doi:10.1111/evo.12758

Historical introgression among the American live oaks and the comparative nature of tests for introgression

Deren A. R. Eaton,[1] Andrew L. Hipp,[2,3] Antonio González-Rodríguez,[4] and Jeannine Cavender-Bares[5,6]

[1]*Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut, 06511*

[2]*The Morton Arboretum, Lisle, Illinois, 60532*

[3]*Department of Botany, The Field Museum, Chicago, Illinois, 60605*

[4]*Centro de Investigaciones en Ecosistemas, Universidad Nacional Autónoma de México, Morelia, Michoacán, 58190, Mexico*

[5]*Department of Ecology, Evolution and Behavior, University of Minnesota, Saint Paul, Minnesota, 55108*

[6]*E-mail: cavender@umn.edu*

Received March 10, 2015
Accepted August 3, 2015

Introgressive hybridization challenges the concepts we use to define species and infer phylogenetic relationships. Methods for inferring historical introgression from the genomes of extant species, such as ABBA-BABA tests, are widely used, however, their results can be easily misinterpreted. Because these tests are inherently comparative, they are sensitive to the effects of missing data (unsampled species) and nonindependence (hierarchical relationships among species). We demonstrate this using genomic RADseq data sampled from all extant species in the American live oaks (*Quercus* series *Virentes*), a group notorious for hybridization. By considering all species and their phylogenetic relationships, we were able to distinguish true hybridizing lineages from those that falsely appear admixed. Six of seven species show evidence of admixture, often with multiple other species, but which is explained by introgression among a few related lineages occurring in close proximity. We identify the Cuban oak as the most admixed lineage and test alternative scenarios for its origin. The live oaks form a continuous ring-like distribution around the Gulf of Mexico, connected in Cuba, across which they could effectively exchange alleles. However, introgression appears highly localized, suggesting that oak species boundaries and their geographic ranges have remained relatively stable over evolutionary time.

KEY WORDS: **Admixture, Cuba, hybridization, phylogeny, *Quercus*, RADseq.**

Introgressive hybridization is a common phenomenon among biological organisms, including our own species (Green et al. 2010). It impacts how we understand the nature of species and infer their historical relationships, with important implications for conservation and biodiversity research (Rhymer and Simberloff 1996). Because introgression between divergent lineages can give rise to genetically admixed individuals and populations that are heterogeneously distributed in space and/or time (Avise 2000; Petit and Excoffier 2009), sampling such individuals will generally bias estimates for the order and timing of species divergences (Leaché et al. 2014). Yet phylogenetic studies rarely sample a sufficient number and variety of individuals to detect whether admixture is present, or variable within species. Similarly, the common practice of excluding apparent hybrid individuals from phylogenetic studies prevents researchers from evaluating their influence on phylogeny. To the extent that introgression is common, the practice of sparse sampling in phylogenetics will underestimate its frequency, and in doing so infer an inflated role for stochastic processes, such as incomplete lineage sorting (Maddison and Knowles 2006), in explaining discordant genealogical relationships.

© 2015 The Author(s).

EVOLUTION
INTERNATIONAL JOURNAL OF ORGANIC EVOLUTION

Recent years have seen the development of new methods for inferring admixture from the genomes of extant species (Green et al. 2010; Durand et al. 2011), the results from which are often interpreted as evidence of introgression between their ancestors. Connecting pattern (admixture) and process (introgression) in this way is a difficult problem, however, and one that similarly suffers from the effects of sparse taxon sampling. To account for such effects, we highlight two important considerations that should generally be taken into account. First, the problem of missing samples: when the true source of introgression is not sampled (i.e., it is a ghost lineage), the source will usually be incorrectly attributed to the sampled population most closely related to the ghost lineage (Durand et al. 2011; Eaton and Ree 2013; Rogers and Bohlender 2015). In practice, the extent to which truly spurious conclusions would be drawn from sampling a closest available (or extant) lineage will generally depend on the size of the clade to which hybridizing lineages belong, and their rate of ecological or morphological divergence. Diverse clades would require sampling many or all species to identify that a species which appears admixed does not have a close relative harboring a yet stronger signal of admixture.

A second and related consideration is that even when all relevant lineages are sampled in a study, it still remains difficult to distinguish a history of introgression between two populations from a signal of admixture between those populations that can arise when one species harbors introgressed alleles from a close relative of the other (Eaton and Ree 2013). To distinguish true introgression from such secondary genomic admixture, introgression must be considered in an explicitly hierarchical (phylogenetic) context, rather than on a species-by-species basis. For example, suppose there are two species, A and D, which exchanged alleles at some time in the past. Species A is member of a clade including several other species (B and C) with which it shares many derived alleles since their divergence from D. As a consequence of their relatedness, introgression from species A into D will necessarily introduce alleles that A also shares with its close relatives, which can give the appearance (admixture) that B and C also hybridized with D. To identify whether the relatives of A independently introgressed into D, versus whether they simply share ancestry with the true hybridizing lineage, requires not only sampling all relevant lineages in the clade but also accounting for their phylogenetic structure.

Oaks (*Quercus*) are notorious for hybridization (Burger 1975; Hardin 1975) to the extent they have been dubbed a "worst case scenario for the biological species concept" (Coyne and Orr 2004). For this reason, they also provide a compelling case study for investigating introgression at the clade level, among multiple interacting species. Within *Quercus*, the American live oaks (*Quercus* subsection *Virentes* Nixon) form a young clade of seven ecologically divergent species that span a range of climatic regimes from the seasonal dry tropics to the temperate zone (Muller 1961; Nixon 1985; Cavender-Bares et al. 2011, 2015). They include both narrow endemics and widespread species that collectively cover the southeastern United States, eastern Mexico, southern Baja, Central America, and Cuba (Fig. 1 A). The species are all diploid and interfertile, and many occur in sympatry throughout all or parts of their range. A complex history of hybridization has likely contributed to difficulties in resolving their phylogenetic relationships (Cavender-Bares and Pahlich 2009; Gugger and Cavender-Bares 2013).

The live oaks are part of a predominately American oak clade (Pearse and Hipp 2009; Hipp et al. 2014) comprising sections *Quercus* (the white oaks sensu stricto, including the live oaks of the Americas and roburoids of Eurasia), *Lobatae* Loudon (the red or black oaks), and *Protobalanus* (Trelease) A. Camus (the intermediate or golden oaks). Although hybrids are commonly observed within each major section (Hardin 1975), hybrid swarms are uncommon, as is hybridization between major sections (Muller 1961). The red and white oak clades diverged approximately 40 Ma (Borgardt and Pigg 1999), and the live oaks split from the remaining white oaks 27–31 Ma (Cavender-Bares et al. 2015). Because they are phylogenetically distant and isolated from all other oak species, the live oaks provide a manageable system in which to reconstruct a clade-level history of introgression.

Here, we utilize restriction-site associated DNA sequencing (RADseq) (Baird et al. 2008) to sample thousands of genomic regions across a large number of samples for phylogenetic inference, and to test introgression between lineages. A recent study demonstrating high conservation of RAD sequences across a phylogenetic scale spanning more than 40 Mya in the American clade oaks (Hipp et al. 2014) motivates our current study. Although genetic admixture has been previously described in the live oaks between focal species pairs (Cavender-Bares and Pahlich 2009; Gugger and Cavender-Bares 2013), this is the first study to bring genome-scale data to bear on the question, and more importantly, to investigate introgression among all extant species in the clade simultaneously and within a phylogenetic context.

We focus particular attention to resolving the phylogenetic placement of the Cuban oak species, *Q. sagraeana*. The origin of this isolated and distinct taxon has long puzzled systematists: its origin has been variously ascribed to one or more species in the southeastern United States, to a Central American species, or to hybridization among other live oaks (Muller 1961; Nixon 1985; Gugger and Cavender-Bares 2013). Chloroplasts are commonly exchanged between sympatric oak species (Whittemore and Schaal 1991; Petit et al. 1997), and consequently chloroplast DNA (cpDNA) haplotypes exhibit little species specificity compared to nuclear markers (Dumolin-Lapegue et al. 1999; Petit and

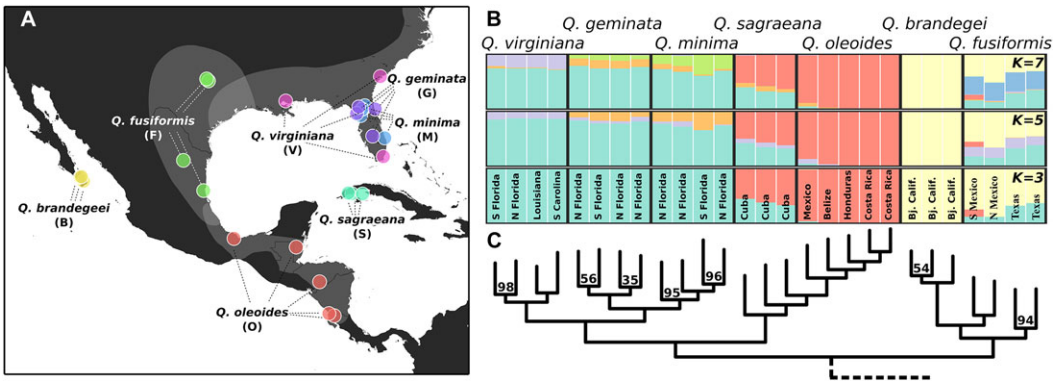

Figure 1. **Sampling locations spanning the geographic ranges of each of the seven live oak taxa. The approximate ranges of the three most wide-ranging species are shown in gray. (B) Population clustering inferred with admixture at three values of *K*. Sampling locations are indicated. (C) Rooted ML phylogeny inferred from the largest (Allmin4) concatenated RADseq dataset. Only ingroup taxa are shown. Bootstrap support is 100 except where indicated.**

Excoffier 2009). The cpDNA haplotype common in Cuba is also shared with both of its hypothesized parent lineages, and is thus inconclusive about the biogeographic origins of the species (Gugger and Cavender-Bares 2013). Using >70K RAD loci sequenced from multiple individuals across the geographic ranges of all seven extant species of live oaks, we ask the following: (1) Which lineages have experienced hybrid introgression? (2) How does admixture affect phylogenetic inference? (3) Can we tease apart nonindependent signals of admixture among multiple closely related species? and (4) What is the origin of the Cuban oak?

