' (Warren et al. 2017) indicated qualitatively satisfactory convergence, but the approximate ESS statistic was between 30 and 100 per replicate. Combining post-burnin trees across replicates yielded ESS values of 209 and 336 for 0-fold and 4-fold data sets, respectively, which exceeds the "rule of thumb" of 200 widely cited for convergence (Nascimento et al. 2017).

We did not find any effect of choice of the rooting of the starting tree in our experiment in which 10 possible rerooted starting trees were used in turn. Because SNAPP is a Bayesian method, we also looked for possible sensitivity to priors. There are few priors that can be modified in SNAPP. We reran the tree search analyses using a uniform prior on the theta value (advocated due to ascertainment bias by Stange et al., 2018), rather than the informative prior based on observed heterozygosity. The inferred trees were identical to those described above except for small differences in some clade posterior probabilities.

Rooting

Both SNAPP and PP analyses provided estimates of the root of the population tree, but the marginal

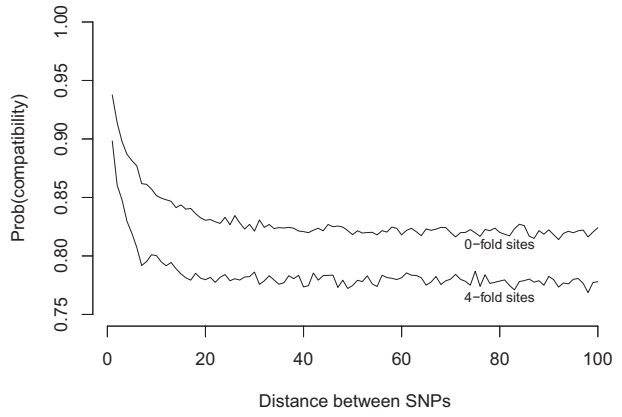

FIGURE 2. Decay of pairwise genotype compatibility (autocorrelation) versus distance between SNPs in data sets.

likelihood estimates from SNAPP provided a quantitative perspective. Irrespective of which method was used to compute them, marginal likelihoods differed by about 500 log likelihood units between the highest and lowest values, indicating very substantial information about the root (Table 1). Moreover, different rootings displayed a strong geographic structure (Fig. 1). Rerootings along edges of the unrooted tree in the southern part of the range have very high marginal likelihoods, especially in the area east of Kino Bay toward San Marcial and Orégano. Likelihoods decline slightly further south toward the southern range limit in Masiaca and Guásimas, and slightly again just to the north. Then, there is a dramatic decrease of several hundred log likelihood units north of El Dipo, associated with a northern clade of Caborca, Joyita, Tucson, and Wickenburg. Thus, we find two discretely different geographical regions, a southern one containing the probable root location, and a northern one in which the root is highly improbable, rather than a smooth latitudinal transition.

These findings are highly consistent between methods of computing the marginal likelihood. For example, stepping stone and harmonic mean rankings of populations are almost identical under the assumption of either narrow or wide informative gamma priors on population size (Tables S1 and S2 of the Supplementary material available on Dryad). Under a very broad uninformative uniform prior we did observe some numerical instabilities in stepping stone estimates (a few positive or infinite log likelihood values), but even there the ranking for the harmonic mean estimates was consistent with the other results.

The PP results agree with the SNAPP results in rooting the tree in south central Sonora, Mexico, with a northern clade of four populations that is consistently and strongly supported across both data partitions and the complete spectrum of thinned data sets (Fig. 5). However, PP infers a root within a grade consisting of the two San Marcial individuals, geographically close to but slightly southeast of where SNAPP infers the root. The two San Marcial individuals have much more missing data than other samples, but in a downsampling experiment to test whether these missing data might explain the rooting, PP returned the same rooted trees as in Figure 5 (see Supplementary material available on Dryad for detailed methods and results).

Migration/Gene Flow

As noted above, allowing migration in TreeMix runs had very little impact on inferences about the topology of the underlying population tree; it effectively just superimposed migration events upon a stable tree. Note that the program's heuristics do perform rearrangements to the tree following addition of migration edges, so it was possible for migration to alter the underlying tree (Pickrell and Pritchard 2012), but the effect was slight.

TreeMix did add migration edges, but the import of these additions was relatively minor and difficult to summarize. TreeMix uses composite likelihoods, so a comparison of these for different network topologies does not provide a direct significance test (Pickrell and Pritchard 2012). Instead, it uses a resampling estimate of the standard error of the weight of these added migration edges, which is a measure of the relative impact of migration on gene frequency changes. In the 0-fold data in searches allowing up to five added migration edges, at most two edges were added that had weights above 10% (maximum of 24%: Table S3 of the Supplementary material available on Dryad). Both were between neighboring populations in the central Sonoran part of the range. In the 4-fold data, the same pair of populations was involved in added edges, but with more inconsistent results as the number of migration edges allowed was changed, and generally with only one of the two having weights above 10% (maximum of 27%: Table S4 and Fig. S1 of the Supplementary material available on Dryad).

DISCUSSION

Origin and Diversification of Saguaro

Saguaro is common on the rocky hillsides, outwash slopes, and occasional sandy flats throughout its range in the Arizona, Sonora, and California portions of the Sonoran Desert (Shreve 1964; Turner et al. 1995). Its geographic range has evidently been influenced by climatic changes since the last Pleistocene glacial maximum in North America (18,000 ^{14}C yr. BP: Thompson and Anderson, 2000). Evidence from packrat middens containing preserved plant macrofossils indicates that saguaro reached southwestern Arizona and southeastern California by 10,500 years ago and had spread to the northern part of its present range by the mid-Holocene, 6000–8000 ^{14}C yr. BP (Van Devender 1987; McAuliffe and Van Devender 1998). This is consistent with radiocarbon data that indicates the southwestern deserts had attained their modern-day northern and western limits in Arizona and California by 6000 ^{14}C

