# Mutational load and adaptive variation are shaped by climate and species range dynamics in *Vitis arizonica*

**Christopher J. Fiscus[1], Jonás A. Aguirre-Liguori[2,3], Garren R. J. Gaut[4] and Brandon S. Gaut[1]**

[1]Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92697, USA; [2]Departamento de Ecología Tropical, Campus de Ciencias Biológicas y Agropecuarias, Universidad Autónoma de Yucatán, Mérida, 97000, Mexico; [3]Laboratorio Nacional de Biología del Cambio Climático, Mexico City, 04530, Mexico; [4]Material+, Los Angeles, CA 90025, USA

Author for correspondence:
*Brandon S. Gaut*
*Email: bgaut@uci.edu*

Received: *20 February 2025*
Accepted: *26 April 2025*

*New Phytologist* (2025) **247:** 998–1014
**doi**: 10.1111/nph.70238

**Key words:** adaptation, climate change, crop wild relatives, grapes, mutational load, species distribution models, species' range.

## Summary

- Genetic load can reduce fitness and hinder adaptation. While its genetic underpinnings are well established, the influence of environmental variation on genetic load is less well characterized, as is the relationship between genetic load and putatively adaptive genetic variation. This study examines the interplay among climate, species range dynamics, adaptive variation, and mutational load – a genomic measure of genetic load – in *Vitis arizonica*, a wild grape native to the American Southwest.
- We estimated mutational load and identified climate-associated adaptive genetic variants in 162 individuals across the species' range. Using a random forest model, we analyzed the relationship between mutational load, climate, and range shifts.
- Our findings linked mutational load to climatic variation, historical dispersion, and heterozygosity. Populations at the leading edge of range expansion harbored higher load and fewer putatively adaptive alleles associated with climate. Climate projections suggest that *V. arizonica* will expand its range by the end of the century, accompanied by a slight increase in mutational load at the population level.
- This study advances understanding of how environmental and geographic factors shape genetic load and adaptation, highlighting the need to integrate deleterious variation into broader models of species response to climate change.

## Introduction

Mutations provide the raw material for evolutionary change, impacting fitness in various ways. A small fraction of mutations confers a fitness advantage, and many are evolutionarily neutral or nearly neutral. However, the vast majority are deleterious (Eyre-Walker & Keightley, 2007) and, as a result, are likely to face selective pressure to be purged. Nevertheless, deleterious mutations can persist within populations and thus influence evolutionary outcomes (Hedrick & Kalinowski, 2000; Robinson *et al.*, 2023). Surveying the number, prevalence, and fitness effects of deleterious mutations is crucial for understanding a wide array of biological phenomena, including rates of adaptation, the incidence and severity of genetic diseases (Kryukov *et al.*, 2007), and the conservation status of wild populations (Kyriazis *et al.*, 2021).

Given their importance and ubiquity, deleterious variants have been the focus of many theoretical and empirical studies (Robinson *et al.*, 2023). These studies have shown that the frequency and number of deleterious variants within a population are shaped by several evolutionary parameters, particularly effective population size ($N_e$; Charlesworth & Charlesworth, 1998). Populations with lower $N_e$ tend to accumulate more deleterious mutations than populations with higher $N_e$ because increased genetic drift and reduced selection efficacy limit the removal of deleterious variation (Bertorelle *et al.*, 2022). Accordingly, the number and frequency of deleterious variants often increase through population bottlenecks and other demographic events (Lohmueller, 2014). The accumulation of deleterious variants also depends on dominance coefficients (Simons *et al.*, 2014), the distribution of fitness effects, and the duration and timing of a bottleneck (Brandvain & Wright, 2016; Bortoluzzi *et al.*, 2020). (In fact, bottlenecks of sufficient duration and severity can purge highly deleterious, homozygous mutations (Grossen *et al.*, 2020; Femerling *et al.*, 2023).) Demography helps explain the increased number and frequencies of deleterious variants in domesticated species (Renaut & Rieseberg, 2015; Liu *et al.*, 2017; Moyers *et al.*, 2018) and in small populations of conservation concern (Femerling *et al.*, 2023). The accumulation of deleterious variants can also be influenced by hitchhiking and gene flow. Gene flow can either raise or lower genetic load, depending on the burden of deleterious variants carried by introgressed haplotypes (Kim *et al.*, 2018; Zhang *et al.*, 2020; Xiao *et al.*, 2023).

The evolutionary processes shaping genetic load also vary across geographic landscapes. Populations at range edges often experience environmental marginality – approaching the limits of their ecological tolerance – potentially leading to reduced $N_e$

and the accumulation of deleterious variation. However, this pattern can differ between 'leading' and 'trailing' edges of a range shift (Angert *et al*., 2020). Populations at the leading (i.e. expanding) edge of a range shift are likely to experience serial founder events, assortative mating, and strong selection pressure to adapt to new biotic and abiotic environments. Once established, deleterious mutations can propagate with an expanding range front via gene surfing (Travis *et al*., 2007; Excoffier *et al*., 2009). By contrast, populations on trailing edges are more likely to have had large historical population sizes with high genetic diversity that can erode as the climate shifts and population sizes crash (Hampe & Petit, 2005). Consistent with these predictions, increased numbers and frequencies of deleterious variants have been found at range edges in both laboratory (Weiss-Lehman *et al*., 2017; Bosshard *et al*., 2017) and natural populations (González-Martínez *et al*., 2017; Willi *et al*., 2018, 2022; Rougemont *et al*., 2020; Takou *et al*., 2021).

Because populations at the leading edge of range expansion often encounter new or unique environmental challenges, it seems reasonable to posit that the number and frequency of deleterious variants covary with environmental and climatic markers. Yet, there have been few attempts to relate distributions of deleterious variants to climatic variation (Willi *et al*., 2022). We believe that exploring the connection between deleterious variants and climate is worthwhile for at least two reasons. First, such an exploration is likely to provide additional insights into the evolutionary processes that shape the spatial distribution of genetic variation. Second, since deleterious variants affect fitness, understanding their patterns can be informative about the probability of population persistence (Aguirre-Liguori *et al*., 2021) and adaptation in the context of climate change (Sánchez-Castro *et al*., 2022). Recently, substantial attention has focused on modeling the fate of putatively adaptive mutations to predict the fate of populations under predicted climate change (Fitzpatrick & Keller, 2015; Waldvogel *et al*., 2020; Capblancq *et al*., 2020a). Except in the special case of antagonistic pleiotropy, these approaches typically overlook deleterious variants, representing a potentially major conceptual gap in the field of climate genomics (Aguirre-Liguori *et al*., 2021).

Many of the arguments about population size and edge effects also apply to adaptive variants, although often with opposite trends. For example, increased drift in small, edge populations can counteract selection, leading to the expectations of fewer adaptive alleles in these populations (Willi *et al*., 2006). Gene flow from large, central populations may contribute to this phenomenon by swamping the complement of new adaptive alleles in edge populations (Sexton *et al*., 2009). This framework generally posits that mutational load for small $N_e$ edge populations should be negatively correlated with the complement of adaptive variants, perhaps especially at the leading edge of expansion (Sánchez-Castro *et al*., 2022). However, some have argued that adaptive alleles are more likely to be found in leading-edge populations (Macdonald *et al*., 2017) due to strong selection in marginal habitats. More generally, the accumulation of adaptive alleles depends also on factors like the severity of environmental gradient across the species' range, the temporal pace of range expansion, population connectivity, historical $N_e$, and the strength of selection (Hedrick & Garcia-Dorado, 2016; Polechová, 2025). Given these complex dynamics, it is not surprising that the empirical literature is mixed (Willi *et al*., 2006). Some studies find fewer putatively adaptive variants in edge populations (Sánchez-Castro *et al*., 2022), while others identify more climate-associated (and putatively adaptive) alleles in edge populations (Aguirre-Liguori *et al*., 2017). Still, others find evidence for increased load in edge populations but without reduced fitness (Willi *et al*., 2006; Takou *et al*., 2021). Additional empirical work that incorporates some of the complexities of time, population size, environmental variation, and range shifts may help establish general relationships that may, in turn, be useful for predicting the ecological limits of species (Sexton *et al*., 2009).

Here, we examine the influence of climatic variation and species range dynamics on deleterious genetic variation in *Vitis arizonica* (Engelm.), a perennial crop wild relative (CWR) of domesticated grapevine (*V. vinifera* ssp *vinifera*). Crop wild relatives have multi-billion dollar impacts on the global economy (Bohra *et al*., 2022) and provide agronomically beneficial traits to domesticate as rootstocks or through hybrid breeding. As such, they are considered to be of urgent conservation concern (Khoury *et al*., 2020), particularly as climate change initiates species' range shifts. *Vitis arizonica* is an interesting candidate for study because it is native to a wide environmental range encompassing Northern Mexico and the Southwest United States (Heinitz *et al*., 2019), where extreme heat and drought are common. It thus has the potential to contribute adaptations – such as drought tolerance, salinity tolerance, and pathogen resistance – that may be useful to viticulture. In fact, *V. arizonica* segregates for resistance to *Xylella fastidiosa* (Riaz *et al*., 2018, 2020; Morales-Cruz *et al*., 2023), a plant pathogen that causes Pierce's disease (PD) in grapevines and also infects major crops, such as almonds, olives, and coffee. Consequently, *V. arizonica* has already been utilized in breeding programs to introduce PD resistance into *V. vinifera* varieties (Quinton, 2019).

In this study, we estimated mutational load, a measure of genetic load, and identified putatively deleterious and adaptive variants in 162 resequenced *V. arizonica* individuals. We interrogated these variant classes in relation to genomic, geographic, and bioclimatic factors to address four key questions: (1) How are putatively deleterious variants distributed among *V. arizonica* individuals, and how does this distribution vary across the species' range? (2) Which aspects of population history and environmental conditions best predict the spatial distribution of deleterious variants across the landscape? (3) To what extent are mutational load and putatively adaptive variants associated, and do their relationships differ between leading-edge and trailing edge populations? (4) What insights do predictive climate models yield about potential trends for deleterious and adaptive variation?

## Materials and Methods

### Variant calling

Illumina whole-genome sequencing reads from 172 individuals (Supporting Information Table S1), obtained from NCBI BioProjects PRJNA731597 (Morales-Cruz *et al*., 2021) and

PRJNA842753 (Morales-Cruz *et al.*, 2023), were processed using TRIMMOMATIC v.0.39 (Bolger *et al.*, 2014) to remove adapters and low-quality sequence using the following arguments: 'ILLUMINACLIP:'$ADAPTERSPE':2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:60'. Processed reads were then mapped to the *Vitis arizonica* (Engelm.) B40-14 v.2.0 genome assembly (Morales-Cruz *et al.*, 2023) with BWA-MEM 0.7.12r1039 (Li, 2013) using the default parameters. Sequence Alignment/Map format (SAM) alignments were sorted and converted to indexed Binary Alignment/Map (BAM) using SAMTOOLS 1.17 (Danecek *et al.*, 2021). Sequencing duplicates were marked in the alignments using the picard MarkDuplicates module included in GATK v.4.2.6.1 (Van der Auwera & O'Connor, 2020).

Single-nucleotide polymorphisms (SNPs) were called per sample using GATK HaplotypeCaller in GVCF mode followed by joint calling across all samples with GenotypeGVCFs. BCFTOOLS v.1.17 (Danecek *et al.*, 2021) was used to filter SNPs, keeping only biallelic SNPs with quality of 20 or greater that also passed the GATK 'best practices' hard filters: excluding sites with 'QD $< 2$ | FS $> 60$ | SOR $> 3$ | MQ $< 40$ | MQRankSum $< -12.5$ | ReadPosRankSum $< -8.0$'. Sites with minor allele frequency $< 0.01$, with site depth greater than the mean plus SD of depth across all sites, and sites that had $> 5\%$ missing calls between individuals were filtered. Annotation for predicted SNP effects was done with SIFT-4G (Vaser *et al.*, 2016) using a custom database based on the *Vitis arizonica* (Engelm.) B40-14 v.2.0 genome assembly and annotation (Morales-Cruz *et al.*, 2023). SNPs were polarized by including six outgroup samples: three individuals each from *Vitis girdiana* (Munson) (individuals SC11, SC33, and SC51) and *Vitis monticola* (Buckley) (individuals C20-93A, T_03-02_S01, and T40) in the call set. These outgroup samples were mapped to the reference genome and jointly genotyped with the focal samples. Subsequent analyses were based on sites that had no missing data across outgroup samples and were also consistently homozygous in the outgroups; for sites that fit these criteria, the outgroup genotype was assigned as the ancestral state.

As a final filtering step, principal component analysis (PCA) on SNPs was done using only the *V. arizonica* samples using PLINK v.2.0 (Chang *et al.*, 2015). Variants were pruned for linkage disequilibrium in 50 variant windows, with a step size of 10, and $R^2$ threshold of 0.20 before the PCA. Based on the 1.5 interquartile range rule, 10 samples were in the extremes of PC1 and visibly distinct from the remainder of the samples (Fig. S1). Upon further examination, several of the 10 samples were collected from disjunct geographic locales, raising concerns about their provenance. Furthermore, all 10 samples differed from the remaining samples by containing evidence of admixture from multiple other *Vitis* species, based on preliminary analyses of an unpublished, multispecies dataset. As a result, these 10 samples were removed from the dataset, leaving 162 individuals that represent much of the predicted species distribution (Fig. 1).

