Influence of late Quaternary climate change on present patterns of genetic variation in valley oak,
*Quercus lobata* Née

PAUL F. GUGGER,* MAKIHIKO IKEGAMI†‡ and VICTORIA L. SORK*‡

*Department of Ecology and Evolutionary Biology, University of California, Box 957239, Los Angeles, CA 90095, USA,
†Donald Bren School of Environmental Science and Management, University of California, Santa Barbara, CA 93106-5131, USA,
‡Institute of the Environment and Sustainability, University of California, Box 951496, Los Angeles, CA 90095, USA

Abstract

Phylogeography and ecological niche models (ENMs) suggest that late Quaternary glacial cycles have played a prominent role in shaping present population genetic structure and diversity, but have not applied quantitative methods to dissect the relative contribution of past and present climate vs. other forces. We integrate multilocus phylogeography, climate-based ENMs and multivariate statistical approaches to infer the effects of late Quaternary climate change on contemporary genetic variation of valley oak (*Quercus lobata* Née). ENMs indicated that valley oak maintained a stable distribution with local migration from the last interglacial period (~120 ka) to the Last Glacial Maximum (~21 ka, LGM) to the present compared with large-scale range shifts for an eastern North American white oak (*Quercus alba* L.). Coast Range and Sierra Nevada foothill populations diverged in the late Pleistocene before the LGM [104 ka (28–1622)] and have occupied somewhat distinct climate niches, according to ENMs and coalescent analyses of divergence time. In accordance with neutral expectations for stable populations, nuclear microsatellite diversity positively correlated with niche stability from the LGM to present. Most strikingly, nuclear and chloroplast microsatellite variation significantly correlated with LGM climate, even after controlling for associations with geographic location and present climate using partial redundancy analyses. Variance partitioning showed that LGM climate uniquely explains a similar proportion of genetic variance as present climate (16% vs. 11–18%), and together, past and present climate explains more than geography (19%). Climate can influence local expansion--contraction dynamics, flowering phenology and thus gene flow, and/or impose selective pressures. These results highlight the lingering effect of past climate on genetic variation in species with stable distributions.

Keywords: ecological niche modelling, genetic diversity, phylogeography, Pleistocene, *Quercus*, redundancy analysis

Received 10 May 2012; revision received 28 February 2013; accepted 6 March 2013

Introduction

Historical processes leave imprints on the genetic structure of present populations, especially of long-lived, sessile organisms. Thus, the present genetic structure of a variety of taxa has been used to infer historical vicariance associated with geological change (Avise 2000), dispersal history (Petit et al. 1997) and episodes of expansion and contraction associated with global climate change (Hewitt 2000). Recent research has integrated ecological niche models (ENM) based on the climate niche to independently infer historical demographic processes (Carstens & Richards 2007; Waltari et al. 2007; Knowles & Alvarado-Serrano 2010). Major insights include independent support for postglac-
climatic and genetic variables to the principal relationships and can be used to partition variation among multiple classes of explanatory variables, for example, by partialing out other purely neutral forces associated with geographic structure (e.g. phylogeographic structure and spatial autocorrelation). In concert with other approaches to studying the influence of past climate on genetic variation, such as the combination of ecological niche modelling with phylogeographic analysis, we can begin to more precisely understand how past and present climates and climate change shape genetic variation.

To investigate the extent to which current and past climate and late Quaternary climate change affect contemporary genetic variation in tree populations, we examined present genetic variation in relation to the past and present climate niche of a declining, endemic California oak species, valley oak (Quercus lobata Née). Valley oak is found scattered throughout lowland parts of the California Floristic Province, which is the most biodiverse region of North America (Myers et al. 2000). One hypothesis for the exceptional biodiversity of California suggests that topographic and local climate heterogeneity (topoclimate, Thornthwaite 1953; Dobrowski 2011) combined with regional climate stability generate opportunity for genetic diversity to accumulate within populations and for genetic divergence to arise among regions through drift and local adaptation (Stebbins & Major 1965; Calsbeek et al. 2003). Previous studies in valley oak find patterns consistent with this idea. High diversity and strong population differentiation at maternally inherited chloroplast microsatellites (cpSSRs) suggest historically persistent local seed dispersal (Grivet et al. 2006), rather than dramatic range expansions found for oaks in other temperate regions (Petit et al. 2002a; Svenning et al. 2008). Nuclear microsatellites (nSSRs) reveal genetic differentiation among populations in the inland foothills of the Coast Range and the western foothills of the Sierra Nevada, with higher gene flow north–south along climatically similar mountain corridors than east–west across the Central Valley (Grivet et al. 2008; Sork et al. 2010), suggesting a role for environmentally mediated historical vicariance. Those studies identified centres of genetic diversity in the San Francisco Bay area and Transverse Ranges of southern California, which coincide with some glacial refugia in other taxa (Gugger & Sugita 2010). Given the inferred stability of local populations and expectations for stable populations without major bottlenecks under genetic drift, we hypothesize that niche stability leads to higher levels of genetic diversity (Carnaval et al. 2009; Ortego et al. 2012). Moreover, the predicted stability of valley oak populations offers a unique opportunity test for and quantify the association of present genetic variation with past climate, beyond present climate (Sork et al. 2010), consistent with
a lingering influence of past climate adaptation and/or past climate-mediated demographic responses.

