Understanding population history of range expansion and environmental adaptation by visualizing gene–environment–trait associations

Gene–environment–trait associations

Reiichiro Nakamichi[1*], Shuichi Kitada[2], and Hirohisa Kishino[3*]

[1] Japan Fisheries Research and Education Agency, Yokohama 236-8648, Japan

[2] Tokyo University of Marine Science and Technology, Tokyo 108-8477, Japan

[3] Graduate School of Agriculture and Life Sciences, The University of Tokyo, Tokyo 113-8657, Japan

[*]Corresponding authors: Reiichiro Nakamichi, nakamichi@affrc.go.jp

Hirohisa Kishino, kishino@g.ecc.u-tokyo.ac.jp

Abstract

A population adapts to various environmental stresses by changing its geographical distribution or genomes. Therefore, population genomic data provide insight into adaptation in the form of genetic population structure and association of SNP allele frequencies with measured environmental variables. In this paper, we propose an exploratory approach that captures the essential events throughout a population's history. The input data consisted of SNP allele frequencies, environmental variables, and average trait values for each local subpopulation. The core component of our approach is correspondence analysis that maps the subpopulations in the cloud of SNPs, environments, and traits; their relative locations represent the population structure. Population-specific F_{ST} values enable inference of the history of geographical distributions. By highlighting significant associations, we obtained insight into the environmental stresses that the population encountered, traits that were affected by the stresses, and genes that enabled the population to adapt to the stresses. The potential usefulness of our exploratory approach was tested by a detailed analysis of simulations and a public dataset of wild poplar in northwestern America. The wild poplar analysis depicted the history of range expansion from inland to coastal areas and northern extension and identified the genetic background underlying adaptation to the local environments. The R codes for our representation method and simulations of population colonization used in this study are available in the Supporting Information.

Keywords

adaptation, correspondence analysis, gene–environment–trait associations, history of range expansion, population structure

1 | INTRODUCTION

Individuals adapt to new environments through selection on pre-existing alleles and/or new mutations in adaptation-related genes (Barrett & Schluter, 2008). Therefore, adaptation of populations of a species to novel environments changes allele frequencies of loci under selection. Standing genetic variation can shape the basis of evolutionary responses to environmental changes (Capblancq et al. 2020). Environmental adaptation processes can also create significant differences in phenotypes and traits among populations of a species. For example, populations of sockeye salmon exhibit diversity with regard to life history traits such as spawning time and habitat, adult body size and shape, rearing time in freshwater and seawater, and adaptation to local spawning and rearing habitats within complex lake systems (Hilborn, Quinn, Schindler, & Rogers, 2003). Moreover, populations of walking stick insects diverged in body size, shape, host preference, and behavior in parallel with the divergence of their host plant species (Nosil, Crespi, & Sandoval, 2002). When correlated with variation in environmental factors over local populations, such variation in traits and phenotypes can provide insight into natural selection processes (Coop, Witonsky, Di Rienzo, & Pritchard, 2010).

Adaptation to environmental factors can change traits and phenotypes of a species, thereby creating population structure underpinned by functional loci. Geographical isolation, which can lead to reproductive isolation and consequent differences in allele frequencies of neutral loci, also contributes to population structuring (Wright, 1965). Divergent selection in an environmental gradient may affect genome-wide population structure (Nosil, Funk, & Ortiz-Barrientos, 2009; Orsini, Vanoverbeke, Swillen, Mergeay, & De Meester, 2013). Empirical studies showed that aridity gradients caused geographically structured populations of Poaceae characterized by cytotype segregation of diploids and allotetraploids (Manzaneda et al., 2012), and geographic distance and habitat differences between subpopulations impacted population structure (Bradbury & Bentzen, 2007; Jorde et al., 2015; Kitada, Nakamichi, & Kishino, 2017).

Therefore, population structure needs to be considered when analyzing correlations among genes, traits, and environmental factors across population samples taken from a wide range of geographical regions.

Genome-wide association studies (GWASs) are widely used to identify associations between genes and traits/environments (Visscher et al., 2017). When data are obtained from a metapopulation exhibiting population structure, the effect of genotypes can be inferred by eliminating population structure effects (Devlin & Roeder, 1999) to avoid spurious associations (Pritchard & Rosenberg, 1999). One representative software program, TASSEL (Yu et al., 2006; Bradbury et al., 2007), performs this type of analysis using a unified mixed model. Alternatively, a structured population can be decomposed into Hardy–Weinberg subpopulations, and the associations tested for each subpopulation (Pritchard, Stephens, Rosenberg, & Donnelly, 2000). Future challenges for large-scale GWASs from wild populations (wild GWASs) include development of methods that take population structure into account (Santure & Garant, 2018).

Genotype-environment associations (GEAs) analyze the allele frequencies of SNPs in sampling locations and test their associations with the environmental variables (Capblancq et al. 2020). So-called "genome scan methods" consider geographically structured populations and detect SNPs related to environmental variables, traits, and phenotypes (De Mita et al., 2013; De Villemereuil, Frichot, Bazin, François, & Gaggiotti, 2014). For example, BayeScan (Foll & Gaggiotti, 2008) detects SNPs that create major differentiation in terms of locus-specific global F_{ST} values over a metapopulation. Additionally, Bayenv (Coop, Witonsky, Di Rienzo, & Pritchard, 2010) and the latent factor mixed model (Frichot, Schoville, Bouchard, & François, 2013) can detect SNPs that are highly correlated with environmental factors and traits on the basis of allele frequencies. Notably these methods essentially do not require phenotypic data. Hence, they are valid especially when the life history is complex and cannot be appropriately measured by a few trait variables or the environmental selection on the phenotypes are not

characterized (Capblancq et al. 2020). During the evolutionary history of range expansion, the frequencies of existing and derived alleles in a population vary stochastically, and various pressures of environmental selection affect the allele frequencies of related genes and phenotypes. To fully understand underlying mechanisms of population structuring and local adaptation, an integrated understanding of population history and structure that incorporates relationships between SNPs, the environment, and phenotypes could be useful.

Here, we propose a method that captures the whole history of range expansion and environmental adaptation of a species by visualizing the data of gene–environment–trait interactions. Our proposed method uses the latent factor mixed model framework. The R codes for our representation method and simulations of population colonization used in this study are available in the Supporting Information. This approach also accepts microsatellite data and SNP genotype data, and reads Genepop format (Raymond & Rousset, 1995; Rousset, 2008).

2 | MATERIALS AND METHODS

2.1 | Colored correspondence analysis: history of range expansion and environmental stress

On the basis of the values of the environmental factors, the mean values of the traits, and the allele frequencies at SNPs in each subpopulation, correspondence analysis (Benzécri, 1973; Hayashi, 1953) generates a biplot that visualizes the correspondence between the subpopulations and the variables of traits, environments, and genes (SNPs) (see Figure 3 and Figure 6 in RESULTS section). For the allele frequencies at SNPs, we used the frequencies of derived alleles to infer signatures of environmental adaptation. However, it is difficult to know which allele is derived in the actual data. Therefore, if we could not identify derived alleles, we used the frequencies of minor alleles instead. This simple assignment for derived alleles may have errors, but we hoped that it would capture the SNPs that enhanced the allele frequencies

to adapt to the local environments. To distinguish the association by positive and negative correlations, we introduced two types of environmental variables: the original environmental value itself, and the sign-reversed value of the original value. Genes and traits that had a positive/negative correlation with the original environmental factors were connected to the original/sign-reversed environmental variables. To understand the subpopulations in the context of the evolutionary change in the distributional range, we assigned a gradient of colors to the subpopulations. The colors represent population-specific F_{ST} (Weir & Goudet, 2017). Population-specific F_{ST} estimates the genetic deviation from the ancestral population on the basis of the difference between the heterozygosity of the entire population and the heterozygosity of each subpopulation. We extended the population-specific F_{ST} estimator to overall loci as

$$\mathrm{ps}\hat{F}_{\mathrm{ST}}^{i} = \frac{\sum_{l=1}^{L}(\widetilde{M}_{W,l}^{i} - \widetilde{M}_{l}^{B})}{\sum_{l=1}^{L}\left(1 - \widetilde{M}_{l}^{B}\right)},$$

where $\widetilde{M}_{W,l}^{i}$ is the unbiased within-population matching of two distinct alleles of locus l ($l = 1 \sim L$) in subpopulation i ($i = 1 \sim K$), and $\widetilde{M}_{l}^{B}$ is the between-population-pair matching average over pairs of subpopulations (Buckleton et al., 2016).

2.2 | Gene–environment–trait association

The environment–gene and environment–trait pairs that have significant correlations are highlighted by connection lines in the biplot (see Figure 3 and Figure 6 in RESULTS section). Regarding the significance of the correlation between an environment and a gene (SNP), we accounted for the correlation structure of the residuals. At each locus, the variance matrix of the observed allele frequencies reflects the genetic drift and gene flow and the sampling variance (Nicholson, Smith, Jónsson, Gústafsson, & Stefánsson, 2002; Coop, Witonsky, Di Rienzo, & Pritchard, 2010). Here, we adopted the frequentist approach to choose significant pairs given a

value of false discovery rate (FDR) (1% for the simulation and 5% for the poplar data).

