Climatic niche predicts the landscape structure of locally adaptive standing genetic variation

Vikram E. Chhatre[1,2], Karl C. Fetter[1,6], Andrew V. Gougherty[3,7], Matthew C. Fitzpatrick[3], Raju Y. Soolanayakanahally[4], Ronald S. Zalesny Jr.[5], and Stephen R. Keller[*1]

[1]Department of Plant Biology, University of Vermont, Burlington, VT 05405, USA

[2]Wyoming INBRE Bioinformatics Core, University of Wyoming, Laramie, WY 82071, USA

[3]University of Maryland Center for Environmental Science, Frostburg, MD 21532, USA

[4]Indian Head Research Farm, Agriculture and Agri-Food Canada, Indian Head, SK S0G 2K0, Canada

[5]USDA Forest Service, Northern Research Station, Institute for Applied Ecosystem Studies, Rhinelander, WI 54501, USA

[6]Present address: Department of Biological Sciences, University of Calgary, Calgary, AB, Canada T2N 1N4, Canada

[7]Present address: Department of Botany, University of British Columbia, Vancouver, BC, Canada V6T 1Z4,

*Author for correspondence (**srkeller@uvm.edu**)

Keywords: rear edge, local adaptation, population adaptive index, landscape genomics, gene-environment association

Abstract

Within a species' range, intraspecific diversity in the form of adaptive standing genetic variation (SGV) may be non-randomly clustered into different geographic regions, reflecting the combined effects of historical range movements and spatially-varying natural selection. As a consequence of a patchy distribution of adaptive SGV, populations in different parts of the range are likely to vary in their capacity to respond to changing selection pressures, especially long-lived sessile organisms like forest trees. However, the spatial distribution of adaptive SGV across the landscape is rarely considered when predicting species responses to environmental change. Here, we use a landscape genomics approach to estimate the distribution of adaptive SGV along spatial gradients reflecting the expansion history and contemporary climatic niche of balsam poplar, *Populus balsamifera* (Salicaceae), a widely distributed forest tree with a transcontinental distribution in North America. By scanning the genome for signatures of spatially varying local adaptation, we estimated how adaptive SGV has been shaped by geographic distance from the rear range edge (expansion history) versus proximity to the current center of the climatic niche (environmental selection). We found that adaptive SGV was strongly structured by the current climatic niche, with surprisingly little importance attributable to historical effects such as migration out of southern refugia. As expected, the effect of the climatic niche on SGV was strong for genes whose expression is responsive to abiotic stress (drought), although genes upregulated under biotic (wounding) stress also contained SGV that followed climatic and latitudinal gradients. The latter result could reflect parallel selection pressures, or co-regulation of functional pathways involved in both abiotic and biotic stress responses. Our study in balsam poplar suggests that clustering of locally adaptive SGV within ranges primarily reflects spatial proximity within the contemporary climatic niche – an important consideration for the design of effective strategies for biodiversity conservation and avoidance of maladaptation under climate change.

Introduction

Adaptive standing genetic variation (SGV) is of fundamental importance to the response of populations to environmental change. R.A. Fisher[1] first identified that the response to selection is proportional to the genetic variance in fitness, but also assumed that most populations exist at or near their current fitness optima such that selection primarily acts on new mutations as opposed to SGV[2]. However, several processes are known to maintain SGV over broader temporal and spatial scales. For example, fluctuating or balancing selection may favor the long-term persistence of multiple alleles within populations, and this has been particularly well documented for genes associated with mating type[3,4] or immunity[5–7]. Similarly, spatial population structure and local selection can maintain adaptive SGV at range-wide scales, for example due to polygenic selection shifting adaptive allele frequencies along environmental gradients[8] or when hard or soft selective sweeps act locally within different populations[9].

For many species, the current spatial distribution of adaptive SGV will in part reflect historical events such as range contractions and expansions in response to past glaciation, as well as effects of environmental selection during or since these events[10–13]. Hampe & Petit[14] argued that populations distributed along the rear edge of a species' range may be closely descended from refugial populations, and hence are predicted to be older and harbor higher levels of SGV compared to younger populations in areas of recent expansion[14,15]. Reduced SGV with distance from the rear range edge is also a predicted consequence of serial founder effects during range expansion[16], which can lower adaptive potential with distance from rear edge source populations[17]. Moreover, rear edge populations at lower latitudes may be subject to increased levels of abiotic stress, hybridization with congeners, or experience higher disease or pathogen incidence compared to populations in the core of the species' range[14,15,18–20], suggesting that the strength or nature of selection may vary with proximity to the rear range edge[21,22]. However, the effect of range limits on adaptive genetic variation remains poorly understood[13,23], and only a handful of experimental studies have quantified how SGV under selection varies along a continuum from the core to the edge of a species' range (e.g.[24–28]).

While proximity to the rear range edge may predict if SGV for ancestral polymorphism has been reduced by colonization bottlenecks, it may not adequately capture how SGV has been shaped by spatially-varying selection after expansion[29]. In this context, populations inhabiting ecologically marginal environments are of particular interest[13,30]. Marginal environments may impose intense selection pressures on populations to

locally adapt, but such adaptation may be constrained by demographic effects of genetic drift[23,31] or the immigration of maladapted alleles from more abundant populations elsewhere in the range[32]. On the other hand, gene flow of beneficial alleles may facilitate adaptation to marginal environments[33,34], and could contribute substantially to the global pool of adaptive SGV present across a species' range. Identifying local populations containing unique adaptive variation relative to the rest of the range has gained increased importance in the context of contemporary global change[22], as efforts are needed to identify and conserve adaptive germplasm *in situ* for crop breeding[35], or to forecast the adaptive response of natural populations to climate change[36,37].

North American tree species provide excellent study systems for addressing the effects of range context on locally adaptive SGV, as their expansion history into formerly glaciated regions is well documented in the fossil pollen record, and is known to have tracked the changing climate during the Holocene[38,39]. Due to their long individual lifespans, frequently spanning hundreds of years, trees occur in populations characterized by overlapping generations in which SGV for fitness traits can be maintained by demographic or selective processes over eons[40]. Furthermore, many tree species show strong clines across their ranges for functional traits and adaptive allele frequencies that occur across complex spatial gradients of expansion history and climatic environment[27,41–45]. This makes forest trees important systems in which to understand the spatial landscape of adaptive variation[46], and the roles of history and selection in shaping where SGV is concentrated within the range[39].

In this study, we report a landscape genomics analysis investigating the importance of range context for the distribution of locally adaptive genomic variation in the widely distributed boreal tree, *Populus balsamifera* L. (Salicaceae). The occurrence of *P. balsamifera* is spread across 30 degrees of latitude and over 100 degrees of longitude[47], and as such populations span a massive range in climate and growing season length[48]. Previous genetic studies place the likely refugial location of *P. balsamifera* during the last glacial maximum (LGM) in the central Rocky Mountains[49,50], where *P. balsamifera* currently exists in relatively isolated populations along its southern range edge[19,48]. The current center of abundance of the species lies to the north of these rear edge populations across central Canada, where population genetic and species distribution models suggest the presence of a large effective population size (N_e) with high historic levels of gene flow to the northern and eastern peripheries of the range[49]. *Populus balsamifera* is also known to exhibit strong clinal variation for phenology and ecophysiological traits[51–53] and for nucleotide diversity and divergence in flowering time genes, suggesting pervasive local adaptation to climate and photoperiod across its range[43,54].

Here, we use a combination of approaches to scan the genome for signals of local adaptation, conditioned on the spatial pattern of neutral population structure, and show that adaptive SGV varies strongly with proximity to the species' core climatic niche, with weaker effects of distance from the rear range edge. The predominant influence of the contemporary climatic niche shaping adaptive SGV is largely robust across genes involved in defense vs. drought stress, suggesting similar processes have shaped variation in genomic pathways involved in response to biotic and abiotic stress. Our landscape genomics approach gives spatially-explicit insights into where and why SGV is clustered within a species' range, and serves as a guide for developing strategies aimed at identifying and conserving areas of unique genetic diversity and mitigating loss of adaptation under environmental change.

Results

Our GBS libraries for 508 individuals generated 3.9 billion single-end reads passing filter, of which 1.35 billion passed QC in the GBS pipeline as "good, barcoded reads" *sensu* Glaubitz et al.[55], consisting of a perfect match to one of the barcodes and no ambiguous base calls. Reads passing QC clustered into 3,755,656 sequence tags that mapped uniquely to 220.9 Mb of the *P. trichocarpa* reference genome v3.0. Variant calling produced 1,107,538 sites which, after filtering, resulted in 167,324 biallelic SNPs for inference of population structure and selection, with a median (across individuals) per-site depth of 21.8X.

Population Structure

Inference of population structure via discriminant analysis of principal components (DAPC) revealed support for 3 to 4 genetically distinct clusters, with mean BIC over the 1000 iterations minimized at $K=3$ (Fig. S1 A, B). Clustering assignments at $K=3$ separated individuals into a northwestern cluster in Alberta and British Columbia, a central cluster throughout the range core from Saskatchewan to Ontario including the Great Lakes region, and an Atlantic Canada cluster along the eastern coast (Fig. 1C inset). In about 10% of the iterations, DAPC favored a model of $K=4$ that showed a slightly higher BIC averaged across iterations compared to $K=3$. At $K=4$, the dominant solution split the range core into eastern and western clusters that formed distinct groups but were only weakly differentiated ($F_{ST} = 0.008$) (Fig. 1C).
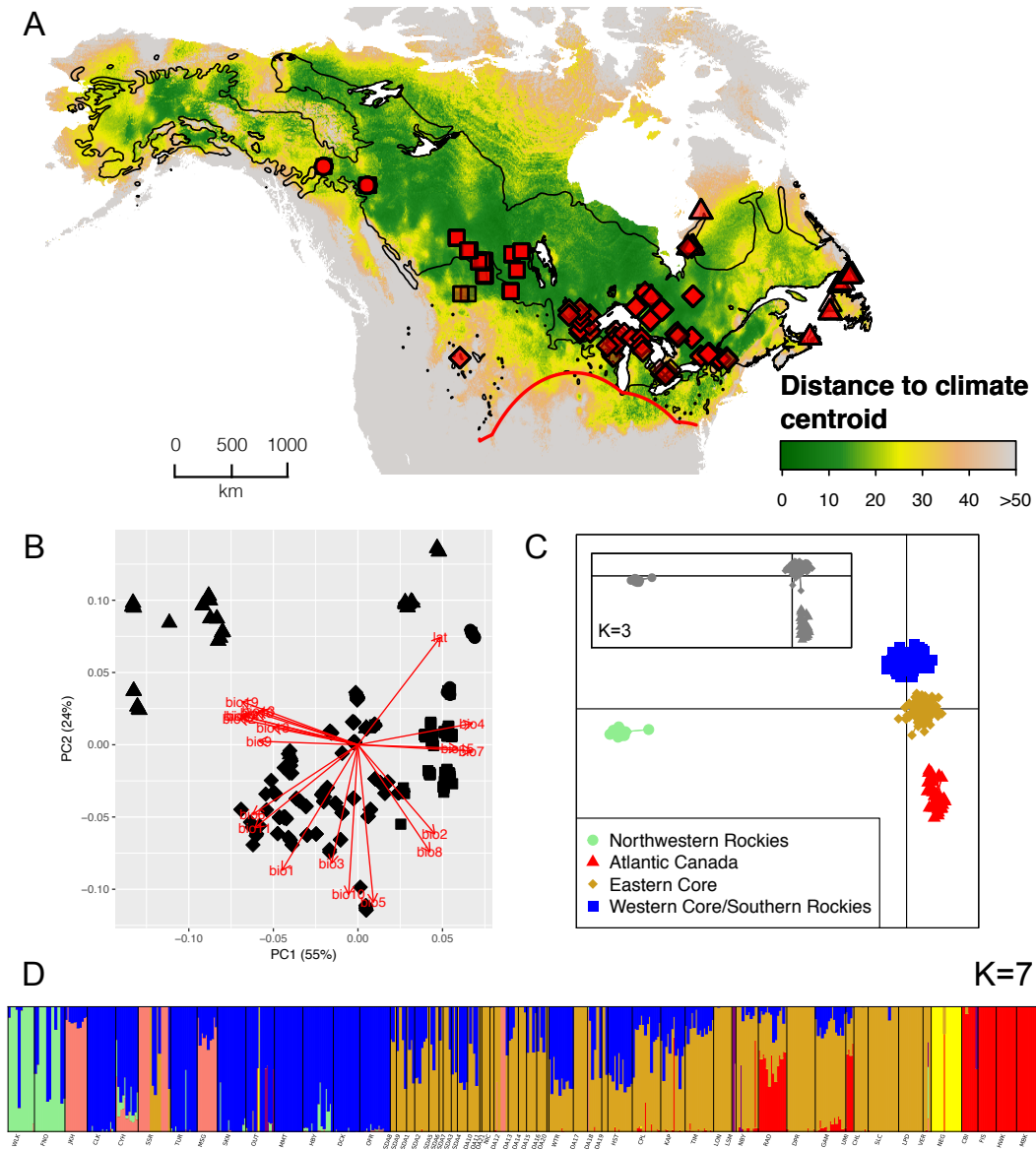
Figure 1 Sample collection map of the 437 individuals analyzed in selection scans. The red line represents the estimated location of the southern range edge based on occurrence records (A). The range boundary of *P. balsamifera* is denoted by the black line and is adapted from[47]. 19 BIOCLIM variables and latitude were included in a PCA and PC1 and PC2 were used in gene-environment analyses (B). Discriminant analysis of principal components (DAPC) at K=4 separates Northwestern Rocky Mountains from populations east of Saskatchewan along DAPC axis 1. DAPC axis 2 separates an Atlantic Canada clusters and two range core clusters. Shapes in the sample map, PCA plot correspond to the DAPC (C). ADMIXTURE clustering of all 508 individuals sequenced in the study at K=7, the level of K with the lowest cross validation error. Colors represent the following clusters: blue & goldenrod: range core, lightgreen: Rocky Mountain region, red: Atlantic Canada, yellow: New Glasgow, NS, Canada with introgression from *P. deltoides*, pink: admixture from Northwestern cluster reported by Keller *et al*[49]. Individuals with less than 90% *P. balsamifera* ancestry were removed prior to performing selection scans.