Using phylogeographic inference, climate-based ENMs and multivariate statistical approaches, we test the following specific hypotheses for how present patterns of genetic structure and diversity arose: (i) the distribution of valley oak was stable through late Quaternary climate cycles and migration was local; (ii) population genetic diversity is higher in sites with more stable climate niches; (iii) past climate-induced isolation explains regional genetic structure and (iv) past climate shaped present genetic variation.

Materials and methods

DNA preparation

We compiled previously published range-wide data for six cpSSR markers from 190 individuals from 64 sites and six nSSR loci from 267 individuals from 65 sites that were generated using standardized methods across studies (Grivet et al. 2006, 2008; Sork et al. 2010) (Dryad doi:10.5061/dryad.g645d/3). We added nuclear DNA sequences (nDNA) for two commonly sequenced low-copy genes: elongation factor 1-α (ef1α; 21 individuals from 11 sites, 736 bp) and glyceraldehyde 3-phosphate dehydrogenase (g3pdh; 29 individuals from 29 sites, 606 bp). DNA extraction and PCR conditions followed Werth & Sork (2008) using the following primers with 58 °C annealing temperature: efa-pF.deg = 5'-ACCCCAAARTACTCCAGGC-3' and efa-pR.deg = 5'-GACGCTTGGTACCTTG-3' for ef1α, and gpd-cgF = 5'-TGAATTGTGAGGGTCTCAT-3' and gpd-cgR = 5'-TGCTGTCAACAATGAGTCG-3' for g3pdh. Only the forward primer was used in sequencing reactions.

Sequences were aligned and edited using SEQUENCHER 4.8 (Gene Codes, Ann Arbor, MI). Haplotypes were resolved using the PHASE 2.1 algorithm (Stephens et al. 2001) implemented in DNASP 5.0 (Librado & Rozas 2009) allowing for recombination (Li & Stephens 2003). Haplotype networks were constructed in TCS 1.21 (Clement et al. 2000). We estimated standard diversity statistics: number of haplotypes, Nh; number segregating sites, S; gene diversity, H'; nucleotide diversity, π. We tested for selective neutrality (and constant population size) using D (Tajima 1989) and for recombination with a four-gamete test (Hudson & Kaplan 1985) in DnaSP to ensure these new molecular data met the assumptions of the IMA2 model (below).

Population structure

To describe regional and local population genetic structure for subsequent analyses, we used Bayesian clustering methods. For nSSRs, we estimated population structure with admixture and correlated allele frequencies using a Bayesian Markov Chain Monte Carlo (MCMC) clustering analysis in STRUCTURE 2.3.4 (Pritchard et al. 2000; Falush et al. 2003). We ran 10 replicates of 200 000 steps after 100 000 burn-in steps for 1–10 population clusters (K). We used the AK method (Evanno et al. 2005) to select the best-supported number of clusters, as implemented in STRUCTURE HARVESTER (Earl & vonHoldt 2012). For cpSSRs, we estimated population structure using the ‘clustering with linked loci’ analysis in BAPS 5.2 (Corander et al. 2008). We ran the model three times for K = 1–40 to ensure convergence and chose the K with the highest log likelihood. We also generated a median-joining haplotype network coloured by geographic region using NETWORK 4.6 (Bandelt et al. 1999). Results were visualized using ARCGIS 9.3 (ESRI, Redlands, CA, USA).

To quantify variation in nSSRs and cpSSRs among populations and genetic clusters, we performed analyses of molecular variance (AMOVA) (Excoffier et al. 1992) using Φ-statistics in GENALEX 6.1 (Peacock & Smouse 2006).

Divergence time, gene flow and population size changes

To set regional patterns of population structure in an historical context, divergence time among coastal and Sierran population clusters (see Results), effective sizes of ancestral and extant populations and bidirectional gene flow rates (as population migration rate, 2Nem) were estimated using an isolation-with-migration model in IMA2 (Hey 2010). We compared three replicate runs (ESS > 50) of 40 geometrically heated chains (la = 0.96, hb = 0.9) with a 10 000 000-step burn-in followed by >20 000 000 steps sampled every 100 steps to ensure convergence. We chose the maximum values for priors drawn from a truncated uniform distribution on the basis of short preliminary runs (Won & Hey 2005): t = 5, q1 = q2 = qA = 80, and m1 = m2 = 9. We assumed an infinite sites model (Kimura 1969) of evolution for DNA sequences and a stepwise mutation model for nSSRs (Ohita & Kimura 1973). We excluded cpSSRs from the analysis because long runs (>3 months) with 200 heated chains did not reach stationarity when cpSSRs were included.

To convert parameter estimates to demographic units, we considered a range of mutation rates and generation times. For nDNA, we applied rates calculated for other nuclear loci in Fagaceae of 1.6 × 10^{-10} to 2.3 × 10^{-9} per bp per year (Cavender-Bares et al. 2011). For nSSRs, we considered rates ranging from 1.76 × 10^{-4} per gen in tomatoes (Solanum lycopersicum; Azaiez et al. 2006) to
8.8 × 10⁻⁴ per gen in Arabidopsis thaliana (Marriage et al. 2009). To convert these estimates to the per-locus, per-year rates required in IMa2, per-generation nSSR mutation rates were multiplied by a mean generation time in oaks of 100 years (Cavender-Bares et al. 2011; Gugger & Cavender-Bares 2013), and per-base pair nDNA rates were multiplied by the length of the locus in base pairs.