We consider K subpopulations derived from a common ancestral population and L loci of biallelic neutral markers. Let p_i^l and p_A^l be the derived allele frequency of marker l ($l = 1 - L$) in subpopulation i ($i = 1 - K$) and the (unobserved) ancestral population. Given the samples from the subpopulations, the allele frequencies are estimated by the observed counts as $\hat{p}_i^l = n_i^l / n_i$, where n_i is the number of the samples.

The among-population mean allele frequencies vary largely among neutral loci. Therefore, we incorporate the contribution of variable allele frequencies in the ancestral population to estimate the among-population correlation that is shared among loci. Given the allele frequency in the ancestral population, the variance–covariance matrix of the allele frequencies $\boldsymbol{p}^l = \left(p_1^l, \dots, p_K^l\right)'$ at locus l is formulated as

$$\mathbf{E}(\boldsymbol{p}^l) = p_A^l \mathbf{1}$$

$$\mathbf{V}(\boldsymbol{p}^l) = p_A^l \left(1 - p_A^l\right) \boldsymbol{\Omega}.$$

$v_{ij} = \boldsymbol{\Omega}_{ij}$ $(i, j = 1, \dots, K)$ represents the among-subpopulation covariance (Weir & Hill, 2002; Coop, Witonsky, Di Rienzo, & Pritchard, 2010). The variance and covariance of the observed allele frequencies are

$$\begin{aligned}
\mathrm{V}\left(\hat{p}_i^l\right) &= \mathrm{V}\left(p_i^l\right) + E\left[p_i^l (1 - p_i^l)/n_i\right] \\
&= p_A^l \left(1 - p_A^l\right) v_{ii} + \frac{1}{n_i}\left(p_A^l - \left(p_A^l \left(1 - p_A^l\right) v_{ii} + {p_A^l}^2\right)\right) \\
&= p_A^l \left(1 - p_A^l\right)\left(\left(1 - \frac{1}{n_i}\right) v_{ii} + \frac{1}{n_i}\right)
\end{aligned}$$

$$\mathrm{COV}\left(\hat{p}_i^l, \hat{p}_j^l\right) = \mathrm{COV}\left(p_i^l, p_i^l\right) = p_A^l \left(1 - p_A^l\right) v_{ij} \ (i \neq j).$$

From this, we obtained the moment estimator $\boldsymbol{\Omega}$ as

$$\hat{v}_{ii} = \frac{1}{L}\sum_{l=1}^{L}\left(\frac{\frac{\hat{p}_i^{l\,2}}{\hat{\bar{p}}_{\cdot}^{l}(1-\hat{\bar{p}}_{\cdot}^{l})} - \frac{1}{n_i}}{1-\frac{1}{n_i}}\right)$$

$$\hat{v}_{ij} = \frac{1}{L}\sum_{l=1}^{L}\left(\frac{\hat{p}_i^l\hat{p}_j^l}{\hat{\bar{p}}_{\cdot}^{l}(1-\hat{\bar{p}}_{\cdot}^{l})}\right) \qquad (i \neq j),$$

where $\hat{\bar{p}}_{\cdot}^{l} = n^l/n$. Because most SNPs rarely have alleles in equilibrium (Wright 1931), these estimates are accurate when many neutral loci are available. With this estimated variance–covariance matrix of the allele frequencies, we obtained the variance–covariance matrix of the observed counts, $\hat{\boldsymbol{p}}^l$, as

$$\widehat{\mathrm{V}}(\hat{p}_i^l) = \hat{\bar{p}}_{\cdot}^{l}(1-\hat{\bar{p}}_{\cdot}^{l})\hat{v}_{ii} + \hat{\bar{p}}_{\cdot}^{l}(1-\hat{\bar{p}}_{\cdot}^{l})/n_i$$

$$\widehat{\mathrm{COV}}(\hat{p}_i^l, \hat{p}_j^l) = \hat{\bar{p}}_{\cdot}^{l}(1-\hat{\bar{p}}_{\cdot}^{l})\hat{v}_{ij}.$$

The p-values were calculated by the standard generalized least squares (GLS) method. Out of environment–gene (SNP) pairs and environment–trait pairs, we selected the significant pairs with an FDR of 0.05 using the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995).

2.3 | Neutral markers vs adaptive markers

To improve the power of detecting associations between genes and environments, we focused on the SNPs that were differentiated among subpopulations. However, neutral markers provide unbiased estimates of population-specific F_{ST} and population genetic spatial structure $\boldsymbol{\Omega}$. We defined these two sets of SNPs on the basis of the global F_{ST} values of the SNPs. We obtained the maximum likelihood estimates of the locus-specific global F_{ST} values (Beaumont & Bolding, 2004) using R package `FinePop2` in CRAN. The distribution of the locus-specific global F_{ST} values was fitted by a gamma distribution. We used the SNPs with F_{ST}

values larger than those in the 10th percentile of this gamma distribution for the correspondence analysis. However, we excluded the outlier SNPs with extremely large global F_{ST} values when estimating the population genetic parameters, population-specific F_{ST}, and population genetic spatial structure $\mathbf{\Omega}$. We adopted the threshold of $100/(10 \times number\ of\ SNPs)$ percentile of the gamma distribution.

2.4 | Simulation of range expansion and adaptation

To show the general framework and performance of our method, we conducted a simulation of the colonization model (Austerlitz, Jung-Muller, Godelle, & Gouyon, 1997) and included K (= 25) linearly arrayed subpopulations. Starting from subpopulation 1 located at the edge, subpopulations were successively colonized in every 10 generations and exposed to the local environments (Figure S1). The allele frequencies of the SNPs in each deme varied stochastically by genetic drift, migration from the neighboring subpopulations, and selection pressure of the environments. The environmental factor had two states, *high* and *low*, and the values 1 and 0 were assigned to the states. The environmental factor had the value of 0 in most subpopulations. Only subpopulations 9 and 15 were exposed to *high* environments. The environmental factor does not affect the allele frequencies at the neutral SNPs but affects the allele frequencies of the SNPs that contribute the traits, most of which are often unobserved. Adaptation to an environmental stress is often accompanied by the cost of reduced activity in the normal environment (e.g., Baucom & Mauricio, 2004); therefore, the derived alleles can adapt to the *high* environment at the expense of cost in the *low* environment (see Appendix 1 for details). The sample size was set to 50 for each of the 25 subpopulations. We examined the performance of our method under four values of the genetic–environmental parameter s (see equations A4 and A5 in Appendix 1). We performed 100 simulation trials under each s to examine the performance of our method.

206

207 **2.5 | Application to wild poplar data in North America**

208 As an empirical example, we analyzed publicly available data that included genetic and trait
209 information of 445 individuals of wild poplar (*Populus trichocarpa*), which were collected from
210 various regions over a range of 2,500 km near the Canadian–US border at a latitude of 44′ to
211 59′ N, a longitude of 121′ to 138′ W, and an altitude of 0–800 m (McKown et al., 2014a;
212 McKown et al., 2014b; Geraldes et al., 2013). The data included geographical information of
213 sampling locations, genotypes of 34,131 SNPs (3,516 genes), and values of stomatal anatomy,
214 leaf tannin, ecophysiology, morphology, and disease. We performed a clustering analysis using
215 the geographical distribution and divided the 445 individuals into groups. We applied model-
216 based clustering (Fraley & Raftery, 2016) with three types of spatial information—latitude,
217 longitude, and altitude—as the explanatory variables. Using the Bayesian information criterion
218 on the basis of the `mclustBIC` function in the R package `mclust` (Scrucca, Fop, Murphy, &
219 Raftery, 2016) under the VEV model (ellipsoidal, equal shape), we obtained 22 demes (hereafter,
220 subpopulations): 5 in northern British Colombia (NBC), 11 in southern British Colombia (SBC),
221 3 in inland British Colombia (IBC), and 3 in Oregon (ORE). We calculated the averages of the
222 environmental values at the sampling locations and phenotypic values of the individuals, and
223 we considered them representative values of the trait and environment for each subpopulation.
224 We plotted the estimated longitudes and latitudes for the subpopulations on the map (Figure 1).
225 Because our major concern was identifying correlations between among-population
226 differentiations of genes, traits, and environmental factors, we selected the SNP with the highest
227 global F_{ST} value over 22 subpopulations from each of the 3,516 gene regions. Here, we focused
228 on the four trait variables, adaxial stomata density (ADd), abaxial stomata density (ABd), and
229 two measurements of leaf rust disease morbidity in 2010 and 2011 (DP10 and DP11,
230 respectively). Each sampling location of a subpopulation was described by nine

environmental/geographical variables: altitude (ALT), day length in summer time (DAY), frost-free days (FFD), mean annual temperature (MAT), mean warmest month temperature (MWMT), mean annual precipitation (MAP), mean summer precipitation (MSP), annual heat–moisture index (AHM, ~MAT/MAP, an indicator of drought), and summer heat–moisture index (SHM, ~MWMT/MSP). In addition, four soil conditions, the ratio of clay, silt, sand, and gravel in the topsoil, were obtained from The Unified North American Soil Map (Liu et al., 2014) and used as environmental values of the sampling locations.