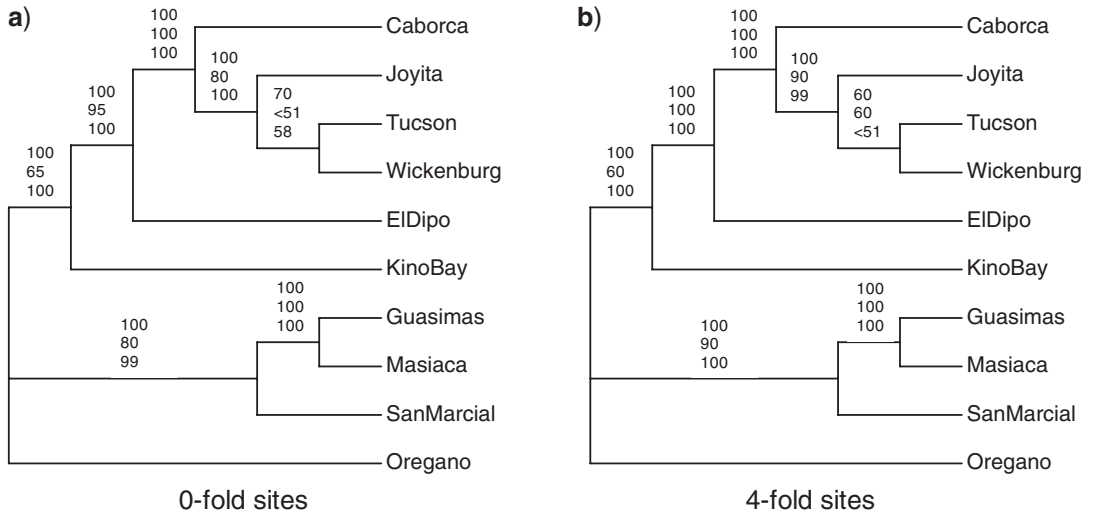

FIGURE 3. Unrooted population trees reconstructed with three methods, all returning the same majority rule tree. Support values are shown for each edge. Top to bottom are: (i) PoMo, with support values across thinned data sets; (ii) SVDquartets, with support values across thinned data sets; and (iii) TreeMix, with block bootstrap support values. a) Biallelic 0-fold sites; b) biallelic 4-fold sites. Outgroups were selected based on rooted trees in Figure 4.

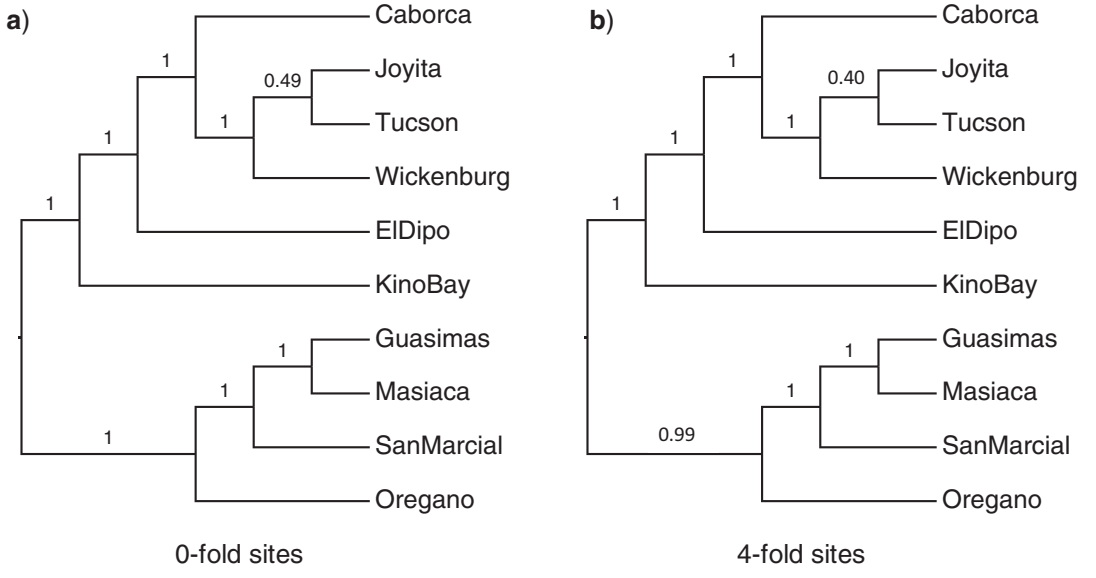

FIGURE 4. Rooted population trees reconstructed with SNAPP in BEAST2. These are maximum clade credibility trees for a single compatibility-thinned data set for a) biallelic 0-fold sites; b) biallelic 4-fold sites. Posterior samples of trees were combined from post-burnin trees in five independent runs of 100,000 generations. Posterior probabilities are indicated on edges.

yr. BP (Thompson and Anderson 2000). Given that the species split from its closest relatives (e.g., *Pachycereus pringlei* (S. Watson) Britton & Rose) perhaps several million years ago (Copetti et al. 2017), it seems likely that glacial era refugia must have existed somewhere near the southern end of saguaros current range in mainland Mexico (southern Sonora/northern Sinaloa). This was then followed by postglacial range expansion northward beginning in the early Holocene as the region warmed and became more arid, a pattern inferred for other Sonoran Desert plants (*Encelia*: Fehlberg and Ranker, 2009; *Melampodium*: Rebernig et al., 2010) and other North American columnar cacti (*Lophocereus*: Nason et al., 2002). We find evidence to support this hypothesis in both our genetic diversity data and phylogenomic results.

Pooled estimates of neutral genetic diversity for saguaro as a whole, near 0.0025 (Table 3), are low among plants, even compared to other long lived perennials (e.g., spruce, which has about 2.5 times the diversity of saguaro: Chen et al., 2019). Genetic diversity was highest at sea level at Kino Bay, Sonora, Mexico, along the coast of the Gulf of California, and a bit south of the approximate center of the species' range (Fig. 1). The genetic diversity

