MOLECULAR ECOLOGY

Molecular Ecology (2012) **21**, 5512–5529 doi: 10.1111/j.1365-294X.2012.05709.x

Characterizing genomic variation of *Arabidopsis thaliana*: the roles of geography and climate

JESSE R. LASKY,* DAVID L. DES MARAIS,* JOHN K. McKAY,† JAMES H. RICHARDS,‡ THOMAS E. JUENGER* and TIMOTHY H. KEITT*

*Section of Integrative Biology, University of Texas at Austin, 1 University Station A6700, Austin, Texas 78712-0253, USA, †Bioagricultural Sciences and Pest Management, Colorado State University, Campus delivery 1177, Fort Collins, Colorado 80523, USA, ‡Land, Air and Water Resources, University of California, Davis, One Shields Avenue, Davis, California 95616, USA

Abstract

***Arabidopsis thaliana* inhabits diverse climates and exhibits varied phenology across its range. Although *A. thaliana* is an extremely well-studied model species, the relationship between geography, growing season climate and its genetic variation is poorly characterized. We used redundancy analysis (RDA) to quantify the association of genomic variation [214 051 single nucleotide polymorphisms (SNPs)] with geography and climate among 1003 accessions collected from 447 locations in Eurasia. We identified climate variables most correlated with genomic variation, which may be important selective gradients related to local adaptation across the species range. Climate variation among sites of origin explained slightly more genomic variation than geographical distance. Large-scale spatial gradients and early spring temperatures explained the most genomic variation, while growing season and summer conditions explained the most after controlling for spatial structure. SNP variation in Scandinavia showed the greatest climate structure among regions, possibly because of relatively consistent phenology and life history of populations in this region. Climate variation explained more variation among nonsynonymous SNPs than expected by chance, suggesting that much of the climatic structure of SNP correlations is due to changes in coding sequence that may underlie local adaptation.**

Keywords: biogeography, eigenanalysis ordination, population structure, principal components of neighbourhood matrices

Received 27 February 2012; revision received 10 May 2012; accepted 28 May 2012

Introduction

Spatial selective gradients can drive local adaptation such that local genotypes have greater fitness than non-local genotypes. Local adaptation may underlie a substantial portion of genotypic and phenotypic variation among populations of a species. However, the aspects of environmental variation that drive selective gradients are poorly known for most species. The importance of selective gradients and local adaptation may vary across spatial scales and within a species range (Manel *et al.* 2010; Lee & Mitchell-Olds 2011; Urban 2011), although such patterns are also poorly understood. Selective gradients and local adaptation may leave footprints on spatial genomic variation that contains rich information about environmental interactions (Manel *et al.* 2010; Sork *et al.* 2010; Lee & Mitchell-Olds 2011; Salathé & Schmid-Hempel 2011). The increasing availability of genomic marker data combined with data on geographic variation in environment can complement traditional approaches to understanding genotypic and phenotypic variation (Hancock *et al.* 2008; Fournier-Level *et al.* 2011).

Arabidopsis thaliana (Brassicaceae) is a key model dicot plant species and is a leading study system for understanding how evolution and ecology shape genomic variation (Mitchell-Olds 2001). *Arabidopsis thaliana* was

Correspondence: Jesse R. Lasky, Fax: 512 471 0961;
E-mail: jesserlasky@utexas.edu

the first plant to have its genome fully sequenced and thousands of scientists have investigated its development, physiology and molecular biology. The existing knowledge of *A. thaliana* is an immense resource for studying mechanisms of natural variation because of the unparalleled opportunities to link population ecology to molecular biology (Mitchell-Olds 2001; Tonsor *et al.* 2005; Metcalf & Mitchell-Olds 2009). Determining the ecological context of genomic variation is essential to predicting evolutionary trajectories of *A. thaliana* (Bergelson & Roux 2010). Despite these opportunities, researchers have only recently begun to connect spatial environmental variation in the field to potentially adaptive genomic variation in *A. thaliana* (Fournier-Level *et al.* 2011; Hancock *et al.* 2011).