3 | RESULTS

3.1 | Analysis of simulated data and performance of our method

Figure 2 shows the population-specific F_{ST} values and reveals that the history that population expansion started around subpopulation 1 and extended through subpopulation 25. Figure 3 shows the biplot of the correspondence analysis for the genetic–environmental parameter $s = 0.16$; that is, the homozygote of the derived allele is 1.16 times more advantageous than the homozygote of the ancestral allele under the *high* environmental condition $(E = 1)$ and 1.16 times less advantageous under the *low* environmental condition $(E = 0)$ (see Appendix 1 for details). The gradient of the colors assigned to the subpopulations in the biplot indicates the history of range expansion and the environmental stresses that the population experienced during the course of range expansion. The environmental factor and environmentally adaptive genes showed positive correlations (connection between green node +E and purple nodes). However, the environmental factor and trait showed negative correlation (connection between green node -E and the orange node). Most of the subpopulations were characterized by neutral loci; however, subpopulations 9 and 15 were arranged near the environmental factor. This indicates that these two subpopulations were characterized by both neutral and environmentally adaptive loci.

We examined the performance of our GLS detection of correlation between the environment and genes/traits by receiver operating characteristic curve (ROC) and area under curve (AUC) analysis (Figure 4). The effect of parameter s showed that even a difference in viability less than 10% can be effectively detected by this method.

Next, we showed the subpopulations in which environmentally adaptive alleles emerged in 100 simulation trials (Figure 5). The horizontal axis indicates the subpopulations, and the vertical axis indicates the frequencies of the loci that retained their polymorphism after 260 generations. In the cases of low selection pressure (small s), loci that emerged in deme 1 were highly observed, because subpopulation 1 had large *N*e and a high chance of newly derived alleles. However, in the cases of high selection pressure (large s), loci that emerged in subpopulations 9 and 15 were observed because of adaptation to the environment.

3.2 | Analysis of wild poplar data

Our generated 2D plot of correspondence and correlation analysis identified the global structure of genetic differentiation and adaptation (Figure 6a). The placement of subpopulations and coloration by population-specific F_{ST} provide an interpretation of habitat expansion of three directions, from the inland to the coast and to northern and southern areas (Figure 6b).

The color gradient of population-specific F_{ST} values on the deme labels indicated that the ancestral population might inhabit the inland area (IBC), which is characterized by high altitude (+ALT) and dry conditions year-round (+AHM, -*MAP*), as shown in the center of Figure 6a. Dry conditions (+AHM) were correlated with genes associated with drought and osmotic regulation (Figure 6c): CBF4 (response to drought and cold stress; Haake et al. 2002, Hussain et al. 2018), XERICO (response to osmotic stress, response to salt stress; Ko, Yang, & Han, 2006), SAL1 (response to water deprivation and salt stress; Wilson et al. 2009), MYB85 (cell wall biogenesis responding water deprivation and salinity; Winter et al. 2007), MKK2 (water

transport, response to osmotic stress; Danquah, de Zelicourt, Colcombet, & Hirt, 2014), and iqd21 (stomatal morphogenesis; Wang et al., 2011). This indicates that poplar was initially adapted to the dry and cold uplands.

Slightly larger population-specific F_{ST} values than those of IBC indicated that the population expansion might have then occurred to the coastal area (SBC), which was characterized by relatively short day length in summer (-*DAY*); this means that the seasonal variation of day length is small in the southern area, with mild temperatures (+MAT, +FFD) and wet conditions year-round (-*AHM*, +MAP), as plotted in top left of Figure 6a. The small seasonal variation of day length (-*DAY*) was correlated with abaxial stomata density, which indicated that strong southern sunlight stimulates photosynthesis and requires many stomata (see Supplementary Figure S2). Mild temperatures (+MAT, +FFD) were correlated with genes associated with body growth (Figure 6c): LAS (shoot formation; Greb et al. 2003), GSL12 (signaling during growth and development; Yadav et al. 2014), and iqd2 (leaf growth regulator; Nikonorova et a. 2018). The year-round wet environment (-*AHM*) was correlated with a gene related to water conditions (Figure 6d), HRA1 (response to hypoxia, Giuntoli et al. 2014). The results indicated that subpopulations in SBC were adapted to warmth and oxygen deprivation due to excessive water. After adaptation in SBC, wild poplar might have expanded to the southern area (ORE), which is characterized by warm and dry conditions, particularly in the summer (+SHM), as plotted in the bottom left of Figure 6a. Dry summer (+SHM) was correlated with stress response (Figure 6c) in the gene NAC090 (salt and drought tolerance; Zang et al. 2019), which revealed that the population adapted to hot and dry summer conditions.

The largest population-specific F_{ST} values in NBC revealed that wild poplar finally expanded to the northern area, which is characterized by long day length in summer (+DAY), which means that day length varies greatly from season to season, and low temperatures (-*MAT*, -*MWMT*, -*FFD*), as described in Figure 6a. These variables were correlated with adaxial stomata

density and leaf rust disease (DP10, DP11). This finding supports the preceding knowledge that the adaxial stomata compensates for reduced photosynthetic efficiency in the northern area; however, there is a risk of pathogen invasion (Melotto et al. 2006, see Supplementary Figure S2). DAY was correlated with genes associated with light conditions (Figure 6e): ACT7 (response to light stimulus; McDowell et al. 1996), PRR7 (circadian rhythm; Alabadí et al. 2001), and APX2 (response to high light intensity; Karpinski et al. 1997). These results indicated that the population adapted to the light conditions, which vary greatly among seasons.

Finally, we summarized our results on a map, which visualized the history of range expansion of the population during adaptation to environmental stresses (Figure 7a) to provide an interpretation of population expansion in the distribution range and environmental adaptation of genes. These genes showed a population-specific allele frequency gradient as they adapted to the environment in each area (figure not shown). For example, the allele frequency of CBF4, which was correlated with +AHM, was high inland (year-round dry), but almost zero in the coastal area (year-round wet) (Figure 7b). This indicated that the drought adaptive allele was lost through expansion into wet coastal areas.

4 | DISCUSSION

A population evolves in space and time and responds to variable environments. From its birth, a population may continuously change its distribution range, and the initial localities may have occasionally been exposed to unprecedented environmental stress. In these localities, a population can adapt to such environmental stresses by changing its geographical distribution or genomes. We developed an exploratory population genomic approach that visualizes the whole life history of a population. We propose a novel graphical representation of significant associations between genes, environments, and traits that consists of three approaches: 1) correlation analysis between environments and genes using the GLS method, 2) correspondence

analysis of population structure characterized by gene–environment–trait differentiation, and 3) trace of population history using population-specific F_{ST}.

Geographic gradients of genomic diversity and genomic population structure reflect the history of colonization and range expansion. Natural selection mostly affects the allele frequencies of genes that contribute to stress resistance. Therefore, the SNP allele frequencies were the primary features for analysis. Because little is known about the elements of natural selection, we developed an exploratory approach. Rather than presetting the environmental values as dependent variables, we labeled significant associations between the environments, traits, and genes in the biplot of the correspondence analysis. The locations and colors of the subpopulations represent the pattern of population structure and the ages of the subpopulations.

Using our method, the labeled environments, traits, and genes provide insight into the environmental stresses that the population encountered, locations of the stresses, traits that were affected by the stresses, and genes that enabled the subpopulations to adapt to the stresses. In many cases, we only had fragmented information on the environmental stresses and traits that potentially affected the life history of the population. Analysis of the wild poplar data revealed that investigating the functions of the significant genes elucidated the effects of the environmental stresses on the physiological states of the population members.

Our simulations and analysis of wild poplar data revealed the potential of our method for comprehensive understanding of ecological and genetic mechanisms underlying environmental adaptation and population structuring. Although conventional GWAS and genome scanning effectively search for genes related to some given factors or traits, our method captures the overall picture of the relationship among genes, environmental factors, and traits in association with population structure.

Our analysis uses subpopulations as units of the analysis. However, subpopulations are often defined by post-stratification of the sample. In the case of wild poplar data, we defined the

subpopulations on the basis of longitude, latitude, and altitude of the samples. The power of our exploratory approach depends on the accuracy of the assignment of the individuals to local subpopulations.

Our method describes relationships between population means of allele frequencies, trait values, and environmental/location factors; therefore, like Bayenv, it analyzes sample means among measured data. We derived a moment estimator for the among-population covariance matrix of allele frequencies. The estimates were accurate when many neutral loci were available. In such cases, our procedure can provide an unbiased picture of gene–environment–trait associations and the history of range expansion and adaptation.

ACKNOWLEDGEMENTS

This study was supported by Japan Society for the Promotion of Science Grants-in-Aid for Scientific Research KAKENHI nos. 16H02788 and 19H04070 to H.K. We thank Mallory Eckstut, PhD, from Edanz Group (https://en-author-services.edanzgroup.com/ac) for editing a draft of this manuscript.

REFERENCES

Alabadí, D., Oyama, T., Yanovsky, M. J., Harmon, F. G., Más, P., & Kay, S. A. (2001). Reciprocal regulation between TOC1 and LHY/CCA1 within the Arabidopsis circadian clock. *Science*, *293*, 880–883. https://doi.org/10.1126/science.1061320

Austerlitz, F., Jung-Muller, B., Godelle, B., & Gouyon, P.-H. (1997). Evolution of coalescence times, genetic diversity and structure during colonization. *Theoretical Population Biology*, *51*, 148–164. https://doi.org/10.1006/tpbi.1997.1302

Barrett, R.D., & Schluter, D. (2008). Adaptation from standing genetic variation. *Trends in*

Ecology and Evolution, *23*, 38–44. https://doi.org/10.1016/j.tree.2007.09.008

Baucom, R. S. & Mauricio, R. (2004). Fitness costs and benefits of novel herbicide tolerance in a noxious weed. *Proceedings of the National Academy of Sciences, USA*, 101, 13386–13390. https://doi.org/10.1073/pnas.0404306101

Beaumont, M. A., & Balding, D. J. (2004). Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, *13*, 969–980. https://doi.org/10.1111/j.1365-294X.2004.02125.x

Benjamini, Y., & Hochberg, Y. (1995). Controling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Benzecri, J. P. (1973). *Data Analyses. Volume II. Correspondence Analysis.* Paris, France: Dunod.