## Estimation of mutational load

Mutational load (load$_M$) was estimated for each individual similarly to Willi *et al.* (2018), but focusing on single individuals and considering counts of alleles instead of counts of homozygous sites. Briefly, load$_M$ was calculated as $P_n/(P_n + P_s)$, where $P_n$ corresponded to the proportion of derived alleles across all nonsynonymous sites and $P_s$ corresponded to the proportion of derived alleles across all synonymous sites. A complementary measure of mutational load based on the subset of nonsynonymous SNPs (nSNPs) predicted to be putatively deleterious by SIFT-4G was also calculated as: $P_d/(P_n + P_s)$, where $P_d$ was the proportion of derived deleterious alleles across all putatively deleterious sites. Sites with missing data for an individual were excluded from the calculations of $P_n$, $P_s$, and $P_d$, effectively making these statistics relative rates of nonsynonymous, synonymous, and deleterious mutations, respectively.

## Species distribution modeling

Species distribution models (SDMs) for *V. arizonica* were calculated for both the present and the Last Glacial Maximum (LGM) following the procedure described in Aguirre-Liguori *et al.* (2022). To assemble the data for constructing the SDMs, the WorldClim 2 bioclimatic variables (mean of observations from 1970 to 2000) (Fick & Hijmans, 2017) (Table S2) were extracted at 2.5 min resolution for all available geographic references of *V. arizonica* from the Global Biodiversity Information Facility (GBIF) (Occdownload Gbif.Org, 2020) (accessed 6 July 2022) and for the sampling locations (Table S1) using the RASTER 3.6–26 R package (Hijmans, 2023). The COORDINATECLEANER v.3.0.1 R package (Zizka *et al.*, 2019) was used to remove duplicate references; outliers (references that were in the top 5% of mean distance to all other locations); records in bodies of water, records at GBIF headquarters facilities, and records near country centroids or capitals.

To build SDMs, the correlated bioclimatic variables were first pruned based on variance inflation factor, retaining variables with $R < 0.8$. Next, the background area was set by selecting the overlap between the pruned occurrence records and the terrestrial ecoregions defined by (Olson *et al.*, 2001). The model was then built using the BIOMOD2 4.2–4 R package (Thuiller *et al.*, 2009) and the Maxent algorithm (Phillips *et al.*, 2006; Phillips & Dudík, 2008) using 20 bootstrap replicates, utilizing 70% of occurrences as the training data and retaining 30% as a test dataset. The final distribution model was selected by evaluating true skill statistics among 10-fold internal cross-validation. To estimate how the distribution of *V. arizonica* has changed from the past to the present and evaluate how it will change from the present to the future, the SDM was projected using the same set of bioclimatic variables for the LGM layer (*c.* 22 000 years ago) and to the future bioclimatic layers described below. For all SDM models, all range areas were calculated using the *expanse* function from the TERRA 1.7–55 R package (Hijmans, 2024), and the distance between centroids was calculated using the *distHaversine* function from the GEOSPHERE 1.5–18 R package (Hijmans, 2022).

## Calculation of features for predictions

The features for statistical modeling consisted of 24 variables and are summarized in Table 1; Fig. S2. The 19 WorldClim 2

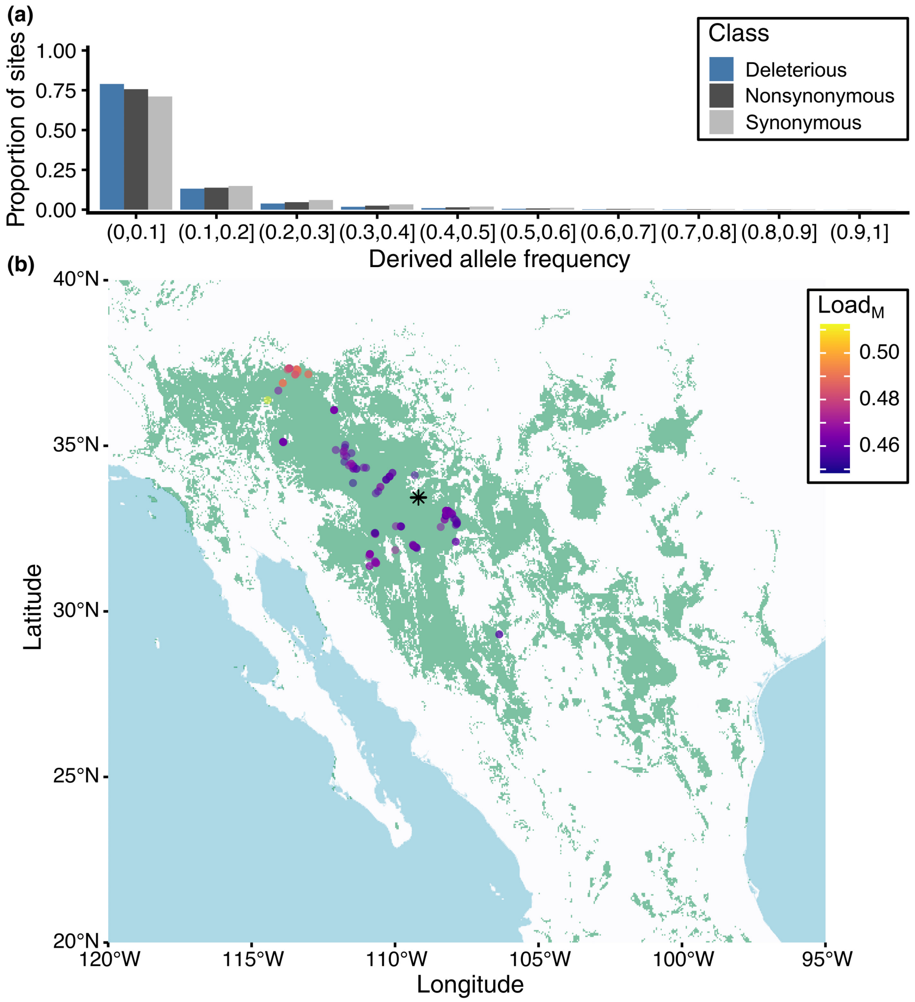

**Fig. 1** Site frequency spectra and mutational load in *Vitis arizonica*. (a) The derived allele frequency for three SNP classes: dSNPs (deleterious), nSNPs (nonsynonymous) and sSNPs (synonymous). (b) The species distribution model in the present (green) based on WorldClim data that summarize bioclimatic averages from 1970 to 2000 and GBIF species occurrence data. The points represent sampling locations for individuals used in genetic analyses and are colored according to the load$_M$ estimate per individual calculated with nonsynonymous variants. The black asterisk indicates the geographic centroid of the predicted range.

**Table 1** Summary of features used in modeling mutational load in *Vitis arizonica*.

| Feature | Class | Description |
|---|---|---|
| $H_o$ | Genetic | Observed heterozygosity for each individual based on SNPs detected across the entire sample |
| bio1–bio19 | Bioclimatic | WorldClim bioclimatic variables |
| $d_{geo}$ | Geography | Distance from geographic centroid based on species distribution modeling (SDM) |
| $d_{edge}$ | Geography | Distance to nearest species distribution edge based on SDM |
| $d_{niche}$ | Bioclimatic | Distance from the centroid of a bioclimatic PCA |
| $d_{dispersal}$ | Dispersal | Retroactive predicted dispersal distance between SDMs (e.g. from Last Glacial Maximum to the present) |

bioclimatic variables (Fick & Hijmans, 2017) were extracted for each individual by collection site coordinate using the RASTER 3.6–26 R package (Hijmans, 2023). Observed heterozygosity ($H_o$) was calculated from variable sites after pruning for linkage disequilibrium (50 SNP windows, 10 SNP step size, and $R^2$ threshold 0.20) using PLINK 2.0 (Chang *et al.*, 2015). The SDMs were used to generate the remaining features. The geographic centroid was defined as the median coordinate of the present SDM. For each of the samples, the distance to the geographic centroid was then calculated as the Euclidean distance between each sampling location and the centroid. The distance to geographic range edge was calculated between each sampling location and the nearest edge of the species range using the *distGeo* function from the GEOSPHERE 1.5–18 R package (Hijmans, 2022). The range boundaries were defined using the *boundaries* function implemented in the TERRA 1.7–55 R package (Hijmans, 2024). The distance to the niche centroid for each individual was calculated following the approximation of Lira-Noriega & Manthey (2014). Briefly, the 19 WorldClim 2 bioclimatic variables (Fick & Hijmans, 2017) were extracted for all pixels in the present SDM. Next, a PCA of the bioclimatic data was performed and the first six principal components (PCs) (explaining 95% of the variation in the dataset) were retained. The niche centroid was defined as the mean value among all observations along the six PCs. Finally, the distance to the niche centroid was then calculated as the Euclidean distance in multidimensional PC space between the niche centroid and the observation for each individual.

The final predictor was the estimated dispersal distance from the LGM to the present. The assumption of this calculation was that individuals with collection sites in the present but not past range must have dispersed to their present locations since the last epoch, while individuals present in the overlapping range did not necessarily need to disperse to their present location (dispersal = 0). As such, the geographic dispersal distance for each individual was calculated using the *distGeo* function from the GEOSPHERE 1.5–18 R package (Hijmans, 2022) between the collection site and the closest boundary of the past distribution.

## Statistical modeling

Random forest (RF) regression models (Breiman, 2001) were built using the TIDYMODELS 1.1.1 framework (Kuhn & Wickham, 2020) utilizing the RANGER 0.16.0 engine (Wright & Ziegler, 2017) running in the R programming environment (R Core Team, 2023). The dataset was split, allocating 75% of sample observations to training and reserving 25% for testing. Hyperparameter optimization for mtry (the number of randomly sampled predictors used to split the decision trees), min_*n* (the number of observations required for a tree node to be split again after segmentation), and trees (the number of decision trees to be included in the ensemble) was conducted over the respective ranges of 1–20, 1–10, and 500–1000 through Latin hypercube sampling, selecting 100 unique combinations for evaluation. Optimal hyperparameters were determined via 10-fold cross-validation on the training data and were used for the final model fits (Fig. S3). The permutation approach was used to calculate predictor importance.

The transformation to reduce collinearity among variables followed Johnson's method (Johnson, 2000). An orthogonal approximation ($\boldsymbol{Z}$) of the original data matrix ($\boldsymbol{X}$) was generated and used to train the RF model. The resulting model coefficients (i.e. variable importance) were then transformed back into the original data space for interpretation. Specifically, the training data were first centered and scaled by subtracting the mean and dividing by the SD so that each variable had a mean of 0 and a SD of 1. After standardization, singular value decomposition (SVD) was performed on the dataset. The data were then orthogonally approximated by the following transformation: $\boldsymbol{Z} = \boldsymbol{P}\,\boldsymbol{Q}^{T}$, where $\boldsymbol{P}$ and $\boldsymbol{Q}$ represent the left and right singular vectors from the SVD, respectively. A transformation matrix, $\boldsymbol{\lambda}$, was then calculated as: $\boldsymbol{\lambda} = \boldsymbol{Q}\,\boldsymbol{D}\,\boldsymbol{Q}^{T}$, with $\boldsymbol{D}$ being a diagonal matrix containing the singular values from the SVD. After the RF model was fitted using the orthogonalized data, the resulting importance values were approximated back into the original data space (i.e. the original predictors) by multiplying $\boldsymbol{\lambda}^{2}$ by the matrix of importance values. To use the models trained on orthogonally approximated data for predictions, test data ($\boldsymbol{X}_1$) were first scaled using the column means and SD of the original training dataset before being projected into the SVD space using the following formula: $\boldsymbol{Z}_{\textbf{test}} = \boldsymbol{X}_{\textbf{1 scaled}}\,\boldsymbol{Q}\,\boldsymbol{D}^{-1}\,\boldsymbol{Q}^{T}$.

Associations between load$_{\mathrm{M}}$ and each independent variable were calculated using univariate linear mixed models fit using the *lmekin* function from the COXME 2.2–18.1 R package (Therneau, 2022). All predictors were scaled using the *base::scale* function in R (R Core Team, 2023) before model fitting. A standardized relatedness matrix was calculated with GEMMA 0.98.5 (Zhou & Stephens, 2012) and included as a random effect in the linear mixed models.