The genetic clusters from DAPC aligned along major gradients of climate and growing season length, as captured by the first two principal components of the climate PCA (Fig. 1B). Large values of PC1 represented dry environments with high temperature and precipitation seasonality, and were primarily occupied by the western core and northwestern Rockies groups, while small values of PC1 (wet, low seasonality) were occupied by individuals with ancestry in the Atlantic Canada cluster. PC2 represented a north-south gradient of growing season length, with large values representing high latitude, cool sites occupied by the northwestern Rockies and a subset of the Atlantic Canada clusters, while small values of PC2 (low latitude, warm sites) were dominated by members of the eastern core cluster (Fig. 1B). Ancestry clusters were also non-randomly distributed with respect to distance from the rear range edge and proximity to the species-wide climatic niche centroid (Fig. 1A). Notably, the Atlantic Canada cluster was farthest from the climatic centroid while spanning a range of distances from the rear edge. In contrast, the western and eastern core clusters were close to the climatic centroid, with many eastern core populations also proximate to the rear range edge (Fig. 1A).

ADMIXTURE analysis largely corroborated the population structure found by DAPC, but also revealed extensive mixed ancestry among *P. balsamifera* clusters, as well as introgression with other *Populus* species (Fig. 1D). At lower levels of K (2–3), introgression was especially prevalent with *P. trichocarpa* along the rear range edge and in the northwestern Rockies (Fig. S2; see also Chhatre *et al*[19]). At K=4, a separate source of introgressed ancestry was apparent in the Atlantic Canada population of NEG, representing hybridization with *P. deltoides*[20]. Scattered introgression was also apparent at a low frequency throughout other portions of the range. At higher levels of K (5–6), ancestry within *P. balsamifera* separated into an Atlantic Canada cluster and a subdivided central range core split into eastern and western clusters, as observed in DAPC (Fig. 1D). Individual ancestry assignments to the eastern and western range core revealed a longitudinal cline, indicating extensive shared ancestral polymorphism and/or gene flow between these two weakly differentiated groups. At K=7, the best model supported by ADMIXTURE cross-validation, an additional distinct ancestry cluster was evident in the far northwestern Rockies (WLK and FNO; Fig. 1C).

Genomic Signals of Local Adaptation

After eliminating individuals with less than 90% *P. balsamifera* ancestry (N=26 individuals; Fig. 1D), we retained 437 individuals and 129,251 SNPs to test for genomic signatures of local adaptation, with 67% of SNPs within 5 kb of an annotated gene model. Genome scans identified 397 outlier SNPs in tests for elevated
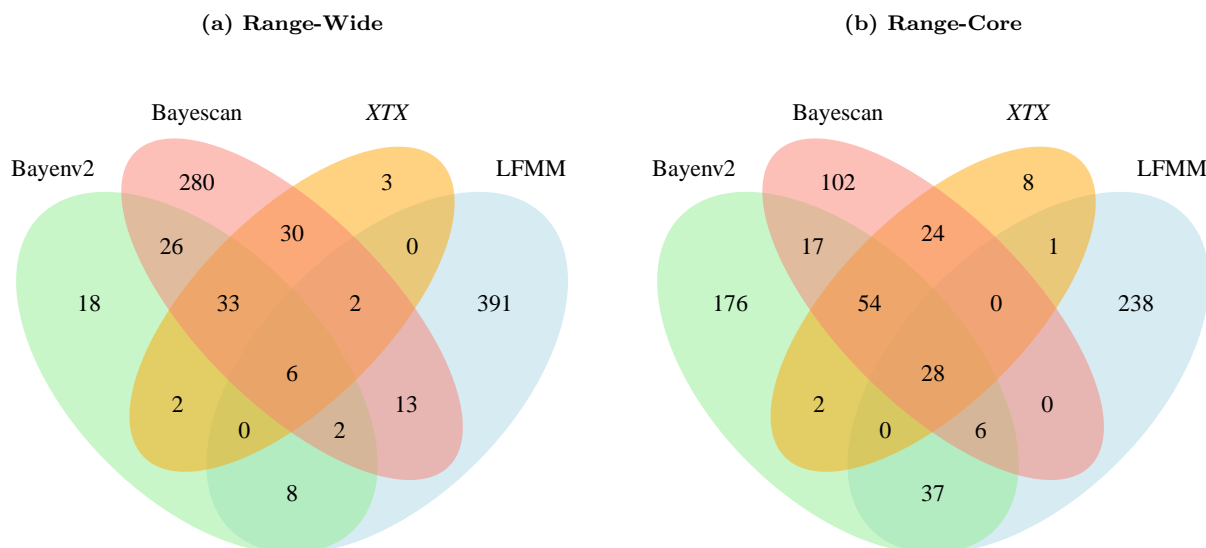
Figure 2 Venn diagrams showing overlap of local adaptation outliers among four selection scans using the Range-Wide and Range-Core samples. Bayenv2 and LFMM test for gene-environment associations (GEA), whereas Bayescan and $X^T X$ test for elevated population differentiation.

population differentiation (Bayescan and $X^T X$) that exceeded the empirical Type I error rate based on the distribution of intergenic SNPs ($\alpha=0.0006$) – an enrichment of more than 2.5X over neutral expectations. Most differentiation outliers were significant only in Bayescan (321/397 loci, 81%) or to a lesser extent, in $X^T X$ (5 loci, 1%), while 71 outliers (18%) overlapped between the two differentiation tests (Fig. 2a). Gene-environment association (GEA) tests (LFMM and Bayenv2) returned 535 outliers associated with either climate PC1 or PC2, more than 1.7X greater than neutral expectations. Similar to differentiation tests, most outliers in GEA tests were sensitive to the method used to detect them, with 16 of 535 loci (3%) in the range-wide set shared between Bayenv2 and LFMM (Fig. 2a). Combining across all four genome scan approaches, we identified 814 loci within 5kb of 724 genes that showed evidence of either elevated differentiation or gene-environment association relative to the empirical neutral distribution, with most loci unique to a single analysis (85%) and only 6 loci (0.7%) shared across all four methods.

Outliers unique to a single method may be false positives caused by inadequate control of background structure due to residual relatedness or departure from the assumed demographic history. Alternatively, these loci may be true positives that are not detected by the complimentary GEA or differentiation method, for example if the statistical modeling of background genetic structure over-corrects the allele frequencies and dilutes the

selection signal. To explore this possibility further, we performed each selection scan again on just those populations belonging to the eastern and western range core clusters, where background population structure was minimal (Fig. 1; N=336 individuals; $F_{ST} < 0.01$). This greatly increased the frequency of shared outliers among differentiation methods to 44% (106/242 loci; Fig. 2b), and confirmed several prominent genomic regions of high differentiation in both the range-wide and core sets (Fig. 3). These genomic regions included a large peak encompassing ca. 168 kb on the distal arm of chromosome 3, a second prominent peak spanning ca. 369 kb on chromosome 4, and an outlier marked by a single SNP on chromosome 8 (Fig. 3).
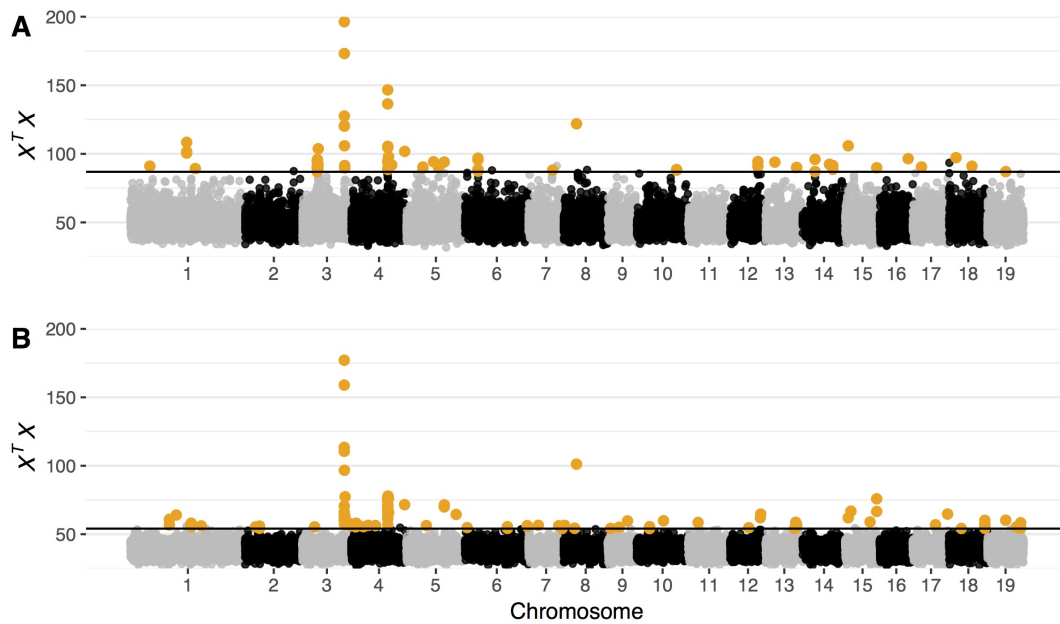

Figure 3 Manhattan plots of $X^T X$ outlier scan for the set of range-wide individuals (A) and the range core individuals (B). Highlighted in each panel are outlier loci that were also identified in Bayescan; 106 loci in panel A and 71 in panel B. The horizontal black line indicates the -$\log_{10} P$-value significance threshold based on the empirical neutral distribution

A more modest increase in the frequency of shared outliers was evident between GEA methods performed on the range core set (13%; 71/559 loci); however, the number of outliers identified by Bayenv2 increased dramatically, suggesting the selection signal in the range-wide set likely contained false negatives due to over-correction of population structure (Fig. 2b). Genomic regions of shared outliers in the range-wide set were limited to a few scattered chromosomal regions, whereas in the range core, shared outliers were clustered into prominent peaks on chromosomes 3, 4, 8, 12, 15, and 17 (Fig. 4).
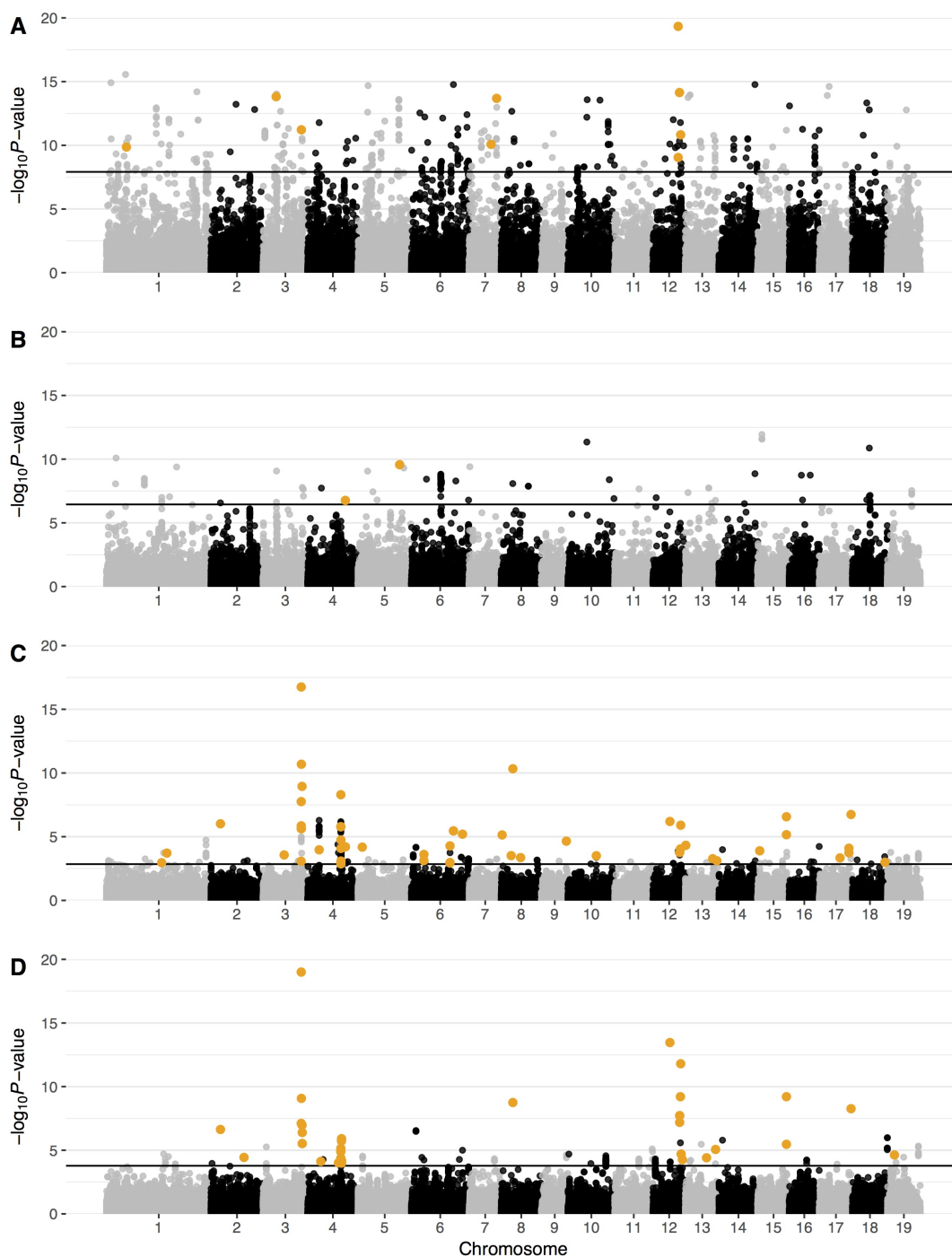
Figure 4 LFMM results for gene-environment association tests for the range-wide (A,B) and range core (C,D) sample sets with climate PC1 (A,C) and PC2 (B, D). Loci in orange were also significant in the Bayenv2 gene-environment test. The horizontal black line indicates the -$\log_{10}P$-value significance threshold based on the empirical neutral distribution