Bradbury, I. R. & Bentzen, P. (2007). Non-linear genetic isolation by distance: implications for dispersal estimation in anadromous and marine fish populations. *Marine Ecology Progress Series*, *340*, 245–257. doi:10.3354/meps340245

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*, 2633–2635. https://doi.org/10.1093/bioinformatics/btm308

Buckleton, J., Curran, J., Goudet, J., Taylor, D., Thiery, A., & Weir, B. S. (2016). Population-specific F_{ST} values for forensic STR markers: A worldwide survey. *Forensic Science International: Genetics*, *23*, 91–100. https://doi.org/10.1016/j.fsigen.2016.03.004

Capblancq, T., Fitzpatrick, M. C., Bay, R. A., Exposito-Alonso, M., & Keller, S. R. (2020). Genomic prediction of (mal) adaptation across current and future climatic landscapes. *Annual Review of Ecology, Evolution, and Systematics*, 245–269.

https://doi.org/10.1146/annurev-ecolsys-020720-042553

Coop, G. D., Witonsky, D., Di Rienzo, A., & Pritchard, K. J. (2010). Using environmental correlations to identify loci underlying local adaptation. *Genetics 185*, 1411–1423. https://doi.org/10.1534/genetics.110.1148

Danquah, A., de Zelicourt, A., Colcombet, J., & Hirt, H. (2014). The role of ABA and MAPK signaling pathways in plant abiotic stress responses. Biotechnology Advances, *32*, 40–52. https://doi.org/10.1016/j.biotechadv.2013.09.006

De Mita, S., Thuillet,A. C., Gay, L., Ahmadi, N., Manel, S., Ronfort, J., & Vigouroux, Y. (2013). Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. Molecular Ecology, *22*, 1383–1399. https://doi.org/10.1111/mec.12182

De Villemereuil, P., Frichot, É., Bazin, É., François, O., & Gaggiotti, O. E. (2014). Genome scan methods against more complex models: when and how much should we trust them? Molecular Ecology, *23*, 2006–2019. https://doi.org/10.1111/mec.12705

Devlin, B., & Roeder, K. (1999). Genomic control for association studies. *Biometrics*, *55*, 997–1004. https://doi.org/10.1111/j.0006-341X.1999.00997.x

Foll, M., & Gaggiotti, O. (2008). A genome scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics*, *180*, 977–993. https://doi.org/10.1534/genetics.108.092221

Fraley, C., & Raftery, A. E. (2016). Model-based clustering, discriminant analysis and density estimation. *Journal of the American statistical Association*, *97,* 611–631. https://doi.org/10.1198/016214502760047131

Frichot, E., Schoville, S. D., Bouchard, G., & François, O. (2013). Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular Biology and Evolution*, *30*, 1687–1699. https://doi.org/10.1093/molbev/mst063

Geraldes, A., Difazio, S. P., Slavov, G. T., Ranjan, P., Muchero, W., … & Tuskana, G. A. (2013). 34K SNP genotyping array for Populus trichocarpa: Design, application to the study of natural populations and transferability to other Populus species. *Molecular Ecology Resources*, *13*, 306–323. https://doi.org/10.1111/1755-0998.12056

Giuntoli, B., Lee, S. C., Licausi, F., Kosmacz, M., Oosumi, T., …& Perata, P. (2014). A trihelix DNA binding protein counterbalances hypoxia-responsive transcriptional activation in Arabidopsis. *PLOS Biology*, *12*, e1001950. https://doi.org/10.1371/journal.pbio.1001950

Greb, T., Clarenz, O., Schäfer, E., Müller, D., Herrero, R., … & Theres, K. (2003). Molecular analysis of the LATERAL SUPPRESSOR gene in Arabidopsis reveals a conserved control mechanism for axillary meristem formation. *Genes & Development*, *17*, 1175–1187. https://doi.org/10.1101/gad.260703

Haake, V., Cook, D., Riechmann, J. L., Pineda, O., Thomashow, M. F., & Zhang, J. Z. (2002). Transcription factor CBF4 is a regulator of drought adaptation in Arabidopsis. *Plant Physiology*, *130*, 639–648. https://doi.org/10.1104/pp.006478

Hayashi, C. (1953). Multidimensional quantification: With applications to analysis of social phenomena. *Annals of the Institute of Statistical Mathematics*, *5*, 121–143. https://www.ism.ac.jp/editsec/aism/pdf/005_2_0121.pdf

Hilborn, R., Quinn, T. P., Schindler, D. E., & Rogers, D. E. (2003) Biocomplexity and fisheries sustainability. *Proceedings of the National Academy of Sciences, USA*, *100*, 6564–6568. https://doi.org/10.1073/pnas.1037274100

Hussain, H.A., Hussain, S., Khaliq, A., Ashraf, U., Anjum, S.A., Men, S., & Wang, L. (2018). Chilling and Drought Stresses in Crop Plants: Implications, Cross Talk, and Potential Management Opportunities. *Frontiers in Plant Science*, *9*, 393. https://doi.org/10.3389/fpls.2018.00393

Jorde, P. E., Søvik, G., Westgaard, J. I., Albretsen, J., André, C., …. & Jørstad, K. E. (2015). Genetically distinct populations of northern shrimp, *Pandalus borealis*, in the North Atlantic: adaptation to different temperatures as an isolation factor. *Molecular Ecology*, *24*, 1742–1757. https://doi.org/10.1111/mec.13158

Karpinski, S., Escobar, C., Karpinska, B., Creissen, G., & Mullineaux, P. M. (1997). Photosynthetic electron transport regulates the expression of cytosolic ascorbate peroxidase genes in Arabidopsis during excess light stress. *Plant Cell*, *9*, 627–640. https://doi.org/10.1105/tpc.9.4.627

Kitada, S., Nakamichi, R., & Kishino, H. (2017). The empirical Bayes estimators of fine-scale population structure in high gene flow species. *Molecular Ecology Resources*, *17*, 1210–1222. https://doi.org/10.1111/1755-0998.12663

Ko, J. H., Yang, S. H., & Han, K. H. (2006) Upregulation of an Arabidopsis RING-H2 gene, XERICO, confers drought tolerance through increased abscisic acid biosynthesis. *Plant Journal*, *47*, 343–355. https://doi.org/10.1111/j.1365-313X.2006.02782.x

Manzaneda, A. J., Rey, P. J., Bastida, J. M., Weiss-Lehman, C., Raskin, E., & Mitchell-Olds, T. (2012). Environmental aridity is associated with cytotype segregation and polyploidy occurrence in Brachypodium distachyon (Poaceae). *New Phytologist*, *193*, 797-805. https://doi.org/10.1111/j.1469-8137.2011.03988.x

McDowell, L. M., An, Y., Huang, S., McKinney, E. C., & R. B. Meagher, R. B. (1996). The Arabidopsis ACT7 actin gene is expressed in rapidly developing tissues and responds to several external stimuli. *Plant Physiology*, *111*, 699–711. https://doi.org/10.1104/pp.111.3.699

McKown1, A. D., Guy, R. D., Klápště, J., Geraldes, A., Friedmann, M., … & Douglas, C. J. (2014a). Geographical and environmental gradients shape phenotypic trait variation and genetic structure in *Populus trichocarpa*. *New Phytologist*, *201*, 1263–1276.

https://doi.org/10.1111/nph.12601

McKown. A. D., Guy, R. D., Quamme, L., Klápště, J., La Mantia, J., … & Azam M. S. (2014b). Association genetics, geography and ecophysiology link stomatal patterning in *Populus trichocarpa* with carbon gain and disease resistance trade-offs. *Molecular Ecology*, *23*, 5771–5790. https://doi.org/10.1111/mec.12969

Melotto, M., Underwood, W., Koczan, J., Nomura, K., & He, S. Y. (2006). Plant stomata function in innate immunity against bacterial invasion. *Cell*, *126*, 969–98. https://doi.org/10.1016/j.cell.2006.06.054

Nicholson, G., Smith, A. V., Jónsson, F., Gústafsson, O., Stefánsson, K., & Donnelly, P. (2002). Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*, 695–715. https://doi.org/10.1111/1467-9868.00357

Nikonorova, N., Van den Broeck, L., Zhu, S., van de Cotte, B., Dubois, M., … & De Smet, I. (2018) Early mannitol-triggered changes in the Arabidopsis leaf (phospho) proteome reveal growth regulators. *Journal of Experimental Botany*, *69*, 4591–4607. https://doi.org/10.1093/jxb/ery261

Nosil, P., Crespi, B. J., & Sandoval, C. P. (2002) Host-plant adaptation drives the parallel evolution of reproductive isolation. *Nature*, 417, 440–443. https://doi.org/10.1038/417440a

Nosil, P., Funk, D. J., & Ortiz-Barrientos, D. (2009). Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, *18*, 375–402. https://doi.org/10.1111/j.1365-294X.2008.03946.x

Orsini, L., Vanoverbeke, J., Swillen, I., Mergeay, J., & De Meester, L. (2013). Drivers of population genetic differentiation in the wild: isolation by dispersal limitation, isolation by adaptation and isolation by colonization. *Molecular Ecology*, 22, 5983–

5999. https://doi.org/10.1111/mec.12561

Pritchard, J. K., Stephens, M., Rosenberg, N. A., & Donnelly, P. (2000). Association mapping in structured populations. *American Journal of Human Genetics*, *67*, 170–181. https://doi.org/10.1086/302959

Pritchard, J. K., & Rosenberg, N. A. (1999). Use of unlinked genetic markers to detect population stratification in association studies. *American Journal of Human Genetics*, *65*, 220–228. https://doi.org/10.1086/302449

Raymond, M., & Rousset, F. (1995). GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity*, *86*, 248–249.