Both climate and ecologically important traits vary extensively across the native Eurasian range of *A. thaliana* (Hoffmann 2002; Beck *et al.* 2008) and some phenotypic variation likely represents local adaptation to climate. Common garden experiments with inbred lines of wild origin, or accessions, have demonstrated phenotype correlations to latitude (Li *et al.* 1998; Stinchcombe *et al.* 2004; Lempe *et al.* 2005), altitude (Picó 2012) and climate of origin (Hannah *et al.* 2006; Christman *et al.* 2008; McKay *et al.* 2008; Hancock *et al.* 2011; Montesinos-Navarro *et al.* 2011). However, by necessity, common garden studies sample incomplete portions of climate space and are limited in their ability to reveal which climatic gradients have the strongest association with local adaptation. Although greenhouse experiments are able to isolate causal factors, they do so at the expense of limited fidelity to environmental regimes. Natural selective gradients are likely not comprised of single climate variables devised by scientists, but rather complex, possibly nonlinear, combinations of climate variables (Whittaker *et al.* 1973). Hence, analysis of large field data sets plays an important role complementary to experimental approaches.

The challenge of phenology in studying local adaptation

Intraspecific phenological variation presents a significant challenge to identifying environmental gradients underlying local adaptation because not all individuals experience climate in the same way. Climate driven natural selection depends on the timing of climate events relative to temporal life cycle patterns (Stenseth & Mysterud 2002; Helmuth *et al.* 2005; Korves *et al.* 2007). Some *A. thaliana* plants are rapid-cycling annuals, completing their life cycle within a growing season and over-wintering as seeds (Picó 2012). Other plants germinate in the fall and overwinter as rosettes, flowering in the spring (i.e. winter annuals). This life history variation is caused by both genetic and environmental variation (Koornneef *et al.* 1998; Michaels & Amasino 1999; Johanson *et al.* 2000; Stinchcombe *et al.* 2004; Wilczek *et al.* 2009).

Among other factors, *A. thaliana* phenology is affected by temperature (Wilczek *et al.* 2009), water availability and day length (Corbesier & Coupland 2005; Lempe *et al.* 2005). Growing conditions across northern Eurasia typically occur in spring, summer and fall, whereas in southern Europe and Central Asia, summer conditions are typically too dry and hot for growth. In southern Europe, growing conditions primarily occur in winter and spring (e.g. Montesinos *et al.* 2009). Ignoring this variation could lead to omission of important climate-genome correlations owing to temporal misalignment between proposed selective gradients and actual growth periods. Unfortunately, the natural phenology of most accessions is unknown.

Here, we develop models of vegetative growth phenology based on climate. Previously, Wilczek *et al.* (2009) used a model of variation at several flowering time loci to closely predict flowering time in common gardens of varied environment, based on the accumulation of photothermal units since germination. However, this model does not account for seasonal water limitation and considers limited environmental variation. Knowledge of natural germination period would be required to use such a model for our purposes. In the absence of phenological data from sites where accessions were collected, we used climate diagram models of potential growing periods based on temperature and precipitation (Walter & Lieth 1960). Climate diagrams are a valuable ecological tool to identify likely growing seasons for specific plant species and populations (e.g. McKay *et al.* 2008; Huston & Wolverton 2009).

Characterizing selective gradients

Our goal is to identify which particular gradients in a high-dimensional environment create selective gradients along which local populations are adapted. Correlation between allele frequencies and environmental gradients is evidence for local adaptation (Endler 1986; Hancock *et al.* 2008, 2011). Environmental gradients underlying local adaptation may be identified by finding multivariate gradients along which many loci show correlated variation (Manel *et al.* 2010; Sork *et al.* 2010; Lee & Mitchell-Olds 2011). Candidate environmental gradients can then be tested for genotype-dependent effects on fitness in common garden experiments.

Simple correlations can be misleading, however, because of the difficulty in separating adaptive genetic variation from variation caused by population structure, both of which frequently exhibit spatial autocorrelation. Methods commonly used in community ecology to

model simultaneous environmental and dispersal effects can be applied to genomics. We extend previous approaches to control for population structure (e.g. Hancock *et al.* 2008) using redundancy analysis (RDA), a multivariate regression technique often employed by community ecologists when both predictors and responses are multivariate (Legendre & Legendre 1998). We use RDA to disentangle the association of spatial structure (a proxy for population structure) and climate with genomic variation (Urban 2011). Additionally, we use RDA to study how: (A) the spatial structure of genomic variation changes with spatial scale, for comparison with previous studies of population structure (e.g. Sharbel *et al.* 2000) and (B) the climatic structure of genomic variation changes across regions of the range of *A. thaliana*, which has not been previously characterized.