Across all four methods, inference of loci under local adaptation appeared to be sensitive to the genome scan approach (differentiation vs. GEA) and the specific method used. Despite multiple controls in place for background genetic structure, false positives are inevitably part of these differences. However, the genetic architecture of local adaptation is likely to be dominated by many loci of small effect that exhibit weak allele frequency (co)variance among populations, making it difficult to distinguish these loci from the genomic background.

Additionally, some loci likely experienced selection along portions of the climatic gradients that run parallel with the neutral genetic structure (e.g, climate PC2 and the Atlantic Canada cluster; Fig. 1). Accordingly, we retained two sets of outliers for further investigation: (1) an inclusive set of all 814 outlier loci identified by one or more selection scan methods in the range-wide analyses, and (2) only those outliers corroborated by both methods within a given genome scan approach (differentiation or GEA), resulting in a reduced set of 203 loci. The former is more likely to retain true positive cases of weakly selected loci or those spatially confounded with neutral structure at the expense of including more false positives, while the latter is less likely to contain fewer false positives but will be biased towards loci of large effect (e.g., large allele frequency differences) that are not closely aligned with neutral population structure.

Distribution of adaptive SGV along geographic and climatic range edges

We investigated the spatial turnover in SGV by determining the relative contribution of individual populations to the adaptive variation we observed across the entire set of populations. For each population, we estimated the population adaptive index (PAI) of Bonin *et al.*[56]:

$$PAI_j = \sum_{i=1}^{n} |p_{ij} - \bar{p}_i|$$

where p_{ij} is the minor allele frequency of the ith outlier locus in the jth population, and $\bar{p}_i$ is the mean frequency across all populations. Populations with high values of PAI contribute disproportionately to the pool of adaptive allele frequencies; while populations with low PAI have allele frequencies closer to the range-wide average. To assess how this among-population component of SGV varied across the landscape, we modelled PAI as a function of distance from the rear edge and the climatic niche centroid. When considering the inclusive set of 814 selection outliers as an index of genomic beta diversity, over 75% of the variability in PAI was

explained by jointly accounting for geographic distance from the rear edge and the euclidean distance from the climatic niche centroid (Table 1). Contrary to predictions based on expansion history[12,14], PAI showed little effect of distance from the rear range edge except through an interaction with climate distance, which revealed a dramatic increase in PAI with combined distance away from both the rear edge and climatic niche centroid, consistent with divergent local selection in climatically marginal environments that are geographically located far from the current rear range edge (Fig. 5a). Driving this trend were large values of PAI for populations in Atlantic Canada that inhabited the easternmost portion of the species' range – a region climatically distinct from the majority of the species' niche (Fig. 1A, B; Fig. 5a). The overall trend in PAI was similar for the large effect outliers (N=203 loci), although this set of loci showed some rear edge populations exhibiting slightly elevated levels of PAI but still far below levels seen in marginal climates (Table 1; Fig. 5b).

Table 1 Response of population adaptive index (PAI) to distance from the rear range edge and from the climatic niche centroid. (a) The inclusive set of significant outliers identified across all four selection scan methods; (b) large effect outliers corroborated by both GEA-based tests and/or both F_{ST}-based tests; (c,d) outliers proximate to (¡5kb) genes differentially expressed in response to (c) biotic (beetle or mechanical) or (d) abiotic (drought) stress. Values are coefficients (SE) from multiple regression.

	(a) Inclusive	(b) Large effect	(c) Biotic stress	(d) Abiotic stress
N SNPs	814	203	48	174
Rear edge	-1.71e-08 (1.26e-08)	-3.10e-08 (2.12e-08)	2.30e-08 (1.51e-08)	-1.64e-08 (1.31e-08)
Climate centroid	-2.16e-03 (6.53e-04)**	-3.26e-03 (1.10e-03)**	-7.40e-04 (7.75e-04)	-2.33e-03 (6.89e-04)**
Rear edge:climate centroid	3.46e-09 (5.67e-10)***	4.40e-09 (9.51e-10)***	1.20e-09 (6.73e-10)†	3.35e-09 (5.91e-10)***
Model R^2	0.772	0.593	0.547	0.733

$^\dagger P < 0.1$, $^* P < 0.05$, $^{**} P < 0.01$, $^{***} P < 0.001$

The historic and selective processes shaping the geographic distribution of SGV may differ for adaptive variation underlying responses to abiotic vs. biotic stress[13,18]. To test this prediction, we calculated PAI separately for outlier SNPs that tagged genes showing differential expression under controlled drought (abiotic) or beetle/mechanical wounding (biotic) based on the gene expression atlas on the Populus Genome Integrative Explorer (PopGenIE)[57]. Among the 724 genes associated with the set of inclusive outliers, 105 genes were differentially expressed under drought (N=174 outlier SNPs), and 30 were differentially expressed under beetle or mechanical wounding (48 outlier SNPs). Both types of stressor showed a strong increase in PAI towards

climatically marginal range edges, although the effects were more pronounced for loci associated with abiotic stress response (Table 1, Fig. 5c,d). Interestingly, PAI for biotic stress outliers increased consistently with distance from the rear edge in a trend that was not evident among drought-associated loci (Fig. 5c).
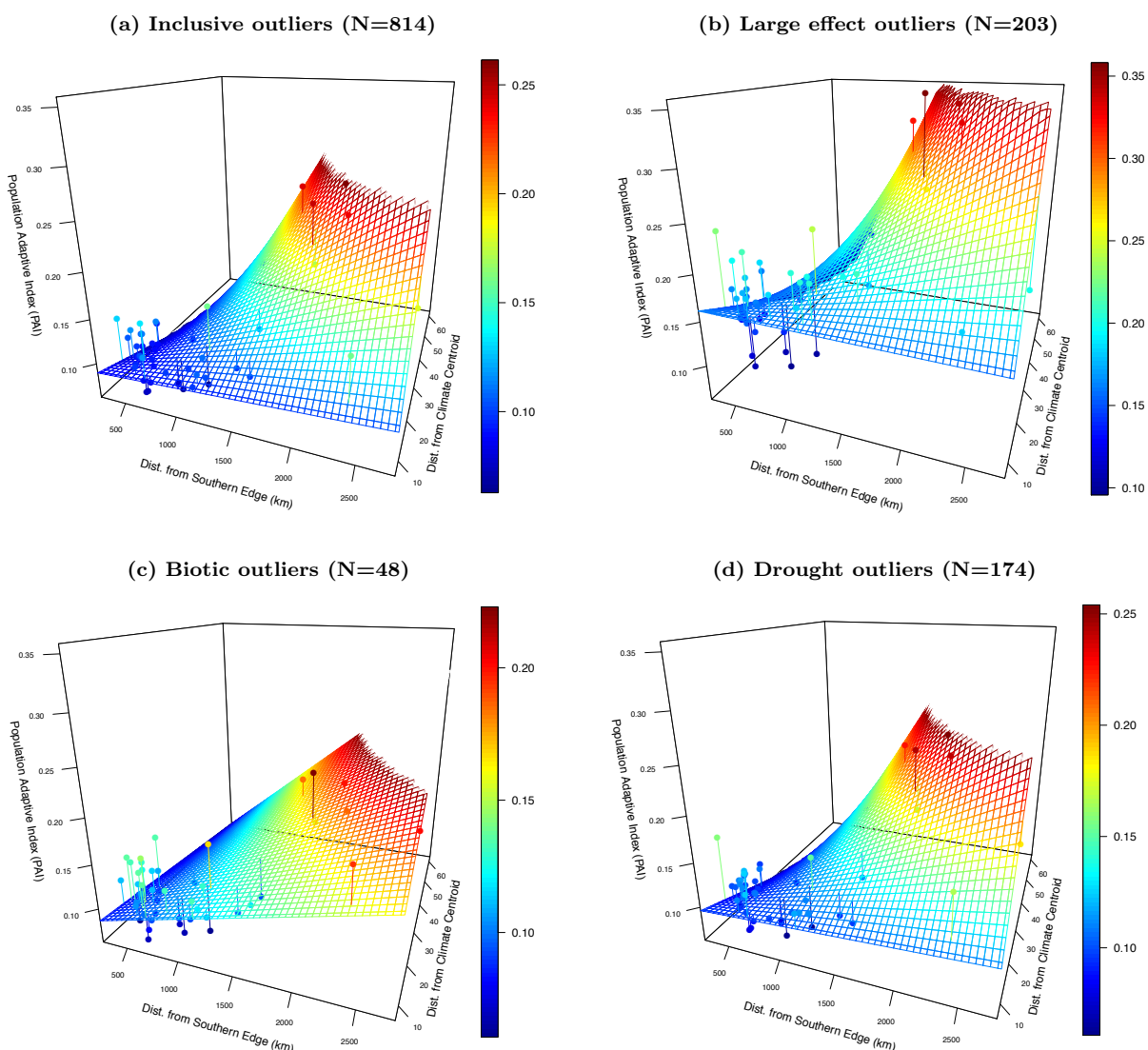
Figure 5 Predicted model fit surface of the relationship between population adaptive index (PAI) and two predictor distances (distance from the rear range edge and distance from the climate centroid). PAI estimates are summed over all outliers and normalized for comparison between data sets (see definitions in Table 1). Surface colors show PAI magnitude from low (blue) to high (red).

As an index of genetic beta diversity, PAI identifies populations whose allele frequencies deviate most from

the sample-wide mean, and thus this estimator is inherently on the *among*-population component of SGV. To investigate the effects of range position on *within*-population SGV (i.e., alpha diversity), we estimated the variance in allele frequencies at biallelic sites ($pq=p^*(1-p)$) for each population averaged across outlier loci. The inclusive set of 814 outliers showed a highly significant interaction between rear edge and climatic distances (Table 2), with pq for locally-adaptive alleles highest with increasing distance from the rear edge but close to the center of the climatic niche (Fig. 6a).

Table 2 Response of allele frequency variance (pq) averaged over outlier loci to distance from the rear range edge and from the climatic niche centroid. Variable definitions and outlier groups are as described in Table 1

	(a) Inclusive	(b) Large effect	(c) Biotic stress	(d) Abiotic stress
N SNPs	814	203	48	174
Rear edge	2.02e-08 (4.10e-09)***	9.79e-09 (8.79e-09)	1.41e-08 (5.46e-09)*	2.13e-08 (4.20e-09)***
Climate centroid	5.46e-04 (2.11e-04)*	4.17e-04 (4.55e-04)	-1.22e-04 (2.83e-04)	3.52e-04 (2.17e-04)
Rear edge:climate centroid	-7.63e-10 (1.83e-10)***	-6.67e-10 (3.95e-10)†	-2.57e-11 (2.46e-10)	-7.31e-10 (1.89e-10)**
Model R^2	0.386	0.111	0.388	0.466

$^{†}P < 0.1$, $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$

For large effect outliers and those associated with drought stress, allele frequency variance was lowest on average for climatically distant populations where PAI was high, suggesting selection has depleted SGV locally (pq) while increasing the genetic distinctiveness (PAI) of climatically marginal populations (Fig. 6). In contrast, within-population SGV for outliers associated with biotic stress increased significantly with distance from the rear edge (Table 2), but showed no consistent effect of position along the climate gradient (Fig. 6).

