ECOLOGY LETTERS

Ecology Letters, (2015) **18**: 1–16 doi: 10.1111/ele.12376

IDEA AND PERSPECTIVE

Ecological genomics meets community-level modelling of biodiversity: mapping the genomic landscape of current and future environmental adaptation

Matthew C. Fitzpatrick[1,*] **and Stephen R. Keller**[1,2]

[1]*Appalachian Lab University of Maryland Center for Environmental Science Frostburg, MD, USA*
[2]*Present address:Department of Plant Biology University of Vermont Burlington, VT, USA*

**Correspondence:*
E-mail: mfitzpatrick@umces.edu

Abstract

Local adaptation is a central feature of most species occupying spatially heterogeneous environments, and may factor critically in responses to environmental change. However, most efforts to model the response of species to climate change ignore intraspecific variation due to local adaptation. Here, we present a new perspective on spatial modelling of organism–environment relationships that combines genomic data and community-level modelling to develop scenarios regarding the geographic distribution of genomic variation in response to environmental change. Rather than modelling species within communities, we use these techniques to model large numbers of loci across genomes. Using balsam poplar (*Populus balsamifera*) as a case study, we demonstrate how our framework can accommodate nonlinear responses of loci to environmental gradients. We identify a threshold response to temperature in the circadian clock gene *GIGANTEA-5* (*GI5*), suggesting that this gene has experienced strong local adaptation to temperature. We also demonstrate how these methods can map ecological adaptation from genomic data, including the identification of predicted differences in the genetic composition of populations under current and future climates. Community-level modelling of genomic variation represents an important advance in landscape genomics and spatial modelling of biodiversity that moves beyond species-level assessments of climate change vulnerability.

Keywords

Biodiversity, climate change, generalised dissimilarity modelling, gradient forests, intraspecific variation, landscape genetics, local adaptation, *Populus balsamifera*, Single-nucleotide polymorphism, species distribution modelling.

Ecology Letters (2015) **18**: 1–16

INTRODUCTION

A major challenge in predicting the impacts of climate change on biodiversity is to move beyond species-level models and towards a greater consideration of intraspecific variation in climatic tolerances due to local adaptation (Jump & Penuelas 2005; Jay *et al*. 2012). While most models assume uniformity of climate responses below the species level, in reality local adaptation of different populations is common among many plants and animals (Kawecki & Ebert 2004; Savolainen *et al*. 2007; Leimu & Fischer 2008), and intraspecific variation in climatic tolerances and local adaptation has been documented for many physiological and life-history traits (Savolainen *et al*. 2007; Mimura & Aitken 2010; Keller *et al*. 2011b). In the current era of environmental change, key objectives for the future will thus be translating intraspecific variation in climate responses into landscape models of local adaptation, and identifying geographic regions that are predicted to be most sensitive to disruption of standing patterns of local adaptation under climate change (Jay *et al*. 2012; Weinig *et al*. 2014).

Increasingly, local adaptation to climate is being studied at the molecular level using high-throughput sequencing methods, with applications spanning both model and non-model organisms (Hancock *et al*. 2011; Jones *et al*. 2013; Savolainen *et al*. 2013; Sork *et al*. 2013). These ecological genomic studies are providing unparalleled, genome-wide insights into the genetic basis of local adaptation to climate at landscape scales, especially in forest trees (Eckert *et al*. 2010; Holliday *et al*. 2012a; Keller *et al*. 2012). However, translating these new genomic insights into spatially explicit predictions of adaptive variation requires the development of new spatial modelling frameworks (Schoville *et al*. 2012). Ideally, such frameworks would be capable of: (1) linking genetic and environmental data to characterise how the frequencies of locally adaptive alleles vary along multiple, often correlated environmental gradients, and (2) projecting these gene–environment relationships across space and through time to create landscape predictions of how local adaptation may be disrupted under scenarios of environmental change.

Here, we demonstrate how two relatively new, but contrasting biodiversity modelling techniques based on the concept of community-level compositional turnover functions – Generalised Dissimilarity Modelling (GDM; Ferrier *et al*. 2007) and Gradient Forests (GF; Ellis *et al*. 2012) – can be powerfully applied to the problem of analysing and mapping genomic variation. Using these methods in tandem, we explore range-wide climate adaptation (current and future) in balsam poplar

(*Populus balsamifera*), a widespread forest tree, and draw on a history of using trees as model systems to understand adaptive genetic variation in relation to geography and climate.

Scaling from molecules to landscapes

There are two main challenges to translating ecological genomic data sets, typically consisting of thousands to millions of single-nucleotide polymorphisms (SNPs), into spatial predictions of adaptive genetic diversity. The first challenge is discovering SNPs that are candidates for local adaptation, while avoiding the occurrence of false positives due to genome-wide background variability arising from demographic history (e.g. changes in population size, gene flow and/or range expansion). Two widespread approaches to this challenge are identifying SNPs that show: (1) extreme allele frequency divergence among populations (F_{ST}) (Lewontin & Krakauer 1973; Foll & Gaggiotti 2008; Excoffier *et al*. 2009), and (2) strong associations between allele frequencies and environmental gradients (Joost *et al*. 2007; Coop *et al*. 2010; Manel *et al*. 2012; Frichot *et al*. 2013). Both approaches currently are undergoing intense theoretical development, with refinements emphasising increases in model sensitivity while minimising false positives (e.g. Narum & Hess 2011; De Villemereuil *et al*. 2014). The second major challenge involves transforming a set of candidate loci identified by one of these outlier methods into landscape-level mapped predictions of adaptive variation. In contrast to the effort afforded to the development of models for identifying locally adapted SNPs from the rest of the genome, tools for translating genomic data into spatial inferences about local adaptation remain in their infancy. We know of only one such approach to landscape mapping of intraspecific genetic variation under current and projected environments (Jay *et al*. 2012), which emphasised the prediction of how genetic clusters from Bayesian clustering of ancestry may shift in ecological niche space, rather than explicitly considering variation in locally adaptive loci.

To date, efforts to incorporate adaptive variation below the species level into spatial biogeographical modelling have focused mainly on species distribution models (SDMs; Elith & Leathwick 2009). Studies have explored the use of SDMs in predicting the geographic distribution of subspecies (Pearman *et al*. 2010), lineages (Espíndola *et al*. 2012; D'Amen *et al*. 2013; Yannic *et al*. 2014), genotypes or genetic clusters (Banta *et al*. 2012; Jay *et al*. 2012), climate zones (Sork *et al*. 2010) or molecular marker loci (Fournier-Level *et al*. 2011). While within-taxon SDMs represent a step towards predictive mapping of intraspecific variation, spatial modelling of genetic diversity presents a number of challenges that are difficult to overcome within a classical SDM framework. Foremost, even when used at a finer taxonomic division than species level, SDMs still require identification of a discrete unit (e.g. taxonomic, phylogenetic and genetic cluster) upon which an individual SDM must be built and evaluated. Not only are SDMs therefore unable to account for the continuous, multidimensional nature of genomic variation within populations and across space, but identifying and individually modelling numerous genetic loci represents a severe computational challenge. For example, whole-genome resequencing or genotyping by sequencing methods commonly generate thousands to millions of SNPs (Elshire *et al*. 2011; Narum *et al*. 2013). While in theory it would be possible to individually model the distribution of each SNP using SDMs, in practice the sheer number of loci renders it impractical to analyse and interpret individual models for each. In addition to computational burden, each locus would require dozens of occurrences to ensure robust statistical inference (Wisz *et al*. 2008). Therefore, low frequency or poorly sampled alleles would be excluded entirely in an SDM framework.

In many ways, the inherent challenges of spatial modelling of genetic variation within species are similar to those of spatial modelling of highly diverse assemblages of species. Under these circumstances, 'community-level' modelling strategies (Ferrier *et al*. 2002; Ferrier & Guisan 2006) may offer significant advantages. In a community-level modelling framework, all species in an assemblage are simultaneously modelled as a function of a common set of environmental predictors. A familiar category of such methods include constrained ordination techniques (e.g. canonical correspondence analysis; ter Braak 1986) and redundancy analysis; Legendre & Legendre 2012) already commonly used in population genetics to analyse compositional variation in dozens to thousands of SNPs as a *linear* function of a set of environmental covariates (Sork *et al*. 2010; Lasky *et al*. 2012). This class of techniques rarely has been used for predictive mapping. A second class of community-level methods remains, to our knowledge, unexplored in landscape genomics, and includes regression-based models that analyse and map patterns of turnover in biological composition using nonlinear functions of environmental gradients. Such turnover functions underlie both GDM and GF.

We suggest community-level modelling of turnover in allele frequencies along environmental gradients offers a powerful, but largely unexplored means of scaling from individual- or population-level genomic variation to landscape scale predictions of ecological adaptation and the impacts of environmental change. These new techniques can be applied to large genomic data sets, either on their own or in conjunction with genome scan approaches (F_{ST} outliers and gene–environment correlations), to gain inference on the spatial distribution of locally adapted genetic variation. When applied to genetic data, GDM and GF can: (1) accommodate pronounced nonlinearities in the exploration of gene–environment relationships, (2) handle large genomic data sets that include numerous rare, low-frequency alleles, (3) provide insights into regions of the genome ostensibly under local selection and (4) generate maps of how adaptive genomic diversity is predicted to vary across the landscape. An especially powerful feature of GDM and GF is that models can be projected to scenarios of future climates to estimate the potential impacts of climate change on biodiversity at the genetic level and how these impacts vary spatially. Most importantly, community-level modelling of genomic data offers a feasible solution for moving beyond SDMs, which assume all populations within a species respond identically to environmental gradients, and towards predictive mapping of adaptive genetic variation at the whole-genome level in response to changes in climate (Hickerson *et al*. 2010; Hancock *et al*. 2011).

COMMUNITY-LEVEL MODELLING APPLIED TO LANDSCAPE GENOMICS

Ferrier & Guisan (2006) described community-level models as techniques that combine data from multiple species to analyse and map geographic patterns of biodiversity at a collective community level instead of, or in addition to, the level of individual species. In other words, rather than modelling species individually using SDMs and then assembling the individual spatial predictions into communities or macroecological properties (i.e. a 'predict first, assemble later' strategy), community-level models 'assemble and predict together' all species within a single integrated process. Community-level modelling algorithms differ in their specific execution of the 'assemble and predict together' strategy, but most take as inputs a site-by-species matrix or some derivative thereof (e.g. GDM uses a site-by-site compositional dissimilarity matrix) and a corresponding site-by-environment predictor matrix. Simultaneous modelling of the entire site-by-species matrix confers several benefits (Ferrier & Guisan 2006), among which are an ability to: (1) rapidly analyse large numbers of species, (2) detect shared patterns of response to environmental gradients for rarely recorded species and (3) assess the relative importance of environmental predictors in explaining overall patterns of biological variation.