Recent studies have identified an enrichment of nonsynonymous variants among single nucleotide polymorphisms (SNPs) associated with individual climate gradients (Hancock *et al.* 2011) and found that favoured alleles in common gardens had nonrandom climatic signatures (Fournier-Level *et al.* 2011). We build upon this work by studying climate associations with allele frequency variation in a multivariate context to identify specific climate gradients explaining genomic variation. We model the association between multivariate predictors (climate and spatial gradients) with multivariate responses (SNP allele frequencies). Modelling the multidimensional association between environment and genetic variation may more accurately capture selective gradients that are combinations of multiple climate variables (Whittaker *et al.* 1973) and their effect on allele frequencies across many loci.

Here, we characterize the association of geography and climate with genomic variation across the range of *A. thaliana*. The remainder of this paper is organized as follows. First, in novel analyses we quantify the proportion of genome-wide SNP variation explained by climate and spatial gradients and the regional change in SNP associations across the species range. Next, we identify specific climate variables that may be the strongest selective agents affecting local adaptation in *A. thaliana*. Third, we test for enrichment of climate associations among different classes of polymorphisms with varying phenotypic effects. Last, we identify outlier loci with the strongest associations to multivariate climatic gradients.

Methods

Data

Genome data. We used published data on 1307 *Arabidopsis thaliana* accessions that were genotyped at 214 051 single nucleotide polymorphisms (SNPs) (Kim *et al.* 2007; Atwell *et al.* 2010; Hancock *et al.* 2011; Horton *et al.* 2012). On average, one SNP occurred every ~500 bp in the data set, giving sufficient marker coverage to resolve variation among most genes (Kim *et al.* 2007). SNP categories of synonymous, nonsynonymous and intergenic were identified using TAIR10 as implemented by Hancock *et al.* (2011).

The SNP data set included latitude—longitude coordinates of origin for 1302 accessions. Collection locations were unknown for five accessions and these were discarded from climate analyses. Sampling was global but most dense in northern and western Europe and relatively sparse in eastern Europe and central Asia (Horton *et al.* 2012). Samples were collected by dozens of researchers over the last several decades, sometimes collecting more than one individual per population. Analyses were restricted to accessions found in the Eurasian native range (Hoffmann 2002). We also eliminated accessions that likely do not originate from their reported collection location (Anastasio *et al.* 2011), leaving a total of 1003 accessions from 447 locations across Eurasia.

Climate data. We compiled climate data for each accession collection location. Climate data sources were global in coverage and publicly available, but varied in spatiotemporal resolution and parameters.

WorldClim data were spatially interpolated from 1950 to 2000 weather station data and resolved to 30 arc-second grid squares by Hijmans *et al.* (2005). Mean monthly minimum, mean and maximum temperatures and mean monthly precipitation averaged across years of the time period were estimated by Hijmans *et al.* (2005). WorldClim additionally contains derived variables of biological importance. A measure of aridity (mean annual precipitation divided by mean annual potential evapotranspiration) using WorldClim data was also included (CGIAR-CSI Global-Aridity database; Zomer *et al.* 2007, 2008).

We used Climate Research Unit (CRU) data to estimate vapour pressure deficit (VPD). VPD is the difference between water vapour partial pressure and maximum potential pressure at a given air temperature and reflects evaporative demand on plants (Johnson & Ferrell 1983). Variation in VPD predicts variation among accessions in an important trait related to water use and stomatal conductance, suggesting this is an important selective gradient for local adaptation (Christman *et al.* 2008). CRU data are 1961–1990 weather station data interpolated to 10′ resolution (New *et al.* 2002). We took mean monthly relative humidity and temperature from CRU and calculated VPD at mean conditions (Murray 1967).

A third database was used to estimate inter-annual variability in precipitation, which may select for phenotypic plasticity in the form of drought acclimation. NCEP reanalysis data were generated on a T62 grid (resolution ~210 km) for the years 1948–2009 (*data provided by NOAA/OAR/ESRL PSD, http://www.esrl.noaa.gov/psd/*). Reanalysis is a global climate model using a variety of inputs, including both locally and remotely sensed surface and atmospheric data (Kalnay *et al.* 1996). The wide array of data used in the reanalysis model makes it less susceptible to error than climate data that is interpolated from sparse weather station data. We took monthly surface precipitation rate for grid cells and calculated each calendar month's coefficient of variation (CV) across years and the CV of annual precipitation.