## Projections of climate and mutational load in 2100

The future SDMs were built using the bioclimatic data for four Earth System Models (ESMs) (IPSL-CM6A-LR, MPI-ESM1-2-HR, MRI-ESM2-0, and UKESM1-0-LL) and four shared socioeconomic pathways (SSPs) (SSP126, SSP245, SSP370, and SSP585) for the period 2081–2100, downloaded from the CMIP6 project (Eyring *et al*., 2016). These SSPs represent alternative greenhouse gas trajectories, ranging from low emissions and strong mitigation (SSP126) to high emissions and continued fossil fuel use (SSP585; Riahi *et al*., 2017). The distance to geographic centroids and the distance to range edge for each model were calculated as described previously. To calculate forecasted dispersal, each sampling location was first assessed to determine whether it fell within the present range and was predicted to remain within the forecasted range. If so, the future dispersal distance was set to zero. If not, the geographic distance to the nearest forecasted range edge was estimated from the current location. Bioclimatic data were then extracted for all individuals using the present sampling coordinates or the nearest forecasted range edge (for individuals predicted to disperse by 2100).

The RF models trained with present-day bioclimatic data were used for all projections of load$_{\mathrm{M}}$ in the future. As the independent variables, WORLDCLIM 2.1 bioclimatic variable projections for 2081–2100 (2.5-min resolution) (Fick & Hijmans, 2017) for all 16 ESM:SSP combinations were used, along with the projections of future geographic centroids, distance to geographic range edges, and distance to niche centroid. For consistency in the model, dispersal was calculated from the LGM to the future as the sum of predicted dispersal from LGM to present and predicted dispersal from present to future. Observed heterozygosity was kept constant.

## Identifying adaptive alleles and performing GF projections

Climate-associated SNPs were identified for each of the 19 bioclimatic variables using latent factor mixed models (LFMMs) broadly following the methods described in Morales-Cruz *et al*. (2023). Missing genotypes were imputed using the LEA 3.10.2 R package (Frichot & François, 2015) by first estimating the *K* ancestral populations in the dataset ($K = 6$) using the *snmf* function and then using the *impute* function to set the missing genotype to the most common genotype of the individual's assigned ancestral population. We selected $K = 6$ based on a PCA of all SNPs and a scree plot of the latent factor variance; we also confirmed that results with $K = 6$ were conservative because they identified fewer SNPs that were a subset of those identified with lower $K$ values (e.g. $K = 4$). SNPs with minor allele frequencies below 0.05 were filtered from subsequent analysis, leaving 1012 352 SNPs. Latent factor mixed models were fit using the LFMM 1.1 R package (Caye *et al.*, 2019). In the LFMM, population structure was accounted for using $K = 6$ latent factors, a ridge penalty was applied to prevent overfitting, and test statistic inflation was controlled for using a genomic inflation factor. Finally, SNPs with Bonferroni-adjusted $P < 0.05$ were considered to be associated with climate and putatively adaptive.

To estimate the number of adaptive SNPs per individual ($N_G$), ordinal logistic regression (OLR) models were fit for each putatively adaptive SNP using the *polr* function from the MASS 7.3-60 R package (Venables & Ripley, 2002). Each OLR modeled genotype (encoded 0, 0.5, and 1 for 0, 1, and 2 alternative alleles, respectively) as a function of the respective associated bioclimatic variable. *P*-values were calculated from *z*-values of each OLR, and SNPs were considered for further analysis if they had a Bonferroni-corrected $P < 0.05$. SNPs in which the genotype predicted from the model matched the observed genotype were considered as adaptive and added to the adaptive genotype count ($N_G$). If a SNP was associated with multiple bioclimatic variables, it was only counted once toward $N_G$. Similar results were obtained when using a less stringent set of fitted SNPs, such as the complete set of putatively adaptive SNPs, with $N_G$ highly correlated (Pearson's $r > 0.980$, $P < 0.001$; Fig. S4) between the two treatments.

Finally, the genetic turnover of individuals across the present landscape was modeled using gradient forest (GF) models implemented in the GRADIENTFOREST 0.1–37 R package (Ellis *et al.*, 2012; Fitzpatrick & Keller, 2015). In the GF models, the unique set of SNPs identified as outliers by LFMM, representing putatively selected loci, was used as the response. The bioclimatic variables were used as the independent variables, and we performed 500 bootstrap iterations of GF analyses. The fitted GF model was then used to predict the expected genomic composition of individuals in the 16 future climate models. No migration was assumed for individuals present in both present and forecasted range (according to the SDM), while individuals expected to disperse by 2100 were considered to have migrated to the nearest predicted range edge, as described previously.

The genetic offset for each individual in each future climate scenario was calculated as the Euclidean distance between the present and future expected genetic composition (Fitzpatrick & Keller, 2015). Since the genetic offsets per individual were found to be highly correlated for the 16 climatic models (Spearman's rho > 0.7), the mean genetic offset per individual was reported and used for subsequent analyses.

### Phenotypic analyses

Linear mixed models to detect associations between phenotype and load$_M$ were built using the *lmekin* function from the COXME 2.2–18.1 R package (Therneau, 2022) using a standardized relatedness matrix calculated with GEMMA 0.98.5 (Zhou & Stephens, 2012) as a random effect as described previously. The likelihood-ratio pseudo $R^2$ was estimated using the *r.squaredLR* function from the MUMIN 1.48.4 R package (Bartoń, 2024).

## Results

### Mutational load estimates across the landscape

We called variants in a cohort of 172 resequenced individuals sampled across much of the native range of *V. arizonica* (Morales-Cruz *et al.*, 2023) (Table S1). After site and sample filtering (see the Materials and Methods section; Fig. S1), the final dataset represented genotypes for 162 individuals across 1320 747 biallelic SNPs with a minor allele frequency of 1% or greater. The filtered dataset had low levels of missing data, averaging 0.39% missing calls both per site and per sample. We annotated these SNPs for predicted effects on protein function, identifying 140 801 synonymous SNPs (sSNPs) and 146 830 nSNPs. The latter included 40 145 putatively deleterious SNPs (dSNPs), as predicted by SIFT-4G (Vaser *et al.*, 2016). We polarized sSNPs, nSNPs, and dSNPs using six individuals from two congeneric species (*V. girdiana* and *V. monticola*). After applying our polarizing criteria (see the Materials and Methods section), we assigned ancestral and derived alleles for 53.02% (74 656/140 801) of sSNPs, 54.01% (79 301/146 830) of nSNPs, and 55.89% (22 436/40 145) of dSNPs.

We expected both nSNPs and dSNPs to have been subjected to purifying selection and thus segregate at lower frequencies than sSNPs. We tested this hypothesis by calculating the site frequency spectrum (SFS) for each SNP class across the entire sample (Fig. 1a). As expected, the SFS for nSNPs was significantly different than that of sSNPs (Kolmogorov–Smirnov test, $D = 0.047$, $P < 2.2\text{e}{-16}$), reflecting an enrichment for low-frequency-derived nSNP alleles (chi-squared test, $\chi^2 = 367.96$, df = 1, $P < 2.2 \times 10^{-16}$). The SFS for dSNPs was also significantly different than the SFS for nSNPs (Kolmogorov–Smirnov test, $D = 0.033$, $P = 6.66 \times 10^{-16}$), with an even greater enrichment of low-frequency variation than of nSNPs (chi-squared test, $\chi^2 = 96.03$, df = 1, $P < 2.2 \times 10^{-16}$). Thus, nSNPs and dSNPs segregated as expected, with the SFS dominated by low-frequency variants.

We next calculated the proportion of derived alleles at either deleterious or nonsynonymous sites, which we called $P_d$ and $P_n$ (see the Materials and Methods section). These measures are analogous to counts of the number of derived dSNPs or nSNPs per genome, corrected for missing data. Similar measures have been used commonly to approximate mutational load because they can be insensitive to demographic history under some conditions (Simons *et al.*, 2014). $P_d$ and $P_n$ were highly correlated across the 162 individuals (Pearson $R^2 = 0.99$; $P < 0.001$), with the highest values clustered in the southernmost part of the species' range (Fig. S5). While this pattern suggests a higher mutational load in these samples, the proportion of derived synonymous alleles ($P_s$) was also elevated in the South (Fig. S6) and strongly correlated with both $P_n$ ($R^2 = 0.99$; $P < 0.001$) and $P_d$ ($R^2 = 0.97$; $P < 0.001$). Since $P_s$ likely reflects segregating neutral genetic variation – and thus is unlikely to be an accurate measure of the load of deleterious variants – we interpreted the strong correlations among $P_d$, $P_n$, and $P_s$ as reflecting the distribution of genetic diversity across samples rather than mutational load per se.

We therefore turned to an alternative measure of mutational load ($\text{load}_\text{M}$), based on previous work (Willi *et al.*, 2018). Briefly, $\text{load}_\text{M}$ was estimated as $P_n/(P_n + P_s)$. This measure is conceptually similar to well-used measures of selection like $d_n/d_s$ or $\pi_n/\pi_s$ that are normalized by presumably neutral genetic variation. $\text{Load}_\text{M}$ varied by 11% among individuals and was unevenly distributed across sampling locations, with higher values concentrated in the northernmost part of the predicted species range (Fig. 1b; Table S1). To complement $\text{load}_\text{M}$, we also calculated an analogous measure (i.e. $P_d/(P_n + P_s)$ based on dSNPs. The two measures were highly correlated ($R^2 = 0.73$, $P < 2.2 \times 10^{-16}$; Fig. S7) and produced qualitatively identical results in downstream analyses. Therefore, for simplicity and consistency with previous work (Willi *et al.*, 2018), we report $\text{load}_\text{M}$ based on nSNPs while briefly mentioning results from dSNPs, $P_d$ and $P_n$ where applicable.

## Compilation of features to predict $\text{load}_\text{M}$

One of our goals was to evaluate the relative contributions of genetic diversity, the environment, and species' range dynamics to $\text{load}_\text{M}$. With this objective in mind, we compiled a set of 24 features related to genomic diversity, climatic diversity, and species' range for each individual (Table 1; Fig. S2). The climatic features included the 19 WorldClim 2 bioclimatic variables, which summarize historical climate variation from 1970 to 2000 (Fick & Hijmans, 2017) at each collection site. To represent genomic variation, we estimated observed heterozygosity ($H_\text{o}$) for each individual, reasoning that this measure reflects aspects of historical $N_\text{e}$ (Crow & Kimura, 1970). The remaining four geographic features were related to the estimated species' range and relied on SDMs. One SDM estimated species' range in the present based on species occurrence data (Occdownload Gbif.Org, 2020) and recent (averaged from 1970 to 2000) bioclimatic data. This present-day SDM was used to calculate the minimum geographic distance of each individual to the nearest range edge ($d_\text{edge}$), measured in kilometers, and the Euclidean distance to the geographic ($d_\text{geo}$) and niche ($d_\text{niche}$) centroids (Table S1).

We also calculated a fourth geographic parameter: the distance of potential historical dispersion events ($d_\text{dispersal}$). To do so, we constructed an SDM from the LGM, reflecting climatic conditions *c.* 22 000 years ago. By comparing the present-day SDM to the LGM SDM, we determined whether the sampling site of an individual remained within the species' geographic niche over time or required migration since the LGM. The two SDMs suggested that the species distribution moved and expanded over geologic time, with a shift in the geographic centroid by 736 km (34 degrees west of north) and a 40.71% expansion by area from the LGM to the present (Fig. 2a). However, the two SDMs overlapped for only 28.08% of the estimated present range, suggesting a substantial history of migration and dispersal. To estimate $d_\text{dispersal}$, we calculated the minimum distance from the present-day sampling location to the LGM SDM range edge (see the Materials and Methods section). For individuals collected at locations that were present in both SDMs, we made the simplifying assumption that $d_\text{dispersal}$ was 0.0 km. Using this approach, 84 sampling locations represented potential historical dispersions since the LGM, with a median distance of 89.97 km. $d_\text{dispersal}$ was particularly pronounced for individuals sampled in more northern latitudes, suggesting that this region represents a leading edge of range expansion.

The collection of 24 genetic, geographic, and climatic features constitutes a multivariate dataset that can be used to predict $\text{load}_\text{M}$. As is common with such datasets, however, the features exhibited a complex pattern of correlated relationships (Figs 2b, S8). For example, bioclimatic variables related to temperature (e.g. bio1, bio5, bio6, bio8, bio9, bio10, and bio11; definitions for each bioclimatic variable are provided in Table S2) were positively correlated with each other, as were precipitation variables (e.g. bio13, bio17, and bio18). Other bioclimatic variables had negative correlations, including temperature variability (bio4 and bio7) relative to precipitation measures. Geographic measures also exhibited complex relationships; for example, the dispersal distance and the distances to geographic and niche centroids were significantly correlated with one another but negatively correlated with most precipitation variables. In short, few of the features were statistically independent, and thus, evaluation of potential predictors of $\text{load}_\text{M}$ must account for complex correlative relationships.

## Using random forest regression models to predict $\text{load}_\text{M}$ in the present

We built RF regression models to predict $\text{load}_\text{M}$ using the genetic, bioclimatic and geographic features as predictors. We chose RF for its ability to infer nonlinear relationships between predictors and response variables, recognizing both that the use of linear models has been criticized in these contexts (Benestan *et al.*, 2016; Fouqueau *et al.*, 2024) and that nonlinear relationships are commonly inferred between climatic and genetic variation (Aguirre-Liguori *et al.*, 2017; Capblancq & Forester, 2021). Random forest also provides a straightforward measure of each feature's importance to the model's predictive performance. However, correlations between predictors can distort these importance scores. For instance, if two features are perfectly correlated, the importance attributed to their shared information is assigned to only one, making the other appear irrelevant since it does not contribute additional information to the model (Breiman, 2001). Consequently, complex correlative relationships between features make it difficult to infer each feature's relevance to the response from importance scores.