The conceptual leap from community-level modelling of species assemblages to spatial modelling of genomic variation is a small one. Rather than assembling and predicting together numerous species within communities, we assemble and predict together any number of SNPs (from one to potentially millions) within a genome sampled across many geographic sites. Instead of a site-by-species matrix, we employ a site-by-SNP matrix, which facilitates analysis and predictive mapping of either genome-wide patterns, as might be governed by ecological factors controlling movement and gene flow across the landscape, or alternatively, an *a priori* subset of SNPs thought to control ecologically important functions – i.e. those underlying local adaptation.

We focus on GDM (Ferrier *et al*. 2007) and GF (Ellis *et al*. 2012), two technically contrasting methods founded on the common idea of modelling compositional turnover using nonlinear functions of environmental gradients. These turnover functions provide a means to transform environmental variables to a common biological scale of compositional turnover (in this case, turnover in allele frequencies), thereby allowing conversion from multidimensional environmental space to multidimensional genetic space while selecting and weighting variables such that they best summarise genomic variation. This provides a major advance over univariate approaches to gene–environment associations that consider relationships in isolation of other covariates. By applying the associated turnover function to each mapped environmental variable, these functions can be used to map expected patterns of genomic variation, including under scenarios of future climate to assess potential impacts to locally adapted genetic variants in a spatially explicit manner. The commonalities in the predictive outputs from GDM and GF (mapped patterns of biological variation) and their unique, but complementary, strengths and limitations make them well suited for comparison. For these reasons, their use in tandem may enable more robust understanding of the factors driving adaptive variation than achievable with either method in isolation.

GENERALISED DISSIMILARITY MODELLING

GDM is a nonlinear extension of permutational matrix regression that models pairwise biological dissimilarity between sites as a nonlinear function of pairwise site differences in environmental and geographic variables. To accommodate nonlinearities, GDM fits a generalised linear model of the form:

$$-\ln(1 - d_{ij}) = a_0 + \sum_{p=1}^{n} |f_p(x_{pi}) - f_p(x_{pj})|, \quad (1)$$

where i and j are sites, d is any distance measure constrained between 0 and 1, a_0 is the intercept, p is the number of covariates and $f_p(x)$ are I-spline transformed versions of the predictor variables (Box 1). For genomic applications, the measure of genetic distance used as the response can be flexibly chosen by the user so long as it is constrained between 0 and 1, including traditional population-level genetic distances (e.g. F_{ST}, Nei's D, Jost's D) and individual distance metrics such as (1-minus) a measure of genetic co-ancestry (Oliehoek *et al*. 2006). GDM produces a unique monotonic I-spline turnover function for each predictor that, in a genomics context, describes the rate and magnitude of turnover in genetic distance along that gradient while holding all other variables constant (Box 1). GDM uses per cent deviance explained as a measure of model fit. See Ferrier *et al*. (2007) for details and Fitzpatrick *et al*. (2013) for a species-level application.

A considerable and unique strength of GDM's distance-based approach is the ability to account for spatial patterns in genetic data caused by demographic processes, such as isolation-by-distance, isolation-by-ecological resistance, or founder effects. In addition to Euclidean distance, GDM can accommodate most any measure of geographic or ecological separation as a predictor, including organism-specific representations of barriers to dispersal, or cost of movement/gene flow through unfavourable habitat (e.g. Spear *et al*. 2010; Thomassen *et al*. 2010). Accounting for such spatial effects remains a major challenge in landscape genomics (Manel *et al*. 2010a). Second, because the response variable is a pairwise genetic distance matrix, in theory GDM can accommodate any number of SNPs of any frequency of occurrence, with computational limits being set by the number of sites (which determines the dimensions of the response and predictor matrices). For these reasons, GDM may be most useful when there are a very large number of SNP loci and the question is one of how landscape patterns of environmental variation affect compositional turnover of the entire genome while accounting for geographic separation between sites. GDM is also appropriate for analysing sets of loci grouped according to their genomic context (haplotypes within a particular gene or gene network). In this context, GDM could also be applied iteratively to each of many loci one at a time by calculating the locus-specific F_{ST} between all pairs of populations.

Box 1 Understanding and interpreting GDM and GF

To conceptualise how GDM and GF work, imagine collecting biodiversity data (either species in the community ecology case, or genotype data in the genomics case) and environmental data at many sites along a transect. The goal is to understand why allele frequencies (or species assemblages) change between sites, while: (1) identifying the environmental gradients associated with biological variation, and (2) determining where along each gradient turnover is slow/rapid. For example, what is the relative importance of changes in temperature vs. soil moisture in explaining allele frequency turnover and does turnover occur more or less rapidly near the warm (dry) or cold (wet) end of the gradient? These are questions that cannot readily be answered using univariate or linear models.

Similar to a Mantel test, GDM models biological variation using distance matrices – specifically by relating dissimilarity in species or genetic composition (biological distance) between all site pairs to how much sites differ in their environmental conditions (environmental distance) and how isolated they are from one another (geographical or resistance distance). Unlike a Mantel test or other linear method, GDM uses splines and a GLM to accommodate two types of nonlinearity: (1) variation in the rate of compositional turnover (non-stationarity) along environmental gradients, and (2) the curvilinear relationship between biological distance and environmental and geographical distance.

Variation in the rate of biological change along and between gradients arises partially because environmental variables are measured on arbitrary scales from a biological perspective. For example, a small change in soil moisture will have a larger effect on species composition in a desert than in a rainforest. To accommodate non-stationarity, splines are fit to the environmental variables themselves (these are the I-spline turnover functions described in Methods), rather than to environmental distances derived from these variables, while ensuring that a model incorporating the scaled environmental distances measured from each predictor's spline provides the best fit between observed and predicted biological distance. The splines provide two key pieces of information. The maximum height indicates the magnitude of total biological change along that gradient and thereby corresponds to the relative importance of that predictor in contributing to biological turnover while holding all other variables constant. The spline's shape indicates how the rate of biological change varies with position along that gradient. Thus, the splines provide insight into the total magnitude of biological change as a function of each gradient and where along each gradient those changes are most pronounced.

The curvilinear relationship between biological distance and environmental and geographic distance arises because most measures of biological distance are constrained between 0 and 1 and therefore asymptote at a maximum value of 1 despite continued increases in environmental and geographical distance. This nonlinearity is handled using a GLM with an exponential link function that relates the scaled environmental distances (from the splines) and measures of geographic isolation to biological distance.

GF approaches the problem of modelling biological variation and building turnover functions in a very different way from GDM. Foremost, GF is not a distance-based, curve-fitting approach. Rather, GF uses a machine-learning algorithm to divide the biological data into different bins (e.g. different values of allele frequencies), with partitions occurring at numerous split values along each environmental variable. Moving along certain portions of a gradient, we might observe few changes in allele frequencies despite changes in environment. At other places along the gradient, we may find large changes in allele frequencies. The question then becomes: how well does a given split value (e.g. between 26 and 27°C) explain biological variation across that split? The amount of variation explained is known as 'split importance', which GF cumulatively sums along each gradient to construct turnover functions. The process can be envisioned as building a staircase. One end of the gradient is the ground floor at an importance value of zero. As we move along the gradient, steps are added (i.e. importance values are cumulatively summed), with step height being proportional to the importance of the split value at that location. Places with many large steps in a row are thresholds where biological change is rapid. Gradients strongly associated with biological variation will have more and/or larger steps and therefore attain a greater maximum height, and therefore overall importance, than other gradients. Thus, the heights and shape of GF turnover functions provide the same inference as the splines from GDM, with the caveat that GF builds a function for each allele, which are aggregated to also provide an overall, genome-wide turnover function.

GRADIENT FORESTS

GF is a community-level extension of the nonparametric, machine-learning regression tree approach known as random forests (Breiman 2001). Random forests is often used to develop SDMs and also has been implemented to test for genotype–phenotype associations (Holliday *et al.* 2012b), including sliding window approaches for handling many SNPs (Jiang *et al.* 2009). Random forests combine numerous individual decision trees (>500) into a single ensemble (i.e. a 'forest') to produce a highly flexible model capable of fitting complex relationships with both accuracy and high predictive performance. Rather than modelling turnover indirectly using dissimilarity between sites as GDM does, GF uses three primary outputs from random forests to model compositional turnover directly, including: (1) the overall goodness of fit (R^2) of the forest for each response (in this case, responses are individual SNPs), (2) the accuracy importance for each environmental predictor within a forest and (3) the importance of that predictor at a given split value in determining changes in

SNP frequency in a particular tree. These measures are used to construct monotonic turnover functions for each SNP (as in a 'predict, then assemble' strategy) and an aggregate, community-level turnover function across all SNPs analogous to GDM's I-splines (Box 1). Only SNPs with random forest models having $R^2 > 0$ are included in the aggregate turnover functions for each variable, using weighting that accounts for variable importance and the goodness of fit of the random forest model for each SNP. See Ellis *et al*. (2012) for details and Pitcher *et al*. (2012) for a species-level application.

Unlike GDM, GF can handle both complex relationships and interactions between predictors, but has no means of incorporating geographic distance. The underlying random forests algorithm is also highly proficient at quantifying variable importance, and can, to some extent, accommodate correlated predictors (Ellis *et al*. 2012) such as climate variables. GF also provides a means to examine the response of individual SNPs to environmental gradients, although will become computationally limited as the number of SNPs becomes very large. In analyses, we have run on a computer with a 2.95 GHz quad-core processor, GF took 0.5 h to analyse 2314 SNP loci, which is not unreasonable compared to other genetic analysis methods for detecting local adaptation. For very large numbers of SNPs, GF may be better suited to a strategy where either portions of the genome are analysed in a sliding window approach (Jiang *et al*. 2009) or a reduced set of candidate SNPs are mapped after testing for their outlier status using other approaches (F_{ST} or genetic association analyses).

Lastly, we note that neither GF nor GDM account for non-independence in the genetic data arising from linkage disequilibrium (LD), or, in the case of GDM, from the use of pairwise distance matrices. Non-independence arising from LD affects most approaches to analysing local adaptation from genome-wide data, and LD must be taken into account when interpreting the results in a genomic context, such as plotting model outputs along the chromosome when a physical map is available for the species under study. In terms of significance testing in GDM, non-independence arising from the use pairwise distance matrices can be overcome using matrix permutation (e.g. Fitzpatrick *et al*. 2011).