Materials and Methods

SAMPLING

Four to five individuals were sampled from across the geographic range of each of the seven live oak species for RAD sequencing (Fig. 1A), in addition to seven outgroup samples (four non-*Virentes* white oaks: *Q. engelmannii*, *Q. arizonica*, *Q. durata*, *Q. douglasii*; one golden oak: *Q. chrysolepis*; and two red oaks: *Q. nigra*, *Q. hemisphaerica*). Leaf samples were collected from wild plants (live oaks) or plants grown in the University of Minnesota greenhouse (outgroup samples). Identification to species was based on leaf, bark, and stem height characters following Muller (1961); Kurz and Godfrey (1962), and Nixon and Muller (1997). Leaves were collected from wild plants in the field, maintained fresh during transport, and stored at −80°C until extraction. Voucher specimens for all RAD-sequenced individuals are housed in the University of Minnesota Bell Museum of Natural History (Table S1).

RADseq PREPARATION AND SEQUENCING

DNA was extracted from fresh or frozen material using the DNeasy plant extraction protocol (DNeasy, Qiagen, Valencia, CA) as reported in Cavender-Bares and Pahlich (2009). RAD libraries were prepared by Floragenex Inc. (Eugene, OR) using the *Pst*I restriction enzyme and sonication following the methods of Baird et al. (2008). An initial multiplex library was created from 30 barcoded and pooled samples sequenced on an Illumina GAIIx sequencer to generate 100 bp single end reads. To increase coverage, a second library was prepared that included an additional 15 samples, seven of which were technical replicates of samples in the first library, sequenced on an Illumina Hi Seq 2000 to generate 100 bp single end reads. After an initial analysis to check that technical replicates grouped together in phylogenetic analyses, they were combined, except for one replicate that may have been contaminated and was excluded. Two additional samples were discarded during bioinformatic analyses due to low sequencing coverage ("TXVW2" and "CUMM5") resulting in 34 final samples.

RADseq ASSEMBLY

Data were assembled into *de novo* loci using *pyRAD* version 2.13 (Eaton 2014). Quality filtering converted base calls with a score <20 into *N*s and reads with >5 *N*s were discarded. Illumina adapters and fragmented sequences were removed using the filter setting "1" in *pyRAD*. Filtered reads were clustered at two different thresholds for within-sample clustering, 85% and 92%, both of which yielded similar results (not shown), therefore we report only the 85% run. Error rate and heterozygosity were jointly estimated from aligned clusters for each sampled individual and the average parameter values were used when making consensus base calls. Clusters with a minimum depth of coverage <5 were excluded. Loci containing more than two alleles after error correction were excluded as potential paralogs (all taxa in this study are diploid). Consensus loci were then clustered across samples at 85% similarity and aligned. A final filtering step excluded

any loci containing one or more sites that appear heterozygous across more than five samples, as we suspect this is more likely to represent a fixed difference among clustered paralogs than a true polymorphism at the scale of this study. The final assembly statistics appeared robust to the choice of filtering thresholds.

In addition to assembling full datasets, smaller matrices were also assembled in which taxa from one or two major clades were selectively excluded. This allowed phylogenetic inference to be performed separately for each major clade in the live oaks, rooted by the outgroups, but without the influence of shared single nucleotide polymorphism (SNPs) between taxa from distant ingroup clades. The motivation for this approach is that to the extent introgression has introduced synapomorphies between distant relatives, subsampling will censor their effect, making them appear instead as autapomorphies (Eaton and Ree 2013). To explore the effect of missing data (locus dropout or low coverage), we also assembled each dataset with different minimums for sample coverage (the number of samples for which data must be recovered to include a RAD locus in the dataset). A large but incomplete version required at least four samples that have data for a locus (e.g., "Allmin4"), whereas a smaller more complete version was also assembled (e.g., "Allmin20"). In total, 15 datasets were generated. The source of missing data between samples was investigated using Mantel tests (9999 permutations) that measured Spearman's rank correlation between the Jaccard's distance of the proportion of shared loci between samples, pair-wise phylogenetic distance, and number of raw input reads.

PHYLOGENY AND POPULATION CLUSTERING

For each assembled dataset, RAD loci were concatenated and missing data entered as *N*s to create a phylogenetic supermatrix. Maximum likelihood (ML) trees were inferred in RAxML version 7.2.8 (Stamatakis 2014) with bootstrap support estimated from 200 replicate searches from random starting trees using the GTR+Γ nucleotide substitution model.

To better visualize genomic variation within individuals, we inferred population clustering with admixture from SNP frequency data within the program *Structure* version 2.3.1 (Pritchard et al. 2000). To minimize missing data across individuals, we used 14,011 putatively unlinked biallelic SNPs, sampled by selecting a single SNP from each locus in the "Ingroupmin20" dataset (17% missing data), which includes only ingroup samples and requires that a locus contain data for at least 20 samples. Ten replicates were run at each value of K between 2 and 8. Each run had a burn-in of 50K generations followed by 500K generations of sampling. Replicates were permuted in the program *CLUMPP* (Jakobsson and Rosenberg 2007), and the optimal K was inferred using the online resource *StructureHarvester* (Earl and vonHoldt 2012).

We also used the program *TreeMix* (version 1.12; Pickrell and Pritchard 2012) to jointly estimate a tree topology (or graph) with admixture using pooled SNP frequency data. For this, individuals were pooled into populations matching to species designations except for *Q. fusiformis* which was split into separate populations for samples from Mexico and Texas. The four non-*Virentes* white oak samples were pooled as an outgroup population. A single biallelic SNP was randomly sampled from each variable locus that contained data for at least one individual across all populations, yielding a total of 12,061 biallelic SNPs. We inferred a topology without admixture, as well as when allowing between one and five admixture events.

INTROGRESSION ANALYSES

The four-taxon D-statistic (Durand et al. 2011) is a metric for detecting admixture between diverged lineages based on the frequencies of SNPs that are discordant with a hypothesized species tree topology. It was most notably used to demonstrate introgression between Neanderthals and modern humans from full genome data (Green et al. 2010), and has similarly been applied to non-model organisms using RADseq data (The Heliconius Genome Consortium 2012; Eaton and Ree 2013). Given a four-taxon pectinate tree [(((P1,P2),P3),O)] in which the outgroup/ancestral allele is labeled "A," and a derived allele labeled "B," the D-statistic compares the occurrence of two discordant site patterns, ABBA and BABA, representing sites in which an allele is derived in P3 relative to O, and is derived in one but not both of the sister lineages P1 and P2. These discordant sites can arise through the sorting of ancestral polymorphisms, but will generally do so with equal frequency due to the stochastic nature of this process. Alternatively, they may arise if introgression occurs between P3 and either P2 or P1, in which case one site pattern will occur more frequently than the other. The D-statistic provides a test for historical admixture by calculating asymmetry in the relative occurrence of these two discordant site patterns:

$$D(P1, P2, P3, O) = \frac{\Sigma_{i=1}^{n} C_{ABBA}(i) - C_{BABA}(i)}{\Sigma_{i=1}^{n} C_{ABBA}(i) + C_{BABA}(i)}, \quad (1)$$

where $C_{ABBA}(i)$ and $C_{BABA}(i)$ are indicator variables of 0 or 1 depending on whether ABBA or BABA is present at each site. Following Durand et al. (2011), we used SNP frequencies instead of allele counts in this study to allow for the inclusion of heterozygous sites. Thus, D was calculated as:

$$D(P1, P2, P3, O) = \frac{\Sigma_{i=1}^{n}[(1-\hat{p}_{i1})\hat{p}_{i2}\hat{p}_{i3}(1-\hat{p}_{i4}) - \hat{p}_{i1}(1-\hat{p}_{i2})\hat{p}_{i3}(1-\hat{p}_{i4})]}{\Sigma_{i=1}^{n}[(1-\hat{p}_{i1})\hat{p}_{i2}\hat{p}_{i3}(1-\hat{p}_{i4}) + \hat{p}_{i1}(1-\hat{p}_{i2})\hat{p}_{i3}(1-\hat{p}_{i4})]}, \quad (2)$$

where $\hat{p}_{i1}$ is the frequency of the derived allele in taxon P1 at site i. If the sampled individual has both copies of the derived allele at this site $\hat{p}_{i1}$=1.0, if it is heterozygous $\hat{p}_{i1}$=0.5, otherwise $\hat{p}_{i1}$=0.0. We calculated D over all combinations of four taxa

fitting the ML topology as well as alternative topologies of interest. For ingroup taxa, we iterated over each sampled individual separately, but for the outgroup taxon instead used a pooled group of samples to measure the SNP frequency. This was made up of the four non-*Virentes* white oak samples, with $\hat{p}_{i4}$ calculated as the frequency of derived alleles in all $2N$ locus copies for N outgroup individuals containing data for a given site. This allowed us to maximize the use of RADseq data with missing sequences because we could use any locus for which the three sampled ingroup taxa shared data with at least one outgroup. This approach also has the effect of down-weighting D if the ancestral allele is not fixed across multiple outgroup samples, making it a more conservative test.