Rousset, F. (2008). Genepop'007: a complete reimplementation of the Genepop software for Windows and Linux. *Molecular Ecology Resources*, *8*, 103–106. https://doi.org/10.1111/j.1471-8286.2007.01931.x

Santure, A. W., & Garant, D. (2018). Wild GWAS-association mapping in natural populations. *Molecular Ecology Resources*, *18*, 729–738. https://doi.org/10.1111/1755-0998.12901

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E., 2016 mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, *8*, 205–233.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., … & Yang, J. (2017). 10 years of GWAS discovery: biology, function, and translation. *American Journal of Human Genetics*, *101*, 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005

Wang, R. S., Pandey, S., Li, S., Gookin, T. E., Zhao, Z., Albert, R., & Assmann, A. M. (2011). Common and unique elements of the ABA-regulated transcriptome of Arabidopsis guard cells. *BMC Genomics*, *12*, 216. https://doi.org/10.1186/1471-2164-12-216

Weir, B. S., & Goudet, J. (2017). A unified characterization of population structure and relatedness. *Genetics*, *206*, 2085–2103. https://doi.org/10.1534/genetics.116.198424

Weir, B. S, & Hill, W. G. (2002). Estimating F-statistics. *Annual Review of Genetics*, *36*, 721-

https://doi.org/750. 10.1146/annurev.genet.36.050802.093940

Wilkinson, S., & Davies, W. J. (2002). ABA-based chemical signalling: the co-ordination of responses to stress in plants. *Plant, Cell and Environment*, *25*, 195–210. https://doi.org/10.1046/j.0016-8025.2001.00824.x

Wilson, P. B., Estavillo, G. M., Field, K. J., Pornsiriwong, W., Carroll, A. J., …. & Pogson, B.J. (2009). The nucleotidase/phosphatase SAL1 is a negative regulator of drought tolerance in Arabidopsis, The Plant Journal, 58, 299–317. https://doi.org/10.1111/j.1365-313X.2008.03780.x

Winter D., Vinegar B., Nahal, H., Ammar, R., Wilson, G. V., & Provart, N. J. (2007). An "electronic fluorescent pictograph" browser for exploring and analyzing large-scale biological data sets. *PLoS One* 2, e718. https://doi.org/10.1371/journal.pone.0000718

Wright, S. (1931). Evolution in Mendelian population. *Genetics*, *16*, 97-159. PMID: 17246615

Wright, S. (1965). The interpretation of population structure by *F*-statistics with special regard to systems of mating. *Evolution*, *19*, 395–420. https://www.jstor.org/stable/2406450

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., ... & Kresovich, S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, *38*, 203–208. https://doi.org/10.1038/ng1702

Zang, D., Wang, j., Zhang, X., Liu, Z., & Wang, Y. (2019). Arabidopsis heat shock transcription factor HSFA7b positively mediates salt stress tolerance by binding to an E-box-like motif to regulate gene expression. *Journal of Experimental Botany*, *70*, 5355–5374. https://doi.org/10.1093/jxb/erz261

DATA ACCESSIBILITY STATEMENT

The authors affirm that all data necessary for confirming the conclusions of the article are present within the article, figures, and supplementary information. The R codes to perform our representation method and simulations of population colonization are available in the Supporting Information.

AUTHOR CONTRIBUTIONS

H.K. designed the study and developed the theory. R.N. and H.K. analyzed data. R.N. created R package for implementing the method and performed simulations. S.K., H.K., and R.N. provided critical feedback, helped shape the research, and wrote the manuscript.

Appendix 1| Simulation scenario

A1.1| Populations and SNPs

We included $K = 25$ linearly arrayed subpopulations (see Figure S1). Subpopulation 1 accommodated an ancestral population of $N_e = 10^5$. In every generation, the populations exchanged 1% of N_e individuals with adjacent populations. Once in 10 generations, habitat expansion occurred, and 1% of N_e immigrated to the adjacent vacant deme and increased the population size to the capacity N_e in one generation. We introduced an environmental factor that had a *high* (1) state in deme 9 and 15, and *low* (0) state in the other subpopulations:

$$E_i = \begin{cases} 0 & (i \neq 9, 15) \\ 1 & (i = 9, 15) \end{cases}.$$

This did not affect the allele frequencies of neutral alleles but affected the survival of non-neutral alleles.

For the initial population, we generated 50,000 polymorphic loci whose alleles were neutral

against the environment. Their allele frequencies followed the predictive equilibrium distribution, $f(q) \propto q^{-1}(1-q)^{-1}$ (Wright 1931). Then, additional polymorphisms of 50 neutral loci and 50 environmentally adaptive loci were introduced to the existing populations in each generation. The derived alleles had an initial frequency of 0.01 in the birth populations and 0 in the other populations; they mimicked new mutations that survived the initial phase after their birth. Allele frequencies varied with random drift under a binomial distribution.

After 260 generations, we obtained loci that retained their polymorphism; however, we considered them "unobserved" loci. Then, we randomly selected some of them as "observed" loci. In this simulation, we selected 500 initial neutral loci, 50 newly derived neutral loci, and 10 newly derived environmentally adaptive loci. Then, we generated 50 individuals for each population. Genotypes of these 560 loci were randomly generated for each individual following their allele frequencies in the population to which each individual belonged. Finally, we randomly generated 10 unobserved survival traits $T(G, E_i)$ for each individual and obtained "observed" trait T' as the summation of T. As a sample of the 25 populations (50 individuals from each population), we obtained a set of "observations" (genotypes at the 560 loci and one environment-related trait for 25 populations × 50 individuals).

A1.2| Environmental selection and fitness of derived alleles

In a neutral gene, both ancestral and derived alleles are equally affected by the environment, and the allele frequencies are changed by random drift. However, in an environmentally adaptive gene, derived alleles have advantages/disadvantages in *high*/*low* conditions compared with the ancestral allele. Therefore, derived allele frequency increases/decreases in *high*/*low* conditions by natural selection. We inferred the environmental adaptation and the cost on the basis of the survival probability of the relevant trait.

The genotype $G = 0, 1, 2$ of the environmental adaptation gene represents the number of

derived alleles. This gene contributes to an unobserved survival trait $T(G,E)$ with the interaction of genotype G and environmental factor E:

$$T(G,E) = \gamma_G G + \gamma_E E + \gamma_{GE} G \times E + \varepsilon_T \qquad \varepsilon_T \sim N(0,1). \qquad \text{(A1)}$$

The survival probability $S(T)$ of the trait value T is described as the probability that T is positive:

$$S(T) = \mathrm{Prob}(T > 0). \qquad \text{(A2)}$$

The larger the trait, the greater chance of survival. The survival probability of a genotype G under environmental condition E, $S(G|E)$, is given as

$$S(G|E) = T(G,E).$$

In deme i, given the frequency $P_G^{(t)}$ of genotype G at generation t, the allele frequency at the next generation is obtained as

$$P_G^{(t+1)} = P_G^{(t)} S(G|E_i) / \textstyle\sum_G \left\{ P_G^{(t)} S(G|E_i) \right\}.$$

We note that only the relative values of survival probabilities are relevant for the population genetic dynamics.

The derived allele is assumed to be advantageous over the ancestral allele under the *high* environmental condition $E = 1$, whereas it is disadvantageous under the *low* environmental condition $E = 0$. Therefore, we consider the case where $\gamma_G \leq 0$, $\gamma_E \leq 0$, and $\gamma_{GE} \geq 0$.

We set a simulation parameter r and s to control environmental effect γ_E, genetic effect γ_G, and gene-environment interaction γ_{GE}. The parameter r $(0 < r < 1)$ was introduced to represent the stress of *high* environmental condition $E = 1$ and was defined by the ratio of the survival probabilities between the two conditions:

$$r = S(G = 0|E = 1) / S(G = 0|E = 0). \qquad \text{(A3)}$$

The parameter s (≥ 0) represents the fitness of the derived allele in the *high* environmental condition $E = 1$:

$$1 + s = S(G = 2|E = 1)/S(G = 0|E = 1). \quad \text{(A4)}$$

We assumed the cost of adaptation by reversing the fitness in the *low* environmental condition $E = 0$:

$$S(G = 2|E = 0)/S(G = 0|E = 0) = \frac{1}{1+s} \quad \text{(A5)}$$

The coefficients γ_E, γ_G, and γ_{GE} were obtained from r and s. First, we noted that, from equations (A1) and (A2), $S(G = 0|E = 0) = \frac{1}{2}$ and $S(G = 0|E = 1) = \Phi(\gamma_E)$. Φ is the cumulative distribution of the standard normal distribution. Hence, from equation (A3), we obtained $\gamma_E = \Phi^{-1}\left(\frac{r}{2}\right)$. Similarly, we obtained $\gamma_G = \frac{1}{2}\Phi^{-1}\left(\frac{1}{2(1+s)}\right)$ from equation (A5). Finally, we obtained $\gamma_{GE} = \frac{1}{2}\left(\Phi^{-1}\left(\frac{rs}{2}\right) - 2\gamma_G - \gamma_E\right)$ from equation (A4).