Since our goal was to interpret the biological relevance of features for predicting $\text{load}_\text{M}$, we mitigated this challenge by applying Johnson's relative weights (Johnson, 2000) to RF regression models (see the Materials and Methods section). Johnson's method accounts for the correlation structure among predictors by transforming the features into uncorrelated PCs via SVD. The RF model is then trained using these PCs as features, generating

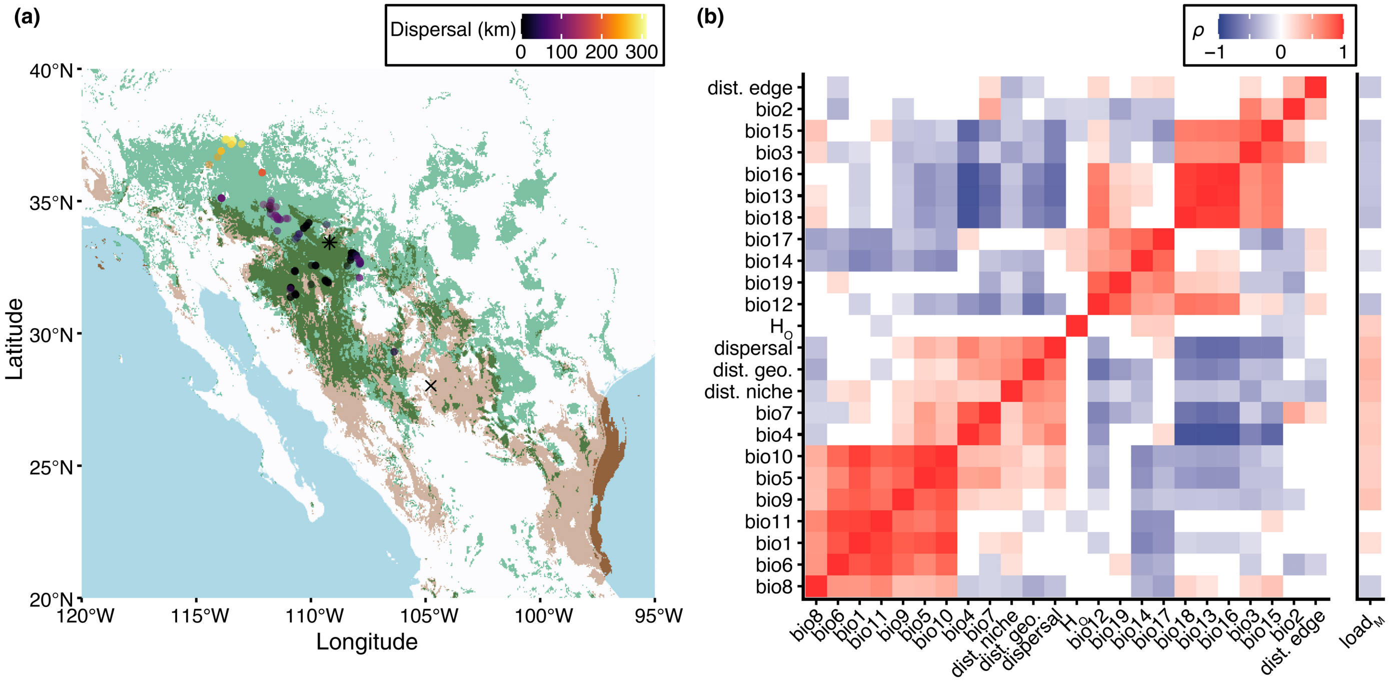

**Fig. 2** Sample dispersal and feature correlations. (a) Projected species distribution models (SDM) *for Vitis arizonica* during the Last Glacial Maximum (LGM, *c.* 22 Kya, brown), for the present (green), and for the overlap between the two SDMs (greenish brown). The dark brown region within the present-day Gulf of Mexico represents areas where the SDM for the LGM overlaps with land that was exposed due to lower sea levels. Each point on the landscape represents the sampling location for an individual and is colored according to predicted dispersal from the LGM to the present, with more distantly dispersed individuals in warmer colors. The asterisk indicates the geographic centroid in the present, while the X indicates geographic centroid during the LGM. (b) Spearman's correlations between pairs of features used in the random forest (RF) model, as well as between load$_M$ and each individual feature. In addition to the 19 WorldClim bioclimatic variables, the features include distance from the present-day SDM edge (dist. edge), observed heterozygosity ($H_o$), the estimated dispersal distance shown in Panel a (dispersal), the distance from the present-day geographic centroid (dist. geo), and the distance from the present-day niche centroid (dist. niche). Only correlation tests with $P < 0.05$ are filled, with the color demonstrating both the magnitude and direction of each correlation ($\rho$). See also Supporting Information Fig. S8.

importance scores for each PC. Johnson's weightings subsequently transform the importance scores back into the original feature space (see the Materials and Methods section). The resulting relative weighted importance of each original feature therefore should reflect its biological relevance to load$_M$ while controlling for correlations between features.

Our RF model to predict load$_M$ yielded an 'out of bag' $R^2$ estimate of 0.59 and an $R^2$ of 0.61 on the withheld test dataset, slightly below the 10-fold cross-validation interval of $R^2 = 0.70 \pm 0.08$ (SE) (Fig. 3a). After transformation, 90% of the cumulative relative weighted importance was explained by 15 predictors (Fig. 3b). Notably, > 50% of the relative cumulative importance was explained by the combination of the mean temperature of wettest quarter (bio8) and $d_{geo}$. $H_o$ and $d_{dispersal}$ were also important predictors of load$_M$. Similar results were obtained using other measures of load, such as $P_d$, $P_n$ or load$_M$ based on dSNPs (Fig. S9). For example, the model to predict $P_n$ inferred bio8 (mean temperature of wettest quarter; Table S2) as the most important variable, with the top 10 variables again, including $H_o$, $d_{geo}$, and $d_{dispersal}$. In addition to bio8, precipitation in the driest quarter (bio17) was important across RF models; indeed, it was the most important variable when $P_d$ was used as the dependent variable. Altogether, RF modeling identified a consistent set of important variables to predict load, including climatic variables (e.g. bio8 and bio17) and predictors commonly thought to be important based on population genetic theory (e.g. $H_o$ and $d_{geo}$).

One concern is that the RF models do not explicitly account for the nonindependence of observations due to relatedness between individuals (i.e. population structure). We addressed this by fitting univariate linear mixed models that included a kinship matrix to account for population structure. In total, 54.2% (13/24) of features were significantly associated with load$_M$ in the linear mixed models (Bonferroni-adjusted $P < 0.05$; Table S3). The 13 associated features included 11 of 14 (78.6%) of the most important features from the RF model. Although it is clear that linear models often do not adequately capture relationships across landscapes (Aguirre-Liguori *et al.*, 2017; Capblancq & Forester, 2021), the two modeling frameworks (i.e. RF and linear modeling) were complementary with respect to identifying explanatory features, even when genetic relatedness was taken into account.

## Species range and load$_M$ at the end of the century

Given the RF models' ability to predict load$_M$ in the present climate, we next applied the trained model to forecast changes in

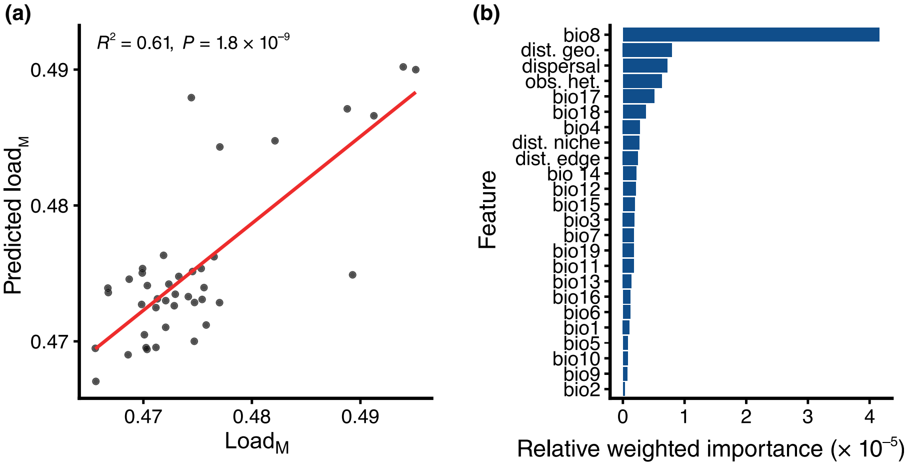

**Fig. 3** Random forest (RF) regression models to predict load$_M$. (a) The performance of the RF model compares the predictions ($y$-axis) to the observed ($x$-axis) values. The red line indicates the linear model fit, and the slope ($R^2 = 0.61$) was highly significant ($P = 1.8 \times 10^{-9}$). (b) The inferred feature importance used to predict load$_M$. The features are ranked by their inferred importance. The distance-related metrics are defined in Table 1; bio8 is the mean temperature in the wettest quarter, and definitions for the remaining bioclimatic variables are provided in Supporting Information Table S2.

load$_M$ by the end of the century. To account for uncertainty in future climates, we evaluated 16 potential climate trajectories, combining four ESMs with four SSP scenarios, using forecasted climate averages from 2080 to 2100.

Overall, the future SDMs predicted that the potential species range is likely to shift northward and expand in size by 2100. Even under the most optimistic emissions scenario (i.e. SSP126, sustainability), the geographic centroid is predicted to move 87.80 to 226.45 km from the present centroid, with the degree of range shift dependent on the ESM (Fig. 4a). Increasing emissions were associated with a greater degree of range shift, ranging from 160.91 to 348.57 km for SSP245, 238.01 to 510.44 km for SSP370, and 226.53 to 697.69 km for SSP585 (Fig. S10). Potential species range area was predicted to increase under all scenarios except for the MPI-ESM1-2-HR ESM at SSP126 and SSP245. The change in range area was greater under increased emissions, with predicted changes of −10.8 to 53.8% by area for SSP126, −10.2 to 84.5% for SSP245, 5.87 to 146% for SSP370, and 17.8 to 209% for SSP585 (Fig. 4b). The predicted range shifts included losses to the present range, particularly in the southernmost extreme. The southern extremes likely contain trailing edge populations that are at risk of extinction.

By comparing the current and projected SDMs, we estimated the minimum dispersal distance ($d_{\text{dispersal}}$) for each sample by 2100. As with our previous analysis, we only calculated dispersal for samples collected in regions occupying present but not future predicted range (e.g. we assumed that samples collected in overlapping regions need not migrate). Across the 16 climate models, we estimated that the lineages of between 1 (MRI-ESM2-0, SSP585) and 17 (UKESM1-0-LL, SSP585) locations will need to disperse by 2100, with a minimum predicted migration between 2.41 and 306.55 km (median, 6.19 km). Finally, we used the trained RF model to predict how load$_M$ might change by 2100. As input to the model, we used the forecasted bioclimatic data along with species' range features (e.g. $d_{\text{geo}}$, $d_{\text{niche}}$, and $d_{\text{dispersal}}$) predicted by future SDMs. In these analyses, we kept $H_o$ constant, because we had no basis to predict its values into the future, representing a limitation to our approach. The RF models yielded two primary insights about load$_M$ in 2100: First, load$_M$ was generally predicted to increase for the majority of individuals and, second, variation in load$_M$ was forecast to decrease markedly, with outliers regressing toward the mean (Fig. 4c). Specifically, load$_M$ was predicted to change by −5.16 to 3.33% per individual, with samples at more northern latitudes (latitude > 36) predicted to have reduced load$_M$, while individuals closer to the present range centroid expected to have increased load$_M$ (Fig. S11; Table S4).

## Measuring the complement of putatively adaptive alleles

To date, most projections of genetic diversity in future climates have focused on alleles that putatively contribute to local adaptation (Palumbi *et al*., 2014; Bay *et al*., 2018; Capblancq *et al*., 2020a), based on methods like GF (Fitzpatrick & Keller, 2015). These methods ignore *most* genetic diversity within populations, much of which may be pertinent for predicting the fate of species and populations (Aguirre-Liguori *et al*., 2021). In this section, we explicitly compare mutational load to the complement of putatively adaptive alleles across *V. arizonica* individuals.