The fourth and final database contained information on spatial variation in photosynthetically active radiation (PAR). The NASA/GEWEX Surface Radiation Budget 3.0 model aggregates a variety of data inputs to estimate radiation on a geographic coordinate grid with 1° cells (data available at http://eosweb.larc.nasa.gov/PRODOCS/srb/table_srb.html). We calculated average seasonal PAR for each accession location for the years 1983–2007. A table of all climate variables is provided (Table S1, Fig. S1, Supporting Information).

Predicting growing season and its climate. We used precipitation and temperature data to model the months of the year when accessions are likely to grow (climate diagram model, Walter & Lieth 1960). Potential growing months were defined as those with abundant soil moisture and mean temperature ≥4 °C (Fig. 1). Soil moisture was considered abundant in a given month if mean precipitation (mm) ≥2 * mean temperature (°C, Walter & Lieth 1960). Kas-2 and Pi-2 are high altitude accessions with no mean monthly temperatures above 4 °C, although summer months did exceed 0 °C. Months

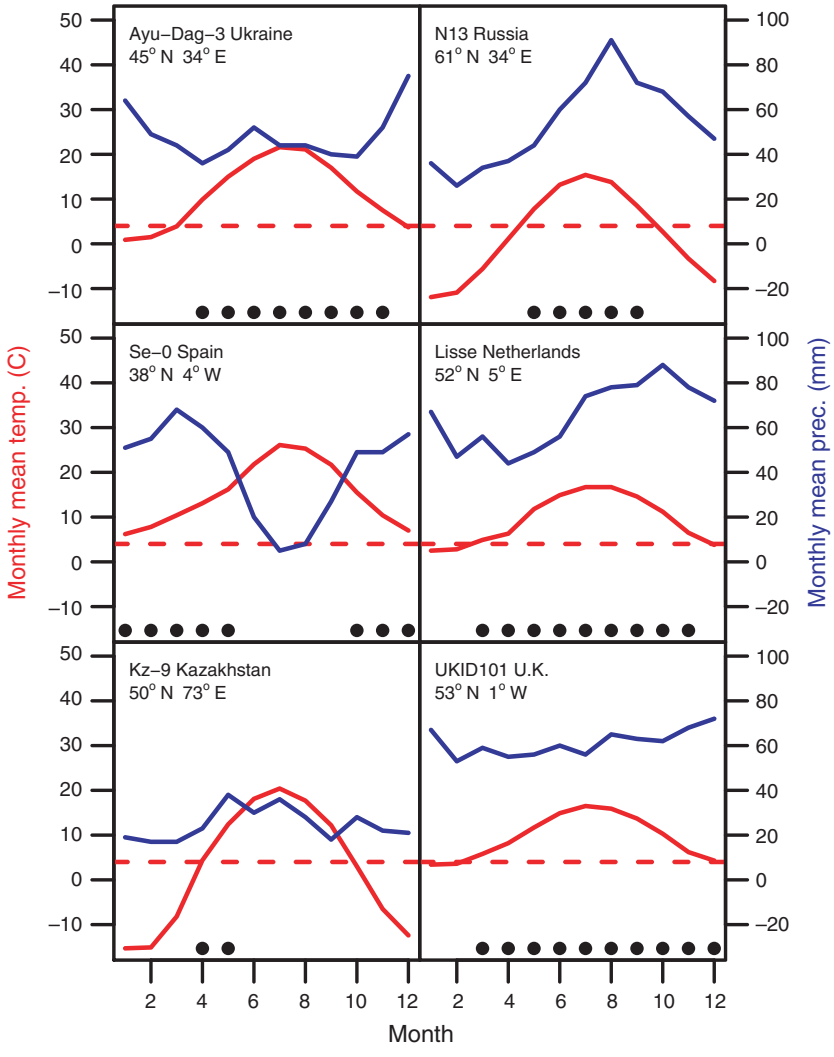

Fig. 1 Climate diagrams representative of monthly mean precipitation and temperature conditions experienced by *Arabidopsis thaliana*. When soil moisture is abundant, the precipitation line (blue) is above the temperature line (red). The dashed red line shows 4 °C, which was the minimum monthly mean temperature required for growing season months. Growing season months are indicated by black dots.

above 0 °C were considered growing season months for these two accessions.