The effects of distance from the rear range edge and climatic niche on components of adaptive SGV are evident when population values of pq and PAI are mapped onto the landscape (Fig. 7). With the exception of the biotic outliers, the overall trend showed higher within-population SGV and lower among-population SGV nearer to the geographic and climatic center of the range, per theoretical expectation of the ecological and evolutionary processes structuring range limits[58]. Two exceptions to this general trend were evident. First was the clearly elevated among-population SGV seen in the Atlantic Canada populations, and that featured prominently in all outlier sets. Second was the unusual combination of SGV associated with biotic stress response genes, in which Atlantic Canada populations harbored high levels of SGV both within- and among-population (Fig. 7).
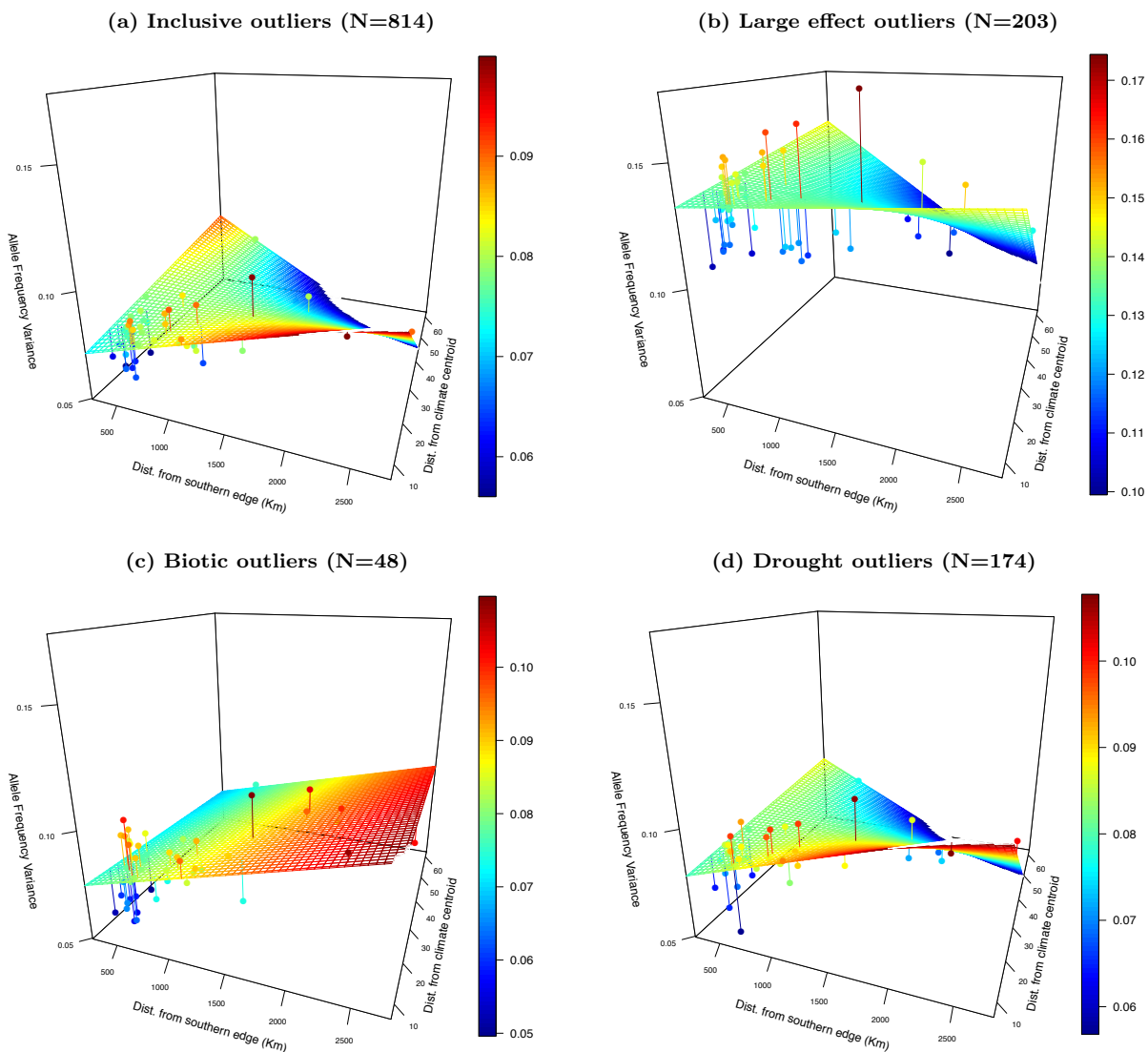
Figure 6 Predicted model fit surface of the relationship between allele frequency variance (*pq*) and two predictor distances (distance from the rear (southern) range edge and distance from the climate centroid). Panels (a) through (d) correspond to categories of outliers decribed in Figure 5. The z-axis is scaled to span the range of allele frequency variance observed across all four comparisons (min=0.04; max=0.17). Surface colors indicate strength of allele frequency variance from low (blue) to high (red).

(a) Inclusive outliers (N=814)

(b) Large effect outliers (N=203)

■ High AFVAR
■ Low AFVAR
○ High PAI
· Low PAI

(c) Biotic outliers (N=48)

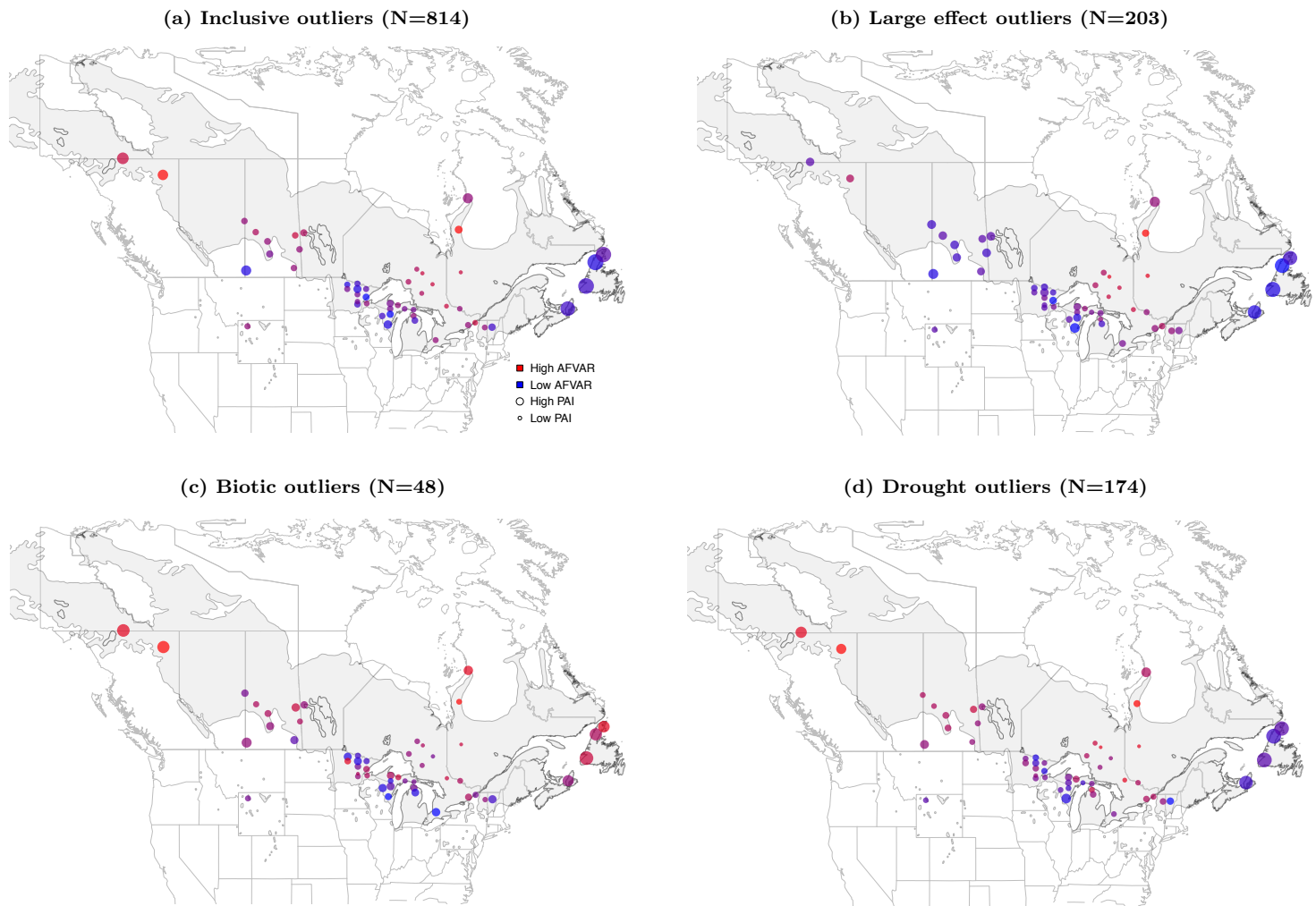**(d) Drought outliers (N=174)**

Figure 7 Landscape level comparison of the relationship between allele frequency variance (color: blue to red) and population adaptive index (size: small to large) among four groups of selection scan outliers. PAI point sizes are not comparable between individual subplots.

Functional enrichment of adaptation outliers

We found 52 gene ontology (GO) terms significantly enriched among genes associated with the 814 outlier set (Supplementary Information). These included clusters of terms relating to enzyme transferase activity, biosynthesis of cell wall components such as cellulose and lipids, signal transduction and ROS activity, response to stress and/or DNA damage repair, and transport across cell membranes (Fig. 8). Genes associated with the 203 outlier set showed significant enrichment for 28 GO terms including many similar functions as the 814 outliers set, but with the addition of several terms associated with vessicle transport within and across cell membranes, and response to foreign bodies (Supplementary File S1).

Several of the enriched GO terms were associated with candidate genes tagged by SNPs that were selection outliers in all four genome scan methods in the range-wide and core sets (Table 3) These included candidate genes within the selected region on chromsome 3: the flavonoid biosynthesis gene CHALCONE SYNTHASE, CELLULOSE SYNTHASE-LIKE D1.1, and two FERRODOXIN genes. Chromosome 12 also contained candidate genes with enriched GO terms associated with DNA repair: an ATPase-like family protein (Potri.012G096300), RESPIRATORY BURST OXIDASE HOMOLOG C, a Tyrosyl-DNA phosphodiesterase, and a UvrD DNA helicase. Chromosomes 15 and 17 each contained a single gene with enriched GO terms: the MADS-box transcription factor TRANSPARENT TESTA 16, and the stress-responsive molecular chaperone HEAT SHOCK PROTEIN 90-1 (Table 3).