Ecological niche modelling

To test the hypothesis that the distribution of valley oak was stable through late Quaternary climate cycles, we compared ENMs for valley oak to ENMs of an eastern North American white oak (Quercus alba L.) expected to have experienced dramatic range shifts in response to glacial–interglacial cycles (a ‘positive control’) (Davis 1981). Specifically, we used MAXENT 3.3.3e (Phillips et al. 2004, 2006) to estimate their distributions during the present, Last Glacial Maximum (LGM, ~21 ka) and last interglacial (LIG, ~120–140 ka). MAXENT is a machine learning-based ENM that uses presence points and performs well with few occurrence points (Elith et al. 2006). The ‘climatic distribution’ of a species represents the modelled statistical association between its known occurrences and current climatic variables. This model was transformed from climate space into geographic space to generate maps showing probability of occurrence. Past distributions were then estimated by projecting this relationship onto scenarios of past climate, assuming that current relationships between climate and distribution are retained. This assumption may be a reasonable approximation for most purposes, but we also built models separately for coastal and Sierran subpopulations to control for observed divergence within valley oak (more below). Final models were constructed using 10 replicate runs. We evaluated model performance by running 10 replicate models divided into 70%/30% train/test partitions of occurrence points and then estimating the area under the receiver operating curve (AUC), where 1 is the maximum prediction and 0.5 suggests a random prediction.

We obtained 247 Quercus lobata presence points from Sork Lab sample sites, others sample sites (Sage et al. 2011) and personal observations representative of the entire presettlement distribution (Griffin & Critchfield 1972; Davis et al. 1998; Whipple et al. 2011) (Dryad doi:10.5061/dryad.g645d/3). We obtained 1947 occurrence points of Quercus alba from the Global Biodiversity Information Facility (GBIF) portal (http://data.gbif.org) and retained 1920 points after removing those outside the known current distribution (http://gec.cr.usgs.gov/data/atlas/little/) (Little 1971). The distribution of valley oak is thought to be primarily limited by growing-season precipitation or water availability, seasonality and temperature extremes (Sork et al. 2010; McLaughlin & Zavaleta 2012). Thus, the following eight climate variables were extracted from the WorldClim database (http://www.worldclim.org, Hijmans et al. 2005) at 2.5-arcmin resolution for present (1950–2000), LGM [the average climate of two commonly used LGM models: MIROC 3.2 (Hasumi & Emori 2004) and CCSM3 (Collins et al. 2006; Braconnot et al. 2007)] and LIG (Otto-Bliesner et al. 2006) climates: mean annual precipitation, precipitation in the warmest quarter, precipitation in coldest quarter, precipitation seasonality, mean annual temperature, maximum temperature of warmest month and minimum temperature of coldest month, temperature seasonality.

To further test for patterns consistent with local migration in valley oak, we estimated migration distance and direction of valley oak and white oak populations. From predicted climate suitability of ENMs, presence/absence maps were created from MAXENT’s logistic probability of occurrence output using ‘maximum sensitivity plus specificity’ threshold (Jiménez-Valverde & Lobo 2007). We calculated the geographic centroids of the distributions for every 60 × 60 km² (i.e. 15 × 15 grid cells in 2.5 arcmin) in each time step. Then, to evaluate the potential source of populations from one time period to the next, we calculated the distance and direction between the centroid of the first time step to the nearest centroid in the second time step (i.e. from LIG to LGM and from LGM to present). The resulting maps display vectors expressing inferred migration distance and direction for each 60 × 60 km² area.

Finally, we built additional ENMs for each nSSR genetic cluster (coastal and Sierran; see Population structure in Results) to determine whether the climate niches were more distinct during the LIG or the LGM than the present. We calculated niche overlap among clusters for each time period by first determining presence/absence with a logistic threshold of equal sensitivity and specificity from test data in MAXENT and overlaying the niche distributions based on each nSSR cluster to calculate the proportion of niche overlap at each time step. A clear separation of projected past distributions among each cluster throughout the past would suggest that past climate-induced isolation explains genetic structure.