MATERIAL AND METHODS

To illustrate our modelling framework, we reanalysed two previously published data sets from a range-wide population collection of balsam poplar, *Populus balsamifera* (Salicaceae). The first data set consists of 412 SNPs genotyped in 474 individuals representing 31 populations (Keller *et al*. 2010). These SNPs, hereafter referred to as 'reference', resulted from randomly sequencing genomic regions without regard to genomic context (coding, non-coding, etc.) and selecting 1 SNP per genomic region (412 regions total) for genotyping in all individuals. Reference SNPs are thus a random sample from the genomic background, and thus have no *a priori* expectation for being involved in ecological adaptation. Reference SNP data are available from Dryad Digital Repository: http://dx.doi.org/doi:10.5061/dryad.1164/1.

The second data set consists of 335 SNPs derived from 27 candidate genes in the flowering time genetic network (Flowers *et al*. 2009), and thus are hypothesized *a priori* to contain some SNPs involved in plant phenological responses to environmental stimuli such as photoperiod and temperature (Keller *et al*. 2012). These SNPs, hereafter referred to as 'candidate', were genotyped in a subset of the individuals and populations from the reference set (443 individuals from 31 populations). Candidate SNP data are available for download from http://www.popgen.uaf.edu/LightGeneSNPs.html.

Our previous work tested for signals of local adaptation to climate in the candidate SNPs using the reference SNPs to control for background variability in the balsam poplar genome due to demographic history (Keller *et al*. 2012; Olson *et al*. 2013). We employed tests for: (1) elevated population structure (F_{ST}) using the hierarchical model in Arlequin (Excoffier *et al*. 2009) and the Bayesian approach in BayeScan (Foll & Gaggiotti 2008), (2) gene–environment associations using the method of Bayenv (Coop *et al*. 2010) and (3) genotype–phenotype associations with bud phenology traits known to be locally adaptive. Across these tests, SNPs from several candidate genes emerged that were significant in multiple tests, thus providing increased confidence that these genes harboured true histories of local adaptation. Most notable of these was the *GIGANTEA-5* (*GI5*) gene that interacts with the plant circadian clock and light perception pathways, and integrates signals to downstream genes controlling meristem development. Two other genes were also repeatedly implicated as outliers in our local adaptation analyses – the vernalisation gene *FRIGIDA* (*FRI*) and the meristem development gene *LEAFY* (*LFY*) (Keller *et al*. 2011a). Here, we use GDM and GF to reanalyse the 412 reference and 335 candidate SNPs, and also separately model SNP variability in the adaptive genes *GI5, FRI* and *LFY* (five SNP data sets total).

Environmental and spatial variables

To characterise environmental conditions at the sampling locations of balsam poplar, we used an uncorrelated set ($r < 0.75$) of six predictor variables describing temperature, precipitation and topography at 5-arcminute resolution (*ca*. 10 × 10 km) from WorldClim (www.worldclim.org; Hijmans *et al*. 2005), including mean annual temperature (BIO1), mean diurnal range of temperature (BIO2), temperature annual range (BIO7), mean summer temperature (BIO10), summer precipitation (BIO18), precipitation seasonality (BIO15) and elevation.

It is well known that spatial patterns of genetic variation and population structure can reflect the influence of historical demographic and other spatial processes, in addition to the possible action of natural selection (Li *et al*. 2012). GDM can account for such processes to some extent using any measure of geographic separation between sites as a predictor. For GF, which cannot directly accommodate spatial effects, we attempted to account for the influence of spatial processes and unmeasured environmental variation using Moran's eigenvector map (MEM) variables (Borcard & Legendre 2002;

Dray *et al*. 2006). Briefly, MEM variables are the eigenvectors of a spatial weighting matrix derived from the geographic coordinates of sampling locations. The resulting uncorrelated spatial eigenfunctions can be used to model geographic structure in genetic patterns across a range of spatial scales. Following previous applications, we used the first half of the MEM eigenfunctions with positive eigenvalues (four in our case), which have been claimed to model broad-scale spatial genetic variation generated by processes such as demographic history and, more likely, unaccounted for environmental variation (Manel *et al*. 2010b, 2012; Sork *et al*. 2013). This approach has analogies to the use of latent factors to account for unobserved sources of genetic variation when testing gene–environment associations (Frichot *et al*. 2013), and given that MEMs may actually reflect unmeasured environmental patterns rather than true isolation-by-distance, they should be implemented and interpreted with caution.

Statistical modelling

We fit GDM and GF to each of the five SNP data sets (Table 1) and used these models to explore environmental and spatial drivers of turnover in allele frequency and to map current and future patterns of genomic variation in relation to climate. For GDM, we used as our response variable a pairwise F_{ST} matrix for each SNP data set between balsam poplar populations, resulting in a 31 × 31 response matrix. For GF, we converted the SNP data into minor allele relative frequencies and removed any SNP that was polymorphic in fewer than five of the 31 populations to ensure robust regression (Table 1). We fit GF using 2000 regression trees per SNP and a variable correlation threshold of 0.5. We used default values for the number of predictor variables randomly sampled as candidates at each split (two in this case) and for the proportion of samples used for training (~0.63) and testing (~0.37) each tree. GF provides a weighted R^2 value to assess relative importance of predictor variables. To estimate relative importance of variables for GDM, we rescaled the maximum value of the fitted I-Splines between 0 and 1, which is proportional to variable importance. We used the R packages gradientForest (Smith & Ellis 2013) and gdm (Manion *et al*. 2014) available from R-Forge (http://r-forge.r-project.org) to fit models. R scripts are available from Dryad Digital Repository: http://doi.org/10.5061/dryad.2s6f9.

Visualising genetic variation

The turnover functions derived from GDM and GF were used to examine changes in allele frequencies along each environmental gradient and to perform biologically informed transformations of the environmental variables into genetic importance values. To visualise the resulting multidimensional genetic patterns in geographic and biological space, we used Principal Components Analysis (PCA) to reduce the transformed environmental variables into three factors. The PCA was centred but not scale transformed to preserve differences in the magnitude of genetic importance among the environmental variables. For each of the five SNP data sets, the difference in genetic composition between grid points was mapped by assigning the first three PCs to a RGB colour palette, with resulting colour similarity corresponding to the similarity of expected patterns of genetic composition. The result was mapped in geographical space and also visualised as a biplot of the first two principal dimensions with labelled vectors indicating the direction and magnitude of major environmental correlates.

To estimate the congruence (or lack thereof) between the mapped genetic patterns for reference SNPs and the four candidate SNP data sets, we performed a Procrustes superimposition on the resulting PCA ordinations, where the matrices were rotated to minimise the sum of square of the distances between the sites in genetic space (Peres-Neto & Jackson 2001). The Procrustes residuals, which in this case measure the absolute distance between sites in genetic space and the rotated ordination space, were mapped to provide a location-specific measure of the difference in genetic composition patterns between reference and the four sets of candidate SNPs. For all visualisations, we constrained predictions to within the geographic range of balsam poplar as defined by Little (1971).

Population-level vulnerability to climate change

To describe a range of potential future climate conditions for 2050, we used six general circulation models and three SRES emission scenarios for a total of nine future climate scenarios (Table S1) obtained from the CCAFS-climate data portal (http://www.ccafs-climate.org/). To estimate the spatial regions where gene–environment relationships will be most disrupted (hereafter, the 'genetic offset') between current and potential

Table 1 Summary of the five SNP data sets used to fit GDM and GF and associated metrics of model performance

SNPs	Total number of SNPs*	SNPs polymorphic in >5 populations†	SNPs with $R^2 > 0$ (%)†	Mean R^2 [range]†	Deviance explained*
Reference	412	360	174 (48.3)	24.40 [0.22, 72.44]	63.26
Candidate	339	293	126 (43.0)	22.63 [0.55, 70.19]	63.08
GI5	22	20	13 (65.0)	30.47 [15.95, 46.86]	63.33
LFY	17	17	8 (47.1)	19.03 [4.22, 28.97]	24.18
FRI	34	32	17 (53.1)	13.77 [0.82, 48.26]	36.87

*GDM only.
†GF only.
All SNPs for each data set were used to calculate F_{ST} for GDM models, whereas GF models were fit only for SNPs polymorphic in >5 populations. GDM, generalised dissimilarity modelling; GF, gradient forests; SNPs, single-nucleotide polymorphisms.

future climates using GF, we first transformed the climate variables from each of the nine future climate scenarios into genetic importance using the turnover functions as described above for current climate. For each grid cell, we then calculated the Euclidean distance between the current and future genetic importance values (Ellis *et al.* 2012), which serves as a metric of genetic offset. Because GDM models genetic distance directly, we simply predicted the genetic offset (which is proportional to the expected F_{ST} between current and future populations) by projecting the GDM developed for current environment to each future climate scenario. Lastly, for each of the five SNP data sets, we mapped the mean genetic offset from the nine scenarios to indicate the spatial distribution of population-level vulnerability to climate change.

The future projection of genetic composition informs us about how much the genetic composition across the landscape would have to change to preserve the gene–environment relationships observed under current environmental conditions. In essence, the genetic offset predicts the magnitude alleles will be perturbed from their adaptive optima (assuming populations reside within this optima) solely as a result of a change in environment shifting populations away from their current position within the gene–environment association. Of course, the actual evolutionary responses of populations to environmental change will be more complex than this simplified projection, and likely to involve interactions between selection, the effective population size, which determines the efficacy of selection response, and the evolutionary processes shaping adaptive variation (e.g. migration, mutation, recombination). However, as a general tool for predicting changes in the selective landscape, the genetic offset approach serves as a useful model for identifying regions of the landscape that are predicted to experience the greatest impacts under future environments if there is no adaptive evolution *in situ* or migration to allow adaptive alleles to track climate change. This near-term assumption is especially plausible for long-lived, sessile organisms such as forest trees.

RESULTS

Community-level models of genetic composition

GDM explained more than 63% of the deviance in turnover in genetic composition of reference, candidate and *GI5* SNPs, with *GI5* slightly exceeding that of reference and candidate SNPs (Table 1). In contrast, GDM explained only 24.2% and 36.9% of the deviance for *LFY* and *FRI* respectively. GF does not provide a metric equivalent to deviance explained, but the mean R^2 of the individual SNP models can serve as an analogous measure (Leaper *et al.* 2011). Using mean R^2, GF model rankings were similar to those from GDM, with *GI5* having the highest mean R^2 of 30.5% and *LFY* and *FRI* having the lowest R^2 of 19.0% and 13.8% respectively.