For each test, we measured the standard deviation of D from 200 bootstrap replicates in which RAD loci were re-sampled with replacement to the same number as in the original dataset, as in Eaton and Ree (2013). The observed D was converted to a Z-score measuring the number of standard deviations it deviates from 0, and significance was assessed from a P-value using $\alpha = 0.01$ as a conservative cutoff after Holm–Bonferoni correction for multiple testing (number of possible sample combinations fitting the given species tree hypothesis).

Partitioned D-statistics (Eaton and Ree 2013) are an extension to this test relevant at deeper evolutionary time scales where the P3 lineage may include multiple distinct sublineages with independent histories of introgression. It measures a five-part allele pattern [(((P1,P2),($P3_1$,$P3_2$)),O)], and contrasts two P3 sublineages at a time by measuring D for three separate pairs of allele counts (ABBBA/BABBA, ABBAA/BABAA, and ABABA/BAABA). These statistics measure asymmetry in the occurrence of derived alleles present in both P3 sublineages (D_{12}), only $P3_1$ (D_1), or only $P3_2$ (D_2), and present in P2 or P1 but not both (Fig. 2A).

$$D_1(\mathrm{P1}, \mathrm{P2}, \mathrm{P3_1}, \mathrm{P3_2}, \mathrm{O}) = \frac{\Sigma_{i=1}^{n} C_{ABBAA}(i) - C_{BABAA}(i)}{\Sigma_{i=1}^{n} C_{ABBAA}(i) + C_{BABAA}(i)} \quad (3)$$

$$D_2(\mathrm{P1}, \mathrm{P2}, \mathrm{P3_1}, \mathrm{P3_2}, \mathrm{O}) = \frac{\Sigma_{i=1}^{n} C_{ABABA}(i) - C_{BAABA}(i)}{\Sigma_{i=1}^{n} C_{ABABA}(i) + C_{BAABA}(i)} \quad (4)$$

$$D_{12}(\mathrm{P1}, \mathrm{P2}, \mathrm{P3_1}, \mathrm{P3_2}, \mathrm{O}) = \frac{\Sigma_{i=1}^{n} C_{ABBBA}(i) - C_{BABBA}(i)}{\Sigma_{i=1}^{n} C_{ABBBA}(i) + C_{BABBA}(i)} \quad (5)$$

As in the four-taxon tests, we used the four non-*Virentes* white oak samples to represent the outgroup, and used a SNP frequency-based version of the test to include data for heterozygous individuals. All D-statistics were measured in pyRAD version 2.13.

In contrast to the four-taxon D-statistic, the partitioned test is polarized by defining P3 as a donor lineage, and P2 or P1 as recipients, which allows D_{12} to act as an indicator of the direction of introgression. Briefly, consider a case where introgression occurred in the reverse direction from how we assign samples to the tips of the tree (e.g., from P2 into $P3_1$); in this case, $P3_2$ would not contain the same derived alleles that P2 shares with $P3_1$ through introgression, and thus the indicator variable D_{12} would be nonsignificant, indicating introgression did not occur in this direction. If we then swap samples across the tips to re-define the P3 lineage, such that introgression occurred from the defined $P3_1$ sublineage into P2, we would now find that $P3_2$ also shares many of the same introgressed alleles that $P3_1$ shares with P2 (significant D_{12}), due to the fact that many of these alleles arose in the ancestor of the two sampled P3 sublineages. In addition to indicating directionality, partitioning ancestral alleles from those that are derived uniquely to either P3 sublineages also allows us to distinguish whether introgression occurred from each P3 sublineage independently into P1 or P2, or if it occurred from only one (Eaton and Ree 2013). We apply this test to two separate cases in the live oaks, involving *Q. fusiformis* and *Q. sagraeana*, in which four-taxon tests show evidence of admixture involving more than two taxa, to test whether each taxon pair hybridized independently.

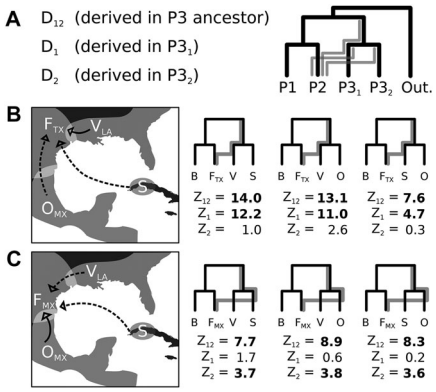

Figure 2. Teasing apart nonindependent signals of admixture. (A) Partitioned *D*-statistics test for directional introgression from the P3 lineage into P2 or P1 and contrast P3 sublineages as introgressive donors. Results are reported as *Z*-scores. (B) Three closely related lineages (S, V, and O; taxon names abbreviated as in Fig. 1) each share alleles with F in Texas to the exclusion of B (significant D_{12}), but when contrasted against each other (D_1 and D_2) only V shares uniquely introgressed alleles with F_{TX} relative to the other two P3 sublineages. (C) A similar test examining F from coastal Mexico shows the opposite result: F_{MX} only shares uniquely introgressed alleles with O, whereas apparent admixture between F_{MX} and S or V is a consequence of the shared ancestry of O with S and V.

DEMOGRAPHIC MODELS

To investigate the origin of the Cuban oak, we compared the joint site frequency spectrum (SFS) generated under three demographic

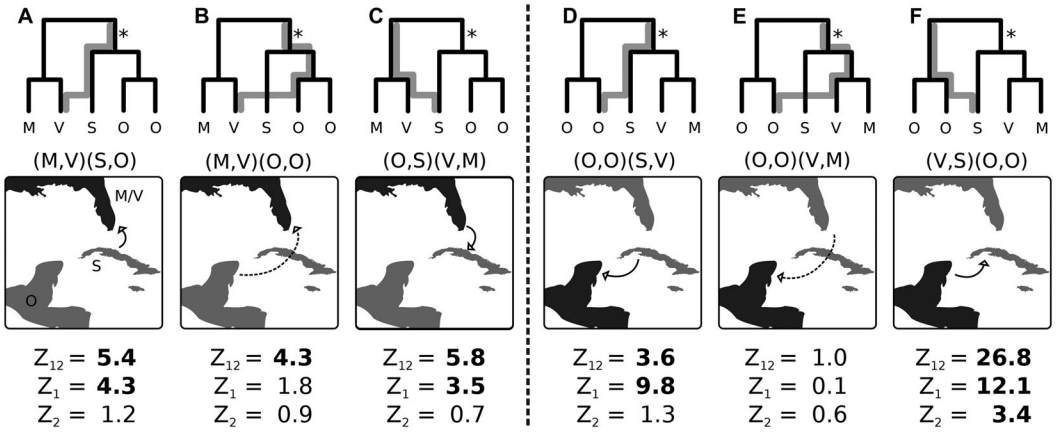

Figure 3. **Partitioned *D*-statistics testing two hypotheses of divergence and gene flow in the Cuban oak. In hypothesis 1 (A–C), S shares an MRCA with O (light gray on map); in hypothesis 2 (D–F), S shares an MRCA with the southeastern U.S. clade (taxon abbreviations are as in Fig. 1). An asterisk marks the hypothesized ancestral relationship of S with either lineage. For each scenario, sampled tips are shown in the following order (P1,P2)(P3$_1$,P3$_2$). The direction of introgression being tested is indicated by an arrow on the map, and a gray line traces the path on the topology through which shared ancestral P3 alleles are introduced into P2 to the exclusion of P1. *D*-statistics are reported as *Z*-scores.**

isolation-migration models (Fig. 4A) to that in our observed data, with a focus on SNPs segregating within and between populations of *Q. oleoides*, *Q. sagraeana*, and the southeastern U.S. clade, using the program ∂a∂i (Gutenkunst et al. 2009). Data were pooled for the three closely related species in the southeastern clade, and the SFS was projected down to require that every locus contain data for at least five individuals in this clade, three individuals in *Q. oleoides*, and three individuals in *Q. sagraeana* (projected chromosomes = [10,6,6]). A single biallelic SNP was randomly selected from each variable locus, yielding 1,626 SNPs from 7,794 usable loci after data projection.