Genetic diversity and niche stability

To test the hypothesis that niche stability explains genetic diversity, we regressed LIG, LGM and present niche suitabilities (i.e. probabilities of occurrence) (NLIG, NFLGM, NPred) and climatic niche stability from ENMs on nSSR genetic diversity in general linear models using R 2.14.1 (http://www.r-project.org/). Niche suitability
was included as a covariate because it could be considered a proxy for population size if more suitable habitat maintains larger populations. We defined climatic niche stability as two variables: \( S_{\text{LGM-Pres}} = 1 - |N_{\text{LGM}} - N_{\text{Pres}}| \)
and \( S_{\text{LGM-LGM}} = 1 - |N_{\text{LGM}} - N_{\text{LGM}}| \). Because the independent variables are related to one another, we present a correlation matrix, which shows low to moderate correlations (0.07 < \( r \) < 0.61, 0 < \( r < 0.54 \)) among predictors (Table S1, Supporting information), and we measured variance inflation factors (VIF), a measure of how much collinearity increases variance in a model. We observed VIF < 2; VIF < 10 is evidence that multicollinearity does not confound the interpretation of individual predictors in our models. For measures of genetic diversity, we used expected heterozygosity (\( H_e \)) and allelic richness (\( A_r \)) and after standardizing with rarefaction in HP-Rare 1.0 (Kalinowski 2005). We tested three nested models and chose the one with the lowest Akaike information criterion (AIC) score. The first model contained all five explanatory variables relating to the LIG, LGM and present; the second contained only variables relating to the LGM and present (\( N_{\text{LGM}} \), \( N_{\text{Pres}} \), \( S_{\text{LGM-Pres}} \)), and the third contained only \( N_{\text{Pres}} \). By comparing these three models, we evaluated how far back in time we could detect the influence of niche suitability and stability on present genetic diversity.

**Genetic variation and past climate**

To test the hypothesis that past climate shape patterns of present genetic variation, beyond the role of present climate and geography (Sork et al. 2010), we performed a series of partial redundancy analyses using the VEGAN 2.0-2 package (Oksanen et al. 2009) in R. Redundancy analysis is a multivariate analogue of linear regression similar to canonical correspondence analysis (ter Braak 1986; Legendre & Legendre 1998). We first tested our hypothesis by testing the association of LGM climate (CCSM, Braconnot et al. 2007) with cpSSR and nSSR variation, in each case partialing out the variation due to present climate and geographic position (i.e. latitude, longitude, squared terms, cubed terms and cross products). Significance was assessed with a permutation test. Following established procedures (Borcard et al. 1992; Liu 1997; Økland 1999), we then partitioned the explainable genetic variance (i.e. inertia, or in this case mean squared contingency coefficients) into percentages of variance explained by LGM climate, present climate, geography and their joint effects. For these analyses, genetic data were transformed to allele frequencies for each individual, removing one allele from each locus (Smouse & Williams 1982; Westfall & Conklin 1992), and climate data for the same eight variables used in ENMs were extracted from the WorldClim database for each sample.

**Results**

**Nuclear DNA sequence diversity**

All g3pdh and ef1x alleles were phased with probability \( P > 0.9 \), although a few specific bases at each locus were assigned with probability 0.5 < \( P < 0.9 \) (unphased data in GenBank: JQ421340–JQ421389; aligned, phased data in Dryad: doi:10.5061/dryad.g645d/1 and doi:10.5061/dryad.g645d/2). Both nDNA loci showed moderate amounts of variation (\( g3pdh: \ n_h = 10, S = 16, H = 0.75, \pi = 0.0067; \ ef1x: \ n_h = 7, S = 7; H = 0.66; \pi = 0.0013 \)) (Fig. S1, Supporting information) (Derory et al. 2010; Gugger et al. 2010). Tajima’s \( D \) was not significantly different from zero for \( ef1x (D = −1.09, P > 0.1) \) nor for \( g3pdh (D = 0.50, P > 0.1) \), and neither showed evidence of intralocus recombination.

**Population structure**

\( \Delta K \) based on STRUCTURE analysis of nSSRs supported two geographically meaningful population clusters (\( ln(PK) = −6066.6, \Delta K = 269.9 \)) (Table S2, Supporting information), a western one mostly contained to the inland foothills of the Coast Range and an eastern one mostly on the western Sierra Nevada foothills (hereforth referred to as ‘coastal’ and ‘Sierran’ clusters) (Fig. 1a). The northernmost part of the distribution shows evidence of interdigitation among population clusters, but not much admixture; most sample sites retain their signature of coastal or Sierran origin despite the potential for long-distance pollen dispersal in highly outcrossing wind-dispersed trees. East of the San Francisco Bay, the coastal population extends eastward into the Sierra Nevada foothills. Seventeen percent of nSSRs showed strong local structure (Griivet et al. 2006) with some intermixing in the northern part of the distribution parallel to that of nSSRs (Fig. 1b). Although the high \( K \) raises suspicions of model overfitting, this result is consistent with the fact that most haplotypes are geographically restricted to a single sample sites or a few neighbouring ones. More important than its specific value, \( K \) is probably large, and structure is local. Among the 25 BAPS clusters \( \Phi_{PT} = 0.92 \), and among the 64 sample sites \( \Phi_{PT} = 0.88 (P = 0.01) \) (Table S3, Supporting information). Hierarchical AMOVA was not performed on cpSSR data because the 64 sample sites were not strictly nested within the 25 BAPS clusters.
Divergence time, gene flow and population size changes

Assuming the mean of published mutation rates for our loci, divergence time (t) among coastal and Sierran population clusters was 104 ka (28–1622) and gene flow was more pronounced from the west to the east [$N_\text{m-CS} = 6.96 (3.3–10.5)$] than east to west [$N_\text{m-SC} = 0.02 (0–2.9)$] (Table 1). The present effective sizes of the coastal and Sierran population clusters were estimated as 18% and 19% of the ancestral effective population size, respectively. Considering only high-end or only low-end mutation rates led to considerable variation in divergence time estimates (Table 1; Fig. S2, Supporting information).