In terms of variable importance, the most prominent pattern for both GF and GDM was the strong relationship of *GI5* with temperature (Fig. 1). However, GDM and GF differed in which aspect of temperature was most important. For GF, mean summer temperature was the most important predictor for *GI5*, with mean annual temperature being the third most important variable after the first eigenfunction (MEM-1). In contrast, GDM found mean annual temperature to be the most important variable, followed by geographic distance, and found virtually no relationship with mean summer temperature. For GF, MEM-1 and MEM-2 were the most important predictors for all SNPs except *GI5,* suggesting either a strong overall spatial component to the genetic differences among populations in each SNP data set or that the MEM variables captured important, but missing environmental predictors. After MEM variables, precipitation seasonality was typically the most important environmental predictor. For GDM, results were less consistent across SNP data sets, although geographic distance ranked among the top two most important predictors for all SNPs except *LFY* and precipitation variables were among the top two predictors for all SNPs except *GI5*.

Patterns of turnover in genetic composition varied by SNP data set and by environmental and spatial gradients (Figs. 2 and 3, Fig. S1). For each environmental and spatial predictor, the aggregate turnover functions from GF were similar in shape and magnitude for all SNP data sets (Fig. 2, Fig. S1), with the most notable exception to this pattern being the prominent response of *GI5* to annual and summer temperature, and the weaker response of *FRI* to MEM-1. *GI5* showed a strong threshold response near values of mean summer temperature between 12 and 15°C, whereas the rate of turnover was more constant in response to mean annual temperature. The I-splines from GDM (Fig. 3) were more varied between SNP data sets than the turnover functions from GF, although tended to differ more in maximum height than in shape. Summer precipitation was one of the few variables for which all SNPs showed a similar response, with the most rapid turnover occurring at the wet end of the gradient. The mean annual temperature I-spline for *GI5* indicated rapid changes in allele frequencies for values less than −3°C and no turnover elsewhere along this gradient, whereas reference SNPs showed nearly the opposite pattern.

In addition to aggregate turnover functions, GF also produces individual turnover functions for each SNP having an $R^2 > 0$. For demonstration purposes, we focus on SNP-level responses to mean summer temperature. For this gradient, the vast majority of SNPs exhibited weak or no response to mean summer temperature (Fig. 4, black lines). In contrast, nearly all modelled SNPs associated with *GI5* exhibited a large, threshold response (Fig. 4, red lines). Similar outlier SNPs were evident in the turnover functions for *LFY* and *FRI* (Fig. S2). However, SNPs within LFY and FRI showed comparatively modest responses to environmental gradients.

Spatial predictions of variation in genetic composition

When mapped in geographic space, the patterns of genetic composition of reference SNPs predicted by GF (Fig. 5a) and GDM (Fig. S3a) were similar, with rapid turnover predicted in the eastern third of the range of balsam poplar and comparatively little elsewhere, consistent with the location of the eastern genetic cluster identified by Keller *et al.* (2010). Both models identified precipitation variables as major environmental correlates of these patterns, with GDM additionally find-

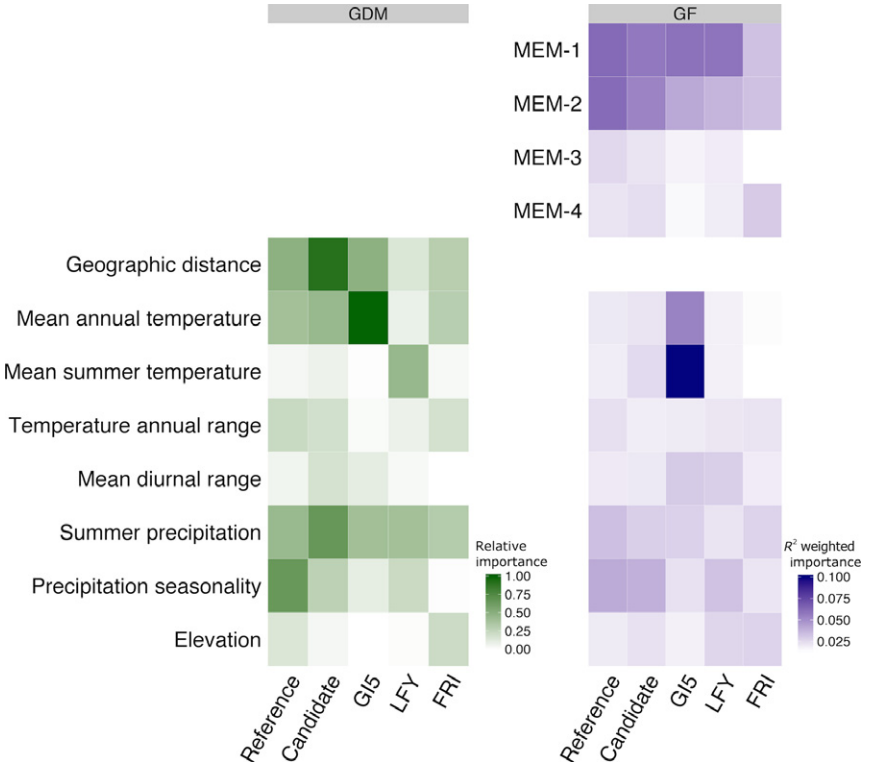

Figure 1 The relative importance of climatic, topographic and spatial predictors used in GDM (left) and GF (right) for the five SNP data sets, with darker shading indicating greater relative importance. The four Moran's eigenvector map (MEM) variables were used as spatial predictors in GF only. Geographic distance was used in GDM only.

ing mean annual temperature as important (biplot insets of Fig. 5a, Fig. S3a). The mapped patterns for *GI5* were less congruent between GF (Fig. 5b) and GDM (Fig. S3b) as compared to reference SNPs, although predicted patterns from both models exhibited greatest turnover in the western and eastern portions of the range and comparatively less in the central region. Spatial predictions for *GI5* from GF were particularly heterogeneous in the western portion of the range and were driven by summer and mean annual temperature.

For both GF and GDM, the difference in predicted patterns of allele turnover between reference SNPs and *GI5* were greatest and most extensive in the south-eastern portion of the range, as indicated by the mapped Procrustes residuals (Fig. 5c, Fig. S3c). GDM additionally identified large differences between reference and *GI5* along the north-western range edge, while GF highlighted two clusters in the northern Rocky Mountains. For the other candidate SNP data sets, the predicted patterns from GDM (Figs. S4–S6) consistently differed most from reference SNPs in the south-eastern portion of the range and, notably, trailing edge populations, and little elsewhere (Figs S4c, S5c and S6c). In contrast, there were comparatively little differences between the predicted patterns for reference SNPs and the three other candidate SNP data sets from GF (Figs S7–S9).

Population-level vulnerability to climate change

GDM and GF both identified regions in the north-western portion of balsam poplar's range as harbouring populations for which the genetic offset for *GI5* is predicted to be greatest under climate change (Fig. 6a and c). Similarly, the central and southern portions of the range were predicted by both models to experience relatively low genetic offset between current and future climates. However, while areas of high genetic offset were restricted to the northwest region for GDM, GF also identified most of the northern range edge and the northeast and southwest regions as also having relatively high genetic offset. For both GDM and GF, areas of large predicted genetic offset corresponded to portions of temperature gradients where small changes in temperature were associated with rapid turnover in allele frequencies (Fig. 6b and d). For GDM, areas of relatively large predicted genetic offset for other SNPs tended to be more widespread than those for *GI5* (Fig. S10). Predictions from GF showed nearly the opposite pattern, with predicted genetic offset for other SNPs being much smaller and less widespread than for *GI5* (Fig. S11).

DISCUSSION

Our case study on balsam poplar illustrates the great potential that GF and GDM possess for identifying, mapping and predicting ecological adaptation from genomic data. Our models were able to identify several prominent gene–environment relationships in balsam poplar, many of which appeared as threshold responses along temperature gradients (Figs 2 and 3). These threshold responses underscore the importance of accommodating nonlinear responses in allele frequency turnover along environmental gradients. Threshold physiological

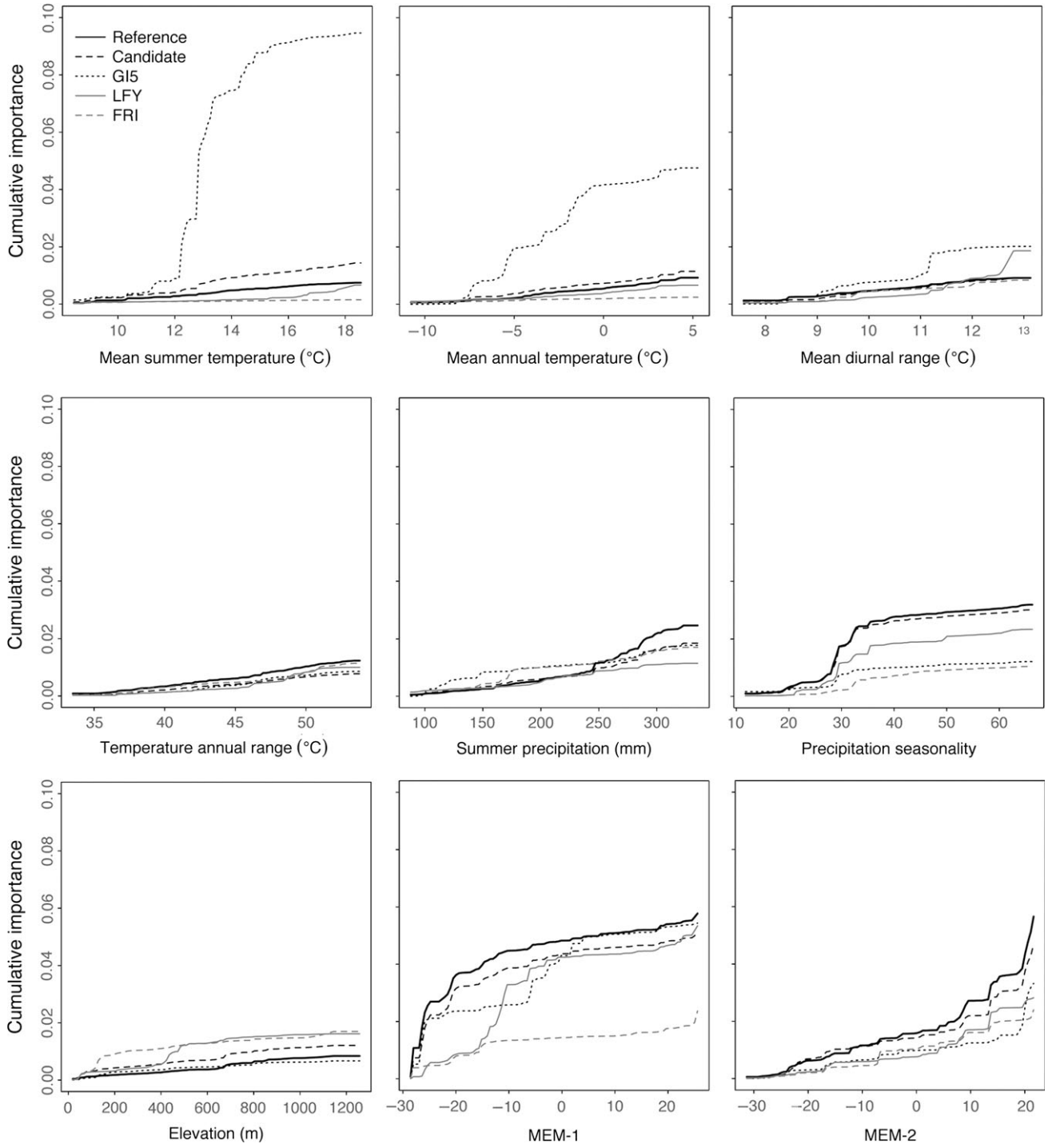

Figure 2 Aggregate compositional turnover functions from GF for each environmental covariate and the first two Moran's eigenvector map (MEM) variables (see Fig. S1 for the other two MEM variables). The maximum height of each curve indicates the total amount of turnover in allele frequencies associated with that variable, and by extension, the relative importance of that variable in explaining changes in allele frequency. The shape of each function indicates how the rate of change in allele frequencies varies along the gradient.