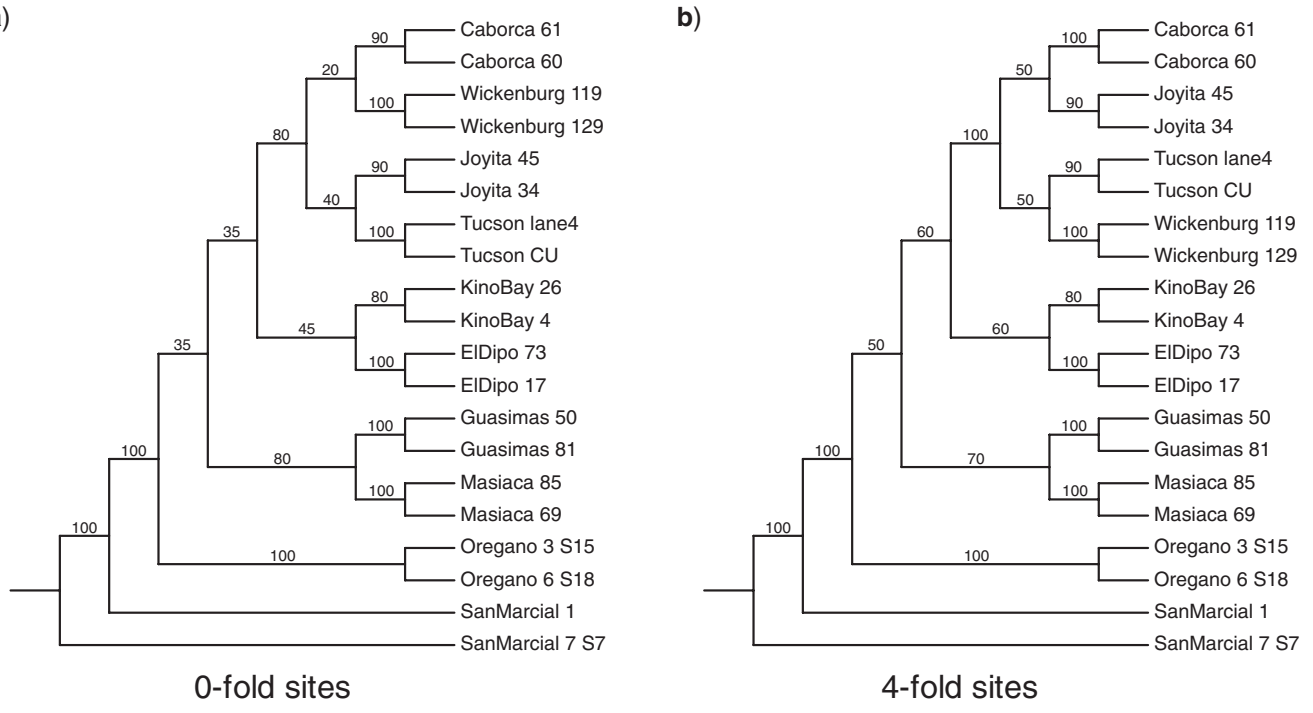

FIGURE 5. Rooted trees of individual diploid genomes reconstructed with polymorphism parsimony in PHYLIP. These are majority rule trees of sets of k_{thin} compatibility-thinned data sets for a) biallelic 0-fold sites ($k_{thin} = 20$); b) biallelic 4-fold sites ($k_{thin} = 10$); Support values are across thinned data sets.

gradient is asymmetric, decreasing quickly to the south and east and much more slowly to the north and northwest. Whether this is a result of historical dynamics or more recent and primarily ecological processes is unclear (Pironon et al. 2016; Lázaro-Nogal et al. 2017).

Despite the low nucleotide diversity, and biotic factors that would tend to foster gene flow (Goldshtein et al. 2020; O'Brien and Swann 2021), our phylogenomic analyses found strong phylogenetic signal among populations within this species. Unrooted trees were highly consistent across five quite different inference methods, two partitions of the gene space of the genome assembly, and 10–20 different thinned subsamples of those partitions (Figs. 3–5). The (unrooted) population trees were essentially identical in TreeMix, PoMo, SVDquartets, and SNAPP analyses, and the PP tree of individuals supported the monophyly of nine of these populations and largely agreed with the unrooted results from the model-based methods, though with weaker support in places. At a minimum there is robust support for a bipartition of populations into a northern group of Arizona and Arizona-Sonora borderland populations, and a southern group south of Caborca, Mexico. Moreover, the unrooted "caterpillar" topology maps remarkably well to the latitudinal positions of these 10 populations (Fig. 1). These phylogenetic conclusions were highly consistent despite the limitations of our sampling design, comprising two individuals sampled at 10 localities. We regarded these operationally as "populations" for the purposes of pooling SNPs into groups to use as terminal taxa in phylogeny (and network) reconstruction. However, sampling was not sufficient to test these provisional hypotheses about population delimitation or infer population genetic processes within these localities.

The two analyses that provide estimates of the root support the hypothesis that the current genetic diversity of saguaro traces back to a geographic origin in central-southern Sonora. Both SNAPP and PP analyses place the root of the saguaro tree to the east and south of Kino Bay toward Orégano and San Marcial populations. This is not far geographically from the center of genetic diversity in Kino Bay, but the gradient toward lower diversity is high to the east, and the preferred rooting consistently appears in these areas of somewhat lower diversity. The SNAPP analyses provided a quantitative assessment of support for different rootings and clearly showed strong support spread across the southern part of the range, but not as strong at the southern boundary population of Masiaca. To the north, there is a very sharp decrease in likelihood near Caborca, where the southwestern tip of the Arizona Upland Sonoran Desert subdivision meets the northern edge of the Plains of Sonora to the east, and the Central Gulf Coast to the west. The likelihood of the root being north of this is plainly very low, which is consistent with the paleoclimatic evidence (Thompson and Anderson 2000) and packrat midden records (McAuliffe and Van Devender 1998).

The distribution of North American columnar cacti is evidently impacted strongly by frost sensitivity, and, for

some species, the extent of the distribution of summer monsoonal precipitation in the northern and northwestern part of their ranges (Turner et al. 1995). In the case of the saguaro, its range is constrained by elevation (towards the Sierra Madre Occidental to the east and the Mogollon Rim to the north), westward by the Gulf of California, and most likely by the very low precipitation of the Gran Desierto where Sonora, California, and Arizona meet (Albuquerque et al. 2018). Unlike some other Sonoran Desert columnar cacti, saguaro is not found in Baja California (Turner et al. 1995). Albuquerque et al. (2018) found that saguaros distribution overall is contracting, but with some isolated areas of expansion in the northern part of its range, particularly in west-central Arizona and eastern California. They estimated a mean loss of almost 7% by 2050 under different climate change scenarios. That contraction is occurring on the western edge of the range from western Arizona to the southernmost extent in Mexico. Our rooted trees largely imply a basal grade of southern and southeastern populations giving rise to several populations to the north and the northwest, which reinforces the idea of a refuge near the southern boundaries of the actual distribution range at the lowest elevations. Further analyses on correlations between SNP variation and climatic, geographic, and ecological variables are necessary to understand these issues more fully.

Methodological Issues

Utility of methods used.—Although phylogenetic results were consistent among all five methods, several significant practical differences were clear. Runtime differences were especially dramatic, requiring different strategies for deployment of runs on high performance computing (HPC) infrastructure. TreeMix was the only program run with the entire 0-fold and 4-fold data sets having ∼100,000 and ∼50,000 sites, respectively. SVDquartets, PoMo, and PP were run with all thinned subsets of each of these having ∼5000 sites each (Table 2). SNAPP was run with only a single thinned data set from each. TreeMix and SVDquartets runs finished in under a few minutes on a single Linux HPC node or even a Mac laptop, while SNAPP runs in BEAST 2 required 2 weeks of HPC time for just a single thinned data set (with multithreading and pooling of enough replicates to obtain ESS values >200). Marginal likelihood calculations for all rerootings and the stepping stone analyses in SNAPP took several weeks of HPC time. PoMo and PP required intermediate running times and benefited greatly from distributing different thinned data sets to different nodes on the HPC cluster, in which case the overall run time was less than a day. These runtime differences stem from both differences in complexity of calculations in the underlying model and different algorithmic and implementation details.