Ecological niche models

The valley oak ENM performed well (AUC = 0.975), whereas the white oak model performed moderately (AUC = 0.876), which can occur when large numbers of presence points are used (Franklin et al. 2009). The ENMs for all valley oak sites show a stable California distribution relative to white oak, which drastically contracted during the LGM and expanded in the Holocene (Fig. 2). Valley oak ENMs show modest local to regional changes in distribution. For example, during current and LIG climate conditions coastal and Sierran populations appear more isolated from each other, whereas they appear more interconnected during the LGM. Nonetheless, some connectivity is apparent in northern California through all time periods (Figs 2 and 3). The Transverse Range populations of southern California appear more connected to coastal rather than Sierran populations.

Migration vector analysis suggested the local movement of valley oak between periods of major climate transition, especially when compared with the dramatic northward expansion of white oak from the LGM to present (Fig. 3). Along the eastern range margin of valley oak, there is some evidence for elevational shift; from the LIG to LGM, vectors point from the Sierras towards the Central Valley and from the LGM to present, the process reverses.

In the southernmost and northernmost parts of valley oak’s distribution, the migration vector analysis implies moderate latitudinal shifts, although those shifts are not nearly as large as those observed in white oak (Fig. 2). This is likely an artefact of the general tendency of the ENMs to overestimate the distribution of valley oak (e.g. compare Fig. 2a–d), suggesting nonclimatic environmental (McLaughlin & Zavaleta 2012), biotic or anthropogenic variables factors not included in our models are important in defining the realized niche of...
valley oak. Nonetheless, the climate variables we included have been shown to be important to valley oak (Sork et al. 2010; McLaughlin & Zavaleta 2012), especially precipitation, which is consistent with the fact that precipitation of the warmest quarter (i.e. during the growing season) had the highest percent contribution and permutation importance among climate variables used to construct the ENMs (Table S4, Supporting information).

Models based on each nSSR genetic cluster show periodic, partial climate niche separation (Fig. 4). For example, the distribution of the coastal genetic cluster was not predicted in the Central Valley nor Sierras during the LIG (Fig. 4b), whereas the Sierran genetic cluster was not predicted on the coast during the LGM (Fig. 4g). Climate niches were most distinct during the LIG (niche overlap = 19.1%) and similarly distinct during the LGM (46.7%) and present (51.1%). The principal area of overlap during the LGM was in the northern Central Valley, similar to today.

Genetic diversity and niche stability

Nuclear SSR heterozygosity ranged from 0.43 to 0.74 and allelic richness averaged across loci after rarefaction to 6 ranged from 2.3 to 4.1 (Table S5, Supporting information). The model that best explained $H_e$ and $A_r$ contained $N_{LGM}$, $N_{Pres}$ and $S_{LGM-Pres}$ but not terms relating to the LIG (Table 2). Both genetic diversity measures were significantly higher as recent niche stability ($S_{LGM-Pres}$) increased ($P < 0.037$). This result was only true when including $N_{LGM}$ and $N_{Pres}$ as covariates (Table 2);
in contrast, simple linear regression of $S_{\text{LGM-Pres}}$ on genetic diversity indicates an insignificant positive correlation ($0.12 < r < 0.18$, $P > 0.15$). This apparent discrepancy can occur when other predictors included in the model are important and thus reduce the residual error variance. Another factor may be that we have low power to detect the pattern in isolation, given the noisy estimates of genetic diversity expected when samples sizes are 3–4 individuals (6–8 alleles) per site. The significant relationship in the general linear model likely reflects a real relationship because we did not detect problematic multicollinearity in tests of variance inflation factors ($\text{VIF} < 2$). The covariates, $N_{\text{LGM}}$ and $N_{\text{Pres}}$, had negative and positive relationships, respectively, with both measures of genetic diversity. Among those, only the relationship of $N_{\text{Pres}}$ with $A_t$ was significant ($P = 0.04$) and all others were marginally significant ($0.05 < P < 0.08$).

**Genetic variation and past climate**

After controlling for associations with present climate and geographic position, partial redundancy analysis suggested that present cpSSR (pseudo-$F = 4.82$, $P < 0.005$) and nSSR variation (pseudo-$F = 1.63$, $P < 0.005$) retain some signature of LGM climate. Minimum temperature during the LGM had the highest loading on the first redundancy analysis axis for both cpSSRs and nSSRs (Fig. S3, Supporting information). For nSSRs, LGM climate explained 16.5% of the present genetic variance, present climate explained 17.8%, geography explained 18.9% and their collinearities explained the rest (Fig. 5). For cpSSRs, LGM climate explained 15.7% of the present genetic variance, present climate explained 11.5%, geography explained 19.4% and their collinearities explained the rest (Fig. 5). Results were similar when using the MIROC instead of CCSM LGM climate data and when summarizing genetic data as allele frequencies per site rather than per individual (although the permutation test for nSSRs analysed by site was not significant, $P = 0.14$) (not shown).