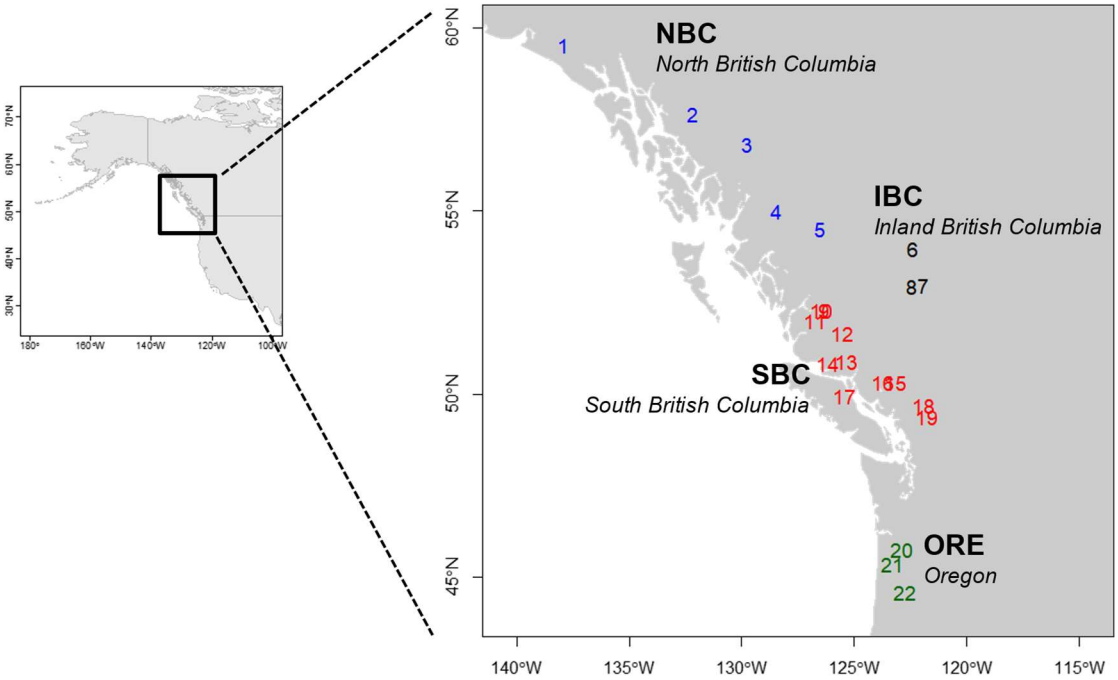

Figure 1 Wild poplar populations in North America (McKown et al., 2014a; McKown et al., 2014b; Geraldes et al., 2013). The individuals were classified into 22 groups (subpopulations) on the basis of longitude, latitude, and altitude of the sampling points. NBC, blue; SBC, red; IBC, black; and ORE, green.

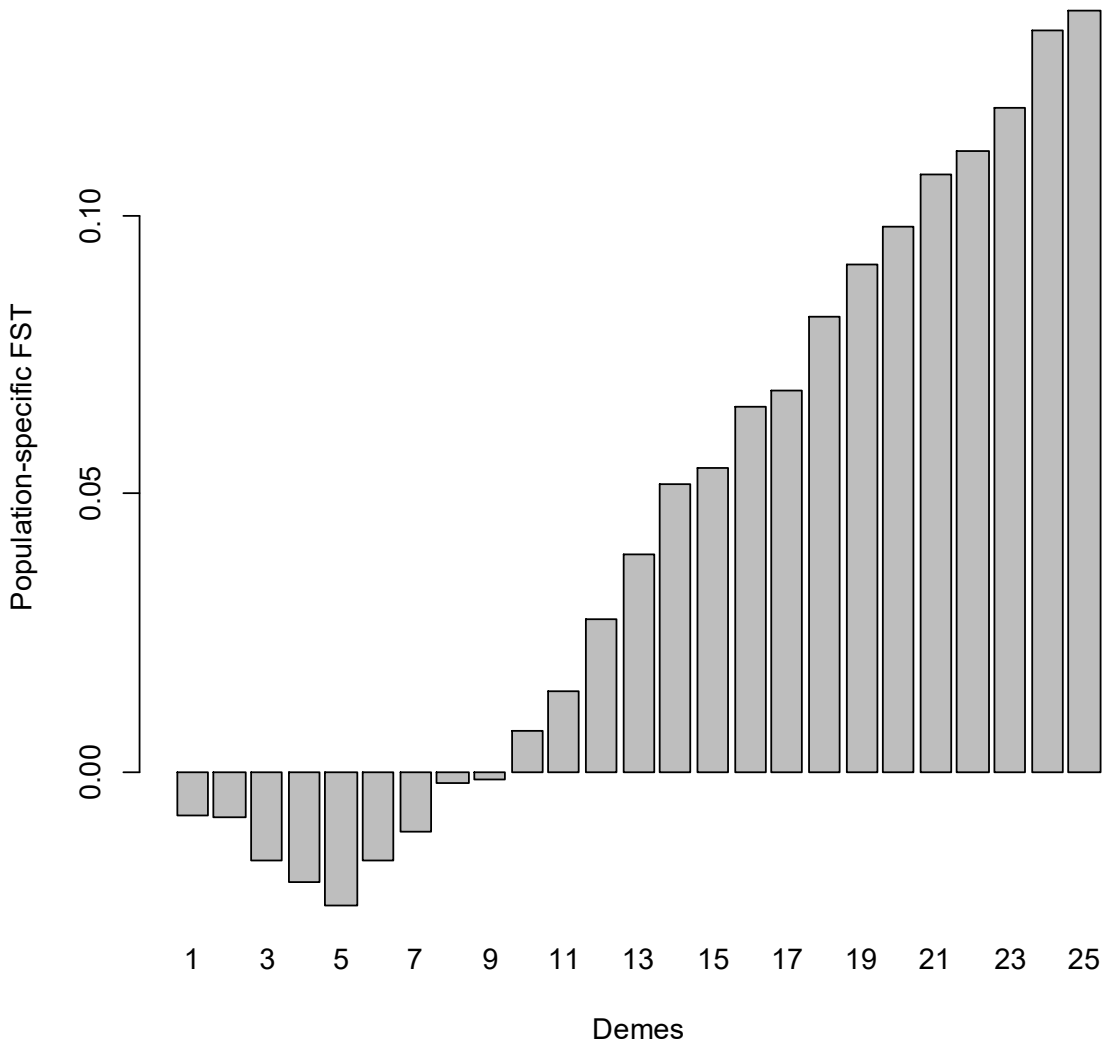

FIGURE 2 Estimated population-specific F_{ST} values for the 25 simulated subpopulations.