We identified putatively adaptive SNP variants by first using LFMM2 (Caye *et al*., 2019) to test associations between SNPs and each of the 19 bioclimatic variables. One bioclimatic variable (bio2, mean diurnal temperature range) had no associated SNPs and was discarded from further analysis. The remaining 18 variables had between two (bio5, max temperature of warmest month) and 770 (bio4, temperature seasonality) associated SNPs, altogether yielding a total of 3225 unique putatively climate-adapted SNPs (Table S5). Although these SNPs were putatively adaptive, it remained unclear which genotypic state was adaptive (or maladaptive) across the climatic range. To further filter for adaptive SNPs and to identify adaptive states for each, we applied OLR. Each regression compared the genotypes at a SNP site to the associated climatic variable (see the Materials and Methods section); this step not only further confirmed genotype–climate associations, but it allowed us to identify genotypes that were adaptive – that is within their expected climate as defined by OLR. After filtering SNPs that were not correlated

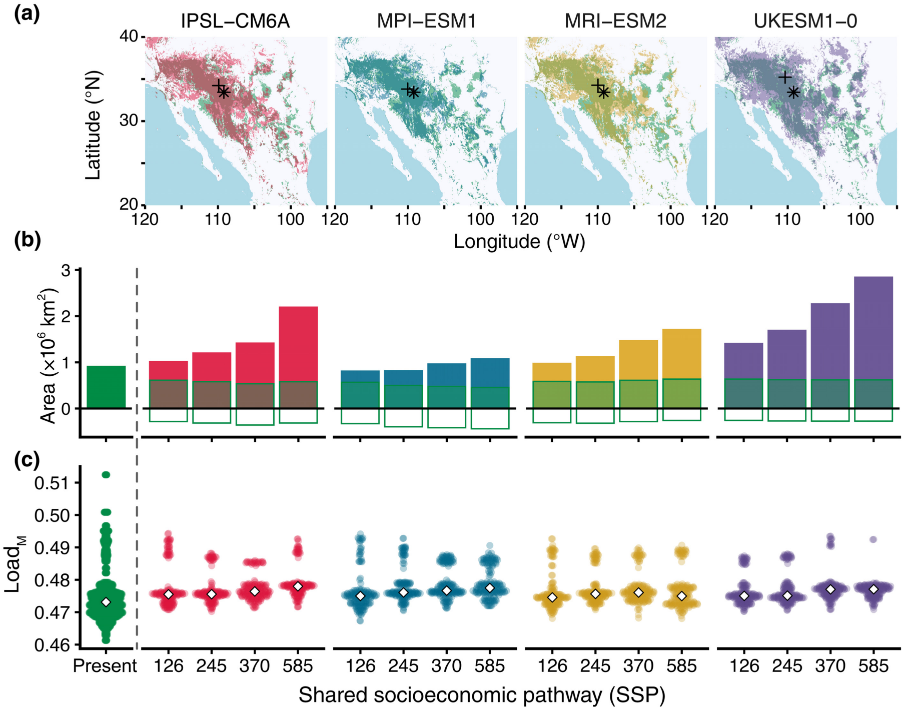

**Fig. 4** Predicted species distribution models (SDM) and load$_M$ in 2100. (a) Predicted SDMs for *Vitis arizonica* in 2100 under Shared Socioeconomic Pathway (SSPs) SSP126 (a sustainability-focused scenario with global warming limited to < 2°C above preindustrial levels) are shown for four Earth Systems Models (ESMs) – IPSL-CM6A (red), MPI-ESM1 (blue), MRI-ESM2 (gold), and UKESM1-0 (purple) – compared to the present-day SDM (green). In each map, the predicted future geographic centroid is denoted by a plus (+), while the present geographic centroid is represented by an asterisk (*). Predicted SDMs for additional SSPs are shown in Supporting Information Fig. S10. (b) The predicted area of each SDM in 2100 compared with the present (left), shown for all combinations of ESM and SSP. The height of each bar indicates the total predicted area subdivided by new areas that were not part of the present SDM (top), overlapping area between the predicted future and present SDM (middle), and the area of the present range predicted to be lost (green outline, negative values). (c) The distribution of load$_M$ in the present (left) compared to the predicted distributions in 2100 for each combination of ESM and SSP, as projected by the random forest model. In each beeswarm, the diamond represents the median values per group.

with climate via OLR, we retained 2162 unique SNPs. For these, we counted the number of sites with putatively adaptive genotypes ($N_G$) across all sites within an individual. $N_G$ varied by as much as 186% among individuals (range: 756–2162) and was markedly lower in the northern ranges of *V. arizonica* (Fig. 5a) than in the samples in the southern extremes (range: 2109–2162). In fact, mean $N_G$ (mean = 1464.94; SD = 338.59) for the 36 northernmost individuals, representing potential leading edges, was significantly lower than that of the 36 southernmost individuals from trailing edges (mean = 2152.08; SD = 12.03; $t = -12.169$, df = 35.09, $P < 0.001$). Furthermore, $N_G$ and load$_M$ were significantly negatively correlated across individuals ($R^2 = 0.53$; $P < 0.001$; Fig. 5b), and $N_G$ was also negatively correlated with geographic measures, most markedly $d_{\mathrm{dispersal}}$ ($R^2 = 0.80$; $P < 0.001$; Fig. S12). In other words, individuals with a history of dispersal had fewer adaptive genotypes, while central and trailing edge samples tended to have higher $N_G$ than leading-edge samples.

Finally, we used GF to measure genetic offsets (*go*) based on the set of putatively adaptive SNPs and the set of 16 end-of-century climate models used to project load$_M$ (as mentioned in the previous section). Higher *go* values reflect individuals that may be more vulnerable to climate change because they require more turnover of adaptive alleles (Fitzpatrick & Keller, 2015; Capblancq *et al.*, 2020a; Gain *et al.*, 2023). We estimated a single *go* for each individual by averaging across SNPs and across the 16 climate models. The *go* estimates were positively correlated with load$_M$ ($R^2 = 0.34$, $P < 0.001$, Fig. 5c) and negatively correlated with $N_G$ ($R^2 = 0.58$, $P < 0.001$, Fig. 5d). It is important to note that correlations between load$_M$ and either $N_G$ or *go* do not appear to be due principally to overlapping SNPs between datasets. For example, of the 3225 SNPs identified as putatively adaptive, only 12.3% (396) overlapped with nSNPs and 3.6% (115) overlapped with dSNPs.

## Load$_M$ and $N_G$ are associated with a key phenotype

One limitation of genomic studies is that they must assume that measures like load$_M$ and *go* are related to fitness. These assumptions are rarely tested (but see Mezmouk & Ross-Ibarra, 2014; Sánchez-Castro *et al.*, 2022), calling into question whether the measures have any relevance. We cannot test fitness directly in *V. arizonica* yet, but our accessions have been phenotyped in the glasshouse for two agronomic traits that may contribute to fitness in the wild: chloride exclusion, a measure of salt tolerance (Heinitz *et al.*, 2020), and quantitative resistance to *Xylella fastidiosa*

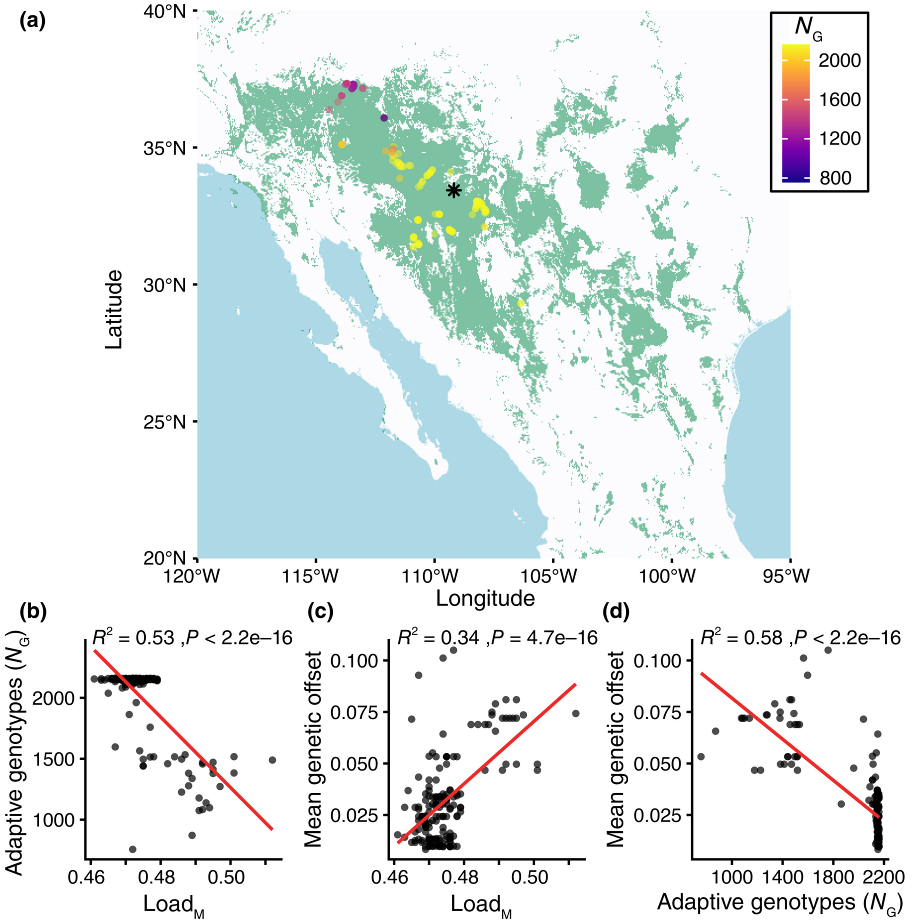

**Fig. 5** Adaptive genotypes on the landscape. (a) As in Fig. 1(b), green represents the species distribution model for *V. arizonica* in the present-day based on WorldClim (bioclimatic averages from 1970 to 2000) and GBIF species occurrence data, with the black asterisk indicating the geographic centroid of the predicted range. The points represent sampling locations for individuals used in genetic analyses and are colored according to the number of adaptive genotypes ($N_G$) estimated per individual, with cooler colors reflecting lower $N_G$. (b) The relationship between observed $\text{load}_M$ and $N_G$ across all individuals. (c) The relationship between $\text{load}_M$ and genetic offsets. (d) The relationship between $N_G$ and genetic offsets. In (b–d), red lines indicate the fit of linear models; in each case, the slope is highly significant, as reflected by the $P$-value and $R^2$.

(Riaz *et al.*, 2020; Morales-Cruz *et al.*, 2021, 2023), the causative agent of PD. We contrasted these quantitative phenotypes to $\text{load}_M$ and $N_G$, finding no evidence that either is related to chloride exclusion phenotypes. However, $\text{load}_M$ was positively associated with assayed concentrations of *X. fastidiosa* postinfection ($P = 1.5 \times 10^{-4}$, $R^2_{LR} = 0.33$, Fig. S13), while $N_G$ was negatively associated ($P = 3.6 \times 10^{-7}$, $R^2_{LR} = 0.38$). These results are consistent with the negative correlation between $\text{load}_M$ and $N_G$. Thus, plants with higher $\text{load}_M$ and fewer adaptive alleles are more susceptible to PD.

## Discussion

Understanding the relationship between climatic and genetic variation is essential for predicting which populations will thrive, persist, or face extinction in the future. Historically, the probability of persistence has been evaluated by estimating SDMs, but they have an important limitation: They ignore the ability for a species to evolve (Garzón *et al.*, 2019; Collart *et al.*, 2020). Newer methods have worked toward incorporating the potential for evolutionary change by using population genomic information in concert with climate predictions (Fitzpatrick & Keller, 2015; Exposito-Alonso *et al.*, 2019; Waldvogel *et al.*, 2020; Capblancq *et al.*, 2020a). The GF method typically relies first on identifying mutations that likely contribute to local adaptation and then on inferring the relationship between bioclimatic variables and the frequency of adaptive variants (Fitzpatrick & Keller, 2015). The inferred relationship can then be used to predict the genetic offset (*go*), providing a quantitative measure of preadaptation to climate shifts. This and similar approaches have been applied to evaluate the potential persistence of wild populations from several species (reviewed in Capblancq *et al.*, 2020a) and also to assess agronomic suitability of specific genotypes in projected climates (Rhoné *et al.*, 2020; Aguirre-Liguori *et al.*, 2022). These approaches do not, however, incorporate evolutionary processes beyond local adaptation, such that they largely ignore genetic drift, gene flow, migration, and dispersal (Waldvogel *et al.*, 2020). Another major limitation is that they usually (although not always; Bay *et al.*, 2018; Ruegg *et al.*, 2018; Rhoné *et al.*, 2020; Booker *et al.*, 2020) focus on a tiny subset of genetic variants – that is putatively adaptive variants associated with climatic measures. They thus ignore most genetic information, including the deleterious variants that contribute to mutational load.

A key challenge is integrating additional categories of genetic variation into climate models and assessing their predictive value for species' persistence amid climate change. Here, we take a step toward this goal by studying relationships among putatively deleterious variants, mutational load, species' range dynamics, climate-associated adaptive SNPs, and climate in *Vitis arizonica*, a wild grape species endemic to the American Southwest. Our investigations are relevant not only for evaluating potential population persistence but also for a number of questions in evolutionary biology, such as differences between edge and nonedge populations (Vucetich & Waite, 2003), relationships between genetic diversity and range centroids (Eckert *et al.*, 2008; Lira-Noriega & Manthey, 2014), and dissimilarities between leading vs trailing edges during range shifts.