**Discussion**

We found evidence that late Quaternary climate has shaped present genetic variation in valley oak in several ways consistent with the effects of drift under relatively
stable climate, the influence of climate on demographic processes, and perhaps the influence of selection by climate. Specifically, ENMs and genetic data support a long history of relative stability in the distribution of valley oak with local, elevational expansion and contraction in response to glacial cycles (Figs 1, 2 and 3). At the site scale, niche stability from the LGM to present could contribute to higher levels of nuclear microsatellite diversity (Table 2). Within the broadly constant distribution, we also found evidence of a late Pleistocene (Table 1) subdivision of coastal and Sierran population clusters (Fig. 4) with subsequent intermixing in regions of more persistent niche overlap (Fig. 3a). Most strikingly, LGM climate explains a substantial portion of nuclear and chloroplast genetic variation (Fig. 5 and Fig. S2, Supporting information), suggesting the residual influence of LGM climate-mediated demographic responses or possibly of LGM climate adaptation on present genetic variation.

**Stable distribution, local migration**

The most prominent feature of the ENMs is that the conditions favourable to valley oak remain relatively constant through major global climate cycles from the LIG to LGM to present (Fig. 2a–d). This pattern is especially clear when compared with models for the eastern North American white oak (Fig. 2e–h) and European white oaks (Petit et al. 2002a; Svenning et al. 2008) that experienced major latitudinal contractions and expansions. General relative climate stability in California (Barry 1983; Bartlein et al. 1998) and local topographic heterogeneity throughout the state probably promoted the regional stability of valley oak's climate niche apparent in ENMs (Tzedakis et al. 2002). Migration vector analyses (Fig. 3) suggest some local migration along elevational gradients at the eastern range margin. For example, from the LIG to LGM, local migration proceeds from the Sierras down slope towards the Central Valley and from the LGM to present the process reverses. These shifts may have been associated with moderate regional population size changes according to coalescent analyses of molecular data (Table 1). Elevational migration and long-term stability have been proposed for California oaks (Byrne et al. 1991) and other taxa (Gugger & Sugita 2010) based on fossil records. These processes are thought to promote high regional genetic diversity and local population structure through drift and local adaptation (Hewitt 1996; Hampe & Petit 2005; Gugger et al. 2011) and help explain the higher
genetic diversity of valley oak (Fig. 1) (Grivet et al. 2006) compared with European oaks (Petit et al. 2002b) and eastern North American trees (Magni et al. 2005; McLachlan et al. 2005) that have undergone large-scale range shifts.

Given the long-term stability of valley oak’s predicted distribution, we further hypothesized that niche stability would promote local genetic diversity. Our general linear models including niche stability and niche suitability support the idea that niche stability from the LGM to present predicts present local genetic diversity (Table 2). The role of LIG to LGM changes remains unclear as these terms were not included in the best model and not significant in the full model. In the best model, present niche suitability is also significantly or marginally significantly positively associated with genetic diversity, which makes sense if sites with more favourable habitat maintain larger population sizes. These results are consistent with the expectation that genetic diversity would be lost more slowly in stable populations.

Late Pleistocene divergence
ENMs suggest that the coastal and Sierran valley oak populations may have maintained partial isolation since at least the LIG. Both niche models and genetic data suggest isolation among coastal and Sierran populations in the southern part of valley oak’s distribution since the LIG. Conversely, the northern part of the distribution

---

**Table 2** General linear models including variables derived from ecological niche modelling that could explain population genetic diversity ($A_r$ and $H_E$). Akaike information criterion (AIC) scores suggest the best model is the LGM/Pres model.

<table>
<thead>
<tr>
<th></th>
<th>$A_r$</th>
<th></th>
<th>$H_E$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>SE*</td>
<td>$t$</td>
</tr>
<tr>
<td>(Intercept)</td>
<td>1.89</td>
<td>0.57</td>
<td>3.30</td>
</tr>
<tr>
<td>$N_{LIG}$</td>
<td>−0.27</td>
<td>0.49</td>
<td>−0.54</td>
</tr>
<tr>
<td>$N_{LGM}$</td>
<td>−0.66</td>
<td>0.43</td>
<td>−1.53</td>
</tr>
<tr>
<td>$N_{Pres}$</td>
<td>1.16</td>
<td>0.58</td>
<td>1.99</td>
</tr>
<tr>
<td>$S_{LIG-LGM}$</td>
<td>−0.03</td>
<td>0.58</td>
<td>−0.06</td>
</tr>
<tr>
<td>$S_{LGM-Pres}$</td>
<td>1.58</td>
<td>0.71</td>
<td>2.22</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE*</th>
<th>$t$</th>
<th>$P^+$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.87</td>
<td>0.53</td>
<td>3.52</td>
<td>0.001</td>
</tr>
<tr>
<td>$N_{LGM}$</td>
<td>−0.75</td>
<td>0.39</td>
<td>−1.94</td>
<td>0.057</td>
</tr>
<tr>
<td>$N_{Pres}$</td>
<td>0.97</td>
<td>0.47</td>
<td>2.09</td>
<td>0.041</td>
</tr>
<tr>
<td>$S_{LGM-Pres}$</td>
<td>1.58</td>
<td>0.61</td>
<td>2.60</td>
<td>0.012</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE*</th>
<th>$t$</th>
<th>$P^+$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>3.12</td>
<td>0.21</td>
<td>14.53</td>
<td>0.000</td>
</tr>
<tr>
<td>$N_{Pres}$</td>
<td>0.41</td>
<td>0.42</td>
<td>0.98</td>
<td>0.334</td>
</tr>
</tbody>
</table>

LGM, Last Glacial Maximum.
*Standard error.
†Significant $P$-values in bold ($P < 0.05$) and marginally significant ($0.05 < P < 0.10$) italicized.