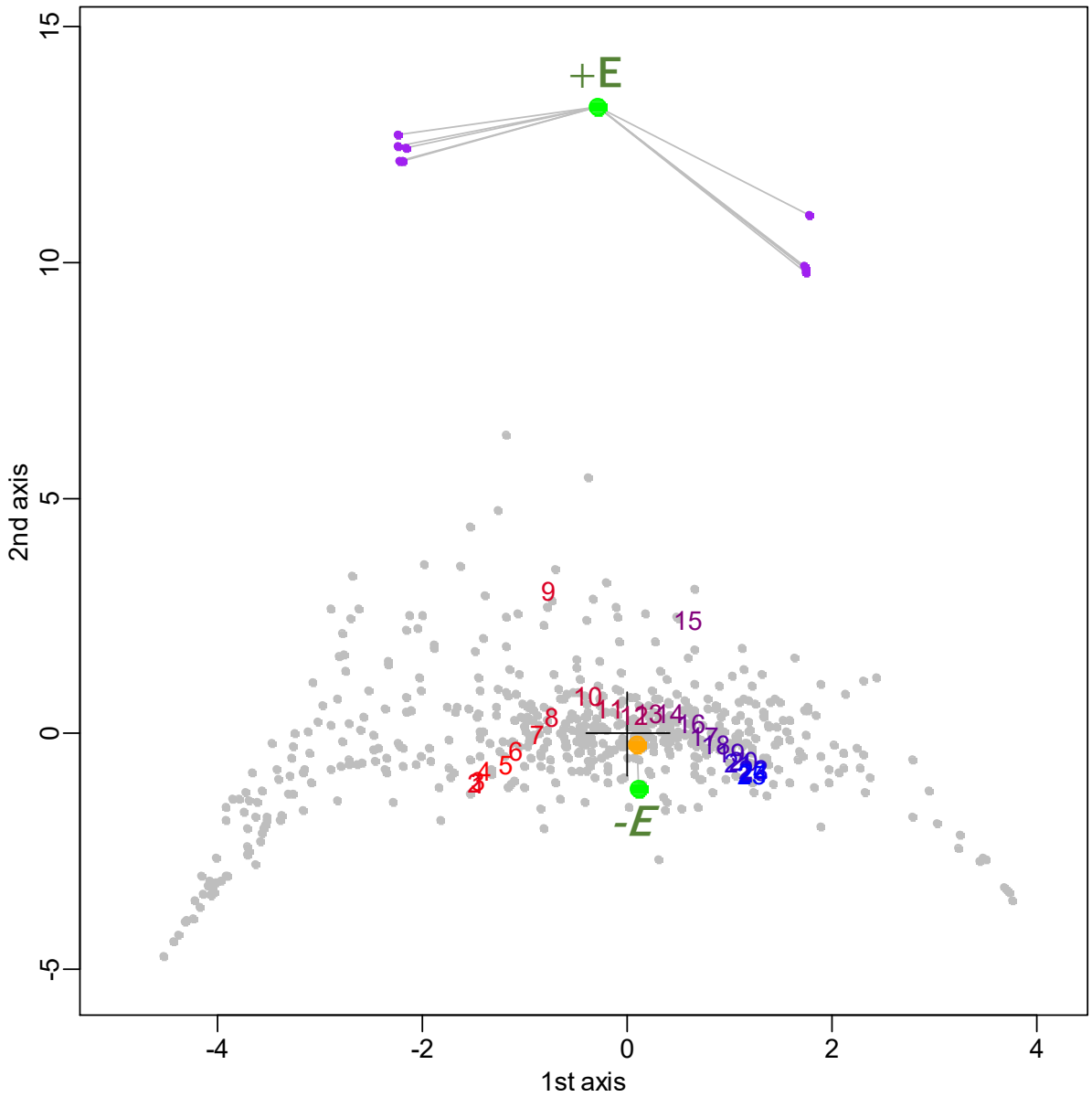

FIGURE 3 Correspondence analysis of data generated by simulation of colonization and adaptation. Green nodes represent environmental factor E. The plus sign (+E) indicates the original environmental value, whereas the minus sign (-E) represents the sign-reversed environmental value. Gray nodes are neutral loci, and purple nodes are environmentally adaptive loci. The orange node is the observed trait T'. Gray lines indicate significant correlations between the environment and gene/trait under FDR = 1%. Each subpopulation label is colored by its population-specific F_{ST} value. Numbers from 1 to 25 represent subpopulations,

and the color gradients on subpopulation labels represent the standardized magnitude of a population-specific F_{ST} value at the sampling point, with colors between blue (for the largest F_{ST}, which represents the youngest subpopulation) and red (smallest F_{ST}, which represents the oldest subpopulation).

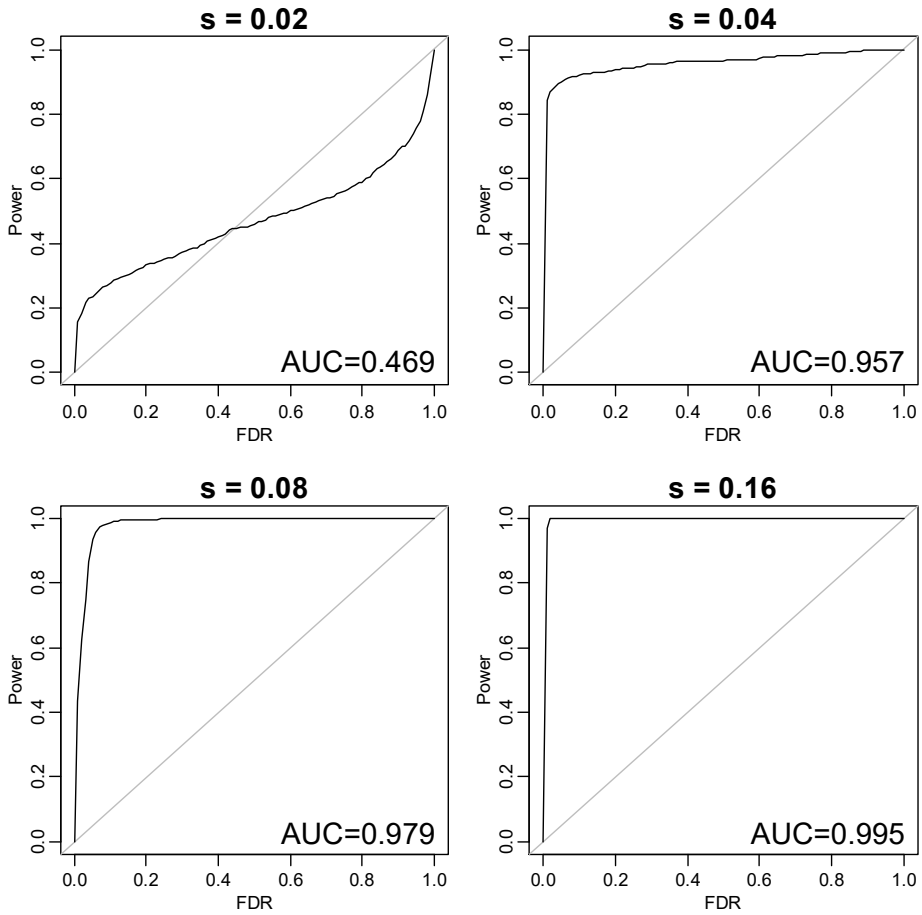

FIGURE 4 ROC curve of gene–environment–trait interaction detection for different values of selection intensity, *s*.

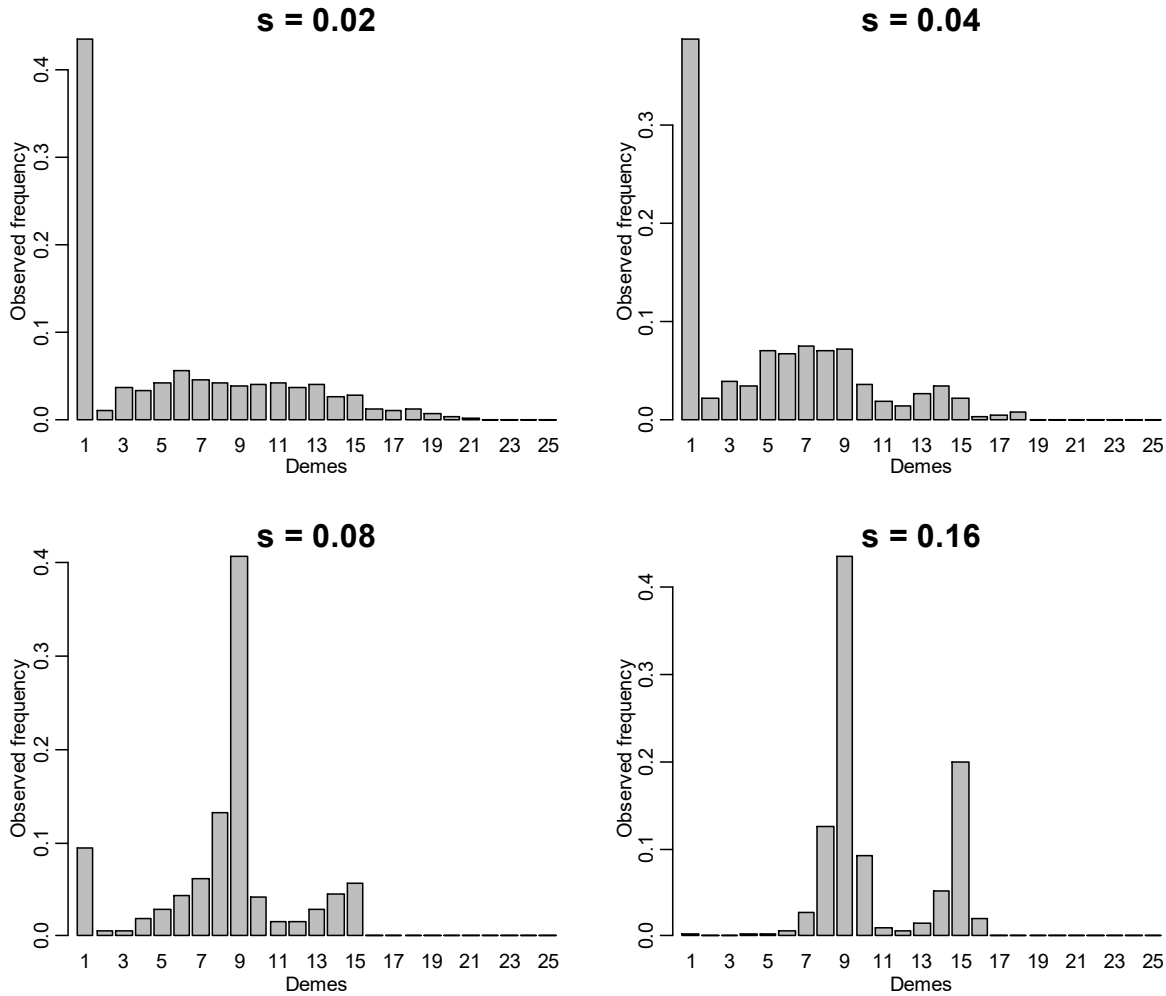

FIGURE 5 Mean numbers of the environmentally adaptive loci that are currently polymorphic. To be consistent with common population genomic studies, we counted loci with minor allele frequencies > 0.01.