Discussion

The current period of rapid global climate change has created an immediate need to understand the genetic basis of climate adaptation in natural populations. For long-lived sessile organisms such as forest trees, adaptation to ongoing climate change will depend in large part on the availability of adaptive standing genetic variation (SGV) within the species, as well as the distribution of SGV across a potentially expansive landscape. Numerous studies have shown the post-glacial biogeography of forest trees shaped SGV over millennia, involving complex cycles of range contraction and expansion, and also local adaptation along climatic gradients[10,14,40,41]. But while forest trees have become important models for understanding the genetic basis of climate adaptation at landscape scales[46,59,60], few studies have explicitly addressed how adaptive SGV is partitioned along spatial gradients reflecting expansion history and current climatic niche[13].
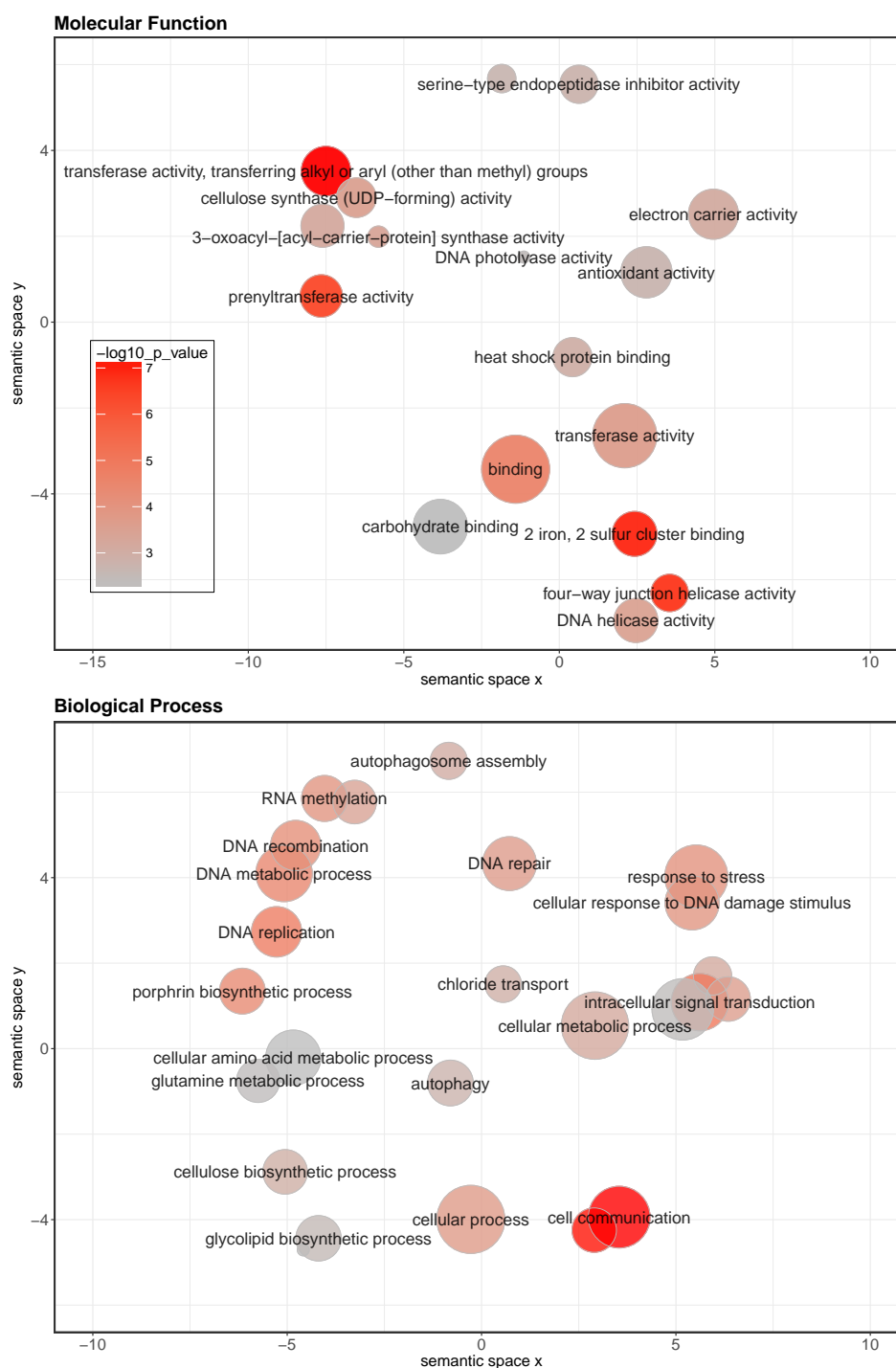
Figure 8 Redundancy clustering of functionally enriched GO terms associated with the 814 loci significant in the range-wide local adaptation scans. Color indicates significance of the enrichment for each GO term, and symbol size is proportional to the frequency of each GO term in the *P. trichocarpa* v3.0 genome annotation

Table 3 Candidate genes for local adaptation in range-wide and range core datasets. Annotated genes are located within 5kb of outlier SNPs identifed by all four selection scan methods.

Gene	Chr	Description	Enriched Gene Ontology (GO) terms	*A. thaliana* homolog
Both sets:				
Potri.003G176500	Chr03	RING/U-box superfamily protein		AT1G04790
Potri.003G176600	Chr03	Dehydratase family protein		AT3G23940
Potri.003G177800	Chr03	CELLULOSE SYNTHASE-LIKE D1	GO:0016760-cellulose synthase activity	AT2G33100
Potri.003G177900	Chr03			AT4G13690
Potri.008G089700	Chr08	PtMYB186 MYB-like DNA binding protein		AT3G01140
Range-wide only:				
Potri.007G082600	Chr07			
Potri.007G082700	Chr07	Cold-regulated protein, similar to LATE EMBRYOGENESIS ABUNDANT (LEA)		AT2G42560
Potri.012G096300	Chr12	AAA family ATPase-like family protein	GO:0009378-Holliday junction helicase activity; GO:0006281-DNA & repair; GO:0006310-DNA recombination	AT4G02480
Potri.012G096400	Chr12	Similar to translational ELONGATION FACTOR 1 SUBUNIT B BETA (EF1B)		AT2G18110
Potri.012G096500	Chr12	Hypothetical protein		AT5G63440
Potri.012G096600	Chr12	Similar to 60S ribosomal protein L17-1		AT1G27400
Range core only:				
Potri.002G068100	Chr02			
Potri.002G068200	Chr02	TUBBY-LIKE PROTEIN 5		AT1G43640
Potri.003G176400	Chr03	2Fe-2S FERREDOXIN-LIKE family protein	GO:0009055-electron carrier activity; GO:0051536-iron-sulfur cluster binding	AT1G32550
Potri.003G176700	Chr03	CHALCONE SYNTHASE 4	GO:0004315-3-oxoacyl-[acyl-carrier protein] synthase activity	AT5G13930
Potri.003G182000	Chr03	NAD(P)-binding Rossmann-fold superfamily protein		AT4G27760
Potri.003G182100	Chr03	FERREDOXIN 2	GO:0009055-electron carrier activity; GO:0051536-iron-sulfur cluster binding	AT4G21090
Potri.004G138000	Chr04	MYB4 transcription factor		AT4G38620
Potri.004G138200	Chr04	Sec14p-like phosphatidylinositol transfer family protein		AT1G72160
Potri.004G138400	Chr04	Ferrochelatase 2 / Protoheme ferro-lyase		AT2G30390
Potri.004G138500	Chr04	Hypothetical protein; homology to UDP-Glycosyltransferase		
Potri.004G140300	Chr04	Protein kinase domain / Subtilase family Peptidase S8		AT3G19300
Potri.006G174800	Chr06			AT5G41980
Potri.006G175100	Chr06			AT5G57700
Potri.008G008000	Chr08	Pt-RPL7.3 60S ribosomal protein		AT3G13580
Potri.008G008100	Chr08	AUTOINHIBITED Ca(2+)-ATPase 9		AT3G21180
Potri.012G111600	Chr12	Similar to RESPIRATORY BURST OXIDASE HOMOLOG C (AtRBOHC)	GO:0009055-electron carrier activity	AT5G51060
Potri.012G111700	Chr12			
Potri.012G111800	Chr12	Tyrosyl-DNA phosphodiesterase	GO:0006281-DNA repair; GO:0005634-nucleus	AT5G07400
Potri.012G114100	Chr12	UvrD DNA helicase	GO:0006281-DNA repair	AT4G25120
Potri.012G114200	Chr12	Protein disulfide isomerase / S-S rearrangase		AT5G61670
Potri.015G125500	Chr15	MADS-box transcription factor, TRANSPARENT TESTA 16	GO:0005634-nucleus	AT5G23260
Potri.015G125600	Chr15	F28K20.6 protein related		AT5G44860
Potri.015G125700	Chr15	Calmodulin-binding family protein		AT3G58480
Potri.017G146600	Chr17	HEAT SHOCK PROTEIN 90-1	GO:0006950-response to stress	AT5G52640
Potri.017G146700	Chr17	RNA binding Plectin/S10 domain-containing protein		AT4G25740, AT5G52650

Our approach fills an important gap in identifying geographic regions where adaptive genomic diversity is concentrated within the range, while also revealing the relative importance of historic and selective processes shaping the current distribution of SGV[13,29,61,62]. Our findings in a widespread forest tree species suggest that adaptive SGV clusters non-randomly along geographic and climatic gradients, with decreasing alpha diversity and increasing compositional turnover (i.e., beta diversity) with distance away from the center of the current climatic niche (Fig. 5 and 6). Contrary to expectations of loss of diversity due to founder effects during expansion from refugia, we observed relatively low SGV in low latitude populations[14]. These findings strongly suggest that regionally common SGV is maintained in the climatic center of the range, while climatically marginal populations are unique in harboring locally reduced but regionally rare adaptive variation.

In many ways, our findings mirror the Centre-Periphery Hypothesis (CPH), that predicts within-population genetic diversity should be highest near the abundant center of the range, while population differentiation should increase towards the range periphery where environments become less optimal and abundance drops[13,58]. Gougherty et al.[63] also found support for the CPH in poplar, using a larger range-wide sample of populations and focusing exclusively on the alpha diversity of putatively neutral SNPs sampled randomly from the genomic background. This suggests that both neutral and adaptive genomic diversity follows the predictions of the CPH because both are jointly influenced by Ne. That is, we might expect higher alpha and lower beta diversity across the genome where the region of largest Ne (which presumably reflects the demographic abundance of the species) is centered with the species' environmental niche. This assumes the major selective influences shaping range limits are climatic in origin, thus aligning climate-adaptive variation with spatial locations of demographic abundance; however, this expectation might break down for genes under strong selection for functional responses not directly related to climate[13]. The majority of the genomic outliers for adaptation studied here fit this prediction, with high alpha diversity (pq) and low beta diversity (PAI) near the geographic and climatic center of the range, consistent with the CPH.

One notable exception occurred for outliers associated with response to biotic stress, which showed alpha diversity increased monotonically away from the rear edge, while beta diversity also increased with distance from the climatic niche (Fig. 6). Thus, genes associated with defense response appear to harbor regionally high levels of diversity in the same climatically marginal region of Atlantic Canada, while also maintaining high levels of within-population polymorphism for these genes. High levels of within-population polymorphism that are also regionally unique could be caused by several evolutionary processes, such as a history of spatial

or temporally varying balancing selection, or selection-driven introgression from congeners. Both processees would be expected to lead to high levels of polymorphism within populations that may also be unique to the region where the selection or introgression is occurring. Regardless of the mechanism, the high alpha and beta diversity for biotic response in Atlantic Canada suggests the selective history on defense-related genes may be unique in this portion of the range.

The Atlantic Canada populations (in our study, populations CBI, HWK, MBK & FIS) occupy a maritime climate quite different from the rest of the range, characterized by cool temperatures and high precipitation amounts with low seasonality[48]. We previously showed that balsam poplar in this region were divergent from the rest of the range for a suite of phenology and ecophysiological traits associated with climate adaptation, and that this phenotypic divergence was in excess of that predicted by background genetic structure, suggesting a unique history of local selection in Atlantic Canada[52]. In a separate study of *P. balsamifera*, Meirmans *et al*[64] similarly reported a genetically distinct Atlantic Canada group that showed little admixture with the nearby range core ancestry group to its west, which they suggested was likely rooted in the local adaptation of the Atlantic Canada group to the unique maritime climate. Consistent with this proposition, the PAI for most of the local adaptation outliers in our study peaked in the Atlantic Canada group, whereas within-population allele frequency variances were lowest in this group, suggesting local selection driving SGV adaptive under these climate conditions towards fixation (Fig 5b, d & 6b, d).

It is important to note that landscape genomic studies of local adaptation invariably capture other selective gradients that are geographically aligned with climate; thus, what we infer as genes under climate-driven selection based on associations with temperature or precipitation gradients may actually reflect other unknown causal agents of selection. In our range-wide analysis, we found outlier genomic regions were significantly enriched for a variety of functions, including oxidoreductase activity, response to stress/DNA damage repair, cell signaling/transport, metal (iron) binding, and biosynthesis of cellulose, lipids, fatty acids, and flavonols (Fig. 8). Many of these enriched functions suggest a shared genetic architecture of response to abiotic and biotic stress, mediated through the action of heme oxygenases such as the diverse and biochemically active family of plant cytochrome P450 monooxygenases (450s)[65]. In plants, P450s are known to be key components of biosynthesis, metabolism, and signaling, involving responses to osmotic and temperature stress and the production of a variety of growth and defense compounds including phenylpropanoids, fatty acids, and hormones such as auxin and jasmonic acid[65,66]. Among the candidate genes repeatedly identified by all four selection

scan methods, several are involved in phenylpropanoid biosynthesis (orthologs of *Arabidopsis thaliana* MYB4, CHALCONE SYNTHASE 4 (CHS4), a UDP-glycosyltransferase (UGT), and TRANSPARENT TESTA 16 (TT16) (Table 3). Phenylpropanoids are physiologically and ecologically important compounds that include lignins, anthocyanins, and condensed tannins (proanthocyanidins). Many of these compounds are known to be inducible in response to light and osmotic stress, and are also active as plant defenses against fungi, insects, and other natural enemies[67–69]. For example, MYB4 and CHS4 are both known to be responsive in *Populus* to infection by the fungal rust *Melampsora medusae*, a common leaf pathogen of poplars[70]. MYB4 expression is also responsive to jasmonic and salicylic acid, and UV-B light treatments in *Arabidopsis thaliana*[71], while in *Populus*, CHS4 expression is induced by wounding[72,73]. A recent study of selection in 25 genes in ten steps of the condensed tannin synthesis pathway that included six CHS genes in *Populus* found elevated levels of nonsynonymous divergence at CHS genes[74]. The transcription factor PtMYB186 presents another intriguing example suggesting the action of local selection on genes with dual roles in both growth and defense (Table 3). This homolog of AtMYB206 is associated with trichome initiation – an established defense against plant herbivores; however, over-expression of PtMYB186 in transgenic poplar results in pleiotropic effects on plant growth rate, even in the absence of herbivory[75]. Overall, the functional enrichment we observed among outlier genomic regions implicates pathways, individual genes, and functional domains associated with responses to environment and plant defense, suggesting the genetic architecture of locally adaptive SGV in *Populus* may have a large pleiotropic component, shared between both abiotic and biotic stress responses.