We calculated climate conditions for the growing season months of each accession, including mean values of monthly precipitation, VPD and minimum, mean and maximum temperature. We calculated CV of mean monthly precipitation within the growing season. Finally, we calculated the inter-annual CV of each growing season month's precipitation and took the mean of monthly inter-annual CVs.

Spatial variables describing geographic variation. Geographic variation among accessions was modelled with principal components of neighbourhood matrices (PCNM), which are variables describing spatial structure (Borcard & Legendre 2002; Manel *et al.* 2010). Anisotropic and non-linear isolation by distance caused by population structure occur in *A. thaliana* (Schmid *et al.* 2006) and can be modelled by PCNM as explanatory variables in regression on genetic variation. PCNM were calculated following Borcard & Legendre (2002). A distance matrix between collection locations was created using great-circle distances along the Vincenty Ellipsoid in the R 'geosphere' package. The distance matrix was truncated above a threshold equal to the minimum distance required to form a network joining all accessions together (i.e. a minimum spanning tree). Distances above the threshold were re-assigned to four times the threshold. This threshold offers a reasonable balance between resolving fine and coarse-scale spatial structure (Borcard & Legendre 2002). We then calculated the eigenvectors of the distance matrix (i.e. PCNM) for use as predictor variables of genomic variation (e.g. Fig. 2B), keeping only eigenvectors of positive eigenvalues.

Explaining genomic variation with geography and climate

We estimated the degree to which genomic variation among accessions was explained by geographic distance and local climate. We employed redundancy analysis (RDA), a multivariate regression technique (van den Wollenberg 1977; Legendre & Legendre 1998). RDA can be used in regression problems with multivariate predictors (here, climate and space) and multivariate responses (here, biallelic SNPs). Like typical partial regression, partial RDA can be conducted on residuals from another set of explanatory variables, allowing us to control for spatial structure.

Redundancy analysis finds linear combinations of multiple explanatory variables that explain linear combinations of multiple response variables, such that the variance explained in response variables is maximized. RDA identifies multiple collinear variables that explain

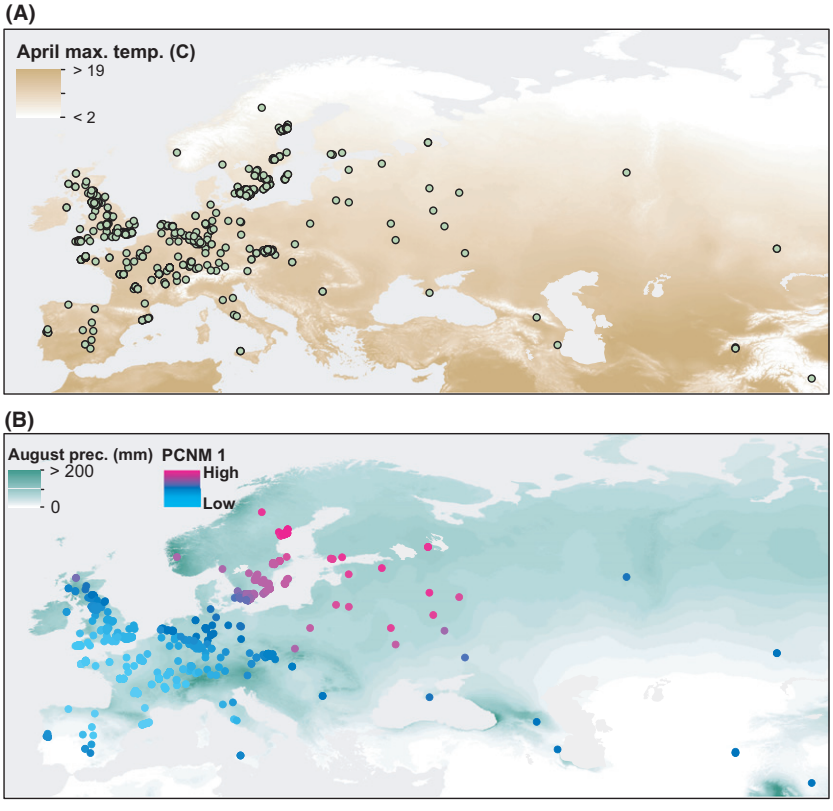

Fig. 2 1003 Eurasian accessions included in redundancy analysis (RDA). (A) Accessions are shown with April maximum temperature, which was the climate variable that explained the most SNP variation in an RDA with all accessions. (B) Accession colour varies to show the first PCNM spatial variable, describing spatial structure among accessions. August precipitation is shown, which was the second most important climate variable after removing PCNM spatial effects.