responses to environmental gradients are common features of organisms, and thus are likely to be reflected in the genetic architecture underlying climate adaptation. To our knowledge, our analyses are the first to both reveal and accommodate such pronounced nonlinearities while mapping gene-environment relationships.

Our results also confirm our previous conclusions of local adaptation in *GI5* using F_{ST} outlier methods (Foll & Gaggiotti 2008) and gene–environment association analyses (Coop *et al.* 2010). However, our previous analyses were methodologically constrained to univariate models with single predictor variables at a time, yielding many significant correlations between

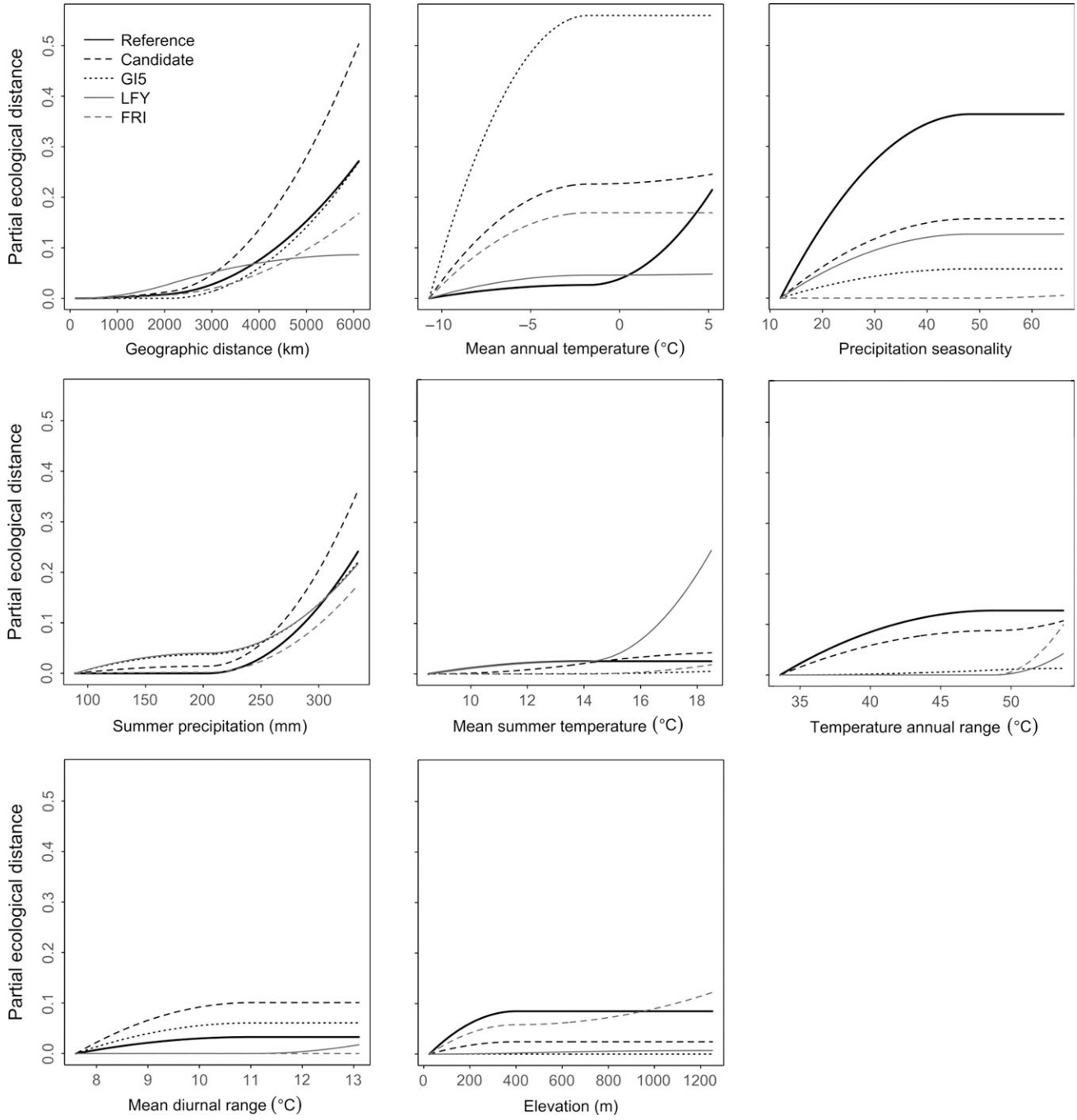

Figure 3 GDM-fitted I-splines for each environmental covariate and geographic distance. The maximum height of each curve indicates the total amount of turnover in allele frequencies associated with that variable, and by extension, the relative importance of that variable in explaining changes in allele frequency, holding all other variables constant. The shape of each function indicates how the rate of change in allele frequencies varies along the gradient.

GI5 SNP frequencies and various autocorrelated aspects of the environment, including latitude, temperature and precipitation variables (Keller *et al*. 2012). The multivariate nature of GF and GDM, and the ability of GF to accommodate correlations among predictors (Ellis *et al*. 2012), provide critical new insight in the analysis of local adaptation in balsam poplar, suggesting that temperature gradients represent the primary driver of turnover in SNP frequencies within *GI5* (Figs 2 and 3). In *Arabidopsis, GIGANTEA* is known to interact with the plant circadian clock to control expression of downstream developmental genes in response to a variety of environmental cues, including cold stress (Cao *et al*. 2005). In *Populus, GI5* has emerged as a strong candidate for genetic control of bud set and bud flush (Olson *et al*. 2013), which are temperature-sensitive traits known to be under local adaptation. Thus, our analyses here confirm *GI5* as a very strong candidate for local adaptation, possibly reflecting the adaptive response of poplar bud phenology to temperature.

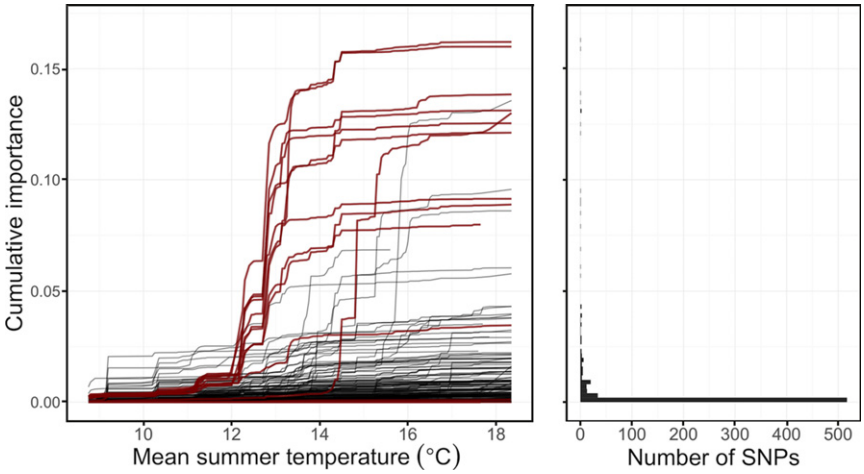

Figure 4 SNP-level compositional turnover functions from GF for (black) all 653 SNPs along the gradient of mean summer temperature. Red highlighted functions indicate the 20 SNPs within *GI5*. The marginal histogram shows the distribution of cumulative importance values across all 653 SNPs, indicating that most SNPs other than those in *GI5* show little or no response to mean summer temperature.

In addition, the analysis of *GI5* revealed new insights into the geographic areas where turnover in SNP frequencies in this candidate gene departs substantially from the genomic background represented by the reference SNPs (Fig. 5). Geographic areas associated with ecological differences in genetic turnover between candidate and reference loci harbour adaptive differences within species, and also suggest via interpolation which unsampled regions may be most profitable for additional sampling of germplasm for prioritisation under conservation management.

Differences between GF and GDM

While both GDM and GF share commonalities in using nonlinear turnover functions to model spatial patterns of biodiversity, they approach this problem in very different ways (Box 1). GDM models biological dissimilarity using a GLM-like framework and a linear combination of environmental and/or geographic distances derived from flexible, monotonic I-splines. In contrast, GF is a nonparametric, machine-learning method that constructs turnover functions piecemeal by summing increases in predictive performance at individual split values along each predictor gradient. Despite these differences, we found that results from GDM and GF analyses were congruent in many aspects and lead to similar inferences, as has been found for species-level comparisons of GDM and GF (Leaper *et al*. 2011; Compton *et al*. 2013). That said, we noted interesting differences between GF and GDM, perhaps the most notable being in the relationship of *GI5* with annual vs. summer temperature. This difference, coupled with other variations between GDM and GF in the magnitudes and shapes of their turnover functions, likely contributed to differences in the mapped patterns, including those of genetic offset under climate change. We speculate different predictive outcomes from GF and GDM may arise due to differences in their abilities to accommodate geographic distance (GDM) and interactions between predictors (GF), as well differences in the nature and shape of the turnover functions. Ongoing comparisons and testing by our group aim to isolate the underlying causes of these disparities. While it will be useful to determine whether GDM or GF most reliably identifies the true nature of gene–environment relationships, which method to use will depend on the goals of the study and the nature of the data. Give their inherent differences, we suggest greatest inference may lie in using GDM and GF in tandem.