The first two demographic models have nine parameters and differ only in their topology: in model 1 the Cuban oak is derived from a taxon in the southeastern United States (abbreviated MGV), whereas in model 2 it originates from Central America (Fig. 4 A). Model parameters include effective population sizes for each population (N_{MGV}, N_O, and N_S) and migration rates between adjacent populations (m_{S-MGV}, m_{MGV-S}, m_{S-O}, m_{O-S}). At time T_2, two ancestral populations diverge (viewed forward in time), and at time T_1 the Cuban population diverges from its sister lineage to maintain a separate constant population size. Model 3 has only seven parameters. In this model, T_2 is again the divergence time for two ancestral populations, but T_1 is now an event in which an independent Cuban population is formed by an instantaneous fusion of a proportion (f) of the southeastern U.S. clade and ($1 - f$) of *Q. oleoides*. There is no further migration between populations.

We used the log L-BFGS-B optimization method to fit parameters for each model. Searches were started from 10 randomly perturbed starting positions, for a maximum of five iterations, followed by a final search using the best-inferred parameters from the previous step as a starting position for a maximum of 20 additional iterations. Extrapolation was performed with a grid size of [12,20,32]. To attain confidence intervals on parameter estimates, we performed parametric bootstrapping by simulating 200 datasets for each of the three models using the program *ms* (Hudson 2002). Bootstrap SFS data were simulated under their ML estimated parameter values and then re-optimized in ∂a∂i to estimate the parameters that would generate these data under the same model by which they were generated.

There exists a multitude of possible demographic models for three populations and it is likely that the true model is more complex than those which we selected to compare (Pelletier and Carstens 2014). Rather than attempt to explore the space of all possible models, we focus instead on assessing the statistical power to distinguish among these three simple scenarios, and whether the best model is concordant with our phylogenetic and *D*-statistic results. For this we used simulated datasets to perform Monte Carlo model selection (Boettiger et al. 2012), in which each dataset was fit to the model under which it was simulated in addition to the two alternative models (three models; 200 bootstrap datasets each; 1800 model fits total). For each comparison, a likelihood ratio ($\delta = -2[\log L_0 - \log L_1]$) was calculated. Larger values for δ indicate more support for model 1 (the alternative) relative to model 0 (the null). Our goal in model selection is to calculate how big δ should be to decide that model 1 is closer to the truth than model 0 (Boettiger et al. 2012). Power to distinguish models, and the sensitivity of our tests, was assessed from the overlap in distributions of δ values from simulated data, and their comparison to δ for our observed data.

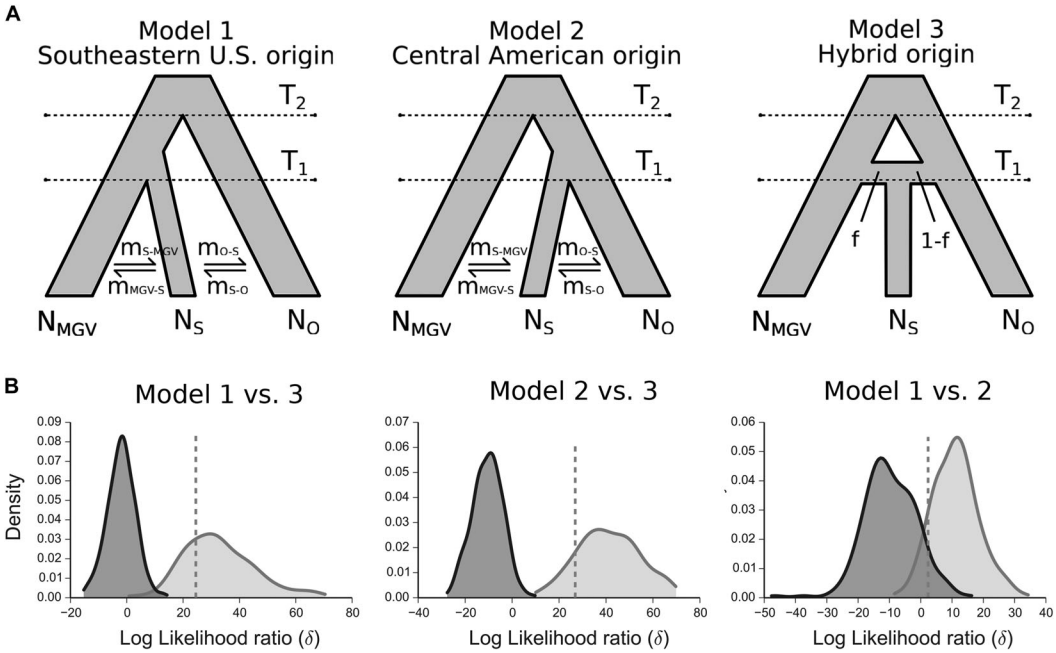

Figure 4. **Three demographic models for the origin of the Cuban oak (S). (A) In models 1 and 2 (nine parameters), S is derived from one mainland taxon or the other (O or MGV; taxon names are abbreviated as in Fig. 1) with subsequent migration between Cuba and either mainland lineage. In model 3 (seven parameters), S forms through instantaneous admixture (hybrid speciation) and remains isolated thereafter. (B) Results of Monte Carlo model comparisons. Distributions of likelihood ratios (δ) show the difference in fit between models when data are simulated under one model or the other. The likelihood ratio fit between models for our observed data is shown in red (δ_{obs}). The proportion of the null model's δ distribution (dark gray) to the right of δ_{obs} measures the false-positive rate, and the proportion of the alternative model's δ distribution (light gray) that overlaps with the null distribution measures the power to reject the null. Model 2 is the best fit to our observed data.**

REPRODUCIBILITY

Scripts to download raw sequence data (NCBI: SRP055977), assemble it, and reproduce all analyses in this study are compiled into IPython notebooks (Pérez and Granger 2007), a tool for reproducible science, available at https://github.com/dereneaton/virentes. Assembled datasets are also archived (DOI 10.528/zenodo.19475).

Results

RAD DATA ASSEMBLY

Following quality filtering and clustering (85% similarity) 77M raw reads (mean ± SD 2.13M ± 1.75M per sample) were reduced to an average of 57K ± 25K high coverage stacks per sample, with a mean depth of 23X. These were further filtered to 52K± 22K consensus sequences per sample (Table S1). Datasets that were assembled with different minimums for sample coverage or with samples excluded had different proportions of missing data: The largest but most incomplete assembled data matrix that includes all loci shared across at least four samples (Allmin4) has 55.5% missing data for 34 individuals across 78,727 loci, whereas all other matrices have fewer missing data (9.6–52.1%; Table 1).

The distribution of missing data did not show strong hierarchical structure, as would be expected if most missing data was caused by locus dropout due to the disruption of restriction recognition sites (Fig. S1). Instead, for the largest dataset (Allmin4), the mean number of raw reads was a better predictor for the number of shared loci between samples than was the phylogenetic distance between samples (Mantel $r_\rho = 0.372$, $P = 0.010$, and $r_\rho = -0.145$, $P = 0.240$, respectively). A similar result was observed in the more complete "Allmin20" dataset (Mantel $r_\rho = 0.479$, $P = 0.002$, and $r_\rho = 0.087$, $P = 0.523$, respectively), suggesting that sequencing effort had a more significant impact on missing data than relatedness.

PHYLOGENY

Missing data (the sparseness of concatenated matrices) had little effect on phylogenetic inference as the larger and more incomplete versions of each dataset yielded similar or identical topologies to the smaller more complete version of that matrix (e.g., Allmin4 and Allmin20; Fig. S2), the latter often with lower bootstrap supports. All phylogenetic analyses recovered perfect support for three major clades: a southeastern U.S. clade (*Q. minima*, *Q. geminata*, and *Q. virginiana*), a southwestern clade (*Q. brandegeei* and

Table 1. Size, completeness, and the number of parsimony informative sites (PIS) in 15 assembled RADseq datasets.

Dataset	*N* samples	*N* loci	PIS	% Missing
Allmin4	34	78,727	251,986	55.47
Allmin20	34	27,369	110,500	26.59
Ingroupmin20	27	15,123	29,957	16.99
MGVmin4	19	72,849	207,713	46.72
MGVmin16	19	9,464	33,829	12.09
OSmin4	15	68,453	182,896	42.22
OSmin13	15	10,845	36,983	9.67
FBmin4	14	69,205	187,949	39.44
FBmin12	14	14,980	51,850	9.60
OSMGVmin4	27	76,839	235,345	52.11
OSMGVmin20	27	15,904	60,873	18.61
FBMGVmin4	26	77,523	239,513	50.93
FBMGVmin20	26	14,925	57,923	16.66
FBOSmin4	22	76,379	230,366	47.63
FBOSmin20	22	21,905	83,516	17.84

[1]Names correspond to abbreviations for the included taxa and the minimum sample coverage for included loci.

Q. fusiformis), and a Central American clade (*Q. oleoides* and *Q. sagraeana*; Fig. 1C). Selectively excluding taxa sometimes yielded different relationships within each major clade, as expected if synapomorphies that are derived from introgression between lineages affect phylogenetic inference (Eaton and Ree 2013). For example, *Q. fusiformis* appears paraphyletic with respect to its putative sister taxon *Q. brandegeei* in datasets that include samples from all three major clades, but monophyly of *Q. fusiformis* is supported when the two other live oak subclades are excluded (Fig. S2E). A similar pattern is observed for the three southeastern U.S. clade oaks, where *Q. virginiana* appears sister to the other two species in full datasets, but *Q. minima* is sister to the other two species when the southwest and Central American clades are excluded (Fig. S2G). The phylogenetic instability of *Q. virginiana* and *Q. fusiformis* is consistent with further evidence below that they have exchanged genes in Texas where they occur in sympatry and that this affects their phylogenetic placement.