Of course, methods differed in what they offer in return for their computational investment. SNAPP integrates over potentially important prior information about demography and diversification dynamics and provides both a rooted tree output and the potential to rank results on the basis of their marginal likelihoods, which respect both the priors and the data. PP also returns a rooted tree and makes few explicit assumptions except for the nontrivial infinite sites assumption, though this is probably not unreasonable for our data. TreeMix provided added value by inferring a limited number of additional migration edges to reconstruct a network and included a built-in confidence estimator appropriate for autocorrelated SNP data (its block bootstrap). SVDquartets was also very fast and returned results comparable to all other methods.

Despite the prevalence of polymorphic character data in the real world, and the obvious connection to diploid genotype data, PP has been used sparingly: for example, for plant morphological data (Baum 1983) and bird retrotransposons (Suh et al. 2015). PP fits within the long-standing framework of gene tree reconciliation (Goodman et al. 1979; Page 1994; Maddison 1997; Nakhleh 2013). The criterion of minimizing the number of polymorphic edges in the individual tree is equivalent to MDC in the "best" allele tree that can be constructed from the genotype data. Though intuitive, this equivalence is not automatic; it holds if the allele tree evolves according to an infinite sites model but not necessarily otherwise (see Theorem 1 of the Supplementary material available on Dryad). This assumption is not made by the other methods used here, though it and the even stronger assumption of the "no mutation model" have been used to reconstruct population trees before (Nielsen 1998; RoyChoudhury et al. 2008). Recently, two methods that also make the infinite sites assumption (Kelleher et al. 2019; Speidel et al. 2019) have been applied to the harder problem of reconstructing the entire ancestral recombination graph (Gusfield 2014) for a set of populations. Although the heuristics used in these methods allow them scale up to very large data sets such as ours, they also require as part of their input two elements not present in our data: phased haplotypes and ancestral state assignments for each SNP.

PP returned more variable results between different thinned partitions of the data than PoMo or SVDquartets did. This might be because the two model-based methods have higher precision, but there are also indications that it might be due to the tree search heuristics in Phylip's "dollop" program. There were numerous cases in which a simple rerooting of the tree returned by the program yielded a tree with better score, which led us to check all rerootings as part of our search protocol. Adding more extensive rooted rearrangement code might help in the heuristic search routines in "dollop." It remains to be seen if PP offers competitive results to model-based methods in terms of accuracy, but at a minimum it could provide a shortcut to finding good candidate starting trees for more computationally intensive likelihood-based searches.

Rooting.—Sequence data from the ingroup alone can be used to infer the root of a phylogenetic tree under certain known conditions. For model-based methods

of inference, these include some or all of the following: nonreversible substitution models (Bettisworth and Stamatakis 2021; Naser-Khdour et al. 2021), constraints on edge lengths (Tria et al. 2017), or assumptions like the MSC (Allman et al. 2011; Yu et al. 2011). For parsimony, an asymmetric cost matrix can do the same (Huelsenbeck et al. 2002). We are not aware of mathematical proofs that simultaneously handle species tree inference, coalescent independent SNP data, and the variety of other model and inference assumptions used by the methods we employed here, but certain conclusions seem clear.

Both PoMo (Schrempf et al. 2016) and TreeMix (Pickrell and Pritchard 2012) use explicitly reversible models and empirically have the same optimality scores under different rootings. SVDquartets (Chifman and Kubatko 2014) evaluates SNPs under the MSC but only with respect to bipartitions of unrooted quartets of the species tree, and these are summarized ultimately in an unrooted species tree. Thus, these three methods cannot be used as is for rooting. On the other hand, SNAPP's (Bryant et al. 2012) expressions for likelihood of a single SNP combine two terms, one for the substitution process, and one for the embedded allele tree history of that SNP. The impact of the latter is governed by the same likelihood calculations as act in pure gene tree methods, and the former assumes a clock, so this method can be used to infer a root. Likewise, though PP is not formulated in terms of a step matrix, asymmetric or otherwise, its score function can return different results depending on rooting, because it effectively penalizes polymorphic edges that span the root (Felsenstein 1979).

Our efforts to root the saguaro population tree with SNAPP and PP provided a largely consistent picture, but they also highlighted several computational issues. Accurate estimation of marginal likelihoods in SNAPP, for all rerootings, using best-in-practice methods such as stepping stone estimators (Oaks et al. 2019), required tens of thousands of hours of walltime on a computing cluster for just one thinned data set. In our data, the harmonic mean estimator returned very similar results despite its known shortcomings (Xie et al. 2011; Oaks et al. 2019), and its running time was an order of magnitude faster. We also encountered some numerical issues using a uniform prior on population size, but results from two gamma distributed priors, one with a narrow variance and one quite broad, were highly similar (Tables S1 and S2 of the Supplementary material available on Dryad).

Even though PP is computationally faster than SNAPP, repeating runs for all rerootings with even a modest number of random heuristic search replicates combined to make these runs moderately computionally intensive as well. However, it was still possible to complete runs on all thinned data sets for both gene partitions. PP inferred a root within the San Marcial population not far to the southeast of the general location of the root found by SNAPP analyses. However, its rooting *within* the population is somewhat puzzling, as the genetic distance between its two individuals is no larger than that between the two samples in any other population. These individuals have the most missing data among the samples, but a downsampling experiment to assess the effect of these missing data recovered the same rooted trees (see Supplementary material available on Dryad for detailed results). The mechanism behind the presumably erroneous rooting by PP remains unknown but seems not to be missing data per se.

Recombination, site autocorrelation, and statistical support.—The decay of statistical dependence of phylogenetic signal with recombination distance along the chromosome (Slatkin and Pollack 2006; Pollard et al. 2006) can be seen in our saguaro SNP data (Fig. 2). Based on a pairwise phylogenetic compatibility assay we estimated that in the complete data sets of ~50,000 4-fold generate sites and ~100,000 0-fold sites autocorrelation disappears at a distance of approximately 10 and 20 SNPs, respectively. On average this corresponds to roughly 5–6 genes in the assembly, which may partly reflect the small size of many scaffolds (scaffold N50 was 61.5 kb: Copetti et al., 2017).

Four of the five methods we used assume statistical independence between SNPs in calculations of likelihoods (SNAPP: Bryant et al., 2012, p. 1919; PoMo: De Maio et al., 2015, p. 1029), SVD split scores (SVDquartets: Chifman and Kubatko, 2014, p. 3317), or the PP score (Felsenstein, 1979, p. 50). TreeMix computes its composite likelihood score in contiguous blocks the size of which can be set by the user (Pickrell and Pritchard 2012). Simulation evidence suggests that SNP-based inference methods are relatively robust to violations of this assumption (Chifman and Kubatko 2014; Zhu et al. 2018), but the impact on sample size seems likely to affect estimates of statistical support.