## Mutational load is predictable and elevated at leading edges

We estimated mutational load (load$_M$), as the proportion of exonic SNPs that were predicted to be nonsynonymous and then related load$_M$ to various bioclimatic and geographic variables. The geographic variables were estimated from SDMs that used recent data (averaged from 1970 to 2000), data from the LGM and bioclimatic projections to 2100. Comparing the LGM to the present day reveals that some of our sampling locations likely represent historical dispersal events, particularly samples from the North that represent a leading edge of expansion (Fig. 2A). Species distribution models also indicate that the range of *V. arizonica* has expanded over the last *c.* 22 000 yr and is likely to continue to expand into the future, principally by continuing to move northward (Figs 4a, S10).

We found that load$_M$ was generally elevated in Northern samples (Fig. 1B), consistent with previous studies reporting increased load and/or reduced $N_e$ in leading-edge populations (Willi *et al.*, 2018, 2022; Takou *et al.*, 2021; Sánchez-Castro *et al.*, 2022; Cisternas-Fuentes & Koski, 2023). These observations are consistent with the concept of 'expansion load', where alleles that are normally purged by selection can 'surf' to high frequency at range fronts due to repeated bottlenecks and drift, thereby reducing fitness at the expansion margin (Travis *et al.*, 2007; Excoffier *et al.*, 2009). By contrast, trailing and central populations tended to have lower load$_M$, perhaps reflecting longer term, historical stability.

To investigate potential causes of these patterns, we used RF models to determine which of 24 potentially informative features – including four geographic descriptors, one genetic descriptor, and 19 bioclimatic variables – contribute to predicting load$_M$. This analysis led to at least two key conclusions. First, although mutational load is expected to be shaped primarily by stochastic genetic drift, it was nonetheless highly predictable. The RF models yielded $R^2 > 0.60$ on withheld test data (Fig. 3a). Second, bioclimatic variables were among the top-ranked predictors (Fig. 3b). We do not conclude from these analyses that climate affects mutational load directly. Instead, we propose that the bioclimatic variables represent ecological conditions that affect population size, population density, or perhaps additional aspects of population history that are not fully captured by our set of genetic and geographic predictors (Willi *et al.*, 2018). Consistent with this view, different genetic and geographic summaries, such as $d_{\text{dispersal}}$, $H_o$, and $d_{\text{geo}}$, were also among the important predictors of load$_M$, suggesting that no single genomic metric aptly summarizes the complexities of range expansion and population dynamics. Not all genomic summaries were valuable, however, because two additional geographic measures – $d_{\text{niche}}$ and $d_{\text{edge}}$ – had generally low predictive importance. Although both have been used extensively to study landscape dynamics (Lira-Noriega & Manthey, 2014), we suspect they are poor predictors because they do not indicate directionality – for example if the niche spans a cold-to-warm gradient, samples at either extreme can have identical values for $d_{\text{niche}}$.

Our RF models are, of course, subject to caveats and assumptions. One concern is that the reported feature importance may not accurately reflect biological significance due to correlations between predictors – an issue often overlooked in the literature. We address this problem by applying Johnson's relative weights, originally developed to adjust beta values from multiple linear regression (Johnson, 2000), to RFs. Johnson's method has been utilized widely across scientific fields but rarely, as far as we know, applied to multifaceted biological and genomic data (e.g. Core *et al.*, 2014; Chen *et al.*, 2018; Ghanipoor-Samami *et al.*, 2018; Shen & Chen, 2020; Li *et al.*, 2022). Another caveat is that we relied on load$_M$ as a measure of mutational load under the assumption it reflects genetic load. Measuring genetic load directly is notoriously difficult because it is a theoretical construct related to the unmeasurable (i.e. a decline in fitness relative to a theoretical fitness optimum) (Crow, 1958). To be thorough, we also explored other well-known measures of mutational load, such as the proportion of deleterious variants ($P_d$ or $P_n$) per genome. Although the measures differed in their distributions across geographic space (Figs 1, S5, S6), they led to similar inferences with respect to predictability and the relative importance of predictors. Needless to say, all of these measures of load share limitations – that is, they are strictly additive, assume that each putatively deleterious allele has the same effect on fitness, and do not account for potential dominance relationships between alleles. A final concern is that individual features may have been predictive because they covary with (and may merely reflect) genetic relatedness. We tested this idea by applying univariate linear mixed models that include genetic relatedness as a cofactor. These results generally confirmed our RF inferences; for example, bio8 was still a highly significant ($P = 2.40 \times 10^{-13}$; Table S3) predictor after correcting for genetic relatedness.

## Forecasting future mutational load and adaptive variation

The ability to model load$_M$ in the present is important because it offers insights into the relative importance of factors that contribute to the accumulation of deleterious genetic variation. It also provides a foundation for projecting load$_M$ into the future. Such projections may seem ill-considered, given that mutational load is largely shaped by random genetic drift and hence likely to be inherently difficult (or even impossible) to predict with precision. Nonetheless, adaptive variants have been widely modeled as a product of the deterministic process of selection despite the fact that adaptive allele frequencies are often influenced by genetic drift (Fitzpatrick & Keller, 2015; Capblancq *et al.*, 2020b; Gain *et al.*, 2023). We predicted $\mathrm{load}_M$ into the future using data from 16 climate projections that represent different levels of global greenhouse gas emissions for the year 2100. These predictions led to consistently higher mean values but decreasing variance across samples (Fig. 4c). The lower variance may reflect that our predictive models do not fully incorporate the evolutionary processes (like drift) that likely contribute to variance in $\mathrm{load}_M$; additional work is necessary to build on this initial effort. Nonetheless, the projection of load is interesting because leading-edge populations are generally predicted to have decreased $\mathrm{load}_M$ in the future, while trailing edge and central populations have increased $\mathrm{load}_M$. These projections match the predicted species' dynamics in which leading-edge populations become more centralized, with the potential for commensurately larger population sizes, but trailing edge populations experience increasingly marginal habitat, fragmentation, and the potential for genetic erosion (Hampe & Petit, 2005; Aitken *et al.*, 2008; Hannah, 2022).

We also studied the relationship between current and predicted $\mathrm{load}_M$ to measure adaptation in the present ($N_G$) and the future (go). $N_G$ was negatively correlated with $\mathrm{load}_M$, as is expected when (for example) small $N_e$ leads to the accumulation of deleterious variants and less efficacious selection for adaptive variants (Willi *et al.*, 2006). The negative correlation between $N_G$ and $\mathrm{load}_M$ represents a potentially potent combination to fuel genetic erosion or extinction. Both $N_G$ and $\mathrm{load}_M$ were also strongly correlated with $d_{\mathrm{dispersal}}$, suggesting again that dispersal history contributes directly or indirectly to mutational load, as shown for deleterious variants in humans after serial bottlenecks (Henn *et al.*, 2016). The significantly positive relationship between *go* and $\mathrm{load}_M$ is further evidence of a strong relationship among adaptation, climate, and mutational load, because it implies that individuals that will be more vulnerable to climate change, as measured by *go*, may have lower fitness due to mutational load. The fate of edge populations is of particular interest in these relationships. Our observations of low $N_G$ in leading-edge populations contravene the argument that adaptive alleles are more likely to be found in leading-edge populations (Macdonald *et al.*, 2017) but are consistent with previous empirical studies suggesting fewer adaptive alleles in leading-edge populations (Willi *et al.*, 2018; Takou *et al.*, 2021). Interestingly, because southern populations harbor more adaptive alleles (i.e. higher $N_G$) for higher temperatures, they could serve as sources of adaptive variation as temperatures rise in northern regions if gene flow is sufficient.

## Lessons for *V. arizonica* and beyond

Superficially, the future looks bright for *V. arizonica*, because it is atypical among CWRs by having a potentially expanding habitat. By contrast, a recent SDM-based study found that *c.* 85% of 600 North American CWRs are either vulnerable or endangered due to shifting climates and shrinking niches (Khoury *et al.*, 2020). We also project only minor increases in $\mathrm{load}_M$ over time and many populations have low *go* values, suggesting that they are relatively well situated to adapt in the face of climate change. Moreover, the fact that $\mathrm{load}_M$ and $\mathrm{N}_G$ are correlated with at least one phenotype, resistance to PD (Fig. S13), suggests that our population genomic measures could have relevance for interpreting phenotypes and perhaps even fitness. All of these observations do not mean, however, that *V. arizonica* will thrive in the future. Much of the projected expansion of the niche is to the North, where leading-edge samples already exhibit higher $\mathrm{load}_M$ (Fig. 1) and lower adaptive complements ($N_G$, Fig. 5). If these same populations are the source for continued northward expansion, these extant effects are likely to be exacerbated by dispersal.

These arguments assume that *V. arizonica* can disperse at all. There is historical precedence to make this assumption because our LGM results suggest the species expanded northward since the LGM. But will it be able to do so in the future? Dispersion is difficult to predict across heterogeneous landscapes; it is an important topic that needs further study and modeling (Razgour *et al.*, 2019; Aguirre-Liguori *et al.*, 2021). *Vitis arizonica* has potential advantages and disadvantages with respect to dispersal. One advantage is that its berries are consumed by birds and mammals that can, in theory, disperse seed. Another is that it readily hybridizes with other *Vitis* species, which can fuel rapid adaptation in new habitats (Pease *et al.*, 2016; Morales-Cruz *et al.*, 2021). Note, however, that our work suggests the pace of shifts in the species' range has changed dramatically. As one example, we have estimated that the geographic centroid is estimated to have moved 736 km over the last *c.* 22 000 yr (*c.* 3.35 km every 100 yr), but it is predicted to move anywhere from 87 to 226 km by 2100 depending on the climate model. Based on this simple heuristic, movement in this century is predicted to be c. 26- to 67.5-fold faster than the average rate in the past, raising questions about whether dispersal and rates of adaptation can keep up with the pace. These results complement studies predicting that adaptation over the next century needs to greatly exceed historical rates of dispersal and adaptation (e.g. Quintero & Wiens, 2013; Cang *et al.*, 2016).

One key consideration for range expansion is that it may favor a shift to a selfing mating system in edge populations, which likely bolsters reproductive success at low population densities (Koski *et al.*, 2019). This implies that species with flexible mating systems are better equipped to expand their range. *Vitis* species possess a highly conserved, dioecious sex determination system. While a simple recombination event can produce hermaphrodites, hermaphrodites are not observed in nature, suggesting that they are strongly selected against (Massonnet *et al.*, 2020). Will the constraints of this dioecious mating system limit the potential for range expansion for *V. arizonica* and, indeed, for other obligately outcrossing plants? The answer is not yet clear. Simulations of edge populations have shown that, under a wide range of conditions, range expansion promotes the evolution of selfing in marginal populations (Encinas-Viso *et al.*, 2020). However, an empirical study of *Arabidopsis lyrata* found no effects on the self-incompatibility system in edge populations, despite accumulated load in these populations (Takou *et al.*, 2021). One useful extension of our RF approaches may be to add genomic measures related to selfing rates – such as the fraction of the genome encompassed in runs of homozygosity – as a predictor when studying species like *A. lyrata* that exhibit variation in selfing rates across populations (Mable *et al*., 2005; Perrier *et al*., 2022). Further studies that compare empirical results across mating systems and dispersal histories are likely to help establish general trends.

It remains a substantial challenge to project species persistence. One heuristic is the FOLDs model (Aguirre-Liguori *et al*., 2021), which layers types of information to find populations that seem particularly well (or poorly) suited to persist given climate projections. For *V. arizonica*, the layers of information amassed in this study include SDM-based geography, mutational load, adaptive variants, and even phenotypes, but the layers are sometimes conflicting. For example, the location of leading-edge Northern samples is projected to become more centrally located within the species' niche as it shifts northward, potentially positioning them reasonably well to face climate change. However, these samples also exhibit higher mutational load, fewer adaptive variants, and reduced resistance to a key pathogen, suggesting that they may be poorly equipped for successful dispersal and persistence under future conditions. Additionally, while our analyses have largely identified interesting features in individuals at the leading edge of expansion, one must also be concerned about trailing edges, where a suitable niche is disappearing. In *V. arizonica*, the individuals from trailing locations seem unremarkable with respect to genomic measures, but their habitat is likely to become increasingly fragmented. For these populations, the key for persistence also likely resides in their ability (or not) to disperse northward (Bell, 2017) and the potential for evolutionary rescue via gene flow.

## Acknowledgements

We thank Edwin Solares for supporting the Gaut lab's high-performance computing infrastructure, Rebecca Gaut for assistance with the initial generation of sequencing data, and Kirk Lohmueller for helpful discussions on genetic load. BG is supported by NSF grants IOS-2323123 and IOS-2414478.

## Competing interests

None declared.

## Author contributions

CJF contributed to the conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, writing – original draft preparation, and writing – review and editing. JAA-L contributed to the conceptualization, formal analysis, investigation, methodology, software, visualization, writing – review and editing. GRJG contributed to the methodology, software, and writing – review and editing. BSG contributed to the conceptualization, data curation, funding acquisition, methodology, project administration, resources, supervision, writing – original draft preparation, and writing – review and editing.