---

**Fig. 5** Venn diagrams showing percentage of genetic ‘variance’ (i.e. inertia or mean squared contingency coefficient) explained by present climate, climate during the Last Glacial Maximum, geography and their joint effects for (a) nuclear microsatellites (nSSR) and (b) chloroplast microsatellites (cpSSR).

© 2013 John Wiley & Sons Ltd
has been an area of periodic niche overlap among populations (Figs 2 and 3a), consistent with the interdigitation of nSSR and cpSSR genetic clusters in the region (Fig. 1). The inland penetration of coastal climate and the presence of riparian corridors may facilitate dispersal across Central Valley east of San Francisco Bay, especially eastward, which is supported by the higher migration rate ($2N_{m}$) estimate from coastal to Sierran populations (Table 1). This phenomenon has been termed the ‘trans-valley leak’ (Peabody & Savage 1958) and has been observed in salamanders (Stebbins 1949; Martínez-Solano et al. 2007; Kuchta et al. 2009), spiders (Satler et al. 2011), lichens (Werth S, and Sork VL, unpublished data) and possibly annual plants (Baldwin et al. 2011).

The formation of the east–west split in valley oak may have coincided with the LIG when niche overlap among nSSR clusters was lowest (Fig. 4), but could have occurred at other times in the Middle to Late Pleistocene before the LGM (Table 1, Fig. S2, Supporting information). Although the divergence time cannot be pinpointed, we can exclude recent LGM or post-LGM divergence as contexts for the major subdivision within valley oak. Increased isolation among coastal and Sierran populations during warm interglacial periods (Fig. 4), causing migration of valley oak woodlands to somewhat higher elevations and perhaps out of parts of the Central Valley (Fig. 3) and reduced effective population sizes since the LGM (Table 1), could have facilitated divergence. In contrast, fossil pollen evidence suggests that oaks, although not necessarily valley oak specifically, were more abundant during interglacial periods across a variety of habitats: in the Coast Range of northern California (Clear Lake, Adam et al. 1981), the Central Valley (Tulare Lake, Davis 1999) and southern California (Santa Barbara Basin, Heusser 1995).

Regardless of the uncertainty in associating the divergence with a particular climate regime or transition, ENMs provide preliminary evidence of niche specialization among coastal and Sierran populations despite the interdigitation of genetic clusters. For example, there is some niche differentiation among genetic groups during interglacial periods, especially during the LIG when the coastal populations appear highly restricted to the coast and the Sierran populations appear restricted to inland areas (Fig. 4). However, niche models indicate substantial areas of overlap in the northern Central Valley during the LGM and present. In addition to geographic separation, niche differences may reinforce gene flow within each genetic cluster, consistent with earlier findings of gene flow along mountain corridors and correlations with climate (Grivet et al. 2008; Sork et al. 2010) (Figs 1 and 5).

Many phylogeographic studies of Californian taxa show the split observed among coastal and Sierran populations in nSSR data and ENMs (Lapointe & Rissler 2005). Transverse Range populations of other species tend to be more closely related to coastal populations (Calsbeek et al. 2003), but in some species, Transverse Range populations are more similar to Sierran populations (e.g. Macy et al. 2001). In valley oak, ENMs suggest that Transverse Range populations share closer ties to coastal populations, whereas Bayesian clustering analysis of nSSRs (Fig. 1a) indicates a closer relationship to Sierran populations. This relationship could reflect ongoing gene flow by pollen dispersal or insufficient time for lineage sorting at nuclear loci. The unique cpSSR composition of the Transverse Range suggests isolation, at least through seed-mediated gene dispersal (Fig. 1b).

Correlations with LGM climate

The correlation of LGM climate with present genetic variation after controlling for correlations with geography and present climate suggests that the effects of past climate on genetic variation can persist for many generations. Moreover, this remaining relationship from the LGM is about as strong as that with present climate or geography for both molecular marker types (Fig. 5). Some of this relationship could be explained by current environmental variables not included in the model that happen to correlate with past climate variables, but it seems unlikely that this explains the entire effect. Moreover, minimum LGM temperature had the highest loading on the first redundancy analysis axis for both cpSSRs and nSSRs (Fig. S3, Supporting information), which could result from post-LGM colonization of habitat that was uninhabitable during the LGM due to low minimum temperatures (i.e. climate-mediated demographic shifts) or selection during the cooler LGM. We hypothesize that genetic variation in tree populations is not in equilibrium with current climate, or that the re-sortment of genetic variation in response to changing climate lags, due to long generation time and thus fewer opportunities per year for recombination, drift and/or selection. Consistent with this hypothesis, present climate explains proportionately more variation in nSSRs compared with cpSSRs, the latter of which represent a single nonrecombining locus. Furthermore, under rapidly changing environments, theory predicts a lag time for adaptive genetic variation (Pease et al. 1989; Bürger & Lynch 1995) that approximately corresponds to temporal scale in generations observed here.