POPULATION STRUCTURE

Population clustering analyses revealed substantial heterogeneity in proportions of admixed ancestry within and between species. The best supported model ($K = 3$) clustered populations into the same three major clades described above. The three oak species of the southeastern U.S. clade are indistinguishable at low values of K (the number of distinct clusters; Figs. 1B and S3), and much of their common ancestry is also shared through apparent admixture with both of their geographically adjacent taxa: *Q. fusiformis* in Texas to the west and *Q. sagraeana* in Cuba to the south. *Quercus sagraeana* also shares significant ancestry with *Q. oleoides* from Central America. In the southwest, *Q. fusiformis* shares ancestry with *Q. brandegeei* and *Q. virginiana*. In contrast, *Q. oleoides* forms a nearly distinct cluster, except for the sample from Mexico which shows slight admixture with different groups at different K values. Only *Q. brandegeei*, endemic to southern Baja California, forms a distinct nonadmixed cluster in all analyses above $K = 2$, suggesting it has remained genetically isolated from all other populations sampled in our study. Within each species, individuals with the greatest proportions of admixed ancestry appear as the earliest diverging in their clade (Fig. 1B and C), suggesting that inferred population-level relationships may reflect admixture proportions to a greater degree than they do historical population divergences—a major concern for phylogeographic studies below the species level.

TreeMix

TreeMix recovered the same topology for population-level relationships as our concatenated ML analyses performed on individuals. With the addition of one admixture edge, approximately 40% admixed ancestry is inferred between *Q. sagraeana* and a lineage from the southeastern U.S. clade, which also changes the backbone topology such that *Q. oleoides* is supported as sister to the remaining live oaks (Fig. S4). Adding a second admixture edge returns a graph similar to that of the original tree topology, but with admixture between *Q. virginiana* and *Q. sagraeana* (47% ancestry), and between *Q. virginiana* and *Q. fusiformis* in Texas (24% ancestry). A notable result of the latter edge is its effect on *Q. brandegeei*, which becomes no longer nested within *Q. fusiformis*. This shows how, despite being completely isolated from admixture itself, introgression occurring into a close relative of *Q. brandegeei* can still affect its phylogenetic placement.

The first admixture edge increases the log-likelihood (LL) by 68.2, the second edge by 60.6, whereas a third edge increases the LL by only 12.2, and all additional edges by less than 5. The first two inferred edges are concordant with *D*-statistic results reported below, and support admixture between *Q. virginiana* and both *Q. fusiformis* in Texas and *Q. sagraeana* in Cuba. The third inferred edge (Fig. S4), which shows admixture between the outgroup population and *Q. minima*, provides only a small improvement to the LL score and is not strongly supported by *D*-statistic results.

D-STATISTICS

Nonparametric *D*-statistics (ABBA-BABA tests) revealed substantial heterogeneity in the presence of admixture within and between species (Table 2). Few tests detected admixture uniformly across all iterations of sampled individuals. Significant results were largely limited to samples that occurred in close geographic proximity. For example, among the three sympatric oaks

Table 2. Four-taxon *D*-statistic tests for admixture.

Test	P1	P2	P3	Range Z^2	nSig/N^3
1	G	G	M	(0.0, 2.3)	0/23
2	M	M	G	(1.3, **6.8**)	12/23
3	G	G	V	(0.2, 2.4	0/17
4	M	M	V	(0.2, **4.7**)	7/17
5	M	G	V	(0.1, **7.9**)	28/47
6	V	V	M	(0.0, 1.6)	0/11
7	V	V	G	(0.1, 2.5)	0/11
8	V	G	M	(0.0, **3.9**)	1/11
9	O	S	(MGV)	(**3.1**, **16.2**)	164/164
10	(MGV)	S	O	(**14.7**, **36.4**)	164/164
11	(MGV)	O	S	(**6.8**, **25.8**)	164/164
12	O	O	F	(0.0, 1.6)	0/39
13	O	O	B	(0.1, 2.4)	0/29
14	B	F	O	(0.0, **8.1**)	29/59
15	B	F	S	(0.9, **8.1**)	30/35
16	B	F	(MGV)	(1.3, **17.9**)	119/131
17	B	B	F	(0.2, 2.6)	0/11
18	S	S	(MGV)	(0.0, **4.1**)	2/32
19	M	V	S	(1.1, **7.1**)	17/35
20	M	G	S	(0.0, **6.9**)	18/47
21	V	G	S	(0.0, 2.9)	0/35
22	(MG)	V	F_{TX}	(**3.6**, **10.6**)	47/47
23	O	O	F_{MX}	(0.0, 1.7)	0/19
24	S	S	O	(0.0, **4.1**)	2/14
25	O	O	S	(0.1, **7.3**)	10/29

[1]Taxon names are abbreviated as in Figure 1 and arranged such that ABBA > BABA. Outgroups not shown. Tests are referred to by number in the text.
[2]Bold indicates significance at $\alpha = 0.01$.
[3]Significant tests over possible sampled individuals.

species in the southeastern United States, *Q. virginiana* shares derived alleles with *Q. geminata* to the exclusion of *Q. minima* when *Q. minima* is sampled from southern Florida, but not when sampled from northern Florida; an apparent consequence of all three taxa being more homogenized in the north (tests 1–5, Table 2). *Q. virginiana* is the only species in this clade to occur widely outside of Florida; however, it shows the same genetic similarity to the other two species in sympatry as it does in allopatry (tests 6 and 7, Table 2), suggesting that *Q. virginiana* has not received introgression from either species in the very recent past. Under an alternative topology in which *Q. minima* is sister to the other two species in the southeastern clade, we detect negligible admixture between *Q. virginiana* and *Q. geminata*, but admixture of both with the more rare taxon *Q. minima* (tests 1–4 and 6–8, Table 2). The most admixed sample of *Q. minima* groups with *Q. geminata* in several phylogenetic analyses (Fig. S2). Both *Q. geminata* and *Q. virginiana* are admixed with *Q. sagraeana* in Cuba, and *Q. virginiana* is also admixed with *Q. fusiformis* in Texas (tests 16 and 18–22, Table 2). Despite this, the three live oak species in the southeast show little genetic differentiation from each other, and thus for simplicity we refer to them as a single pooled taxon (called the southeastern U.S. clade, or abbreviated MGV) in several further analyses.

The Cuban oak, *Q. sagraeana*, shows clear admixture with one or more species in the southeastern U.S. clade and with *Q. oleoides* in Central America. Of the three possible rooted topologies for these three lineages (tests 9–11, Table 2) admixture is greatest when *Q. sagraeana* is sister to the southeastern U.S. clade (in conflict with our phylogenetic results) and exchanging genes with *Q. oleoides*. Here, we see that *Q. sagraeana* shares more derived alleles, to the exclusion of the southeastern clade, with the southernmost populations of *Q. oleoides* (Costa Rica and Honduras) than with northern populations (Mexico and Belize). The alternative test that is concordant with our phylogenetic results entails less admixture, meaning that *Q. sagraeana* shares more alleles with *Q. oleoides* than it does with the southeastern U.S. clade oaks. We suspect that the third possible topology, in which *Q. sagraeana* diverged first from an ancestor of the other two species, is unlikely because *Q. sagraeana* exhibits little independent ancestry relative to the other two lineages (Fig. 1B).

Quercus fusiformis, which ranges from northern Mexico to eastern Texas, shows evidence of admixture with both of the other two major live oak clades, thus spanning the deepest splits in the tree. In Mexico it occurs in sympatry with *Q. oleoides*, and the two form a clear morphological hybrid zone (Cavender-Bares et al. 2015). We did not directly sample this hybrid zone in our genomic dataset, however, the most geographically proximate samples from each taxon show evidence of admixture, suggesting introgression from *Q. oleoides* into *Q. fusiformis* (tests 12–14 and 23, Table 2). In Texas, the range of *Q. fusiformis* overlaps with *Q. virginiana* and the two appear to have exchanged bidirectional gene flow recently (tests 16 and 22, Table 2), since the divergence of *Q. virginiana* from the other two species in the southeastern U.S. clade.

DISTINGUISHING INDEPENDENT INTROGRESSION EVENTS

Reconstructing the history of introgression among lineages does not translate directly from patterns of shared alleles between them, but instead must be placed in a phylogenetic context. A clear example of this can be seen with *Q. fusiformis*, which appears admixed with respect to every other species of live oak save for its sister taxon *Q. brandegeei* (tests 14–17, Table 2). Of its three potential hybridizing partner lineages it seems least likely to have truly hybridized with *Q. sagraeana*, which is allopatric in Cuba, compared to the other two lineages with which it overlaps in Texas or Mexico. By contrasting these lineages as potential donor lineages using partitioned *D*-statistics, we find that the complex patterns of admixture in *Q. fusiformis* can be explained by a

small number of introgression events. The shared derived alleles between *Q. sagraeana* and *Q. fusiformis* in Texas are nearly entirely composed of alleles that these two taxa also share with *Q. virginiana* (Fig. 2B), and similarly, the shared derived alleles between *Q. sagraeana* and *Q. fusiformis* in Mexico are composed almost entirely of alleles also shared with *Q. oleoides* (Fig. 2C). Only *Q. virginiana* shares uniquely introgressed alleles with *Q. fusiformis* in Texas, and only *Q. oleoides* shares uniquely introgressed alleles with *Q. fusiformis* in Mexico. From this we can infer that introgression occurred separately into *Q. fusiformis* from these two distinct lineages, but not from their close relative *Q. sagraeana*, because *Q. sagraeana* does not share introgressed alleles with *Q. fusiformis* to the exclusion of either of its close relatives.