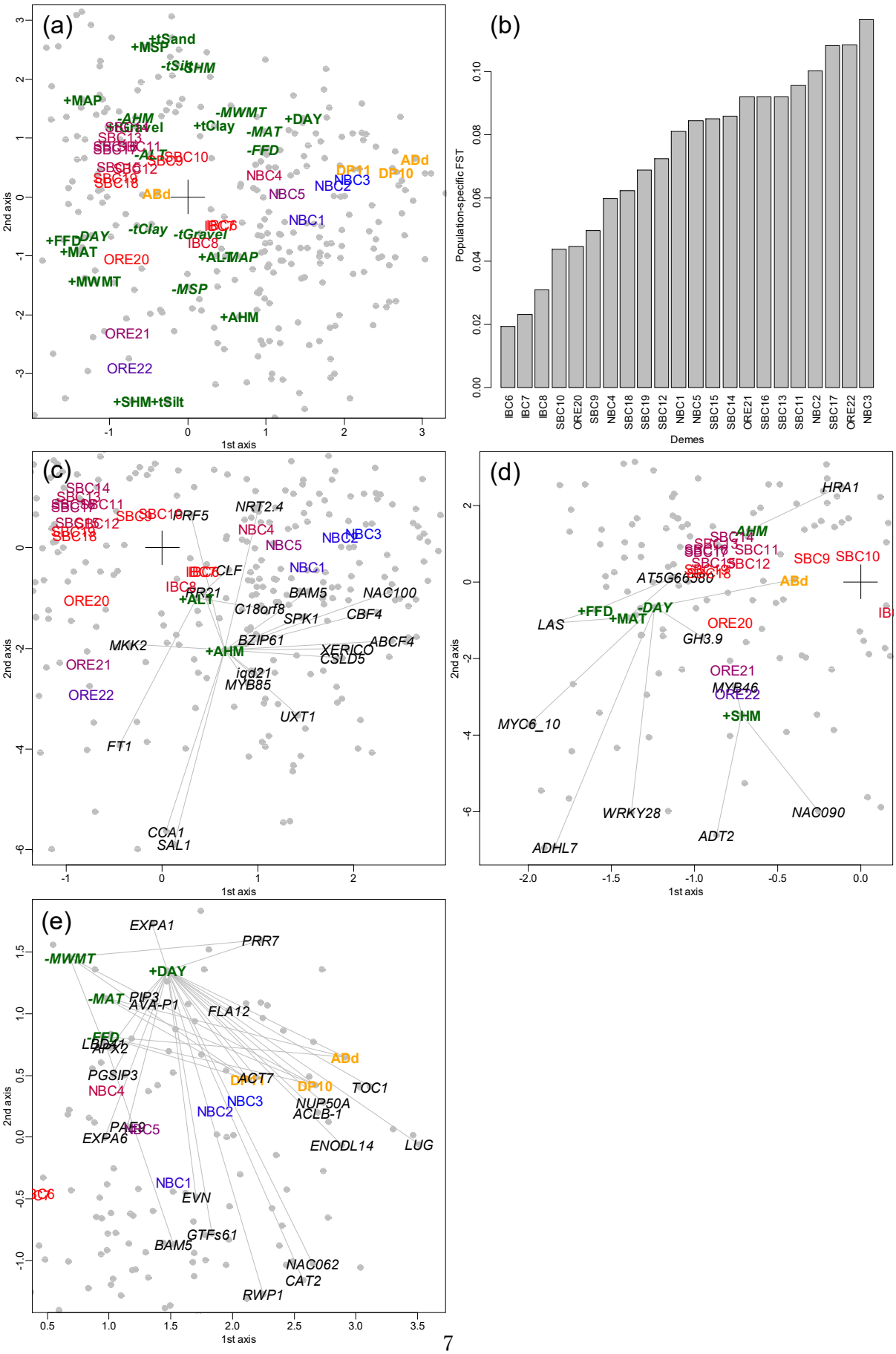
(a)
(b)
(c)
(d)
(e)
1st axis
2nd axis
Population-specific FST
Demes
IBC6
IBC7
IBC8
SBC10
ORE20
SBC9
NBC4
SBC18
SBC19
SBC12
NBC1
NBC5
SBC15
SBC14
ORE21
SBC16
SBC13
SBC11
NBC2
SBC17
ORE22
NBC3
+MSP
+tSand
-tSilt
SHM
+MAP
-AHM
+tGravel
+tClay
-MWMT
+DAY
-MAT
-FFD
ABd
DP11
DP10
NBC4
NBC3
NBC2
NBC5
NBC1
-tClay
-tGravel
IBC8
+FFD
DAY
+MAT
ORE20
+ALT
MAP
+MWMT
-MSP
+AHM
ORE21
ORE22
+SHM+tSilt
SBC14
SBC11
SBC12
SBC18
PRF5
NRT2.4
NBC4
NBC5
NBC3
NBC2
NBC1
CLF
IBC8
RR21
+ALT
BAM5
NAC100
ORE20
C18orf8
SPK1
CBF4
BZIP61
MKK2
+AHM
XERICO
ABCF4
CSLD5
ORE21
iqd21
MYB85
ORE22
UXT1
FT1
CCA1
SAL1
HRA1
AHM
SBC11
SBC12
SBC9
SBC10
AT5G66580
ABd
+FFD
-DAY
+MAT
LAS
ORE20
GH3.9
ORE21
MYB46
ORE22
+SHM
MYC6_10
WRKY28
NAC090
ADT2
ADHL7
EXPA1
PRR7
-MWMT
+DAY
-MAT
PIP3
AVA-P1
FLA12
-FFD
APX2
ABd
PGSIP3
NBC4
ACT7
DP10
TOC1
NBC3
NUP50A
ACLB-1
NBC2
EXPA6
ENODL14
LUG
NBC1
EVN
GTFs61
BAM5
NAC062
CAT2
RWP1

FIGURE 6 Correspondence analysis elucidation of the life history of the western North American wild poplar population. Subpopulations are indicated in the cloud of genes, environments, and traits. Environment labels with plus signs represent the original environmental values, whereas environment labels with minus signs represent the sign-reversed environmental values. Significant associations (FDR = 5%) between the environmental factors and the genes/traits were highlighted by gray lines and labels. The colors of the subpopulations represent low (red) and high (blue) population-specific F_{ST} values. (a) Global picture of population expansion and environmental factors. (b) Population-specific F_{ST} values of the 22 subpopulations. (c–e) Focused picture of the inland habitats (c), coastal habitats (d), and northern habitats (e). ALT (altitude), DAY (longest day length), FFD (frost free days), MAT (mean annual temperature), MWMT (mean warmest month temperature), AHM (annual heat–moisture index), SHM (summer heat–moisture index), tClay (% of clay in top soil), tSilt (% of silt in top soil), tSand (% of sand in top soil), tGravel (% of gravel in top soil). Traits: DP10 (disease progress of leaf rust disease in 2010), DP11 (disease progress of leaf rust disease in 2011), ADd (adaxial stomata density), ABd (abaxial stomata density).

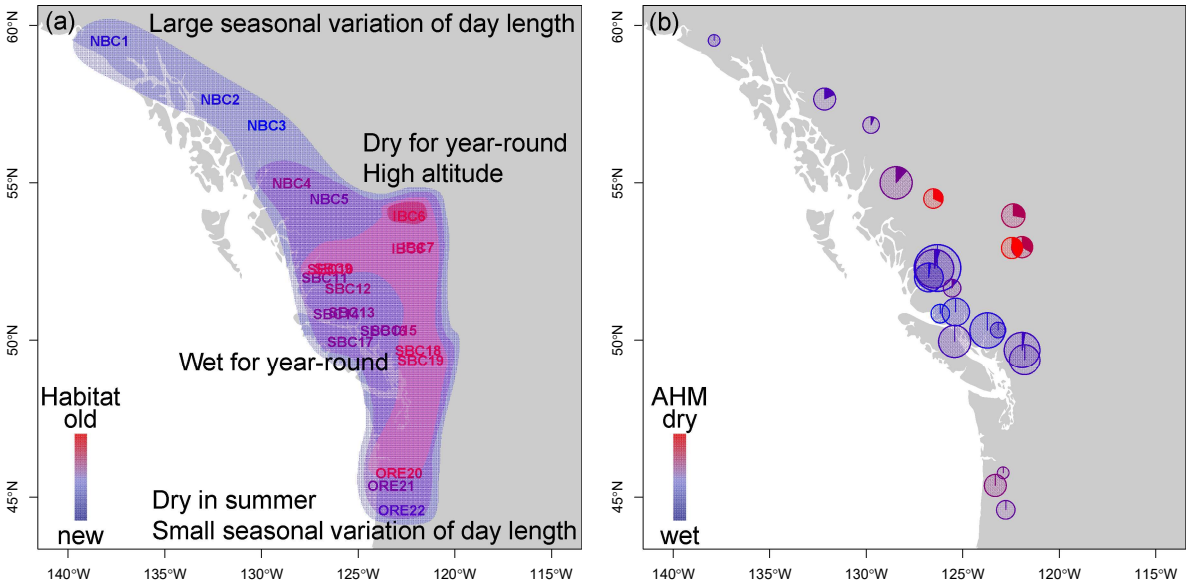

Figure 7 Life history of a population on the geographical map. (a) Range expansion of wild poplar and response to environmental stresses. (b) Response of a drought-tolerance gene (CBF4) as an example. The dark and light parts of the pie charts denote mean frequencies of minor and major alleles, respectively. Gradation of red and blue denotes the annual heat–moisture index. Circle size denotes sample size.