Informing the discovery of adaptive SNPs

Our application of GF and GDM complement existing tools for identifying genomic variation under local adaptation by taking candidate loci and flexibly modelling and mapping the response surface of allele frequency turnover along environmental gradients, including nonlinear and threshold effects. Other computational approaches based on F_{ST} outliers (Foll & Gaggiotti 2008; Excoffier *et al*. 2009), or gene–environment associations (Joost *et al*. 2007; Coop *et al*. 2010; Frichot *et al*. 2013), have seen substantial application in the field of ecological genomics. These methods will continue to provide a robust means of assessing signals of local adaptation, especially as these methods evolve to better control for the difficult problem of false positives due to demographic history (De Villemereuil *et al*. 2014). However, beyond identifying candidate loci for local adaptation, current methods provide little support for spatially explicit inference of adaptive relationships and associated responses to environment change. We argue that this is precisely the area where landscape genomic studies need to grow, and GF and GDM provide a powerful and flexible means of fulfilling this need. Thus, a multistage approach to the analysis of adaptive genetic diversity on the landscape could involve first conducting scans for local adaptation on large genomic data sets using the F_{ST} outlier or gene–environment association approach, and then analysing outliers with GF or GDM to model the functional turnover of adaptive diversity along environmental gradients. These relationships could then be used to map adaptive genetic diversity under current or future environmental conditions.

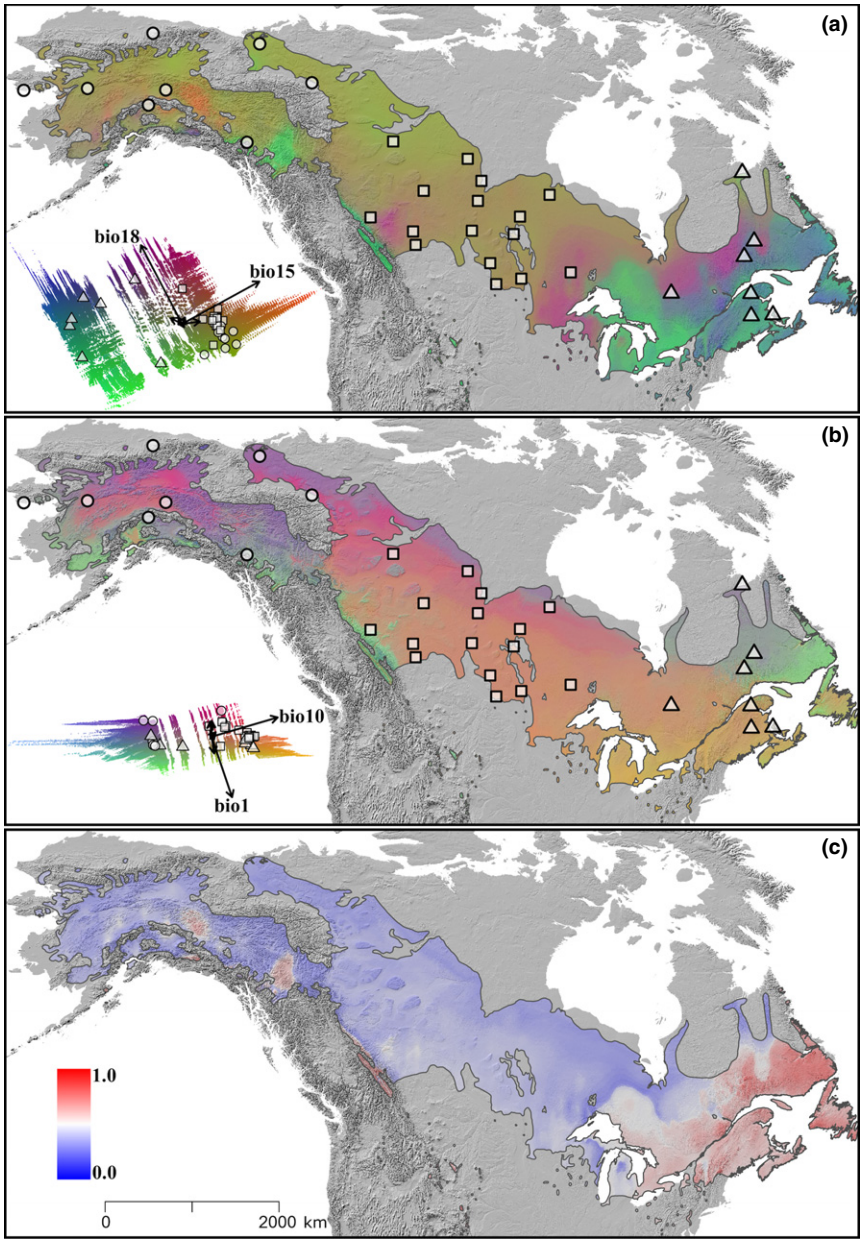

Figure 5 Predicted spatial variation in population-level genetic composition from GF for (a) reference SNPs and (b) *GI5*. The (c) difference between turnover in genetic composition of reference SNPs and *GI5* is based on Procrustes residuals, with distances scaled to the maximum distance found in any comparison between reference and the four candidate SNP data sets from both GF and GDM. Colours in (a) and (b) represent gradients in genetic turnover derived from transformed environmental predictors. Locations with similar colours are expected to harbour populations with similar genetic composition. The associated biplots indicate the contribution of the environmental variables to the predicted patterns of genetic turnover, with labelled vectors indicating the direction and magnitude of environmental gradients with greatest contribution. White symbols in (a) and (b) indicate locations of genotyped balsam poplar populations in geographic (maps) and genetic (biplots) space, with shape indicating three genetically distinct clusters previously identified using STRUCTURE (Keller *et al.* 2010).

Here, we have visualised these patterns as continuous representations of allele composition. However, such predictive outputs also can be discretised using statistical clustering algorithms to map distributions of genotypes.

The strengths of GF and GDM for modelling nonlinear or threshold relationships, and the ability to conduct multivariate analyses on a suite of predictors simultaneously, suggest that these methods themselves may also hold promise for the initial discovery of outlier loci during genome scans. For example, identifying SNPs that show strong responses to environmental gradients may not only indicate loci potentially under selection, but also where along gradients changes in allele frequencies are most pronounced (Fig. 4). The application of multivariate analyses such as these to genomic data may also help control false positives in gene–environment association analyses caused by correlated predictors, compared to the alternative strategy of conducting many univariate analyses on the same data set.

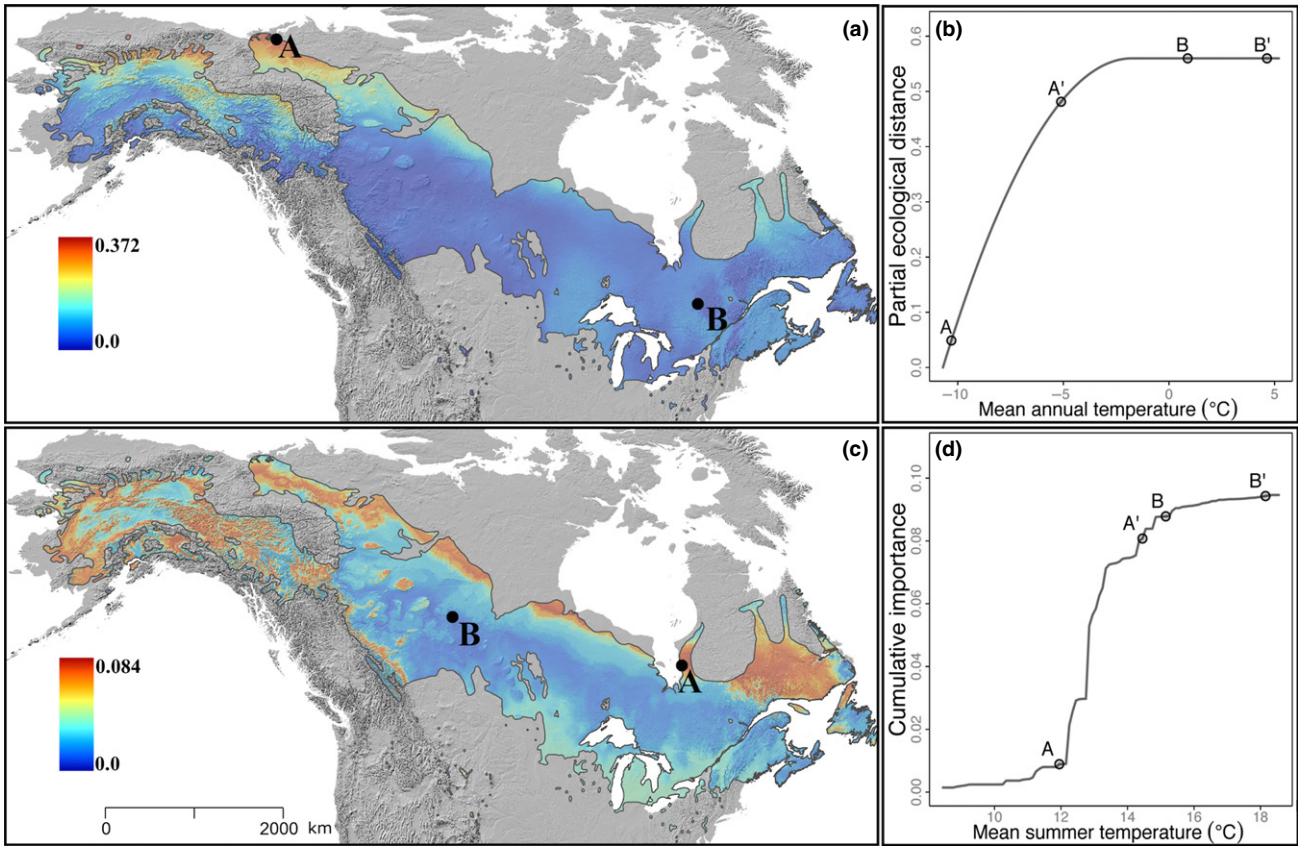

Figure 6 Mean predicted genetic offset for *GI5* from (a) GDM and (c) GF from nine scenarios of 2050s climate (Table S1). Map units in (a) are F_{ST} while those in (c) are dimensionless Euclidean distance between current and future genetic spaces. Locations with large predicted genetic offset occur in portions of temperature gradients where turnover functions from (b) GDM and (d) GF indicate rapid changes in allele frequencies (labelled A in maps and on turnover functions). Correspondingly, future increases in temperature from A to A' on the gradients produce large predicted genetic offset. Conversely, locations with small predictive genetic offset occur in portions of temperature gradients where turnover functions indicate small changes in allele frequencies (labelled B in maps and on turnover functions). Therefore, future increases in temperature from B to B' on the gradients, while similar in magnitude of those occurring at A, produce small predicted genetic offset.