the response, any of which might be causative, in contrast to typical multiple regressions where colinearity among predictors may obscure their role. RDA linear combinations are referred to as canonical axes and are orthogonal. In our case, RDA canonical axes are composed of many covarying loci also correlated with environmental variation; in this sense they are akin to haplotype-environment correlations. Canonical eigenanalyses like RDA are increasingly used to identify environmental factors explaining genomic variation (Manel *et al.* 2010; Sork *et al.* 2010; Lee & Mitchell-Olds 2011; Salathé & Schmid-Hempel 2011). RDA and associated analyses were implemented with the 'vegan' package (Oksanen *et al.* 2011) in R.

We used the RDA framework in four different ways to study the association between climatic, geographic and genomic variation in *A. thaliana*.

Partitioning genomic variation explained by climate and geography. We used variance partitioning to estimate the total proportion of genomic variation explained by all climate and spatial variables and how these correlations change across Eurasia. Variance partitioning uses RDA to estimate how much variation in multivariate responses is explained by the independent contribution of multiple sets of explanatory variables and the contribution of their collinear portion (Peres-Neto *et al.* 2006). SNP variation was partitioned into that explained independently by all climate variables, PCNM and by their collinear portion. Colinearity between climate variables and spatial structure (PCNM) describes spatially autocorrelated climatic variation. We tested the null hypothesis that each set of climate and PCNM variables explained no SNP variation, using variance partitioning conducted on 1000 random permutations. Collection sites were permuted among groups of accessions collected at the same site (Legendre & Legendre 1998).

We also conducted variance partitioning on regional subsamples to assess patterns at smaller scales. We used five groups with sufficient sample size (minimum $n = 96$) of the eight Eurasian groups used in Horton *et al.* (2012). Sample sizes were smaller in the regional groups compared with the full Eurasian panel, so we removed some relatively redundant climate variables to avoid over-parameterizing RDA models. We removed even-numbered month climate variables from regional analyses because they were typically highly correlated with a preceding and following month (Fig. S1, Supporting Information). We conducted a complementary set of analyses where we stratified variance partitioning and RDA into two groups of accessions predicted to have different life histories (i.e. early versus late-flowering accessions, or spring versus winter annuals, see Supporting Information).

The importance of specific climate and spatial variables. After calculating the genomic variation explained by sets of climate and spatial variables (above), we used RDA to: (A) estimate how the spatial structure of genomic variation changes across spatial scales, (B) identify specific climate variables that may be important axes of local adaptation and (C) identify specific climate variables that may be important in local adaptation after correcting for spatial effects. We conducted three separate RDA, where SNP variation among accessions was the response. The predictor variables in the first RDA (A) were PCNM eigenvectors. In the second RDA (B), climate variables were predictors. The third RDA (C) was a partial RDA where we first removed effects of PCNM spatial variables as a method of controlling for population structure (Urban 2011). After removing spatial effects, climate variables were then used to explain SNP residuals (i.e. partial RDA). In RDA, where we compared the importance of specific climate variables (B and C), we only used WorldClim variables because other climate data were of much coarser resolution than WorldClim and their inclusion could have biased our comparisons of the amount of variation explained. In a supplemental analysis, we used RDA to identify climate variables explaining the most SNP variation among accessions stratified by accession flowering time category (early versus late) because life history may reinforce variation in local adaptation to climate (see Supporting Information).

The explanatory contribution of an independent variable (PCNM eigenvector or climate variable) in an RDA, P_x, was calculated using weighted sums of absolute correlations to canonical axes in RDA,

$$P_x = \frac{\sum_k |r_{xk}|\lambda_k}{\sigma^2}$$

where r_{xk} is the correlation coefficient of variable x to canonical axis k, and λ_k is the eigenvalue of axis k, equal to the variance in the SNP matrix explained by axis k. The product is summed across all axes k, giving the total variance explained by variable x and divided by the total variance in SNPs σ^2. Thus, P_x is a measure of the proportion of genomic variation among accessions explained by a predictor variable within RDA. Candidate climate variables underlying local adaptation to climate were considered those with the greatest P_x.