## ORCID

Jonás A. Aguirre-Liguori https://orcid.org/0000-0003-1763-044X
Christopher J. Fiscus https://orcid.org/0000-0001-9569-1809
Brandon S. Gaut https://orcid.org/0000-0002-1334-5556
Garren R. J. Gaut https://orcid.org/0000-0001-5661-6478

## Data availability

All sequencing data referenced in this work are available from NCBI: BioProjects PRJNA731597 (Morales-Cruz *et al*., 2021) and PRJNA842753 (Morales-Cruz *et al*., 2023). The phenotyping data were also previously published (Heinitz *et al*., 2020; Morales-Cruz *et al*., 2023). All other data needed to reproduce the analyses presented here are present within the manuscript and Supporting Information.

## References

**Aguirre-Liguori JA, Morales-Cruz A, Gaut BS. 2022.** Evaluating the persistence and utility of five wild *Vitis* species in the context of climate change. *Molecular Ecology* **31**: 6457–6472.

**Aguirre-Liguori JA, Ramírez-Barahona S, Gaut BS. 2021.** The evolutionary genomics of species' responses to climate change. *Nature Ecology & Evolution* **5**: 1350–1360.

**Aguirre-Liguori JA, Tenaillon MI, Vázquez-Lobo A, Gaut BS, Jaramillo-Correa JP, Montes-Hernandez S, Souza V, Eguiarte LE. 2017.** Connecting genomic patterns of local adaptation and niche suitability in teosintes. *Molecular Ecology* **26**: 4226–4240.

**Aitken SN, Yeaman S, Holliday JA, Wang T, Curtis-McLane S. 2008.** Adaptation, migration or extirpation: climate change outcomes for tree populations. *Evolutionary Applications* **1**: 95–111.

**Angert AL, Bontrager MG, Ågren J. 2020.** What do we really know about adaptation at range edges? *Annual Review of Ecology, Evolution, and Systematics* **51**: 341–361.

**Bartoń K. 2024.** *MuMIn: multi-model inference.* R package, v.1.48.4. [WWW document] URL https://cran.r-project.org/package=MuMIn.

**Bay RA, Harrigan RJ, Underwood VL, Gibbs HL, Smith TB, Ruegg K. 2018.** Genomic signals of selection predict climate-driven population declines in a migratory bird. *Science* **359**: 83–86.

**Bell G. 2017.** Evolutionary rescue. *Annual Review of Ecology, Evolution, and Systematics* **48**: 605–627.

**Benestan LM, Ferchaud A-L, Hohenlohe PA, Garner BA, Naylor GJP, Baums IB, Schwartz MK, Kelley JL, Luikart G. 2016.** Conservation genomics of natural and managed populations: building a conceptual and practical framework. *Molecular Ecology* **25**: 2967–2977.

**Bertorelle G, Raffini F, Bosse M, Bortoluzzi C, Iannucci A, Trucchi E, Morales HE, van Oosterhout C. 2022.** Genetic load: genomic estimates and applications in non-model animals. *Nature Reviews. Genetics* **23**: 492–503.

**Bohra A, Kilian B, Sivasankar S, Caccamo M, Mba C, McCouch SR, Varshney RK. 2022.** Reap the crop wild relatives for breeding future crops. *Trends in Biotechnology* **40**(4): 412–431.

**Bolger AM, Lohse M, Usadel B. 2014.** Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

**Booker TR, Yeaman S, Whitlock MC. 2020.** Variation in recombination rate affects detection of outliers in genome scans under neutrality. *Molecular Ecology* **29**: 4274–4279.

**Bortoluzzi C, Bosse M, Derks MFL, Crooijmans RPMA, Groenen MAM, Megens H-J. 2020.** The type of bottleneck matters: insights into the deleterious variation landscape of small managed populations. *Evolutionary Applications* **13**: 330–341.

**Bosshard L, Dupanloup I, Tenaillon O, Bruggmann R, Ackermann M, Peischl S, Excoffier L. 2017.** Accumulation of deleterious mutations during bacterial range expansions. *Genetics* **207**: 669–684.

**Brandvain Y, Wright SI. 2016.** The limits of natural selection in a nonequilibrium world. *Trends in Genetics* **32**: 201–210.

**Breiman L. 2001.** Random forests. *Machine Learning* 45: 5–32.

**Cang FA, Wilson AA, Wiens JJ. 2016.** Climate change is projected to outpace rates of niche change in grasses. *Biology Letters* 12: 20160368.

**Capblancq T, Fitzpatrick MC, Bay RA, Exposito-Alonso M, Keller SR. 2020a.** Genomic prediction of (Mal)adaptation across current and future climatic landscapes. *Annual Review of Ecology, Evolution, and Systematics* 51: 245–269.

**Capblancq T, Forester BR. 2021.** Redundancy analysis: a Swiss army knife for landscape genomics. *Methods in Ecology and Evolution* 12: 2298–2309.

**Capblancq T, Morin X, Gueguen M, Renaud J, Lobreaux S, Bazin E. 2020b.** Climate-associated genetic variation in *Fagus sylvatica* and potential responses to climate change in the French Alps. *Journal of Evolutionary Biology* 33: 783–796.

**Caye K, Jumentier B, Lepeule J, François O. 2019.** LFMM 2: fast and accurate inference of gene-environment associations in genome-wide studies. *Molecular Biology and Evolution* 36: 852–860.

**Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015.** Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 4: 7.

**Charlesworth B, Charlesworth D. 1998.** Some evolutionary consequences of deleterious mutations. *Genetica* 102: 3–19.

**Chen D, Shi R, Pape J-M, Neumann K, Arend D, Graner A, Chen M, Klukas C. 2018.** Predicting plant biomass accumulation from image-derived parameters. *GigaScience* 7: 1–13.

**Cisternas-Fuentes A, Koski MH. 2023.** Drivers of strong isolation and small effective population size at a leading range edge of a widespread plant. *Heredity* 130: 347–357.

**Collart F, Hedenäs L, Broennimann O, Guisan A, Vanderpoorten A. 2020.** Intraspecific differentiation: implications for niche and distribution modelling. *Journal of Biogeography* 48(2): 415–426.

**Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. 2014.** Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nature Genetics* 46: 1311–1320.

**Crow JF. 1958.** Some possibilities for measuring selection intensities in man. *Human Biology* 30: 1–13.

**Crow JF, Kimura M. 1970.** *An introduction to population genetics theory.* New York, NY, USA: Harper and Row.

**Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM *et al.* 2021.** Twelve years of SAMtools and BCFtools. *GigaScience* 10: 248.

**Eckert CG, Samis KE, Lougheed SC. 2008.** Genetic variation across species' geographical ranges: the central-marginal hypothesis and beyond. *Molecular Ecology* 17: 1170–1188.

**Ellis N, Smith SJ, Pitcher CR. 2012.** Gradient forests: calculating importance gradients on physical predictors. *Ecology* 93: 156–168.

**Encinas-Viso F, Young AG, Pannell JR. 2020.** The loss of self-incompatibility in a range expansion. *Journal of Evolutionary Biology* 33: 1235–1244.

**Excoffier L, Foll M, Petit RJ. 2009.** Genetic consequences of range expansions. *Annual Review of Ecology, Evolution, and Systematics* 40: 481–501.

**Exposito-Alonso M, 500 Genomes Field Experiment Team, Burbano HA, Bossdorf O, Nielsen R, Weigel D. 2019.** Natural selection on the *Arabidopsis thaliana* genome in present and future climates. *Nature* 573: 126–129.

**Eyre-Walker A, Keightley PD. 2007.** The distribution of fitness effects of new mutations. *Nature Reviews. Genetics* 8: 610–618.

**Eyring V, Bony S, Meehl GA, Senior CA, Stevens B, Stouffer RJ, Taylor KE. 2016.** Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development* 9: 1937–1958.

**Femerling G, van Oosterhout C, Feng S, Bristol RM, Zhang G, Groombridge J, Gilbert MT, Morales HE. 2023.** Genetic load and adaptive potential of a recovered avian species that narrowly avoided extinction. *Molecular Biology and Evolution* 40: 1697.

**Fick SE, Hijmans RJ. 2017.** WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology* 37: 4302–4315.

**Fitzpatrick MC, Keller SR. 2015.** Ecological genomics meets community-level modelling of biodiversity: mapping the genomic landscape of current and future environmental adaptation. *Ecology Letters* 18: 1–16.

**Fouqueau L, Reynes L, Tempera F, Bajjouk T, Blanfuné A, Chevalier C, Laurans M, Mauger S, Sourisseau M, Assis J *et al.* 2024.** Seascape genetic study on *Laminaria digitata* underscores the critical role of sampling schemes. *Marine Ecology Progress Series* 740: 23–42.

**Frichot E, François O. 2015.** LEA: an R package for landscape and ecological association studies. *Methods in Ecology and Evolution* 6: 925–929.

**Gain C, Rhoné B, Cubry P, Salazar I, Forbes F, Vigouroux Y, Jay F, François O. 2023.** A quantitative theory for genomic offset statistics. *Molecular Biology and Evolution* 40: msad140.

**Garzón MB, Robson TM, Hampe A. 2019.** ΔTraitSDMs: species distribution models that account for local adaptation and phenotypic plasticity. *New Phytologist* 222: 1757–1765.

**Ghanipoor-Samami M, Javadmanesh A, Burns BM, Thomsen DA, Nattrass GS, Estrella CAS, Kind KL, Hiendleder S. 2018.** Atlas of tissue- and developmental stage specific gene expression for the bovine insulin-like growth factor (IGF) system. *PLoS ONE* 13: e0200466.

**González-Martínez SC, Ridout K, Pannell JR. 2017.** Range expansion compromises adaptive evolution in an outcrossing plant. *Current Biology* 27: 2544–2551.

**Grossen C, Guillaume F, Keller LF, Croll D. 2020.** Purging of highly deleterious mutations through severe bottlenecks in Alpine ibex. *Nature Communications* 11: 1001.

**Hampe A, Petit RJ. 2005.** Conserving biodiversity under climate change: the rear edge matters: rear edges and climate change. *Ecology Letters* 8: 461–467.

**Hannah L. 2022.** Ecological, evolutionary, and biogeographic implications of climate change. In: *Climate change biology.* London, UK: Elsevier, 77–94.

**Hedrick PW, Garcia-Dorado A. 2016.** Understanding inbreeding depression, purging, and genetic rescue. *Trends in Ecology & Evolution* 31: 940–952.

**Hedrick PW, Kalinowski ST. 2000.** Inbreeding depression in conservation biology. *Annual Review of Ecology, Evolution, and Systematics* 31: 139–162.

**Heinitz CC, Riaz S, Tenscher AC, Romero N, Walker MA. 2020.** Survey of chloride exclusion in grape germplasm from the southwestern United States and Mexico. *Crop Science* 60: 1946–1956.

**Heinitz CC, Uretsky J, Dodson Peterson JC, Huerta-Acosta KG, Walker MA. 2019.** Crop wild relatives of grape (*Vitis vinifera* L.) throughout North America. In: Greene SL, Williams KA, Khoury CK, Kantar MB, Marek LF, eds. *North American crop wild relatives, Vol. 2: important species.* Cham, Switzerland: Springer International Publishing, 329–351.

**Henn BM, Botigué LR, Peischl S, Dupanloup I, Lipatov M, Maples BK, Martin AR, Musharoff S, Cann H, Snyder MP *et al.* 2016.** Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *Proceedings of the National Academy of Sciences, USA* 113: E440–E449.

**Hijmans RJ. 2022.** *geosphere: spherical trigonometry.* R package, v.1.5-18. [WWW document] URL https://cran.r-project.org/package=geosphere.

**Hijmans RJ. 2023.** *raster: geographic data analysis and modeling.* R package, v.3.6-26. [WWW document] URL https://cran.r-project.org/package=raster.

**Hijmans RJ. 2024.** *terra: spatial data analysis.* R package, v.1.7-55. [WWW document] URL https://cran.r-project.org/package=terra.

**Johnson JW. 2000.** A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behavioral Research* 35: 1–19.

**Khoury CK, Carver D, Greene SL, Williams KA, Achicanoy HA, Schori M, León B, Wiersema JH, Frances A. 2020.** Crop wild relatives of the United States require urgent conservation action. *Proceedings of the National Academy of Sciences, USA* 117: 33351–33357.

**Kim BY, Huber CD, Lohmueller KE. 2018.** Deleterious variation shapes the genomic landscape of introgression. *PLoS Genetics* 14: e1007741.

**Koski MH, Layman NC, Prior CJ, Busch JW, Galloway LF. 2019.** Selfing ability and drift load evolve with range expansion. *Evolution Letters* 3: 500–512.

**Kryukov GV, Pennacchio LA, Sunyaev SR. 2007.** Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *American Journal of Human Genetics* 80: 727–739.

**Kuhn M, Wickham H. 2020.** *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.* R package, v.1.1.1. [WWWdocument] URL https://www.tidymodels.org.

**Kyriazis CC, Wayne RK, Lohmueller KE. 2021.** Strongly deleterious mutations are a primary determinant of extinction risk due to inbreeding depression. *Evolution Letters* 5: 33–47.