TreeMix employs a built-in block bootstrap method to estimate significance, which was first proposed by Künsch (1989) and discussed occasionally in the context of phylogenetic inference from dependent data (Felsenstein 2005; Holmes 2003; Lemoine et al. 2018). For PoMo, SVDquartets and PP, we used a method proposed slightly earlier (Carlstein 1986), which takes nonoverlapping subseries of the data—the "thinned data sets" we describe above, and then evaluates the agreement between subsets. Overall, support levels for clades or splits using either of these resampling approaches were similar despite being applied to different tree inference algorithms, but support was noticeably higher for PoMo and SVDquartets than it was for PP. Posterior probabilities in the one thinned data set for each data partition were uniformly very high. Further theoretical and simulation work is needed to evaluate statistical performance of these varied confidence estimators in SNP-based methods.

Gene flow and network models.—From a statistical perspective, it would be attractive to perform model comparisons between pure tree models and higher order network models (Kubatko 2009; Stenz et al. 2015). In our previous work on saguaro phylogenomics above the species level (Copetti et al. 2017), for example, we used

PhyloNet's InferNetwork_ML program to infer reticulations in the species tree based on a large set of inferred gene trees, with the help of the AIC statistic, which penalizes the added complexity of a network model (Burnham and Anderson 2010). Absent such a penalty, Wen et al. (2018) argued against using likelihood-based methods for identifying network edges, advocating Bayesian approaches instead. Other gene tree approaches inferring networks have been developed (Yu et al. 2014; Solís-Lemus and Ané 2016; Zhang et al. 2018), and, in addition to TreeMix (Pickrell and Pritchard 2012), several SNP methods have been developed in the last few years that might enable model comparisons, including Ima3 (Hey et al. 2018), and the SNP-based network inference in PhyloNet's MLE_BiMarker and MCMC_BiMarker programs (Zhu and Nakhleh 2018; Zhu et al. 2018).

However, several issues limited their applicability to our data sets. First was scalability. Likelihood or pseudo-likelihood methods that allow reticulations require dramatically more computation (Zhu and Nakhleh, 2018, p. i377). In pilot experiments with PhyloNet's MCMC_BiMarkers code, which can be thought of as "SNAPP for networks" (Zhu et al., 2018, p. 3), we estimated that run times allowing only one reticulation edge and a single cold chain would be 1–2 orders of magnitude longer than for SNAPP, already our most computationally intense tool for these data, taking weeks of run time. This is consistent with the developers' own studies (Cao et al. 2019).

A second issue is the proper method of model selection between a tree model and network model. Relatively fast pseudolikelihood approaches do not admit standard information criterion testing (Zhu and Nakhleh 2018). Resampling methods can be used to gauge support for edges (as in TreeMix, Pickrell and Pritchard, 2012; or PhyloNetwork, Solís-Lemus and Ané, 2016), and a cross-validation procedure like that in Yu et al. (2014) might provide a general solution if its computational overhead were not too large.

Thus, we relied on the consistency of results among a diverse set of population tree inference tools, and TreeMix, which is remarkably scalable and allows discrete migration edges, to draw limited inferences about gene flow. The consistency of phylogenetic results across methods, their individually strong statistical support, and the recovery of just a few low weight migration edges in TreeMix, leads us to conclude that the population trees we inferred are a good reflection of the history of this species. Future work in saguaro population studies is likely to benefit from application of some of the tools that integrate gene flow with phylogenetics (e.g., Ima3: Hey et al., 2018; ∂a∂i: Gutenkunst et al., 2009; or MCMC_BiMarkers: Zhu and Nakhleh, 2018) to more intensively sampled, fewer, and more geographically proximate populations, such as those in the vicinity of Kino Bay, San Marcial, and Orégano. Larger samples from each population would allow better estimation of parameters that depend on the site frequency spectrum and ultimately lead to a more complete understanding of the dynamic evolutionary and ecological history of this species.

Finally, the likely phylogenetic root within saguaro somewhere to the east and south of Kino Bay in Sonora, Mexico, includes areas having some of the highest genetic diversity in the species, as well as the steepest gradients in that diversity. These low elevation areas, likely the location of ancient glacial refuges, are prime candidates for protection aimed at preserving genetic diversity of this iconic species.

Data Availability

All sequencing data have been deposited in the Short Read Archive at NCBI under BioProject number PRJNA767819. VCF files, processed data sets, phylogenetic data sets, and all scripts used to process and analyze data are available at the Dryad Digital Repository (https://doi.org/10.5061/dryad.wdbrv15q4).

Supplementary Material

Data available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.wdbrv15q4.

Funding

This work was supported by the U.S. National Science Foundation [1735604].

Acknowledgments

We thank Joe Felsenstein for fixing a small bug in Phylip's dollop program, Laura Kubatko for helpful comments on SVDquartets, and the editors and three reviewers for comments on the manuscript. We thank University Information Technology Services at the University of Arizona for high performance computing access and the Arizona Genomics Institute for computing resources.

References

Alanzi A., Degnan J. 2017. Inferring rooted species trees from unrooted gene trees using approximate Bayesian computation. Mol. Phylogenet. Evol. 116:13–24.

Albuquerque F., Benito B., Rodriguez M.Á.M., Gray C. 2018. Potential changes in the distribution of *Carnegiea gigantea* under future scenarios. PeerJ 6:e5623.

Allman E., Degnan J., Rhodes J. 2011. Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. J. Math. Biol. 62:833–862.

Andrews S. 2018. FastQC. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Bansal M.S., Burleigh J.G., Eulenstein O. 2010. Efficient genome-scale phylogenetic analysis under the duplication-loss and deep coalescence cost models. BMC Bioinf. 11:S42.