Purely demographic forces could cause the correlation of past and present climate with allele frequency variation if climate acts as a demographic filter, permitting growth under some conditions and contraction under others, or if climate affects gene flow by affecting flowering phenology. Demographic forces have been proposed for similar correlations observed in pine bee-
tiles near an expanding range margin (James et al. 2011). Because we factored out large-scale geographic patterns and their collinearity with climate (Fig. 5) and because no simple pattern of expansion and contraction occurred in valley oak (Figs 2 and 3), such demographic processes would have to be invoked on a local level or along climate gradients without linear, quadratic or cubic associations to latitude or longitude.

It is also possible that both demographic and selective forces account for associations with climate, as both are expected under changing conditions. The ENMs for each nSSR cluster suggest the long-term persistence of somewhat distinct niches for coastal and Sierran clusters and thus the potential for locally adapted populations (Fig. 4). This history could generate linkage disequilibrium between the neutral nSSR loci and genetic variation underlying local adaptation. Selection against nonlocal genotypes could continually reinforce climate-genotype associations in neutral markers, especially if flowering phenology differences or other potential barriers lead to effective, nascent reproductive isolation among the locally adapted clusters (Via et al. 2000; Nosil et al. 2008; Sork et al. 2010; Wang & Summers 2010; Andrew et al. 2012; Ortego et al. 2012). Although we cannot disentangle the effects of selection from climate-mediated demographic responses using these methods (Geffen et al. 2004), such correlations create conditions that would facilitate selection by climate (Endler 1986).

Conclusions

In contrast to the numerous studies of tree populations that experienced major bottlenecks due to glaciation, this investigation explores the impact of climate on the geographic genetic structure in a species that maintained a more stable distribution through recent glacial–interglacial cycles. We found evidence that present genetic variation is shaped by this history of a long-term stable distribution with local migration, late Pleistocene vicariance, climate-mediated demographic responses, and possibly past and present selection by climate. Together, these results support the hypothesis that long-term stability and variable topoclimate explain high biodiversity in the California Floristic Province. By investigating climate associations with present genetic variation at multiple timescales, we provide preliminary evidence for and quantify the endurance of past climate changes on genetic variation and establish an historical context for future genomic studies explicitly examining geographic patterns of adaptive molecular variation. To our knowledge, our study is the first to demonstrate that LGM climate is associated with present genetic variation, raising the question of whether this imprint will shape future responses of tree populations to climate change.

Acknowledgements

We thank K. Lundy and S. Werth for laboratory assistance; E.C. Riordan for helping extract data from WorldClim databases; J. Ortego for helpful conversations; and K. Gaddis, J. Ortego, A. Platt, E.C. Riordan, P. Thompson and four anonymous reviewers for constructive comments on this manuscript. Funding was provided by NSF-DEB-0516529 to V.L. Sork.

References


Ohta T, Kimura M (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetical Research*, 22, 201–204.


© 2013 John Wiley & Sons Ltd


**P.F.G. is a Postdoctoral Scholar interested in combining fossil evidence, molecular data, and models to understand the ecological and evolutionary history of trees. M.I. is a plant ecologist working in the fields of forestry, biogeography, and conservation biology focusing on evaluating species distributions and performances under different environments through modeling and field work. V.S. integrates population genomic, landscape genetic, and molecular ecology approaches to understand how gene flow and natural selection shape the evolutionary potential of natural tree populations.**

### Data accessibility

DNA sequences: GenBank accessions JQ421340–JQ421389 for unphased sequences, and Dryad doi:10.5061/dryad.g645d for aligned, phased sequences.

Sample locations, microsatellite data and occurrence points for ecological niche models: Dryad doi:10.5061/dryad.g645d.

### Supporting information

Additional supporting information may be found in the online version of this article.

**Table S1** Correlation among independent variables used in general linear models to test the role of past niche suitability and stability on present genetic diversity.

**Table S2** Mean log likelihood (lnP(k)), its standard deviation (SD), and ΔK calculated from 10 replicates for each of 1–10 clusters (K) using a Bayesian clustering analysis implemented in STRUCTURE.

**Table S3** Analyses of molecular variance based on ϕ-statistics for nuclear microsatellite (nSSR) sample sites within STRUCTURE clusters, chloroplast microsatellite (cpSSR) sites, and cpSSR BAPS clusters.

**Table S4** Percent contribution and permutation importance of each variable used to construct valley oak (Quercus lobata) ecological niche models in MAXENT.

**Table S5** Nuclear microsatellite (nSSR) allelic richness and heterozygosity, niche suitability, and niche stabilities by site for samples used in genetic analyses (additional sites were used to construct niche models).

**Fig. S1** Maps of genetic variation in valley oak showing haplotype distributions for (a) glyceraldehyde 3-phosphate dehydrogenase (g3phd) and (b) elongation factor 1-α (ef1α).

**Fig. S2** Posterior distributions for per-locus mutation rate-scaled parameter estimates from the IMa2 run with the highest effective sample size.

**Fig. S3** Biplots of partial redundancy analyses of (a) nuclear microsatellite and (b) chloroplast microsatellite variation on Last Glacial Maximum (~21 ka) climate after controlling for present climate and geographic position.