**Li H. 2013.** Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*. doi: 10.48550/arXiv.1303.3997.

**Li X, Wang M, Zhang R, Fang H, Fu X, Yang X, Li J. 2022.** Genetic architecture of embryo size and related traits in maize. *The Crop Journal* **10**: 204–215.

**Lira-Noriega A, Manthey JD. 2014.** Relationship of genetic diversity and niche centrality: a survey and analysis. *Evolution; International Journal of Organic Evolution* **68**: 1082–1093.

**Liu Q, Zhou Y, Morrell PL, Gaut BS. 2017.** Deleterious variants in Asian rice and the potential cost of domestication. *Molecular Biology and Evolution* **34**: 908–924.

**Lohmueller KE. 2014.** The distribution of deleterious genetic variation in human populations. *Current Opinion in Genetics & Development* **29**: 139–146.

**Mable BK, Robertson AV, Dart S, Di Berardo C, Witham L. 2005.** Breakdown of self-incompatibility in the perennial *Arabidopsis lyrata* (Brassicaceae) and its genetic consequences. *Evolution; International Journal of Organic Evolution* **59**: 1437–1448.

**Macdonald SL, Llewelyn J, Moritz C, Phillips BL. 2017.** Peripheral isolates as sources of adaptive diversity under climate change. *Frontiers in Ecology and Evolution* **5**: 1928.

**Massonnet M, Cochetel N, Minio A, Vondras AM, Lin J, Muyle A, Garcia JF, Zhou Y, Delledonne M, Riaz S *et al.* 2020.** The genetic basis of sex determination in grapes. *Nature Communications* **11**: 789.

**Mezmouk S, Ross-Ibarra J. 2014.** The pattern and distribution of deleterious mutations in maize. *G3* **4**: 163–171.

**Morales-Cruz A, Aguirre-Liguori J, Massonnet M, Minio A, Zaccheo M, Cochetel N, Walker A, Riaz S, Zhou Y, Cantu D *et al.* 2023.** Multigenic resistance to *Xylella fastidiosa* in wild grapes (*Vitis* sps.) and its implications within a changing climate. *Communications Biology* **6**: 580.

**Morales-Cruz A, Aguirre-Liguori JA, Zhou Y, Minio A, Riaz S, Walker AM, Cantu D, Gaut BS. 2021.** Introgression among North American wild grapes (*Vitis*) fuels biotic and abiotic adaptation. *Genome Biology* **22**: 254.

**Moyers BT, Morrell PL, McKay JK. 2018.** Genetic costs of domestication and improvement. *The Journal of Heredity* **109**(2): 103–116.

**Occdownload Gbif.Org. 2020.** Occurrence download.

**Olson DM, Dinerstein E, Wikramanayake ED, Burgess ND, Powell GVN, Underwood EC, D'amico JA, Itoua I, Strand HE, Morrison JC *et al.* 2001.** Terrestrial ecoregions of the world: a new map of life on earth: a new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *Bioscience* **51**: 933–938.

**Palumbi SR, Barshis DJ, Traylor-Knowles N, Bay RA. 2014.** Mechanisms of reef coral resistance to future climate change. *Science* **344**: 895–898.

**Pease JB, Haak DC, Hahn MW, Moyle LC. 2016.** Phylogenomics reveals three sources of adaptive variation during a rapid radiation. *PLoS Biology* **14**: e1002379.

**Perrier A, Sánchez-Castro D, Willi Y. 2022.** Environment dependence of the expression of mutational load and species' range limits. *Journal of Evolutionary Biology* **35**: 731–741.

**Phillips SB, Aneja VP, Kang D, Arya SP. 2006.** Modelling and analysis of the atmospheric nitrogen deposition in North Carolina. *International Journal of Global Environmental Issues* **6**: 231–252.

**Phillips SJ, Dudík M. 2008.** Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography* **31**: 161–175.

**Polechová J. 2025.** Evolution of species' range and niche in changing environments. *bioRxiv*. doi: 10.1011/2025.01.16.633367.

**Quintero I, Wiens JJ. 2013.** Rates of projected climate change dramatically exceed past rates of climatic niche evolution among vertebrate species. *Ecology Letters* **16**: 1095–1103.

**Quinton A. 2019.** *UC Davis releases 5 grape varieties resistant to Pierce's disease.* [WWW document] URL https://www.ucdavis.edu/food/news/uc-davis-releases-five-new-wine-grape-varieties [accessed 1 November 2024].

**R Core Team. 2023.** *R: a language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

**Razgour O, Forester B, Taggart JB, Bekaert M, Juste J, Ibáñez C, Puechmaille SJ, Novella-Fernandez R, Alberdi A, Manel S. 2019.** Considering adaptive genetic variation in climate change vulnerability assessment reduces species range loss projections. *Proceedings of the National Academy of Sciences, USA* **116**: 10418–10423.

**Renaut S, Rieseberg LH. 2015.** The accumulation of deleterious mutations as a consequence of domestication and improvement in sunflowers and other compositae crops. *Molecular Biology and Evolution* **32**: 2273–2283.

**Rhoné B, Defrance D, Berthouly-Salazar C, Mariac C, Cubry P, Couderc M, Dequincey A, Assoumanne A, Kane NA, Sultan B *et al.* 2020.** Pearl millet genomic vulnerability to climate change in West Africa highlights the need for regional collaboration. *Nature Communications* **11**: 5274.

**Riahi K, van Vuuren DP, Kriegler E, Edmonds J, O'Neill BC, Fujimori S, Bauer N, Calvin K, Dellink R, Fricko O *et al.* 2017.** The shared socioeconomic pathways and their energy, land use, and greenhouse gas emissions implications: an overview. *Global Environmental Change: Human and Policy Dimensions* **42**: 153–168.

**Riaz S, Huerta-Acosta K, Tenscher AC, Walker MA. 2018.** Genetic characterization of *Vitis* germplasm collected from the southwestern US and Mexico to expedite Pierce's disease-resistance breeding. *Theoretical and Applied Genetics* **131**: 1589–1602.

**Riaz S, Tenscher AC, Heinitz CC, Huerta-Acosta KG, Walker MA. 2020.** Genetic analysis reveals an east-west divide within North American Vitis species that mirrors their resistance to Pierce's disease. *PLoS ONE* **15**: e0243445.

**Robinson J, Kyriazis CC, Yuan SC, Lohmueller KE. 2023.** Deleterious variation in natural populations and implications for conservation genetics. *Annual Review of Animal Biosciences* **11**: 93–114.

**Rougemont Q, Moore J-S, Leroy T, Normandeau E, Rondeau EB, Withler RE, Van Doornik DM, Crane PA, Naish KA, Garza JC *et al.* 2020.** Demographic history shaped geographical patterns of deleterious mutation load in a broadly distributed Pacific Salmon. *PLoS Genetics* **16**: e1008348.

**Ruegg K, Bay RA, Anderson EC, Saracco JF, Harrigan RJ, Whitfield M, Paxton EH, Smith TB. 2018.** Ecological genomics predicts climate vulnerability in an endangered southwestern songbird. *Ecology Letters* **21**: 1085–1096.

**Sánchez-Castro D, Perrier A, Willi Y. 2022.** Reduced climate adaptation at range edges in North American *Arabidopsis lyrata*. *Global Ecology and Biogeography* **31**: 1066–1077.

**Sexton JP, McIntyre PJ, Angert AL, Rice KJ. 2009.** Evolution and ecology of species range limits. *Annual Review of Ecology, Evolution, and Systematics* **40**: 415–436.

**Shen Z, Chen A. 2020.** Comprehensive relative importance analysis and its applications to high dimensional gene expression data analysis. *Knowledge-Based Systems* **203**: 106120.

**Simons YB, Turchin MC, Pritchard JK, Sella G. 2014.** The deleterious mutation load is insensitive to recent population history. *Nature Genetics* **46**: 220–224.

**Takou M, Hämälä T, Koch EM, Steige KA, Dittberner H, Yant L, Genete M, Sunyaev S, Castric V, Vekemans X *et al.* 2021.** Maintenance of adaptive dynamics and no detectable load in a range-edge outcrossing plant population. *Molecular Biology and Evolution* **38**: 1820–1836.

**Therneau TM. 2022.** *Mixed effects cox models (R package COXME v.2.2-18.1).* Khomas, NA, USA: Comprehensive R Archive Network (CRAN).

**Thuiller W, Lafourcade B, Engler R, Araújo MB. 2009.** BIOMOD – a platform for ensemble forecasting of species distributions. *Ecography* **32**: 369–373.

**Travis JMJ, Münkemüller T, Burton OJ, Best A, Dytham C, Johst K. 2007.** Deleterious mutations can surf to high densities on the wave front of an expanding population. *Molecular Biology and Evolution* **24**: 2334–2343.

**Van der Auwera GA, O'Connor BD. 2020.** *Genomics in the Cloud: using Docker, GATK, and WDL in Terra.* Sebastopol, CA, USA: O'Reilly Media.

**Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. 2016.** SIFT missense predictions for genomes. *Nature Protocols* **11**: 1–9.

**Venables WN, Ripley BD. 2002.** *Modern applied statistics with S.* New York, NY, USA: Springer.

**Vucetich JA, Waite TA. 2003.** Spatial patterns of demography and genetic processes across the species' range: null hypotheses for landscape conservation genetics. *Conservation Genetics* **4**: 639–645.

**Waldvogel A-M, Feldmeyer B, Rolshausen G, Exposito-Alonso M, Rellstab C, Kofler R, Mock T, Schmid K, Schmitt I, Bataillon T *et al.* 2020.** Evolutionary genomics can improve prediction of species' responses to climate change. *Evolution Letters* **4**: 4–18.

**Weiss-Lehman C, Hufbauer RA, Melbourne BA. 2017.** Rapid trait evolution drives increased speed and variance in experimental range expansions. *Nature Communications* **8**: 14303.

**Willi Y, Fracassetti M, Zoller S, Van Buskirk J. 2018.** Accumulation of mutational load at the edges of a species range. *Molecular Biology and Evolution* **35**: 781–791.

**Willi Y, Lucek K, Bachmann O, Walden N. 2022.** Recent speciation associated with range expansion and a shift to self-fertilization in North American Arabidopsis. *Nature Communications* **13**: 7564.

**Willi Y, Van Buskirk J, Hoffmann AA. 2006.** Limits to the adaptive potential of small populations. *Annual Review of Ecology, Evolution, and Systematics* **37**: 433–458.

**Wright MN, Ziegler A. 2017.** RANGER: a fast implementation of random forests for high dimensional data in C++ and R. *Journal of statistical software* **77**: 1–17.

**Xiao H, Liu Z, Wang N, Long Q, Cao S, Huang G, Liu W, Peng Y, Riaz S, Walker AM *et al.* 2023.** Adaptive and maladaptive introgression in grapevine domestication. *Proceedings of the National Academy of Sciences, USA* **120**: e2222041120.

**Zhang X, Kim B, Lohmueller KE, Huerta-Sánchez E. 2020.** The impact of recessive deleterious variation on signals of adaptive introgression in human populations. *Genetics* **215**: 799–812.

**Zhou X, Stephens M. 2012.** Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* **44**: 821–824.

**Zizka A, Silvestro D, Andermann T, Azevedo J, Duarte Ritter C, Edler D, Farooq H, Herdean A, Ariza M, Scharn R *et al.* 2019.** COORDINATECLEANER: standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution* **10**: 744–751.

## Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

**Fig. S1** Principal component analysis of SNP dataset.

**Fig. S2** Distributions of all features.

**Fig. S3** Hyperparameter tuning for random forest regression model to predict load$_M$.

**Fig. S4** Correlations between different sets of adaptive genotypes.

**Fig. S5** Proportion of nonsynonymous alleles on the landscape.

**Fig. S6** Proportion of synonymous alleles on the landscape.

**Fig. S7** Relationship between load$_M$ calculated with nonsynonymous compared to deleterious SNPs.

**Fig. S8** Pairwise Spearman's correlations for all variables.

**Fig. S9** Random forest regression models to predict $P_d$, $P_n$, and load$_{M(del)}$.

**Fig. S10** Current vs predicted species distribution models in 2100 under SSP245, SSP370, and SSP585 for four Earth Systems Models.

**Fig. S11** The present distance to geographic centroid is correlated with the mean predicted change in load$_M$ by 2100.

**Fig. S12** Spearman's correlations between $N_G$ and other variables.

**Fig. S13** Load$_M$ and $N_G$ were associated with *Xylella fastidiosa* concentration in a glasshouse experiment.

**Table S1** *Vitis arizonica* individuals with sampling coordinates, genetic metrics, and geographic distances.

**Table S2** WorldClim bioclimatic variables.

**Table S3** Results from linear mixed model used to predict load$_M$.

**Table S4** Predicted loadM in 2100 by climate model for each individual.

**Table S5** *P*-values for SNPs significantly associated with climate.

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

Disclaimer: The New Phytologist Foundation remains neutral with regard to jurisdictional claims in maps and in any institutional affiliations.