HIDDEN ANCESTRY AND THE CUBAN OAK

That *Q. sagraeana* would share ancestry with both *Q. oleoides* and the southeastern U.S. clade to the exclusion of *Q. fusiformis* is consistent with our phylogenetic reconstructions. It is therefore not surprising that introgression from any one of these three related lineages would introduce shared ancestral alleles from all three. By a similar logic, we investigated the origins of the Cuban oak by applying the same test one node lower in the phylogeny—at the first split between a putative ancestor of *Q. oleoides* and the southeastern clade—to test which of these two putative parental lineages shares more ancestral (non-introgressed) alleles with *Q. sagraeana*. Our intention, therefore, was to detect evidence of a putative most recent common ancestor (MRCA) whose historical signature has become obscured, by finding evidence of their shared ancestry in alleles that are introgressed from one or more of their descendant lineages into another.

We compared two competing hypotheses: (1) *Q. sagraeana* shares an MRCA with *Q. oleoides* from Central America but subsequently exchanged alleles with one or more southeastern clade oaks; or (2) *Q. sagraeana* shares an MRCA with (or within) the southeastern clade oaks but subsequently exchanged genes with *Q. oleoides* (Fig. 3). Both scenarios assume that the ancestral lineage established on Cuba through seed and that later introgression occurred infrequently, either through rare long distance dispersal events or wind-dispersed pollen, most likely at times of low sea level when distances between Cuba and the mainland were reduced.

Partitioning shared versus uniquely derived alleles among these three lineages reveals strong support for the Central American origin hypothesis. If we begin by assuming *Q. oleoides* and *Q. sagraeana* are sister species, we find that *Q. sagraeana* shares a set of uniquely derived alleles with *Q. virginiana* (relative to *Q. minima*; significant D_1), and that a set of derived alleles which putatively arose in the ancestor of *Q. oleoides* and *Q. sagraeana* is also shared with *Q. virginiana* (significant D_{12}), but *Q. oleoides* itself does not share a set of uniquely derived alleles with *Q. virginiana* (nonsignificant D_2; Fig. 3 A; tests 26–31, Table S2). This pattern is consistent with a topology in which *Q. oleoides* and *Q. sagraeana* share an MRCA but introgression occurred from only one descendant lineage. It follows then that if this topology were true all populations of *Q. oleoides* should also share with *Q. virginiana* the set of alleles that arose in the ancestor of *Q. oleoides* and *Q. sagraeana*, despite the fact that *Q. oleoides* never hybridized with *Q. virginiana* directly (they are allopatric). This is precisely what we find (Fig. 3B; tests 32–37, Table S2): shared alleles between *Q. oleoides* populations are present in *Q. virginiana*, but no single *Q. oleoides* population shows significantly greater genetic similarity with *Q. virginiana*. Although this result supports our hypothesized scenario, the true history of divergence and gene flow may be more complex; for example, introgression appears to have also occurred in the reverse direction, from the southeastern clade into Cuba, and most likely more than once because both *Q. virginiana* and *Q. geminata* share a different set of uniquely introgressed alleles with *Q. sagraeana* (Fig. 3C; tests 38–43, Table S2) relative to *Q. oleoides*.

The alternative scenario, in which *Q. sagraeana* is derived from the southeastern U.S. clade, yields patterns of admixture that are less consistent with the existence of a hypothetical MRCA. This is apparent first in the overabundance of uniquely shared alleles between *Q. sagraeana* and *Q. oleoides* (D_1), relative to ancestral alleles that should be derived from the hypothetical MRCA of *Q. sagraeana* and *Q. virginiana* (Fig. 3D; tests 44–47, Table S2). It is further apparent because the putative introgression between *Q. sagraeana* and *Q. oleoides* did not introduce any alleles from *Q. virginiana*, or its other southeastern U.S. clade relatives, which are expected to be introduced alongside alleles from *Q. sagraeana* if they shared an MRCA, and if either acted as an introgressive donor (Fig. 3E; tests 48–53, Table S2). Thus, the strong signal of apparent introgression between *Q. sagraeana* and *Q. oleoides* (Fig. 3F; tests 54–57, Table S2) is most likely, rather, a signal of their shared ancestry made apparent by testing for introgression on an incorrect species tree.

DEMOGRAPHIC MODELS

We further compared these two hypotheses with a third model in which the Cuban population was formed by instantaneous admixture from two parent lineages but remained completely isolated thereafter (Fig. 4A)-a scenario akin to hybrid speciation. By fitting the SFS for these three lineages to demographic models in ∂a∂i (Gutenkunst et al. 2009), we found greatest support for a Central American origin (LL = −541.9), followed by a southeastern U.S. origin (LL = −543.1), and hybrid origin (LL = −555.3) models. The least parameter-rich model (hybrid origin) is easily rejected in favor of the two more complex models: the

Table 3. Maximum likelihood (ML) parameter estimates and 95% confidence intervals (CI) for three demographic models for the origin of the Cuban oak.

	Model 1 (SE origin)		Model 2 (CA origin)		Model 3 (Hybrid origin)	
Parameter	ML	95% CI	ML	95% CI	ML	95% CI
$N_{MGV}(\times 10^3)$	89.04	71.48–10.27	88.34	69.72–100.02	90.89	70.39–104.80
$N_O(\times 10^3)$	24.52	18.59–29.47	24.19	17.82–29.34	28.89	22.63–33.64
$N_S(\times 10^3)$	2.73	0.00–5.30	8.44	2.38–13.46	5.76	0.34–10.70
T_2 (Mya)	1.83	1.43–4.00	1.75	1.19–4.00	1.46	0.81–3.54
T_1 (Mya)	0.32	0.00–0.90	0.19	0.04–0.31	0.06	0.00–0.11
$m_{MGV-a}(\times 10^3)$	0.00	0.00–0.01	0.00	0.00–0.00	–	–
$m_{S-MGV}(\times 10^3)$	0.18	0.02–0.34	0.08	0.01–0.09	–	–
$m_{S-O}(\times 10^3)$	0.02	0.00–0.03	0.06	0.02–0.09	–	–
$m_{O-S}(\times 10^3)$	0.30	0.02–0.52	0.00	0.00–0.00	–	–
f_{MGV}	–	–	–	–	0.38	0.34–0.42

SE = Southeastern U.S., CA = Central America.

difference in LL (δ) between models was greater in our observed data than in all simulated datasets generated under the hybrid origin scenario (Fig. 4B). This test was also very sensitive: at a false-positive rate of 5%, we had >99% power to reject the hybrid origin model. There is no clear null when comparing the remaining two models to each other, as they are non-nested, and equal in number of parameters. Thus, a *P*-value of 5% may be considered overly stringent (Boettiger et al. 2012). The observed δ supporting a Central American origin is greater than 93% of simulations generated under the southeastern U.S. origin model ($P = 0.07$), and using this as our test statistic, we have 92% power to reject a southeastern origin if the other model were true. Or, if we use the traditional cutoff of 5%, we have 85% power to correctly distinguish the models (Fig. 4B). Using 2.5×10^{-9} as the average mutation rate per site per generation (inferred from *Populus*, Tuskan et al. 2006), and an average generation time of 30 years, our best model (Central American origin) infers a crown age for these three lineages of 1.75 (1.19–4.00) Ma, with divergence of *Q. sagraeana* occurring 0.19 (0.04–0.31) Ma (Table 3). Introgression occurred predominately into *Q. sagraeana* from the southeastern clade, and to a lesser extent from *Q. oleoides*.

Discussion

Introgressive hybridization is commonly studied at the scale of individual species pairs (Petit et al. 1997), among multiple sympatric species (Whittemore and Schaal 1991), or in a sampling of closely related species (Kane et al. 2009; The Heliconius Genome Consortium 2012; Gugger and Cavender-Bares 2013; Nadeau et al. 2013), but rarely in the context of all extant species within an ecologically and evolutionarily distinct clade. Here, by sampling all relevant populations and comparing them in a phylogenetic context, we were able to reconstruct a clade-level history of introgression, and to correct many potentially misleading signals of admixture. We find that every pair of species occurring in close geographic proximity has exchanged some amount of gene flow, with no evidence of introgression that is not concordant with species present day geographic distributions. This suggests that geographic ranges of the live oaks, at least relative to each other, have likely remained stable through time. Such stasis is consistent with the fact that live oak species exhibit substantial differences in adaptations to climatic niche, particularly with regard to drought and freezing tolerances (Cavender-Bares and Pahlich 2009; Cavender-Bares et al. 2011; Koehler et al. 2012; Cavender-Bares et al. 2015; Ramirez-Valiente et al. 2015. Together they span a nearly continuous range from temperate, to dry desert, and even tropical climates. A classic hypothesis for limits on the spread of introgressed alleles between species is that such alleles may facilitate adaptations to intermediate environments within hybrid zones, but decrease fitness elsewhere (Barton and Hewitt 1985). In the live oaks, genetic exchange is theoretically possible throughout a ring-like complex composing up to six interconnected, interfertile species that effectively encircle the Gulf of Mexico, including a connection through Cuba. However, our results, as well as those from a companion study that used fewer markers but many more individuals (Cavender-Bares et al. 2015), agree in showing that admixture is largely restricted to hybrid zones.