The landscape of natural genetic variation involved in adaptation has become the increasing focus of attention of breeders, conservation biologists, and natural resource managers alike, who are facing the uncertain question of how existing populations are likely to respond to the novel pressures of global change, and where adaptive variation will be most needed in the future. Climate change conservation and mitigation strategies such as restoration or assisted gene flow[76] increasingly require spatially-explicit solutions that rely on identifying and preserving locally adaptive SGV in different parts of a species' range. Thus, understanding local adaptation and the distribution of adaptive SGV in the context of range position has become prerequisite for global biodiversity conservation[15,37,77,78]. Our novel combination of landscape genomics analyses has revealed the current climatic niche is of primary importance in shaping the spatial distribution of locally adaptive SGV across the range of *Populus balsamifera*, and offers a promising general approach that could be used to prioritize regions for the conservation of unique genetic resources in other crop and wild species.

Methods

Biological Material

We targeted southern (rear) edge and nearby regions in the core of *Populus balsamifera*'s range for sampling, spanning from northern BC to Atlantic Canada (Fig. 1). The current sampling scheme therefore covers both the rear range edge, the central core of the range, and the climatically maritime region to the east. During winter 2013, we collected dormant stem cuttings from 508 trees from 57 populations (Fig. 1) spanning 9 Canadian Provinces and 7 US States (Longitudes 55-128°W and Latitudes 39-60°N; Table S1). The cuttings were flushed under greenhouse conditions and fresh foliage was used for extracting whole genomic DNA using DNeasy 96 Plant Mini Kits (Qiagen, Valencia, CA, USA). DNA was quantified using a fluorometric assay (Qubit BR; Invitrogen) and confirmed for high molecular weight using 1% agarose gel electrophoresis.

Estimation of Geographic and Climatic Range Edges

We used two distance metrics to describe the position of each *P. balsamifera* population relative to the broader range of the species: geographic distance from the southern range edge, and a measure of climatic distance from the center of the range in climatic space (Supplementary Information). Both metrics were calculated from a set of *P. balsamifera* occurrences distributed throughout the range, collected from the Global Biodiversity Information Facility and US and Canadian forest inventory programs[79,80]. To reduce spatial bias in the occurrences, we thinned the occurrences in geographic space, such that occurrences were no nearer than 75 km, resulting in 401 unique occurrences throughout the range. Occurrences were thinned using the spThin[81] package in R.

Geographic distances from the southern range edge were calculated by first fitting an alpha hull around the 401 occurrences. Alpha hulls are similar to convex hulls, but are recommended as a way to minimize effects of spatial bias and outliers when defining a species range polygon[82]. The southern range edge was defined as the southern-most line segments of the alpha hull, from which we calculated the minimum great circle distance to each *P. balsamifera* population. Next, we used Mahalanobis distance to quantify the distance of each population from the climatic centroid of the species. We used 19 bioclimatic variables (http://www.worldclim.org/bioclim), elevation, and latitude to calculate the Mahalanobis distance, with the climatic center being defined as the average value of these variables across the 401 *P. balsamifera* occurrences.

Mahalanobis distance is a unit-less, scale-invariant, multivariate distance metric that accounts for correlation among variables by scaling distances by their covariance (thereby ensuring correlated variables do not artificially inflate the measure of distance between locations). We used squared distances calculated with the Mahalanobis function in R[83].

GBS Sequencing

We used genotyping-by-sequencing (GBS)[84] to obtain genome-wide polymorphism data for 508 trees. Genomic sequencing libraries were prepared from 100ng of genomic DNA per sample digested with `EcoT22I` restriction endonuclease followed by ligation of barcoded adapters of varying length from 4–8bp, following Elshire *et al*[84]. Equimolar concentrations of barcoded fragments were pooled and purified with QIAquick PCR purification kit. Purified products were amplified with 18 PCR cycles to append Illumina sequencing primers, cleaned again using PCR purification kit, and the resulting library was screened for distribution of fragment sizes using a Bioanalyzer. These libraries were sequenced at 48plex (i.e., each library sequenced twice) using Illumina HiSeq 2500 to generate 100bp single end reads. Cornell University Institute of Genomic Diversity (Ithaca, NY) performed the library construction and sequencing steps.

We employed the Tassel GBS Pipeline[55] to process raw sequence reads and call variants and genotypes. In order to pass the quality control, sequence reads had to have perfect barcode matches, presence of restriction site overhang and no undecipherable nucleotides (N's). Filtered reads were trimmed to 64bp and aligned to the *P. trichocarpa* reference assembly version 3.0[85] using the Burrows-Wheeler Aligner (BWA)[86]. Single Nucleotide Polymorphisms (SNPs) were determined based on aligned positions to the *P. trichocarpa* reference, and genotypes called under the maximum likelihood framework in Tassel[55]. SNP genotype data along with sequence quality scores were stored in Variant Call Format v4.1 (VCF) files, which were further processed with VCFTools 0.1.11[87]. SNP genotypes were filtered to retain only biallelic sites with no indels, individuals with $<50\%$ missingness, sites with $<20\%$ missingness, genotype quality (GQ) >90, and site depth >5. We further removed SNPs that showed excess heterozygosity in tests of Hardy Weinberg Equilibrium ($P < 0.001$). After filtering, the final dataset consisted of 167,324 SNPs for downstream analyses. Bioinformatic analyses were performed on a local Dell PowerEdge Linux server (24 cpu, 64 Gb RAM) at the University of Vermont and on the Teton computing environment[88] at the University of Wyoming.

Analysis of Population Structure

We characterized the genome-wide background genetic structure within our sample using the maximum likelihood clustering algorithm ADMIXTURE[89]. We evaluated the likelihood models for presence of up to 7 Hardy-Weinberg clusters (K=1 through K=7; Fig S2). Optimal K was inferred as the model that minimized the error rate based on 2000 bootstrap replicates across 10X cross-validation iterations. Cluster membership barplots were visualized using Distruct v2.3 Python script[90,91]. Individuals exhibiting ancestry of 10% or more from other *Populus* sp. were removed from further analysis (N=26).

As a complementary approach to ADMIXTURE, we used Discriminant Analysis of Principal Components (DAPC)[92] to partition individuals along the major axes of genome-wide background genetic variation. The K-means clustering function used in DAPC (`find.clusters()`) can return different results from run to run; thus, we performed 1000 iterations of the function from K=1 to K=10, retaining the best unique solution that was most frequently observed (Fig. S1). Discriminant analyses using the most commonly observed unique solution was performed using 60% of the principal components and retaining two discriminant axes.

Identifying Genomic Regions of Local Adaptation

Genomic signals of natural selection appear as regions of elevated population differentiation across the genomic background. Quantifying background levels of differentiation is necessary for accommodating the highly variable levels of polymorphism found within the genome, and identifying SNPs potentially be under selection. In order to generate a neutral distribution from which to infer putatively locally adapted variants, we annotated the 167,324 SNPs with VCFCodingSNPs v1.5 using the *P. trichocarpa* reference genome annotation v3.0, classifying SNPs as Downstream, Upstream, 3' UTR, 5' UTR, Synonymous, Non-Synonymous and Intronic. SNPs >5kb away from an annotated gene (hereafter called intergenic) were selected as the null set of loci for neutral parameterization of local adaptation genome scans following Lotterhos & Whitlock[93].

For selection scans, we ranked every candidate outlier using its test statistic among the empirical null distribution of intergenic SNPs (N=1,649 for range-wide and N=1,253 for range-core data sets) and determined its empirical P-value. This approach caps the lower end of the empirical P-value proportionate with the size of the null distribution. For example, the lowest possible empirical P-value for the range-wide data set is

$\frac{1}{1649} = 0.0006064281$ and that for the range-core data set is $\frac{1}{1253} = 0.0007980846$. To make the inference robust, we considered an outlier a candidate if its empirical P-value equaled the lowest possible.

To scan for genomic regions showing elevated population differentiation, we applied the F-model implemented in BAYESCAN v2.1[94] to identify SNPs under selection in range-wide as well as core groups of populations. We optimized run conditions with 20 pilot runs of 5000 MCMC sweeps each, followed by 100,000 MCMC sweeps after discarding the first 50,000 sweeps as burn-in. To minimize false positives, we used the recommendation of Lotterhos & Whitlock[93] by setting the prior expectation of selection to one in 10,000. Candidate SNPs were then assigned empirical P-values based on the empirical null distribution of intergenic SNPS at an FDR of 0.01 using the `empPvals()` function in the R package qvalue[95].

Additionally we estimated $X^T X$, analogous to F_{ST} using the correlated allele frequencies model in BAYENV2. The maximum expected value of this statistic assuming neutrality is equal to the total number of populations being tested. To avoid spurious selection candidates, we again performed neutral parameterization to obtain empirical P-values for all $X^T X$ candidate loci and only retained those with lowest empirical P-value.

We tested for association between SNPs and the environment using latent factor mixed models[96] with the R package LFMM[97]. For environmental predictors, we ordinated each individual in multivariate climate space based on a PCA of the 19 bioclimatic variables and latitude. Climatic PC1 and PC2 were then used as separate predictors in association tests. Analyses were conducted on the full set of individuals with >90% *P. baslamifera* ancestry (N=437) and on a subset of individuals from the range core (N=336) that showed minimal background genetic structure (Fig. 1). To determine appropriate values for the number of latent factors and genomic inflation factors for our datasets, we performed a cross-validation analysis across a range of latent factors (K=1 to K=10) and at three values of the genomic inflation factor (1e-10, 1, and 1e+20). We retained parameter values that minimized the cross-validation error, and used them to construct LFMMs with the ridge penalty on a SNP-wise basis. The statistical tests of each model produced lambda-adjusted z-scores, which we used to generate empirical P-values based on the empirical null distribution of intergenic SNPs. We accepted only candidate SNPs with the most extreme empirical P-values as significant.

To allow for robust assessment of candidate outliers obtained from LFMM, we also performed gene-environment association analysis with BAYENV2[98] on the same set of populations and climate PCs as above. First, a co-

variance matrix of allele frequencies was estimated using the intergenic SNPs for range-wide and range-core data sets. Two hundred matrices were drawn (once every 5000 iterations) and the final matrix (200th after 100,000 iterations) was used as a control for population relatedness in our sample. We then ran BAYENV2 on a SNP-wise basis to estimate Bayes Factors for each SNP, and ranked the outliers within the neutral distribution of intergenic SNPs to obtain empirical P-values. Similar to LFMM, we only retained those outliers as candidates with the lowest empirical P-value.

Testing the Landscape Distribution of SGV

To investigate how locally adaptive standing genetic variation (SGV) was spatially distributed along landscape gradients of distance from the rear edge and from the climatic niche centroid, we calculated two population-level metrics of SGV based on the set of identified local adaptation outliers. We estimated a genetic index of alpha diversity (i.e., within-population variation) by calculating the allele frequency variance ($p \times q$). We also calculated a genetic index of beta diversity (i.e., compositional turnover across regions) by calculating the Population Adaptive Index (PAI) *sensu* Bonin *et al*[56]. Briefly, PAI is the absolute difference between adaptive allele frequencies of a focal population and that of the entire metapopulation. It therefore represents the among-population component of SGV contributed by each focal population compared to the rest of the range. To test whether the alpha and beta diversity components of SGV varied with proximity to the rear range edge or climatic niche center, we performed a multiple linear regression between each measure of SGV and the two predictor distances. Regression models were weighted so that residual errors were inversely proportional to the sample sizes (numbers of individuals) per population (Table S1).

Functional Enrichment

We tested for functional enrichment of genomic regions implicated in local adaptation based on the Gene Ontology (GO) terms associated with the *P. trichocarpa* reference annotation v3.0. Enrichment of GO terms associated with genes nearby to candidate SNPs was assessed based on the hypergeometric test implemented in the R package SNP2GO[99]. We tested for enrichment within a 50kb window upstream and downstream of each candidate SNP, and determined significance based on 1e+05 permutations and a false discovery rate (FDR) of 10%. Significantly enriched GO terms were then clustered for redundancy with REVIGO[100], using the SimRel semantic similarity metric and a clustering threshold of 0.7.

Data Archiving

Illumina sequencing data are archived under NCBI Bioproject `SAMN04517823` and NCBI SRA Accession Numbers `SRX1605454-68`. SNP genotyping data, climate PCA loadings, and outputs from selection scans and enrichment tests are available at https://github.com/stephenrkeller/Pbalsamifera_SGV.