The major challenge faced by any new method for detecting environmental adaptation is controlling for non-adaptive differences due to demographic history. Unlike GF, GDM is able to accommodate space directly and more tests are needed to resolve the extent to which MEM variables used in our GF analyses and in other landscape genomics studies adequately capture true spatial variability. In addition, it may be useful to explicitly incorporate reference loci into GDM/GF models of candidate loci, for example, using as predictors a population pair-wise F_{ST} matrix (GDM) or ancestry coefficients from Bayesian genotype clustering (GF) to control for demography. The effectiveness of these approaches in controlling the false-positive rate and the statistical power they have for correctly identifying true loci under ecological selection, is an important question that can best be addressed with spatially explicit simulations. Such work is currently underway.

Moving beyond species-level models of climate change vulnerability

Because populations evolve in response to environmental change, not species, a pressing concern for biodiversity conservation is predicting how intraspecific adaptive variation will be impacted by global environmental change, since this represents the evolutionary potential of populations (Reusch & Wood 2007; Lavergne *et al.* 2010; Pauls *et al.* 2013). By transforming environmental space into genetic space while selecting and weighting predictors, both GF and GDM can model the genetic offset expected from changes in climate or other aspects of the environment. In balsam poplar, we see the gene–environment relationship between *GI5* and temperature is most disrupted along the northern range edge (Fig. 6), especially in north-western North America. GF also predicts some genetic offset for *GI5* along the southern range edge, while both methods predict the range core is likely to experience minimal disruption of the existing gene–environment relationship. These predictions require further sampling and experimentation to validate, but they are consistent with theory predicting loss of adaptation at range edge populations (Hampe & Petit 2005; Bridle & Vines 2007). In principle, one could also integrate locally adaptive phenotypes associated with candidate genes into the GF or GDM analysis, such as bud phenology traits which are under strong genetic control in *Populus* and other trees, and have been linked to variation in *GI5* and other circadian clock genes (Ingvarsson *et al.*

2008; Rohde *et al*. 2011; Olson *et al*. 2013). An integrated approach like this to modelling genotype–phenotype–environment associations could provide substantial novel insights for landscape studies of ecological adaptation (Sork *et al*. 2013).

Conclusions and future directions

A grand challenge for ecologists and evolutionary biologists is to integrate genomics with biogeographical modelling to assess the ecological drivers of adaptation and conserve adaptive diversity under global change. Here, we have shown that community-level modelling methods originally conceived for species-level applications provide a clear advance in our ability to model intraspecific diversity on the landscape and uncover complex spatial patterns in adaptive variation. While most applications and comparisons of GDM and GF to date have used taxonomic-level data, GDM has begun to see extensions to modelling spatial variation in functional traits (Thomassen *et al*. 2010), neutral gene frequencies (Thomassen *et al*. 2013), and phylogenetic beta diversity (Rosauer *et al*. 2014). GF possesses similar capabilities, although owing to its recent development, we know of no relevant studies. These tools provide a mechanism for predicting adaptive genomic variability into unsampled geographic regions or times, and as such constitute an important new resource for designing sampling strategies, targeting regions for conservation or reserve design, and predicting regional impacts of environmental change. For example, these models could help inform the discovery of new genetic resources by revealing locations expected to be most dissimilar in genetic composition. This might be especially useful for prioritising populations of rare or threatened species for conservation, or even to inform the search for novel adaptive germplasm in wild populations of domesticated crops. In addition, predictions of genetic offset under future environments could be applied to neutral genetic data to examine how the environment structures gene flow among populations (e.g. 'isolation-by-resistance'), and thus how population connectivity may be expected to change under future environments.

As the environmental consequences of global change become better known, our ability to translate scenarios of environmental change into biological consequences at landscape scales is still limited. Chief among the current needs are ways to model the most fundamental component of biodiversity–genetic variation that reflects both the landscape of local adaptation and the evolutionary potential of geographically distributed populations to future change. Community-level modelling of ecological genomic data sets using methods such as GDM and GF provides a promising path forward.

ACKNOWLEDGEMENTS

We thank C. Roland Pitcher, Simon Ferrier, and Vikram Chhatre for helpful discussion and encouragement, Matthew Lisk for assistance with figures, and three anonymous reviewers for feedback that improved the manuscript. MCF acknowledges funding from UMCES and travel support from the International Biogeography Society, Australian National University, and CSIRO. This research was supported by NSF Plant Genome Research Program award IOS-1238885 to SRK and MCF and NSF DEB-1257164 to MCF. This is UMCES-Appalachian Laboratory Scientific Contribution No. 4950.

AUTHORSHIP

MCF and SRK conceived the ideas and designed the analyses. MCF wrote the R scripts and performed the modelling using data provided by SRK. MCF and SRK wrote the manuscript.

REFERENCES

Banta, J.A., Ehrenreich, I.M., Gerard, S., Chou, L., Wilczek, A., Schmitt, J. *et al*. (2012). Climate envelope modelling reveals intraspecific relationships among flowering phenology, niche breadth and potential range size in *Arabidopsis thaliana*. *Ecol. Lett*., 15, 769–777.

Borcard, D. & Legendre, P. (2002). All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecol. Model*., 153, 51–68.

ter Braak, C.J.F. (1986). Canonical Correspondence Analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67, 1167–1179.

Breiman, L. (2001). Statistical modeling: the two cultures. *Stat. Sci*., 16, 199–215.

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

Cao, S., Ye, M. & Jiang, S. (2005). Involvement of GIGANTEA gene in the regulation of the cold stress response in *Arabidopsis*. *Plant Cell Rep*., 24, 683–690.

Compton, T.J., Bowden, D.A., Roland Pitcher, C., Hewitt, J.E. & Ellis, N. (2013). Biophysical patterns in benthic assemblage composition across contrasting continental margins off New Zealand. *J. Biogeogr*., 40, 75–89.

Coop, G., Witonsky, D., Di Rienzo, A. & Pritchard, J.K. (2010). Using environmental correlations to identify loci underlying local adaptation. *Genetics*, 185, 1411.

D'Amen, M., Zimmermann, N.E. & Pearman, P.B. (2013). Conservation of phylogeographic lineages under climate change. *Global Ecol. Biogeogr*., 22, 93–104.

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? *Mol. Ecol*., 23, 2006–2019.

Dray, S., Legendre, P. & Peres-Neto, P.R. (2006). Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecol. Model*., 196, 483–493.

Eckert, A.J., Bower, A.D., Gonzalez-Martinez, S.C., Wegrzyn, J.L., Coop, G. & Neale, D.B. (2010). Back to nature: ecological genomics of loblolly pine (*Pinus taeda*, Pinaceae). *Mol. Ecol*., 19, 3789–3805.

Elith, J. & Leathwick, J.R. (2009). Species distribution models: ecological explanation and prediction across space and time. *Annu. Rev. Ecol. Evol. Syst*., 40, 677–697.

Ellis, N., Smith, S. & Pitcher, C. (2012). Gradient forests: calculating importance gradients on physical predictors. *Ecology*, 93, 156–168.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S. *et al*. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, 6, e19379.

Espíndola, A., Pellissier, L., Maiorano, L., Hordijk, W., Guisan, A. & Alvarez, N. (2012). Predicting present and future intra-specific genetic structure through niche hindcasting across 24 millennia. *Ecol. Lett*., 15, 649–657.

Excoffier, L., Hofer, T. & Foll, M. (2009). Detecting loci under selection in a hierarchically structured population. *Heredity*, 103, 285–298.

Ferrier, S. & Guisan, A. (2006). Spatial modelling of biodiversity at the community level. *J. Appl. Ecol*., 43, 393–404.

Ferrier, S., Drielsma, M., Manion, G. & Watson, G. (2002). Extended statistical approaches to modelling spatial pattern in biodiversity in