Baum B. R. 1983. A phylogenetic analysis of the tribe Triticeae (Poaceae) based on morphological characters of the genera. Can. J. Bot. 61:518–535.
Bennett M., Leitch I. 2012. Plant DNA C-values database (Release 6.0, December 2012).
Bettisworth B., Stamatakis A. 2021. Root digger: a root placement program for phylogenetic trees. BMC Bioinformatics 22:225.
Blair C., Ané C. 2020. Phylogenetic trees and networks can serve as powerful and complementary approaches for analysis of genomic data. Syst. Biol. 69:593–601.
Bolger A.M., Lohse M., Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120.
Bouckaert R., Vaughan T., Barido-Sottani J., Duchêne S., Fourment M., Gavryushkina A., Heled J., Jones G., Kühnert D., De Maio N., Matschiner M., Mendes F., Müller N., Ogilvie H., Du Plessis L., Popinga A., Rambaut A., Rasmussen D., Siveroni I., Suchard M., Wu C.-H., Xie D., Zhang C., Stadler T., Drummond A. 2019. BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. PLoS Comput. Biol. 15:e1006650.
Brandrud M., Baar J., Lorenzo M., Athanasiadis A., Bateman R., Chase M., Hedrén M., Paun O. 2020. Phylogenomic relationships of diploids and the origins of allotetraploids in *Dactylorhiza* (Orchidaceae). Syst. Biol. 69:91–109.
Bravo G., Antonelli A., Bacon C., Bartoszek K., Blom M., Huynh S., Jones G., Lacey Knowles L., Lamichhaney S., Marcussen T., Morlon H., Nakhleh L., Oxelman B., Pfeil B., Schliep A., Wahlberg N., Werneck F., Wiedenhoeft J., Willows-Munro S., Edwards S. 2019. Embracing heterogeneity: coalescing the tree of life and the future of phylogenomics. PeerJ 7:e6399.
Bryant D., Bouckaert R., Felsenstein J., Rosenberg N.A., RoyChoudhury A. 2012. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. Mol. Biol. Evol. 29:1917–1932.
Burnham K.P., Anderson D.R. 2010. Model selection and multi-model inference. 3rd ed. New York: Springer.
Bustamante E., Búrquez A., Scheinvar E. and Eguiarte L. E. 2016. Population genetic structure of a widespread bat-pollinated columnar cactus. PLoS One 11:e0152329.
Cao Z., Liu X., Ogilvie H.A., Yan Z., Nakhleh L. 2019. Practical aspects of phylogenetic network analysis using PhyloNet. bioRxiv 746362; doi: https://doi.org/10.1101/746362.
Carlstein E. 1986. The use of subseries values for estimating the variance of a general statistic from a stationary sequence. Ann. Stat. 14:1171–1179.
Chen J., Li L.L., Milesi P., Jansson G., Berlin M., Karlsson B., Aleksic J., Vendramin G.G., Lascoux M. 2019. Genomic data provide new insights on the demographic history and the extent of recent material transfers in Norway spruce. Evol. Appl. 12:1539–1551.
Cherlin S., Heaps S., Nye T., Boys R., Williams T., Embley T. 2018. The effect of nonreversibility on inferring rooted phylogenies. Mol. Biol. Evol. 35:984–1002.
Chifman J., Kubatko L. 2014. Quartet inference from SNP data under the coalescent model. Bioinformatics 30:3317–3324.
Chifman J., Kubatko L. 2015. Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites. J. Theor. Biol. 374:35–47.
Choi J.Y., Dai X., Alam O., Peng J.Z., Rughani P., Hickey S., Harrington E., Juul S., Ayroles J.F., Purugganan M.D., Stacy E.A. 2021. Ancestral polymorphisms shape the adaptive radiation of *Metrosideros* across the Hawaiian islands. Proc. Natl. Acad. Sci. USA 118.
Copetti D., Búrquez A., Bustamante E., Charboneau J.L.M., Childs K.L., Eguiarte L.E., Lee S., Liu T.L., McMahon M.M., Whiteman N.K., Wing R.A., Wojciechowski M.F., Sanderson M.J. 2017. Extensive gene tree discordance and hemiplasy shaped the genomes of North American columnar cacti. Proc. Natl. Acad. Sci. USA 114:12003–12008.
Cutter A. 2013. Integrating phylogenetics, phylogeography and population genetics through genomes and evolutionary theory. Mol. Phylogenet. Evol. 69:1172–1185.
Danecek P., Bonfield J., Liddle J., Marshall J., Ohan V., Pollard M., Whitwham A., Keane T., McCarthy S., Davies R., Li H. 2021. Twelve years of SAMtools and BCFtools. GigaScience 10:giab008.
Danielson J., Gesch D. 2011. Global multi-resolution terrain elevation data 2010 (GMTED2010): U.S. Geological Survey Open-File Report 2011–1073. U.S. Department of the Interior, U.S. Geological Survey.
Dasarathy G., Nowak R., Roch S. 2015. Data requirement for phylogenetic inference from multiple loci: a new distance method. IEEE/ACM Trans. Comput. Biol. Bioinformatics 12:422–432.
De Maio N., Schrempf D., Kosiol C. 2015. PoMo: an allele frequency-based approach for species tree estimation. Syst. Biol. 64:1018–1031.
Degnan J.H., Rosenberg N. A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.
Drummond A.J., Ho S.Y.W., Phillips M.J., Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:699–710.
Edwards A.W.F., Cavalli-Sforza L.L. 1964. Reconstruction of evolutionary trees. In: Heywood V.H., McNeill J., editors. Phenetic and phylogenetic classification, vol. 6. London: Systematics Association Publication. p. 67–76.
Farris J.S. 1978. Inferring phylogenetic trees from chromosome inversion data. Syst. Zool. 27:275–284.
Fehlberg S., Ranker T. 2009. Evolutionary history and phylogeography of *Encelia farinosa* (Asteraceae) from the Sonoran, Mojave, and Peninsular deserts. Mol. Phylogenet. Evol. 50:326–335.
Felsenstein J. 1979. Alternative methods of phylogenetic inference and their interrelationship. Syst. Zool. 28:49–62.
Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376.
Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6.
Ferretti L., Raineri E., Ramos-Onsins S. 2012. Neutrality tests for sequences with missing data. Genetics 191:1397–1401.
Flouri T., Jiao X., Rannala B., Yang Z. 2020. A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. Mol. Biol. Evol. 37:1211–1223.
Goldshtein A., Handel M., Eitan O., Bonstein A., Shaler T., Collet S., Greif S., Medellín R., Emek Y., Korman A., Yovel Y. 2020. Reinforcement learning enables resource partitioning in foraging bats. Curr. Biol. 30:4096–4102.e6.
Goodman M., Czelusniak J., Moore G.W., Romero-Herrera A.E., Matsuda G. 1979. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Syst. Zool. 28:132–163.
Gusfield D. 2014. ReCombinatorics: the algorithmics of ancestral recombination graphs and explicit phylogenetic networks. Cambridge, MA: MIT Press.
Gutenkunst R., Hernandez R., Williamson S., Bustamante C. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 5:e1000695.
Hein J., Schierup M.H., Wiuf C. 2005. Gene genealogies, variation and evolution: a primer in coalescent theory. Oxford, UK: Oxford University Press.
Hey J., Chung Y., Sethuraman A., Lachance J., Tishkoff S., Sousa V., Wang Y. 2018. Phylogeny estimation by integration over isolation with migration models. Mol. Biol. Evol. 35:2805–2818.
Holmes S. 2003. Bootstrapping phylogenetic trees: theory and methods. Stat. Sci. 18:241–255.
Huelsenbeck J.P., Bollback J.P., Levine A.M. 2002. Inferring the root of a phylogenetic tree. Syst. Biol. 51:32–43.
Huson D.H., Rupp R., Scornavacca C. 2010. Phylogenetic networks: concepts, algorithms, and applications. Cambridge, UK: Cambridge University Press.
Jones G. 2019. Divergence estimation in the presence of incomplete lineage sorting and migration. Syst. Biol. 68:19–31.
Jouganous J., Long W., Ragsdale A., Gravel S. 2017. Inferring the joint demographic history of multiple populations: beyond the diffusion approximation. Genetics 206:1549–1567.
Junier T., Zdobnov E.M. 2010. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics 26:1669–1670.
Kelleher J., Wong Y., Wohns A., Fadil C., Albers P., McVean G. 2019. Inferring whole-genome histories in large population datasets. Nat. Genet. 51:1330–1338.