Acknowledgements

Supplementary Figures and Tables
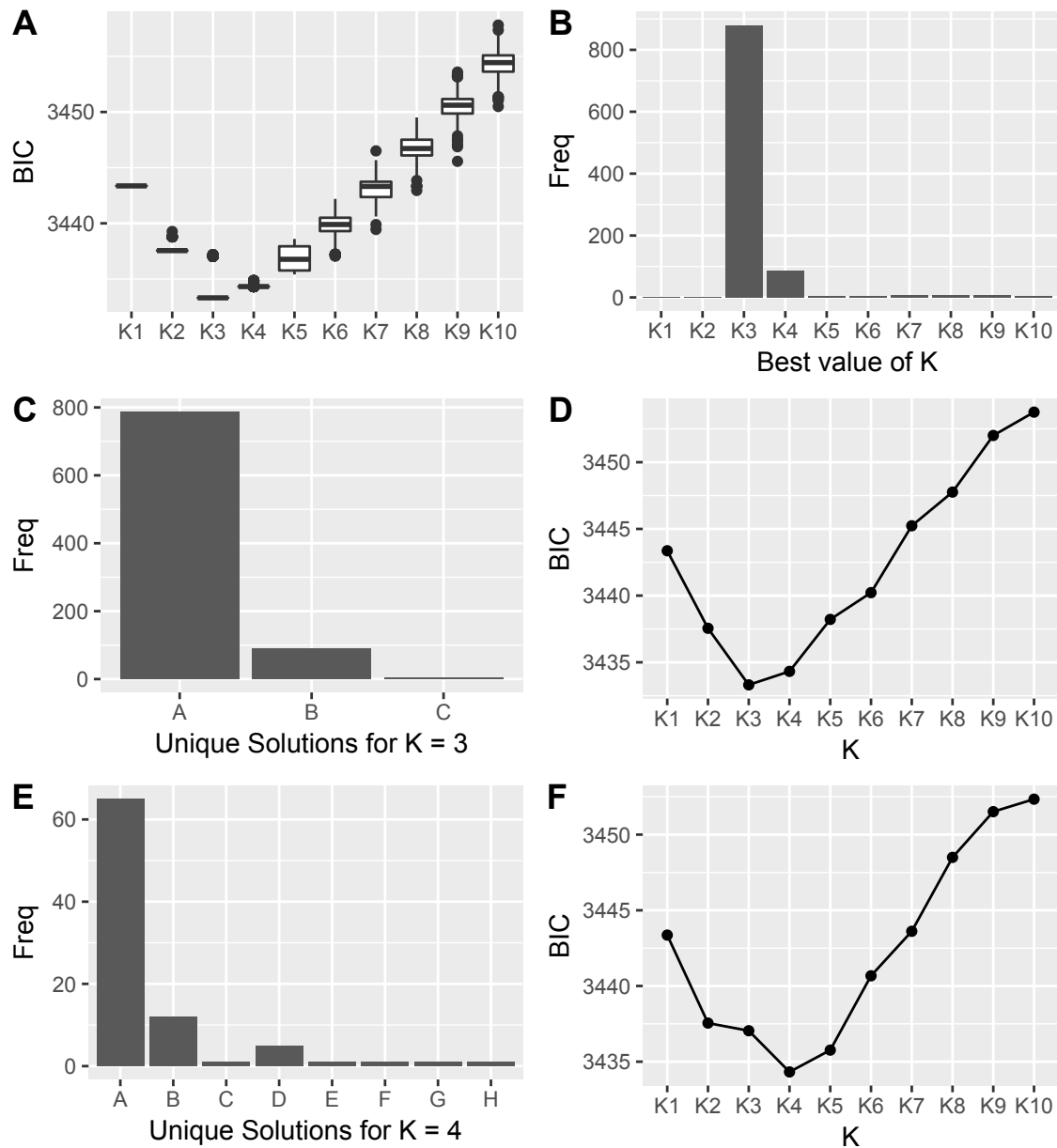
Figure S 1 DAPC model selection output from 1000 iterations of `find.clusters` from $K=1$ to $K=10$ on the range-wide (N=437) data set (A). The best value of K is defined as the cluster with the lowest BIC. $K=3$ most frequently had the lowest BIC score (B). For the $K=3$ clustering results, three unique solutions were recovered (C). The BIC distribution of result A, the most common $K=3$ result (D). For the $K=4$ clustering results, more unique solutions were recovered (E). The BIC distribution of result A, the most common result at $K=4$ (F).

(a) K=2

(b) K=3

(c) K=4

(d) K=5

(e) K=6

Figure S 2 Admixture analysis of 508 individuals collected from across the range of *P. balsamifera* from K=2 to K=6.

Table S 1 Population locality information for 508 *Populus balsamifera* individuals used in this study.

Pop	Location	State/Province	N Individuals	Longitude (°W)	Latitude (°N)
CBI	Cape Breton Island	NS	8	-61.18	46.10
CHL	Chaffey's Lock	ON	7	-76.25	44.58
CLK	Cold Lake	AB	14	-110.07	54.23
CPL	Chapleau	ON	14	-83.26	47.52
CYH	Cypress Hills Prov. Park	SK	11	-109.81	49.64
DCK	Duck Mountain	SK	13	-101.74	51.6
DPR	Deep River	ON	14	-77.51	46.09
FIS	Fischells River	NL	9	-58.41	48.2
FNO	Fort Nelson	BC	15	-122.4	58.51
GAM	Matagami	QC	15	-77.36	49.47
HBY	Hudson Bay	SK	15	-102.39	52.90
HST	Hearst	ON	12	-84.01	49.76
HWK	Hawkes Bay	NL	10	-57.00	50.41
JKH	Jackson Hole	WY	11	-110.48	43.83
KAP	Kapuskasing	ON	12	-83.16	49.38
LON	London	ON	9	-81.16	43.18
LPD	Lake Placid	NY	12	-73.73	44.35
LSM	Limestone Mountain	WV	2	-79.70	39.22
MBK	Main Brook	NL	12	-55.78	51.12
MMT	Moose Mountain Area	SK	14	-102.59	49.87
MSG	Mosier Gulch	WY	10	-106.87	44.32
NBY	North Bay	ON	11	-79.52	46.33
NEG	New Glasgow	NS	15	-62.57	45.32
NIC	Nicolet	WI	2	-88.38	44.62
OFR	Over Flowing River	MB	15	-101.10	53.14
OUT	Outlook	SK	14	-106.25	51.17
RAD	Radisson	QC	14	-77.67	53.43
SKN	Saskatoon	SK	14	-106.58	52.34
SLC	St. Lawrence County	NY	15	-75.17	44.79
SSR	Shoeshone River	WY	16	-109.61	44.46
TIM	Timmins	ON	14	-81.63	48.36
TUR	Turtleford	SK	13	-108.37	53.20
UMI	Umiujaq	QC	4	-76.28	56.34
USDA1	Kettle River	MN	4	-92.98	46.50
USDA10	Iron River	WI	6	-91.58	46.58
USDA11	Marenisco	WI	1	-89.48	46.33
USDA12	Champion	MI	7	-88.03	46.58
USDA13	Sagola	MI	5	-88.03	46.08
USDA14	Pembine	WI	4	-88.03	45.58
USDA15	Shingleton	MI	5	-86.87	46.42
USDA16	Blaney Park	MI	5	-85.85	46.08
USDA17	Hessel	MI	7	-84.50	46.00
USDA18	Afton	MI	7	-84.50	45.42
USDA19	Johannesburg	MI	3	-84.30	45.00
USDA2	Floodwood	MN	7	-92.98	46.75
USDA20	Glen Arbor	MI	1	-85.85	44.92
USDA21	Antigo	WI	4	-89.23	45.42
USDA3	Two Harbors	MN	3	-91.67	47.17
USDA4	Ely	MN	5	-91.67	47.92
USDA5	Hibbing	MN	5	-92.98	47.42
USDA6	Cook	MN	2	-92.98	47.92
USDA7	Kabetogama	MN	4	-92.98	48.42
USDA8	Waskish	MN	3	-94.52	48.33
USDA9	Blackduck	MN	5	-94.52	47.92
VER	Stowe/Waterbury	VT	4	-72.63	44.38
WLK	Watson Lake	YT	13	-128.46	60.05
WTR	White River	ON	12	-85.27	48.59

References

1. Fisher, R. A. *The genetical theory of natural selection* (Oxford University Press, Oxford, UK, 1930).

2. Orr, H. A. The genetic theory of adaptation: a brief history. *Nature Reviews Genetics* **6,** 119 (2005).

3. Charlesworth, B. The evolutionary biology of sex. *Current Biology* **16,** R693–5 (2006).

4. Delph, L. F. & Kelly, J. K. On the importance of balancing selection in plants. *New Phytologist* **201,** 45–56 (2014).

5. Mauricio, R. *et al.* Natural selection for polymorphism in the disease resistance gene Rps2 of *Arabidopsis thaliana. Genetics* **163,** 735–746 (2003).

6. Fumagalli, M. *et al.* A population genetics study of the familial Mediterranean fever gene: evidence of balancing selection under an overdominance regime. *Genes and Immunity* **10,** 678 (2009).

7. Gos, G., Slotte, T. & Wright, S. I. Signatures of balancing selection are maintained at disease resistance loci following mating system evolution and a population bottleneck in the genus *Capsella. BMC Evolutionary Biology* **12,** 152 (2012).

8. Hancock, A. M. *et al.* Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *Proceedings of the National Academy of Sciences* **107,** 8924–8930 (2010).

9. Messer, P. W. & Petrov, D. A. Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology & Evolution* **28,** 659–669 (2013).

10. Hewitt, G. The genetic legacy of the Quaternary ice ages. *Nature* **405,** 907–913 (2000).

11. Excoffier, L., Foll, M. & Petit, R. J. Genetic consequences of range expansions. *Annual Review of Ecology, Evolution, and Systematics* **40,** 481–501 (2009).

12. Hampe, A. & Jump, A. S. Climate relics: past, present, future. *Annual Review of Ecology, Evolution, and Systematics* **42,** 313–333 (2011).

13. Pironon, S. *et al.* Geographic variation in genetic and demographic performance: new insights from an old biogeographical paradigm. *Biological Reviews* **92,** 1877–1909 (2017).

14. Hampe, A. & Petit, R. J. Conserving biodiversity under climate change: the rear edge matters. *Ecology Letters* **8,** 461–467 (2005).

15. Rehm, E. M., Olivas, P., Stroud, J. & Feeley, K. J. Losing your edge: climate change and the conservation value of range-edge populations. *Ecology and Evolution* **5,** 4315–4326 (2015).

16. Slatkin, M. & Excoffier, L. Serial founder effects during range expansion: a spatial analog of genetic drift. *Genetics* **191,** 171–181 (2012).

17. Pujol, B. & Pannell, J. R. Reduced responses to selection after species range expansion. *Science* **321,** 96–96 (2008).

18. Louthan, A. M., Doak, D. F. & Angert, A. L. Where and when do species interactions set range limits? *Trends in Ecology & Evolution* **30,** 780–792 (2015).

19. Chhatre, V. E., Evans, L. M., DiFazio, S. P. & Keller, S. R. Adaptive introgression and maintenance of a trispecies hybrid complex in range-edge populations of *Populus*. *Molecular Ecology* **27,** 4820–4838 (2018).

20. Fetter, K. C., Nelson, D. M. & Keller, S. R. Trade-offs and selection conflicts in hybrid poplars indicate the stomatal ratio as an important trait regulating disease resistance. *bioRxiv,* 814046 (2019).

21. Schoville, S. D. *et al.* Adaptive genetic variation on the landscape: methods and cases. *Annual Review of Ecology, Evolution, and Systematics* **43,** 23–43 (2012).

22. Macdonald, S. L., Llewelyn, J., Moritz, C. & Phillips, B. L. Peripheral isolates as sources of adaptive diversity under climate change. *Frontiers in Ecology and Evolution* **5,** 88 (2017).

23. Antonovics, J. The nature of limits to natural selection. *Annals of the Missouri Botanical Garden,* 224–247 (1976).

24. Bridle, J. R. & Vines, T. H. Limits to evolution at range margins: when and why does adaptation fail? *Trends in Ecology & Evolution* **22,** 140–147 (2007).

25. Mimura, M. & Aitken, S. N. Adaptive gradients and isolation-by-distance with postglacial migration in *Picea sitchensis*. *Heredity* **99,** 224–232 (2007).

26. Sexton, J. P., Strauss, S. Y. & Rice, K. J. Gene flow increases fitness at the warm edge of a species' range. *Proceedings of the National Academy of Sciences* **108,** 11704–11709 (2011).

27. Holliday, J. A., Suren, H. & Aitken, S. N. Divergent selection and heterogeneous migration rates across the range of Sitka spruce (*Picea sitchensis*). *Proceedings of the Royal Society B* **279,** 1675–1683 (2012).

28. Gould, B. *et al.* Local adaptation and range boundary formation in response to complex environmental gradients across the geographical range of *Clarkia xantiana* ssp. *xantiana*. *Journal of Ecology* **102,** 95–107 (2014).