northeast New South Wales. II. community-level modelling. *Biodivs. Conserv*., 11, 2309–2338.
Ferrier, S., Manion, G., Elith, J. & Richardson, K. (2007). Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. *Divers. Distrib*., 13, 252–264.
Fitzpatrick, M.C., Sanders, N.J., Ferrier, S., Longino, J.L., Weiser, M.D. & Dunn, R.R. (2011). Forecasting the future of biodiversity: a test of single- and multi-species models for ants in North America. *Ecography*, 34, 836–847.
Fitzpatrick, M.C., Sanders, N.J., Normand, S., Svenning, J.-C., Ferrier, S., Gove, A.D. *et al*. (2013). Environmental and historical imprints on beta diversity: insights from variation in rates of species turnover along gradients. *P. Roy. Soc. Lond. B Bio*., 280, doi:10.1098/rspb.2013.1201.
Flowers, J.M., Hanzawa, Y., Hall, M.C., Moore, R.C. & Purugganan, M.D. (2009). Population genomics of the *Arabidopsis thaliana* flowering time gene network. *Mol. Biol. Evol*., 26, 2475–2486.
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.
Fournier-Level, A., Korte, A., Cooper, M.D., Nordborg, M., Schmitt, J. & Wilczek, A.M. (2011). A map of local adaptation in *Arabidopsis thaliana*. *Science*, 334, 86–89.
Frichot, E., Schoville, S.D., Bouchard, G. & François, O. (2013). Testing for associations between loci and environmental gradients using latent factor mixed models. *Mol. Biol. Evol*., 30, 1687–1699.
Hampe, A. & Petit, R.J. (2005). Conserving biodiversity under climate change: the rear edge matters. *Ecol. Lett*., 8, 461–467.
Hancock, A.M., Brachi, B., Faure, N., Horton, M.W., Jarymowycz, L.B., Sperone, F.G. *et al*. (2011). Adaptation to climate across the *Arabidopsis thaliana* genome. *Science*, 334, 83–86.
Hickerson, M.J., Carstens, B.C., Cavender-Bares, J., Crandall, K.A., Graham, C.H., Johnson, J.B. *et al*. (2010). Phylogeography's past, present, and future: 10 years after Avise, 2000. *Mol. Phylogenet. Evol*., 54, 291–301.
Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol*., 25, 1965–1978.
Holliday, J.A., Suren, H. & Aitken, S.N. (2012a). Divergent selection and heterogeneous migration rates across the range of Sitka spruce (*Picea sitchensis*). *P. Roy. Soc. Lond. B Bio*., 279, 1675–1683.
Holliday, J.A., Wang, T. & Aitken, S. (2012b). Predicting adaptive phenotypes from multilocus genotypes in Sitka spruce (*Picea sitchensis*) using random forest. *G3: Genes|Genomes|Genetics*, 2, 1085–1093.
Ingvarsson, P.K., Garcia, M., Luquez, V., Hall, D. & Jansson, S. (2008). Nucleotide polymorphism and phenotypic associations within and around the phytochrome B2 locus in European aspen (*Populus tremula*, Salicaceae). *Genetics*, 178, 2217.
Jay, F., Manel, S., Alvarez, N., Durand, E.Y., Thuiller, W., Holderegger, R. *et al*. (2012). Forecasting changes in population genetic structure of alpine plants in response to global warming. *Mol. Ecol*., 21, 2354–2368.
Jiang, R., Tang, W., Wu, X. & Fu, W. (2009). A random forest approach to the detection of epistatic interactions in case-control studies. *BMC Bioinformatics*, 10, S65.
Jones, M.R., Forester, B.R., Teufel, A.I., Adams, R.V., Anstett, D.N., Goodrich, B.A. *et al*. (2013). Integrating landscape genomics and spatially explicit approaches to detect loci under selection in clinal populations. *Evolution*, 67, 3455–3468.
Joost, S., Bonin, A., Bruford, M.W., Després, L., Conord, C., Erhardt, G. *et al*. (2007). A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Mol. Ecol*., 16, 3955–3969.
Jump, A.S. & Penuelas, J. (2005). Running to stand still: adaptation and the response of plants to rapid climate change. *Ecol. Lett*., 8, 1010–1020.
Kawecki, T.J. & Ebert, D. (2004). Conceptual issues in local adaptation. *Ecol. Lett*., 7, 1225–1241.
Keller, S.R., Olson, M.S., Silim, S., Schroeder, W.R. & Tiffin, P. (2010). Genomic diversity, population structure, and migration following rapid range expansion in the balsam poplar, *Populus balsamifera*. *Mol. Ecol*., 19, 1212–1226.
Keller, S.R., Levsen, N., Ingvarsson, P.K., Olson, M.S. & Tiffin, P. (2011a). Local selection across a latitudinal gradient shapes nucleotide diversity in Balsam Poplar, *Populus balsamifera* L. *Genetics*, 188, 941–952.
Keller, S.R., Soolanayakanahally, R.Y., Guy, R.D., Silim, S.N., Olson, M.S. & Tiffin, P. (2011b). Climate-driven local adaptation of ecophysiology and phenology in balsam poplar, *Populus balsamifera* L. (Salicaceae). *Am. J. Bot*., 98, 99–108.
Keller, S.R., Levsen, N., Olson, M.S. & Tiffin, P. (2012). Local adaptation in the flowering-time gene network of balsam poplar, *Populus balsamifera* L. *Mol. Biol. Evol*., 29, 3143–3152.
Lasky, J.R., Des Marais, D.L., McKay, J., Richards, J.H., Juenger, T.E. & Keitt, T.H. (2012). Characterizing genomic variation of *Arabidopsis thaliana*: the roles of geography and climate. *Mol. Ecol*., 22, 5512–5529.
Lavergne, S., Mouquet, N., Thuiller, W. & Ronce, O. (2010). Biodiversity and climate change: integrating evolutionary and ecological responses of species and communities. *Annu. Rev. Ecol. Evol. Syst*., 41, 321–350.
Leaper, R., Hill, N.A., Edgar, G.J., Ellis, N., Lawrence, E., Pitcher, C.R. *et al*. (2011). Predictions of beta diversity for reef macroalgae across southeastern Australia. *Ecosphere*, 2, art73.
Legendre, P. & Legendre, L. (2012). *Numerical Ecology*, 3rd edn.. Elsevier, Oxford, UK, pp. 1–990.
Leimu, R. & Fischer, M. (2008). A meta-analysis of local adaptation in plants. *PLoS ONE*, 3, e4010.
Lewontin, R.C. & Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, 74, 175–195.
Li, J., Li, H., Jakobsson, M., Li, S., Sjödin, P. & Lascoux, M. (2012). Joint analysis of demography and selection in population genetics: where do we stand and where could we go? *Mol. Ecol*., 21, 28–44.
Little, E.L. (1971). *Atlas of United States trees. Volume 1. Conifers and important hardwoods*. United States Department of Agriculture. Forest Service Miscellaneous Publication 1146, Washington D.C.
Manel, S., Joost, S., Epperson, B.K., Holderegger, R., Storfer, A., Rosenberg, M.S. *et al*. (2010a). Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field. *Mol. Ecol*., 19, 3760–3772.
Manel, S., Poncet, B.N., Legendre, P., Gugerli, F. & Holderegger, R. (2010b). Common factors drive adaptive genetic variation at different spatial scales in *Arabis alpina*. *Mol. Ecol*., 19, 3824–3835.
Manel, S., Gugerli, F., Thuiller, W., Alvarez, N., Legendre, P., Holderegger, R. *et al*. (2012). Broad-scale adaptive genetic variation in alpine plants is driven by temperature and precipitation. *Mol. Ecol*., 21, 3729–3738.
Manion, G., Ferrier, S., Lisk, M. & Fitzpatrick, M.C. (2014). *gdm: Functions for Generalised Dissimilarity Modelling*.
Mimura, M. & Aitken, S.N. (2010). Local adaptation at the range peripheries of Sitka spruce. *J. Evol. Biol*., 23, 249–258.
Narum, S.R. & Hess, J.E. (2011). Comparison of FST outlier tests for SNP loci under selection. *Mol. Ecol. Resour*., 11, 184–194.
Narum, S.R., Buerkle, C.A., Davey, J.W., Miller, M.R. & Hohenlohe, P.A. (2013). Genotyping-by-sequencing in ecological and conservation genomics. *Mol. Ecol*., 22, 2841–2847.
Oliehoek, P.A., Windig, J.J., Van Arendonk, J.A. & Bijma, P. (2006). Estimating relatedness between individuals in general populations with a focus on their use in conservation programs. *Genetics*, 173, 483–496.
Olson, M.S., Levsen, N., Soolanayakanahally, R.Y., Guy, R.D., Schroeder, W.R., Keller, S.R. *et al*. (2013). The adaptive potential of *Populus balsamifera* L. to phenology requirements in a warmer global climate. *Mol. Ecol*., 22, 1214–1230.
Pauls, S.U., Nowak, C., Bálint, M. & Pfenninger, M. (2013). The impact of global climate change on genetic diversity within populations and species. *Mol. Ecol*., 22, 925–946.

Pearman, P.B., D'Amen, M., Graham, C.H., Thuiller, W. & Zimmermann, N.E. (2010). Within-taxon niche structure: niche conservatism, divergence and predicted effects of climate change. *Ecography*, 33, 990–1003.

Peres-Neto, P.R. & Jackson, D.A. (2001). How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. *Oecologia*, 129, 169–178.

Pitcher, C.R., Lawton, P., Ellis, N., Smith, S.J., Incze, L.S., Wei, C.-L. *et al*. (2012). Exploring the role of environmental variables in shaping patterns of seabed biodiversity composition in regional-scale ecosystems. *J. Appl. Ecol*., 49, 670–679.

Reusch, T.B.H. & Wood, T.E. (2007). Molecular ecology of global change. *Mol. Ecol*., 16, 3973–3992.

Rohde, A., Storme, V., Jorge, V., Gaudet, M., Vitacolonna, N., Fabbrini, F. *et al*. (2011). Bud set in poplar – genetic dissection of a complex trait in natural and hybrid populations. *New Phytol*., 189, 106–121.

Rosauer, D.F., Ferrier, S., Williams, K.J., Manion, G., Keogh, J.S. & Laffan, S.W. (2014). Phylogenetic generalised dissimilarity modelling: a new approach to analysing and predicting spatial turnover in the phylogenetic composition of communities. *Ecography*, 37, 21–32.

Savolainen, O., Pyhäjärvi, T. & Knürr, T. (2007). Gene flow and local adaptation in trees. *Annu. Rev. Ecol. Evol. Syst*., 38, 595–619.

Savolainen, O., Lascoux, M. & Merilä, J. (2013). Ecological genomics of local adaptation. *Nat. Rev. Genet*., 14, 807–820.

Schoville, S.D., Bonin, A., François, O., Lobreaux, S., Melodelima, C. & Manel, S. (2012). Adaptive genetic variation on the landscape: methods and cases. *Annu. Rev. Ecol. Evol. Syst*., 43, 23–43.

Smith, S.J. & Ellis, N. (2013). *gradientForest: Random Forest functions for the Census of Marine Life synthesis project*.

Sork, V.L., Davis, F.W., Westfall, R., Flint, A., Ikegami, M., Wang, H. *et al*. (2010). Gene movement and genetic association with regional climate gradients in California valley oak (*Quercus lobata* Née) in the face of climate change. *Mol. Ecol*., 19, 3806–3823.

Sork, V.L., Aitken, S.N., Dyer, R.J., Eckert, A.J., Legendre, P. & Neale, D.B. (2013). Putting the landscape into the genomics of trees: approaches for understanding local adaptation and population responses to changing climate. *Tree Genet. Genomes*, 9, 901–911.

Spear, S.F., Balkenhol, N., Fortin, M.-J., McRae, B.H. & Scribner, K.I.M. (2010). Use of resistance surfaces for landscape genetic studies: considerations for parameterization and analysis. *Mol. Ecol*., 19, 3576–3591.

Thomassen, H.A., Buermann, W., Mila, B., Graham, C.H., Cameron, S.E., Schneider, C.J. *et al*. (2010). Modeling environmentally associated morphological and genetic variation in a rainforest bird, and its application to conservation prioritization. *Evol. Appl*., 3, 1–16.

Thomassen, H.A., Freedman, A.H., Brown, D.M., Buermann, W. & Jacobs, D.K. (2013). Regional differences in seasonal timing of rainfall discriminate between genetically distinct east African giraffe taxa. *PLoS ONE*, 8, e77191.

Weinig, C., Ewers, B.E. & Welch, S.M. (2014). Ecological genomics and process modeling of local adaptation to climate. *Curr. Opin. Plant Biol*., 18, 66–72.

Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H. & Guisan, A. (2008). Effects of sample size on the performance of species distribution models. *Divers. Distrib*., 14, 763–773.

Yannic, G., Pellissier, L., Ortego, J., Lecomte, N., Couturier, S., Cuyler, C. *et al*. (2014). Genetic diversity in caribou linked to past and future climate change. *Nat. Clim. Change*, 4, 132–137.

SUPPORTING INFORMATION

Additional Supporting Information may be downloaded via the online version of this article at Wiley Online Library (www.ecologyletters.com).

Editor, Mark Vellend
Manuscript received 6 May 2014
First decision made 17 June 2014
Second decision made 19 August 2014
Manuscript accepted 21 August 2014