Kimura M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61:893–903.
Kingman J.F.C. 1982. On the genealogy of large populations. J. Appl. Prob. 19:27– 43.
Kubatko L. 2009. Identifying hybridization events in the presence of coalescence via model selection. Syst. Biol. 58:478–488.
Künsch H.R. 1989. The jackknife and the bootstrap for general stationary observations. Ann. Stat. 17:1217–1241.
Lanfear R., Hua X., Warren D.L. 2016. Estimating the effective sample size of tree topologies from Bayesian phylogenetic analyses. Genome Biol. Evol. 8:2319–2332.
Langmead B., Salzberg S.L. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9:357–359.
Lartillot N., Philippe H. 2006. Computing Bayes factors using thermodynamic integration. Syst. Biol. 55:195–207.
Lázaro-Nogal A., Matesanz S., García-Fernández A., Traveset A., Valladares F. 2017. Population size, center-periphery, and seed dispersers' effects on the genetic diversity and population structure of the mediterranean relict shrub *Cneorum tricoccon*. Ecol Evol. 7:7231–7242.
Leaché A., Harris R., Rannala B., Yang Z. 2014. The influence of gene flow on species tree estimation: a simulation study. Syst. Biol. 63:17–30.
Lemoine F., Domelevo Entfellner J.B., Wilkinson E., Correia D., Davila Felipe M., De Oliveira T., Gascuel O. 2018. Renewing Felsenstein's phylogenetic bootstrap in the era of big data. Nature 556:452–456.
Li G., Figueiró H., Eizirik E., Murphy W., Yoder A. 2019. Recombination-aware phylogenomics reveals the structured genomic landscape of hybridizing cat species. Mol. Biol. Evol. 36:2111–2126.
Li H., Handsaker B., Wysoker A., Fennell T., RuanJ., Homer N., Marth G., Abecasis G., Durbin R., 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079.
Liu L., Anderson C., Pearl D., Edwards S. 2019. Modern phylogenomics: building phylogenetic trees using the multispecies coalescent model. Methods Mol. Biol. 1910:211–239.
Long C., Kubatko L. 2018. The effect of gene flow on coalescent-based species-tree inference. Syst. Biol. 67:770–785.
Ma B., Li M., Zhang L. 2001. From gene trees to species trees. SIAM J. Comput. 30:729–752.
Maddison W.P. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.
Maddison W.P., Knowles L.L. 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55:21–30.
Mason N., Fletcher N., Gill B., Funk W., Zamudio K. 2020. Coalescent-based species delimitation is sensitive to geographic sampling and isolation by distance. Syst. Biodivers. 18:269–280.
McAuliffe J.R., Van Devender T.R. 1998. A 22,000-year record of vegetation change in the north-central Sonoran desert. Palaeogeogr. Palaeoclimatol. Palaeoecol. 141:253–275.
Meleshko O., M. Martin, T. Korneliussen, C. Schröck, P. Lamkowski, J. Schmutz, A. Healey, B. Piatkowski, A. Shaw, D. Weston, K. Flatberg, P. Szövényi, K. Hassel, H. Stenøien. 2021. Extensive genome-wide phylogenetic discordance is due to incomplete lineage sorting and not ongoing introgression in a rapidly radiated bryophyte genus. Mol. Biol. Evol. 38:2750–2766.
Minh B.Q., Schmidt H.A., Chernomor O., Schrempf D., Woodhams M.D., von Haeseler A., Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. 37:1530–1534.
Müller N., Bouckaert R. 2020. Adaptive metropolis-coupled MCMC for BEAST 2. PeerJ 8:e9473.
Nakhleh L. 2013. Computational approaches to species phylogeny inference and gene tree reconciliation. Trends Ecol. Evol. 28:719–728.
Nascimento F., Reis M., Yang Z. 2017. A biologist's guide to Bayesian phylogenetic analysis. Nat. Ecol. Evol. 1:1446–1454.
Naser-Khdour S., Minh B.Q., Lanfear R. 2021. Assessing confidence in root placement on phylogenies: an empirical study using non-reversible models for mammals. Syst. Biol. syab067. https://doi.org/10.1093/sysbio/syab067.
Nason J.D., Hamrick J.L., Fleming T.H. 2002. Historical vicariance and postglacial colonization effects on the evolution of genetic structure in *Lophocereus*, a Sonoran desert columnar cactus. Evolution 56:2214–2226.
Nielsen R. 1998. Maximum likelihood estimation of population divergence times and population phylogenies under the infinite sites model. Theor. Popul Biol. 53:143–151.
Oaks J., Cobb K., Minin V., Leaché A. 2019. Marginal likelihoods in phylogenetics: a review of methods and applications. Syst. Biol. 68:681–697.
O'Brien K., Swann D. 2021. Three decades of ecological change: the 2020 saguaro census. Part I: changes in the saguaro population 1990-2020. Report to Western National Park Association, Part I of Projects 19-06 and 20-09 Saguaro National Park.
Olave M., Meyer A. 2020. Implementing large genomic single nucleotide polymorphism data sets in phylogenetic network reconstructions: a case study of particularly rapid radiations of cichlid fish. Syst. Biol. 69:848–862.
Page R.D.M. 1994. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. Syst. Biol. 43:58–77.
Pease J., Haak D., Hahn M., Moyle L. 2016. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. PLoS Biol. 14:e1002379.
Pickrell J.K., Pritchard J.K. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genetics 8:e1002967.
Pironon S., Papuga G., Villellas J., Angert A.L., García M.B., Thompson J.D. 2016. Geographic variation in genetic and demographic performance: new insights from an old biogeographical paradigm. Biol. Rev. 92:1877–1909.
Pollard D.A., Iyer V.N., Moses A.M., Eisen M.B. 2006. Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting. PLoS Genetics 2:1634–1647.
Posada D., Buckley T. 2004. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst. Biol. 53:793–808.
Rannala B., Yang Z. 2017. Efficient Bayesian species tree inference under the multispecies coalescent. Syst. Biol. 66:823–842.
Rebernig C., Schneeweiss G., Bardy K., Schönswetter P., Villaseñor J., Obermayer R., Stuessy T., Weiss-Schneeweiss H. 2010. Multiple Pleistocene refugia and Holocene range expansion of an abundant southwestern American desert plant species (*Melampodium leucanthum*, Asteraceae). Mol. Ecol. 19:3421–3443.
RoyChoudhury A., Felsenstein J., Thompson E. 2008. A two-stage pruning algorithm for likelihood computation for a population tree. Genetics 180:1095–1105.
Sackton T. 2014. Identify_4d_sites.pl. Computer program. Available from: https://github.com/tsackton/linked-selection/tree/master/misc_scripts.
Schrempf D., B.Q. Minh, N. De Maio, A. von Haeseler, C. Kosiol. 2016. Reversible polymorphism-aware phylogenetic models and their application to tree inference. J. Theor. Biol. 407:362–370.
Schrempf D., Minh B.Q., von Haeseler A., Kosiol C. 2019. Polymorphism-aware species trees with advanced mutation models, bootstrap, and rate heterogeneity. Mol. Biol. Evol. 36:1294–1301.
Shi C.-M., Yang Z. 2018. Coalescent-based analyses of genomic sequence data provide a robust resolution of phylogenetic relationships among major groups of gibbons. Mol. Biol. Evol. 35:159–179.
Shreve F. 1964. Vegetation of the Sonoran Desert. In: Shreve F. and Wiggins I.L., editors. Vegetation and flora of the Sonoran Desert, vol. I. Stanford, CA: Stanford University Press. p. 1–186.
Slatkin M., Pollack J. 2006. The concordance of gene trees and species trees at two linked loci. Genetics 172:1979–1984.
Solís-Lemus C., Ané C. 2016. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. PLoS Genet. 12:e1005896.
Speidel L., Forest M., Shi S., Myers S. 2019. A method for genome-wide genealogy estimation for thousands of samples. Nat. Genet. 51:1321–1329.