THE COMPARATIVE NATURE OF TESTS FOR INTROGRESSION

Our analyses demonstrate the difficulty of inferring historical introgression over deep evolutionary time scales (crown 8.4–14.1 Ma; Cavender-Bares et al. 2015). In particular, that sparse sampling can lead to false inferences of hybridization when the source of introgressed alleles is unknown, or stems from multiple

sources, as is common for oaks. This is the case for *Q. fusiformis*, which has experienced introgression with two divergent lineages in opposite ends of its geographic range. Because the two lineages with which it hybridized share a common ancestor since their divergence from *Q. fusiformis* each introduced many of the same alleles into it. They also introduced alleles that they share with their other close relatives, including *Q. sagraeana*. Had we failed to sample all extant species, and thus been unable to contrast their patterns of shared versus uniquely derived alleles, we could have easily been misled as to the source of introgression. For example, consider if *Q. oleoides* had not been sampled, in which case only *Q. sagraeana* would appear to share uniquely introgressed alleles with *Q. fusiformis* in the southern part of its range (Fig. 2 C); and similarly, a failure to sample the southeastern U.S. clade would lead us to infer introgression from *Q. sagraeana* into *Q. fusiformis* in the northern part of its range (Fig. 2B). Given that the true result in each of these cases was that introgression occurred from the most geographically proximate taxon such a distinction may seem trivial. However, if we consider that many studies of introgression focus on only a single species pair, the potential for error, especially in highly diverse clades, is clear. The ability to accurately reconstruct a history of hybridization among multiple closely related species from genomic data would provide an invaluable tool for the study of speciation and reproductive isolation (Rabosky and Matute 2013). The case of the American live oaks makes clear that such histories can be highly complex, and teasing them apart requires both fine-scale sampling and careful hypothesis testing.

The time and expenses required to collect and generate sequence data for many biological samples across many species typically limits us from achieving optimal sampling strategies. Here, rather than focus on a specific number of individuals, we stress the importance of attaining "phylogenetically relevant sampling," which we define to include several geographic samples from both within species and across species to allow for contrasts that can reveal the presence and geographic extent of admixture. In the American live oaks, we believe this was achieved with only four to five individuals per species. In a companion study to this one, Cavender-Bares et al. (2015) sequenced microsatellites for 672 individuals across the American live oaks, revealing geographically structured patterns of genetic variation within and between species which confirms that our RADseq sampling strategy has captured most relevant variation.

In some cases even the most optimal sampling strategies will fail to sample sufficient variation to accurately reconstruct historical introgression, as might occur when, for example, a relevant lineage has gone extinct. In this case, the origin of introgressed alleles can often still be approximately inferred by tracing their ancestry to the closest extant relatives. The more important question is then to determine the degree to which an extinct lineage was genetically and ecologically diverged from its closest relatives. The former may require fitting and comparing many complex demographic models (Pelletier and Carstens 2014), whereas the latter would generally be difficult without fossils. Understanding the relative frequency and evolutionary significance of introgression derived from ghost lineages will be necessary to use such historical introgression to study macroscale evolutionary processes such as parallel adaptation and the rate at which reproductive isolation evolves.

INFERRING ADMIXTURE

We explored a range of methods for detecting introgression and admixture, all of which returned complementary results. *Structure* and *TreeMix* share similarities in their underlying parametric models that infer admixture from the distribution of allele frequencies among populations (Pritchard et al. 2000); in the latter case, modeling changes along the branches of a phylogeny (or network) according to genetic drift (Pickrell and Pritchard 2012). The *TreeMix* approach is advantageous over *D*-statistics in that it takes into account the full sampled phylogeny when inferring admixture, as opposed to individual four- or five-taxon subsets of the tree. It thus identifies introgression in the context of all competing hypotheses, and takes into account the nonindependence of introgression events. However, when applied to deeply divergent lineages, as in our data, several assumptions of the model may be violated, such as equal population sizes, and that allelic variation arises from ancestral polymorphisms rather than *de novo* mutations (Pickrell and Pritchard 2012). When allowing more than two admixture edges in the live oaks, *TreeMix* inferred one or more instances of introgression between *Q. minima* and the outgroup "population" (tested as various combinations of the four non-*Virentes* white oak taxa), which we suspect is a false result: it is not supported by *D*-statistics using red oaks as a more distant outgroup (range $Z = [0.25–1.99]$). The simplified assumptions underlying nonparametric *D*-statistics may better facilitate their application for hypothesis testing over deeper evolutionary time scales, however, care must be taken in interpreting results within the context of unsampled phylogenetic relationships.

HYBRID SPECIES

We have focused on reconstructing phylogeny as a representation of the divergence of species through time, assuming that species have remained cohesive lineages despite instances of introgression between them. This view differs from the use of a graph or network to represent truly reticulate histories, or similarly, describing admixed lineages as having arisen through hybrid speciation (Schumer et al. 2014). For the latter case, we explicitly tested a model of instantaneous hybrid speciation for the origin of *Q. sagraeana*, the most admixed lineage in the American live oaks. This model was a poor fit compared to one in which an ancestral population of *Q. oleoides* colonized the island and received

persistent low levels of introgression from one or more species in the southeastern U.S. clade. A similar scenario in which an island population has undergone nuclear "conversion" toward the genomic makeup of another species has been described for ABC Island brown bears off the coast of Alaska (Cahill et al. 2013). Numerous examples of nuclear-chloroplast discordance in mainland oak species suggest this may be a common phenomenon (Petit et al. 2004), perhaps exacerbated by limited seed dispersal but widespread pollen flow in oaks.

INTROGRESSION AND PHYLOGENY

The effects of introgression on phylogenetic inference are often difficult to detect, but is made easier when multiple individuals are sampled from within a species that vary in their proportions of admixed ancestry. The rare and isolated taxon *Q. brandegeei*, from Baja California, provides an interesting example. Phylogenetic analyses suggested that it is nested within *Q. fusiformis*, appearing more closely related to populations from Mexico than from Texas. This finding, it turns out, is not a result of increased similarity between *Q. brandegeei* and *Q. fusiformis* (Mexico), but rather from the decreased relatedness between *Q. brandegeei* and *Q. fusiformis* (Texas); the latter arising from introgression that occurred into *Q. fusiformis* (Texas) from a more distant clade. This is clear from the phylogenetic results of censored datasets excluding the introgressive donor, which recovered strong support for monophyly of *Q. fusiformis* and its sister relationship to *Q. brandegeei* (Fig. S2E). Should we interpret this to mean that *Q. fusiformis* is not truly paraphyletic with respect to *Q. brandegeei*? The answer depends on what we wish our phylogeny to represent. If it is the historical pattern of population splitting, then *Q. brandegeei* clearly does not belong nested within *Q. fusiformis*. If the phylogeny is meant to show the genetic similarity of sampled individuals, then paraphyly of *Q. fusiformis*, which was recovered in most of our analyses, may be the most appropriate representation.

THE NATURE OF OAK SPECIES

The nature of species boundaries in oaks is a long-standing topic of philosophical debate. Burger (1975) and later Van Valen (1976) envisioned oaks as a form of "ecological species" in which populations filling a unique ecological niche remain recognizably distinct through shared adaptations regardless of their genomic makeup. Their classic example involves the widespread and easily recognizable bur oak (*Q. macrocarpa*), which hybridizes with up to seven other species across its range. Van Valen conjectured that it does not matter whether a bur oak population in Quebec is more likely to exchange genes with its local congener than with another bur oak population in Texas. He argued that if a recognizably distinct ecological unit persists across this range, it is sufficient to define the species. In the context of more recent views on ecological speciation (Nosil 2012), and the porous nature of species boundaries (Harrison and Larson 2014), the "ecological species" remains relevant, but with an elevated role for genetics—albeit sometimes very few genes (Wu 2001).

A similar case has been observed in *Heliconius* butterflies, where genomic regions that affect wing patterning involved in defensive mimicry are sometimes exchanged between closely related species (The Heliconius Genome Consortium 2012). In this case, because color patterns also influence assortative mating, introgression can facilitate reproductive isolation and even speciation, such that the spread of introgressed alleles occurs quickly through a small newly isolated population. This is a contrast from the live oaks, where large populations of wind-pollinated, long-lived individuals are likely to evolve reproductive isolation more slowly, with the consequence that introgressed alleles rarely spread across a species' entire geographic range, and thus recent admixture can generally be distinguished more easily from historical speciation events. By identifying which lineages have truly hybridized, we were able to identify how admixture has masked phylogenetic relationships in the live oaks, and in unmasking their effects, we resolved several phylogenetic conflicts, including the long-standing debate about the origin of the Cuban oak.