29. De Lafontaine, G., Napier, J. D., Petit, R. J. & Hu, F. S. Invoking adaptation to decipher the genetic legacy of past climate change. *Ecology* **99,** 1530–1546 (2018).

30. Schiffers, K. *et al.* Landscape structure and genetic architecture jointly impact rates of niche evolution. *Ecography* **37,** 1218–1229 (2014).

31. Polechová, J. & Barton, N. H. Limits to adaptation along environmental gradients. *Proceedings of the National Academy of Sciences* **112,** 6401–6406 (2015).

32. Kirkpatrick, M. & Barton, N. H. Evolution of a species' range. *The American Naturalist* **150,** 1–23 (1997).

33. Holt, R. D. & Barfield, M. Theoretical perspectives on the statics and dynamics of species' borders in patchy environments. *The American Naturalist* **178 Suppl 1,** S6–25 (2011).

34. Chhatre, V. E. & Rajora, O. P. Genetic divergence and signatures of natural selection in marginal populations of a keystone, long-lived conifer, eastern white pine (*Pinus strobus*) from Northern Ontario. *PLoS ONE* **9(5),** e97291 (2014).

35. Bellon, M. R., Dulloo, E., Sardos, J., Thormann, I. & Burdon, J. J. *In situ* conservationharnessing natural and human-derived evolutionary forces to ensure future crop adaptation. *Evolutionary Applications* **10,** 965–977 (2017).

36. Garcia, R. A., Cabeza, M., Rahbek, C. & Araújo, M. B. Multiple dimensions of climate change and their implications for biodiversity. *Science* **344,** 1247579 (2014).

37. National Academies of Sciences, E. & Medicine, *Forest Health and Biotechnology: Possibilities and Considerations* https://doi.org/10.17226/25221 (National Academies Press, 2019).

38. Davis, M. B. & Shaw, R. G. Range shifts and adaptive responses to Quaternary climate change. *Science* **292,** 673–679 (2001).

39. Gavin, D. G. *et al.* Climate refugia: joint inference from fossil records, species distribution models and phylogeography. *New Phytologist* **204,** 37–54 (2014).

40. Petit, R. J. & Hampe, A. Some evolutionary consequences of being a tree. *Annual Review of Ecology, Evolution, and Systematics* **37,** 187–214 (2006).

41. Savolainen, O., Pyhäjärvi, T. & Knürr, T. Gene flow and local adaptation in trees. *Annual Review of Ecology, Evolution, and Systematics* **38,** 595–619 (2007).

42. Eckert, A. J. *et al.* Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L., *Pinaceae*). *Genetics.* https://doi.org/10.1534/genetics.110.115543 (2010).

43. Keller, S. R., Levsen, N., Olson, M. S. & Tiffin, P. Local adaptation in the flowering-time gene network of balsam poplar, *Populus balsamifera* L. *Molecular Biology and Evolution* **29,** 3143–3152 (2012).

44. Sork, V. L. *et al.* Landscape genomic analysis of candidate genes for climate adaptation in a California endemic oak, *Quercus lobata. American Journal of Botany* **103,** 33–46 (2016).

45. Yeaman, S. *et al.* Convergent local adaptation to climate in distantly related conifers. *Science* **353,** 1431–1433 (2016).

46. Sork, V. *et al.* Putting the landscape into the genomics of trees: approaches for understanding local adaptation and population responses to changing climate. *Tree Genetics & Genomes* **9,** 901–911 (2013).

47. Little, E. L. Atlas of United States trees. Volume 1. Conifers and important hardwoods. Miscellaneous publication 1146. *US Department of Agriculture, Forest Service, Washington, DC* (1971).

48. Keller, S. R., Chhatre, V. E. & Fitzpatrick, M. C. Influence of range position on locally adaptive gene–environment associations in *Populus* flowering time genes. *Journal of Heredity* **109,** 47–58 (2017).

49. Keller, S. R., Olson, M. S., Silim, S., Schroeder, W. & Tiffin, P. Genomic diversity, population structure, and migration following rapid range expansion in the Balsam Poplar, *Populus balsamifera. Molecular Ecology* **19,** 1212–1226 (2010).

50. Levsen, N. D., Tiffin, P. & Olson, M. S. Pleistocene speciation in the genus *Populus* (*Salicaceae*). *Systematic Biology* **61,** 401–412 (2012).

51. Soolanayakanahally, R. Y., Guy, R. D., Silim, S. N., Drewes, E. C. & Schroeder, W. R. Enhanced assimilation rate and water use efficiency with latitude through increased photosynthetic capacity and internal conductance in balsam poplar (*Populus balsamifera* L.) *Plant, Cell & Environment* **32,** 1821–1832 (2009).

52. Keller, S. R. *et al.* Climate-driven local adaptation of ecophysiology and phenology in balsam poplar, *Populus balsamifera* L. (*Salicaceae*). *American Journal of Botany* **98,** 99–108 (2011).

53. Olson, M. S. *et al.* The adaptive potential of *Populus balsamifera* L. to phenology requirements in a warmer global climate. *Molecular Ecology* **22,** 1214–1230 (2013).

54. Keller, S. R., Levsen, N., Ingvarsson, P. K., Olson, M. S. & Tiffin, P. Local selection across a latitudinal gradient shapes nucleotide diversity in balsam poplar, *Populus balsamifera* L. *Genetics* **188,** 941–952 (2011).

55. Glaubitz, J. C. *et al.* TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS One* **9,** e90346 (2014).

56. Bonin, A., Nicole, F., Pompanon, F., Miaud, C. & Taberlet, P. Population adaptive index: a new method to help measure intraspecific genetic diversity and prioritize populations for conservation. *Conservation Biology* **21,** 697–708 (2007).

57. Sjdin, A., Street, N. R., Sandberg, G., Gustafsson, P. & Jansson, S. The Populus Genome Integrative Explorer (PopGenIE): a new resource for exploring the Populus genome. *New Phytologist* **182,** 1013–1025 (2009).

58. Eckert, C., Samis, K. & Lougheed, S. Genetic variation across species geographical ranges: the central–marginal hypothesis and beyond. *Molecular ecology* **17,** 1170–1188 (2008).

59. Sork, V. L. Genomic studies of local adaptation in natural plant populations. *Journal of Heredity* **109,** 3–15 (2017).

60. Fetter, K. C., Gugger, P. F. & Keller, S. R. *Landscape Genomics of Angiosperm Trees: From Historic Roots to Discovering New Branches of Adaptive Evolution* in *Comparative and Evolutionary Genomics of Angiosperm Trees* 303–333 (Springer, 2017). https://doi.org/10.1007/7397_2016_19.

61. Travisano, M., Mongold, J. A., Bennett, A. F. & Lenski, R. E. Experimental tests of the roles of adaptation, chance, and history in evolution. *Science* **267,** 87–90 (1995).

62. Keller, S. R. & Taylor, D. R. History, chance and adaptation during biological invasion: separating stochastic phenotypic evolution from response to selection. *Ecology Letters* **11,** 852–866 (2008).

63. Gougherty A, C. V.K. S. & Fitzpatrick, M. Contemporary range position predicts the range-wide pattern of genetic diversity in balsam poplar, *Populus blasmiafera* L.) (unpublished ms).

64. Meirmans, P. G., Godbout, J., Lamothe, M., Thompson, S. L. & Isabel, N. History rather than hybridization determines population structure and adaptation in *Populus balsamifera. Journal of Evolutionary Biology* **30,** 2044–2058 (2017).

65. Schuler, M. A. & Werck-Reichhart, D. Functional Genomics of P450S. *Annual Review of Plant Biology* **54.** PMID: 14503006, 629–667 (2003).

66. Nelson, D. & Werck-Reichhart, D. A P450-centric view of plant evolution. *The Plant Journal* **66,** 194–211 (2011).

67. Lindroth, R. L. & Hwang, S.-Y. *Diversity, redundancy, and multiplicity in chemical defense systems of aspen* in *Phytochemical Diversity and Redundancy in Ecological Interactions* 25–56 (Springer, 1996).

68. Hwang, S.-Y. & Lindroth, R. L. Clonal variation in foliar chemistry of aspen: effects on gypsy moths and forest tent caterpillars. *Oecologia* **111,** 99–108 (1997).

69. Gould, K. S. Nature's Swiss army knife: the diverse protective roles of anthocyanins in leaves. *BioMed Research International* **2004,** 314–320 (2004).

70. Miranda, M. *et al.* The transcriptional response of hybrid poplar (*Populus trichocarpa × P. deltoides*) to infection by *Melampsora medusae* leaf rust involves induction of flavonoid pathway genes leading to the accumulation of proanthocyanidins. *Molecular Plant-Microbe Interactions* **20,** 816–831 (2007).

71. Chen, W. *et al.* Expression profile matrix of *Arabidopsis* transcription factor genes suggests their putative functions in response to environmental stresses. *The Plant Cell* **14,** 559–574 (2002).

72. Sun, Y. *et al.* Isolation and promoter analysis of a chalcone synthase gene PtrCHS4 from *Populus trichocarpa*. *Plant Cell Reports* **30,** 1661 (2011).

73. Tsai, C.-J., Harding, S. A., Tschaplinski, T. J., Lindroth, R. L. & Yuan, Y. Genome-wide analysis of the structural genes regulating defense phenylpropanoid metabolism in *Populus*. *New Phytologist* **172,** 47–62 (2006).

74. Cole, C. T. & Ingvarsson, P. K. Pathway position constrains the evolution of an ecologically important pathway in aspens (*Populus tremula* L.) *Molecular Ecology* **27,** 3317–3330 (2018).

75. Plett, J. M., Wilkins, O., Campbell, M. M., Ralph, S. G. & Regan, S. Endogenous overexpression of *Populus* MYB186 increases trichome density, improves insect pest resistance, and impacts plant growth. *The Plant Journal* **64,** 419–432 (2010).

76. Aitken, S. N. & Whitlock, M. C. Assisted gene flow to facilitate local adaptation to climate change. *Annual Review of Ecology, Evolution, and Systematics* **44.** https://doi.org/10.1146/annurev-ecolsys-110512-135747 (2013).

77. Urban, M. C. *et al.* Improving the forecast for biodiversity under climate change. *Science* **353,** aad8466–1–aad8466–9 (2016).

78. Breed, M. F. *et al.* The potential of genomics for restoring ecosystems and biodiversity. *Nature Reviews Genetics,* 1 (2019).

79. Gillis, M., Omule, A. & Brierley, T. Monitoring Canada's forests: the national forest inventory. *The Forestry Chronicle* **81,** 214–221 (2005).

80. Woudenberg, S. W. *et al.* The Forest Inventory and Analysis Database: Database description and users manual version 4.0 for Phase 2. *Gen. Tech. Rep. RMRS-GTR-245. Fort Collins, CO: US Department of Agriculture, Forest Service, Rocky Mountain Research Station. 336 p.* **245** (2010).

81. Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B. & Anderson, R. P. spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography* **38,** 541–545 (2015).

82. Burgman, M. A. & Fox, J. C. *Bias in species range estimates from minimum convex polygons: implications for conservation and options for improved planning* in *Animal Conservation forum* **6** (2003), 19–28. https://doi.org/10.1017/S1367943003003044.

83. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2018). https://www.R-project.org/.

84. Elshire, R. J. *et al.* A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* **6,** e19379 (2011).

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

86. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

87. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–2158 (2011).

88. ARCC. *Teton Computing Environment: Intel x86_64 Cluster, Advanced Research Computing Center, University of Wyoming, Laramie, WY.* 2018. https://doi.org/10.15786/M2FY47.

89. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19,** 1655–1664 (2009).

90. Raj, A., Stephens, M. & Pritchard, J. K. fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* **197,** 573–589 (2014).

91. Chhatre, V. E. *Distruct 2.3: A modified admixture plotting script* (2018). http://distruct2.popgen.org.

92. Jombart, T., Devillard, S. & Balloux, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics* **11,** 94 (2010).

93. Lotterhos, K. E. & Whitlock, M. C. Evaluation of demographic history and neutral parameterization on the performance of F_{ST} outlier tests. *Molecular Ecology* **23,** 2178–2192 (2014).

94. Foll, M. & Gaggiotti, O. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics* **180,** 977–993 (2008).

95. Storey, J. D., Bass, A. J., Dabney, A. & Robinson, D. *qvalue: Q-value estimation for false discovery rate control* R package version 2.14.1 (2019). https://github.com/jdstorey/qvalue.

96. Frichot, E., Schoville, S. D., Bouchard, G. & François, O. Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular Biology and Evolution* **30,** 1687–99 (2013).

97. Caye, K. & Franois, O. *LFMM: Latent Factor Mixed Models* R package version 0.0 (2017).

98. Günther, T. & Coop, G. Robust identification of local adaptation from allele frequencies. *Genetics* **195,** 205–220 (2013).

99. Szkiba, D., Kapun, M., von Haeseler, A. & Gallach, M. SNP2GO: functional analysis of genome-wide association studies. *Genetics* **197,** 285–289 (2014).

100. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* **6,** e21800 (2011).