Stange M., Sánchez-Villagra M., Salzburger W., Matschiner M. 2018. Bayesian divergence-time estimation with genome-wide single-nucleotide polymorphism data of sea catfishes (Ariidae) supports Miocene closure of the Panamanian isthmus. Syst. Biol. 67:681–699.

Stein J., Yu Y., Copetti D., Zwickl D., Zhang L., Zhang C., Chougule K., Gao D., Iwata A., Goicoechea J., Wei S., Wang J., Liao Y., Wang M., Jacquemin J., Becker C., Kudrna D., Zhang J., Londono C., Song X., Lee S., Sanchez P., Zuccolo A., Ammiraju J., Talag J., Danowitz A., Rivera L., Gschwend A., Noutsos C., Wu C.-C., Kao S.-M., Zeng J.-W., Wei F.-J., Zhao Q., Feng Q., El Baidouri M., Carpentier M.-C., Lasserre E., Cooke R., Rosa Farias D., Da Maia L., Dos Santos R., Nyberg K., McNally K., Mauleon R., Alexandrov N., Schmutz J., Flowers D., Fan C., Weigel D., Jena K., Wicker T., Chen M., Han B., Henry R., Hsing Y.-I., Kurata N., De Oliveira A., Panaud O., Jackson S., Machado C., Sanderson M., Long M., Ware D., Wing R. 2018. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. Nat. Genet. 50:285–296.

Stenz N.W.M., Larget B., Baum D.A., Ané C. 2015. Exploring tree-like and non-tree-like patterns using genome sequences: an example using the inbreeding plant species *Arabidopsis thaliana* (L.) Heynh. Syst. Biol. 64:809–823.

Suh A., Smeds L., Ellegren H. 2015. The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. PLoS Biol. 13:e1002224.

Swofford D.L. 2002. PAUP*. Phylogenetic analysis using parsimony (*and other methods). 4.0 ed. Sunderland, MA: Sinauer.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460.

Than C., Nakhleh L. 2009. Species tree inference by minimizing deep coalescences. PLoS Comput. Biol. 5:e1000501.

Thawornwattana Y., Dalquen D., Yang Z. 2018. Coalescent analysis of phylogenomic data confidently resolves the species relationships in the *Anopheles gambiae* species complex. Mol. Biol. Evol. 35:2512–2527.

Thompson E.A. 1975. Human evolutionary trees. Cambridge, UK: Cambridge University Press.

Thompson R., Anderson K. 2000. Biomes of western North America at 18,000, 6000 and 0 ^{14}C yr BP reconstructed from pollen and packrat midden data. J. Biogeogr. 27:555–584.

Tria F., Landan G., Dagan T. 2017. Phylogenetic rooting using minimal ancestor deviation. Nat. Ecol. Evol. 1:193.

Turner R., Bowers J., Burgess T. 1995. Sonoran desert plants: an ecological atlas. Tucson, AZ: University of Arizona Press.

Van Devender T.R. 1987. Holocene vegetation and climate in the Puerto Blanco mountains, Southwestern Arizona. Q. Res. 27:51–72.

Wang J.R. 2013. Analysis and visualization of local phylogenetic structure within species. [Thesis]. UNC Chapel Hill.

Wang M., Zhang L., Zhang Z., Li M., Wang D., Zhang X., Xi Z., Keefover-Ring K., Smart L., DiFazio S., Olson M., Yin T., Liu J., Ma T. 2020. Phylogenomics of the genus *Populus* reveals extensive interspecific gene flow and balancing selection. New Phytol. 225:1370–1382.

Warren D.L., Geneva A.J., Lanfear R. 2017. RWTY (R we there yet): an R package for examining convergence of Bayesian phylogenetic analyses. Mol. Biol. Evol. 34:1016–1020.

Watterson G. 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul Biol. 7:256–276.

Wen D., Yu Y., Zhu J., Nakhleh L. 2018. Inferring phylogenetic networks using PhyloNet. Syst. Biol. 67:735–740.

Xie W., Lewis P., Fan Y., Kuo L., Chen M.-H. 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. Syst. Biol. 60:150–160.

Xu B., Yang Z.H. 2016. Challenges in species tree estimation under the multispecies coalescent model. Genetics 204:1353–1368.

Yang Z. 1998. On the best evolutionary rate for phylogenetic analysis. Syst. Biol. 47:125–133.

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

Yu Y., Warnow T., Nakhleh L. 2011. Algorithms for MDC-based multi-locus phylogeny inference: beyond rooted binary gene trees on single alleles. J. Comput. Biol. 18:1543–1559.

Zhang C., Ogilvie H., Drummond A., Stadler T. 2018. Bayesian inference of species networks from multilocus sequence data. Mol. Biol. Evol. 35:504–517.

Zhang L.X. 2011. From gene trees to species trees II: species tree inference by minimizing deep coalescence events. IEEE-ACM Trans. Comput. Biol. Bioinf. 8:1685–1691.

Zheng J., Meinhardt L., Goenaga R., Zhang D., Yin Y. 2021. The chromosome-level genome of dragon fruit reveals whole-genome duplication and chromosomal co-localization of betacyanin biosynthetic genes. Hortic. Res. 8:63.

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

Zhu J.F., Nakhleh L. 2018. Inference of species phylogenies from bi-allelic markers using pseudo-likelihood. Bioinformatics 34:i376–i385.