ACKNOWLEDGMENTS

We thank R. Ree, Y. Brandvain, M. Slatkin, and M. Donoghue for helpful discussions and advice, as well as D. Rabosky and three anonymous reviewers for comments that improved the manuscript. This work was partially supported by National Science Foundation grants IOS-0843665 (to JC-B), DEB-1146380 (to JC-B and AG-R), DEB-1146488 (to ALH), and a Lester Armour Graduate Student Fellowship at the Field Museum (to DARE).

DATA ARCHIVING

Raw sequence data are stored on the NCBI sequence read archive: SRP055977; assembled RADseq datasets (DOI 10.528/zenodo.19475) and all code used in our analyses (http://github.com/dereneaton/virentes) are also archived.

LITERATURE CITED

Avise, J. C. 2000. *Phylogeography: the history and formation of species*. Harvard Univ. Press, Cambridge, MA.

Baird, N. A., Etter P.D., Atwood T.S., Currey M.C., Shiver A.L., Lewis Z.A., Selker E.U., Cresko W.A., and Johnson E.A. 2008. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. PLoS ONE 3:7.

Barton, N. H., and G. M. Hewitt. 1985. Analysis of hybrid zones. Annu. Rev. Ecol. Syst. 16:113–148.

Boettiger, C., G. Coop, and P. Ralph. 2012. Is your phylogeny informative? Measuring the power of comparative methods. Evolution 66:2240–2251.

Borgardt, S. J., and K. B. Pigg. 1999. Anatomical and developmental study of petrified *Quercus* Fagaceae) fruits from the middle Miocene, Yakima Canyon, Washington, USA. Am. J. Bot. 86:307–325.

Burger, W. C. 1975. The species concept in *Quercus*. Taxon 24:45–50.

Cahill, J. A. et al. 2013. Genomic evidence for island population conversion resolves conflicting theories of polar bear evolution. PLoS Genet. 9:e1003345.

Cavender-Bares, J., and A. Pahlich. 2009. Molecular, morphological, and ecological niche differentiation of sympatric sister oak species, *Quercus virginiana* and *Q*. geminata (Fagaceae). Am. J. Bot. 96:1690–1702.

Cavender-Bares, J., A. Gonzalez-Rodriguez, A. Pahlich, K. Koehler, and N. Deacon. 2011. Phylogeography and climatic niche evolution in live oaks (*Quercus* series *Virentes*) from the tropics to the temperate zone. J. Biogeogr. 38:962–981.

Cavender-Bares, J., A. Gonzalez-Rodriguez, D. A. Eaton, A. A. L. Hipp, A. Beulke, and P. S. Manos. 2015. Phylogeny and biogeography of the American live oaks Quercus subsection Virentes): a genomic and population genetics approach. Mol. Ecol. 24:3668–3687.

Coyne, J. A., and H. A. Orr. 2004. Speciation. W.H. Freeman, Sunderland, MA.

Dumolin-Lapegue, S., A. Kremer, and R. J. Petit. 1999. Are chloroplast and mitochondrial DNA variation species independent in oaks? Evolution 53:1406–1413.

Durand, E. Y., N. Patterson, D. Reich, and M. Slatkin. 2011. Testing for ancient admixture between closely related populations. Mol. Biol. Evol. 28:2239–2252.

Earl, D. A., and B. M. vonHoldt. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv. Genet. Resour. 4:359–361.

Eaton, D. A. R. 2014. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics 30:1844–1849.

Eaton, D. A. R., and R. H. Ree. 2013. Inferring phylogeny and introgression using RADseq data: an example from flowering plants (*Pedicularis*: Orobanchaceae). Syst. Biol. 62:689–706.

Green, R. E., et al. 2010. A draft sequence of the Neandertal genome. Science 328:710–722.

Gugger, P. F., and J. Cavender-Bares. 2013. Molecular and morphological support for a Florida origin of the Cuban oak. J. Biogeogr. 40:632–645.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 5:e1000695.

Hardin, J. W. 1975. Hybridization and introgression in *Quercus alba*. J. Arnold Arbor. 56:336–363.

Harrison, R. G., and E. L. Larson. 2014. Hybridization, introgression, and the nature of species boundaries. J. Hered. 105:795–809.

Hipp, A. L., D. A. R. Eaton, J. Cavender-Bares, E. Fitzek, R. Nipper, and P. S. Manos. 2014. A framework phylogeny of the American oak clade based on sequenced RAD data. PLoS ONE 9:e93975.

Hudson, R. R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338.

Jakobsson, M., and N. A. Rosenberg. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23:1801–1806.

Kane, N. C., M. G. King, M. S. Barker, A. Raduski, S. Karrenberg, Y. Yatabe, S. J. Knapp, and L. H. Rieseberg. 2009. Comparative genomic and population genetic analyses indicate highly porous genomes and high levels of gene flow between divergent *Helianthus* species. Evolution 63:2061–2075.

Koehler, K., A. Center, and J. Cavender-Bares. 2012. Evidence for a freezing tolerance-growth rate trade-off in the live oaks (*Quercus* series *Virentes*) across the tropical-temperate divide. New Phytol. 193:730–744.

Kurz, H., and R. K. Godfrey. 1962. Trees of Northern Florida. University of Florida Press, Gainesville, FL.

Leaché, A. D., R. B. Harris, B. Rannala, and Z. Yang. 2014. The influence of gene flow on species tree estimation: a simulation study. Syst. Biol. 63:17–30.

Maddison, W. P., and L. L. Knowles. 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55:21–30.

Muller, C. H. 1961. The live oaks of the series *Virentes*. Am. Midl. Nat. 65:17–39.

Nadeau, N. J. et al. 2013. Genome-wide patterns of divergence and gene flow across a butterfly radiation. Mol. Ecol. 22:814–826.

Nixon, K., and C. Muller. 1997. *Quercus* Linnaeus sect. *Quercus* white oaks. Pp. 436–506 *in* Flora of North America Editorial Committee, ed.: Flora of North America North of Mexico. Oxford Univ. Press, New York.

Nixon, K. C. 1985. *A biosystematic study of Quercus series Virentes (the live oaks) with phylogenetic analyses of Fagales, Fagaceae and Quercus, Ph.D. Thesis*. University of Texas, Austin.

Nosil, P. 2012. Ecological speciation. Oxford Univ. Press, Oxford, U.K.; New York.

Pearse, I. S., and A. L. Hipp. 2009. Phylogenetic and trait similarity to a native species predict herbivory on non-native oaks. Proc. Natl. Acad. Sci. 106:18097–18102.

Pelletier, T. A., and B. C. Carstens. 2014. Model choice for phylogeographic inference using a large set of models. Mol. Ecol. 23:3028–3043.

Pérez, F., and B. E. Granger. 2007. IPython: a system for interactive scientific computing. Comput. Sci. Eng. 9:21–29.

Petit, R. J., and L. Excoffier. 2009. Gene flow and species delimitation. Trends Ecol. Evol. 24:386–393.

Petit, R. J., C. Bodénès, A. Ducousso, G. Roussel, and A. Kremer. 2004. Hybridization as a mechanism of invasion in oaks. New Phytol. 161:151–164.

Petit, R. J. et al. 1997. Chloroplast DNA footprints of postglacial recolonization by oaks. Proc. Natl. Acad. Sci. 94:9996–10001.

Pickrell, J. K., and J. K. Pritchard. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 8:e1002967.

Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Rabosky, D. L., and D. R. Matute. 2013. Macroevolutionary speciation rates are decoupled from the evolution of intrinsic reproductive isolation in Drosophila and birds. Proc. Natl. Acad. Sci. 110:15354–15359.

Ramírez-Valiente J.A., Koehler K., and Cavender-Bares J. 2015. Climatic origins predict variation in photoprotective leaf pigments in response to drought and low temperatures in live oaks (Quercus series Virentes). *Tree Physiology*, http://dx.doi.org/10.1093/treephys/tpv03.

Rhymer, J. M., and D. Simberloff. 1996. Extinction by hybridization and introgression. Annu. Rev. Ecol. Syst. 27:83–109.

Rogers, A. R., and R. J. Bohlender. 2015. Bias in estimators of archaic admixture. Theor. Popul. Biol. 100:63–78.

Schumer, M., G. G. Rosenthal, and P. Andolfatto. 2014. How common is homoploid hybrid speciation? Evolution 68:1553–1560.

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

The Heliconius Genome Consortium. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature 487:94–98.

Tuskan, G. A. et al. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. and Gray). Science 313:1596–1604.

Van Valen, L. 1976. Ecological species, multispecies, and oaks. Taxon 25:233–239.

Whittemore, A. T., and B. A. Schaal. 1991. Interspecific gene flow in sympatric oaks. Proc. Natl. Acad. Sci. USA 88:2540–2544.

Wu, C. I. 2001. The genic view of the process of speciation. J. Evol. Biol. 14:851–865.

Associate Editor: D. Rabosky
Handling Editor: M. Servedio

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Figure S1. The distribution of shared RADseq loci between samples across two datasets with different thresholds for the minimum sample coverage.
Figure S2. Rooted ML phylogenies inferred from 15 concatenated RADseq datasets.
Figure S3. Population clustering with admixture for 27 live oak individuals inferred from 14K SNPs.
Figure S4. Population splits and admixtures for pooled population samples inferred by TreeMix, and the corresponding allele frequency covariance matrix.
Table S1. Taxon sampling and summary of RADseq data assembly.
Table S2. Selected results of partitioned *D*-statistic tests investigating the origin of the Cuban oak.