We estimated how geographic spatial structure of genomic variation changes across spatial scales. We compared the proportion of genomic variation explained, P_x, by each PCNM axis to the spatial scale described by that axis to estimate how genomic variation changes across spatial scales. The spatial scale

described by each PCNM axis was estimated with Moran's I, a measure of spatial autocorrelation (e.g. Manel *et al.* 2010). We conducted a nonparametric Spearman's rank correlation test between the values of P_x and I for PCNM axes.

Enrichment of SNP categories for climatic associations. We tested null hypotheses that three different classes of SNPs had similar amounts of variation explained by climate compared with random SNPs. We conducted variance partitioning on each set of SNPs: (i) nonsynonymous, (ii) synonymous, and (iii) intergenic, measuring how much variation was explained by climate and by climate independent of space. The observed portions of variation explained in each set were compared with null distributions generated by permutations of SNP classifications. For each null permutation, the classification of SNPs was shifted a random distance across the genome following Hancock *et al.* (2011). Shifting classifications maintained their order and the linkage disequilibrium of each category. After permuting classifications, we re-calculated the proportion of SNP variation in each category explained by climate to obtain a null distribution.

Identifying outlier loci. We conducted an outlier analysis with RDA results to identify loci most strongly linked with multivariate environmental gradients. Outlier loci identified in RDA can be thought of as indicators for major multi-loci haplotypes most strongly correlated with multi-variate environmental conditions. An advantage of outlier analysis in RDA is that we can identify loci correlated with the multi-variate environmental gradients experienced by plants that may be important to local adaptation, as opposed to testing climate variables individually (e.g. Hancock *et al.* 2011). Outliers were identified as SNPs with the greatest squared scores along the first RDA axis. We also identified outlier SNPs from partial RDA after removing effects of spatial structure on SNPs, because isolation by distance owing to limited dispersal can generate spurious genetic-environmental correlations.

To learn more about the function of outlier loci, we conducted a test for enrichment of gene ontology (GO) terms. GO terms are a set of standardized terms for annotation of gene functional and structural roles compiled from existing molecular literature. We selected the 1000 SNPs with the greatest squared score on the first RDA axis (i.e. those in the ~0.5% tail) of (i) RDA on raw SNPs and (ii) partial RDA on SNPs after removing the effects of spatial structure. We tested for over-representation, or enrichment, of each GO term in the set of all genes within 5 kb of the tail SNPs using the hypergeometric test (agriGO web tool; Du *et al.* 2010). We conducted false discovery rate (FDR) control on the enrichment tests and report all GO terms with FDR < 0.05.

Results

Partitioning genomic variation explained by climate and geography

Climate and space combined explained 22.6% of single nucleotide polymorphism (SNP) variation among all accessions, as determined by variance partitioning (Fig. 3, Table S4, Supporting Information). Climate and space explained less SNP variation among the regional subsets, with the exception of the subset from Scandinavia (33.5%). Climate and space explained a much larger portion of SNP variation among predicted late-flowering accessions (39.5%, see Supporting Information). The observed portions of variation explained by climate variables and by PCNM were greater than the portions explained by each set in all of 1000 permuted data sets (all permutation tests $P < 0.001$). The observed portion of SNP variation explained by growing season variables independent of spatial variables was also greater than the portion explained by each permuted data set (permutation test $P < 0.001$).

The importance of specific climate and spatial variables

In general, the spatial PCNM eigenvectors of higher rank and describing larger spatial scales, that is greater Moran's I, explained greater portions of SNP variation than eigenvectors of lower rank and smaller spatial scales (Figs S4 and S5, Supporting Information). The spatial scales of PCNM (Moran's I) were positively correlated with the SNP variation explained by each PCNM (P_x; Spearman's rank correlation $\rho = 0.69$, $P < 10^{-16}$). Accordingly, the first PCNM eigenvector, which separated northern European accessions from those in western Europe (Fig. 2B), explained the greatest portion of SNP variation, 6%.

Winter and early spring temperatures explained the greatest portions of genomic variation among Eurasian accessions (Table 1; see Fig. S8 for RDA biplot, Supporting Information). After removing the effect of spatial structure, minimum temperatures of the growing season and summer precipitation explained the greatest portions (Fig. S9, Supporting Information).

Enrichment of SNP categories for climatic association

The proportion of both nonsynonymous (NS) and synonymous (S) SNP variation explained by climate was significantly greater than the SNP variation explained

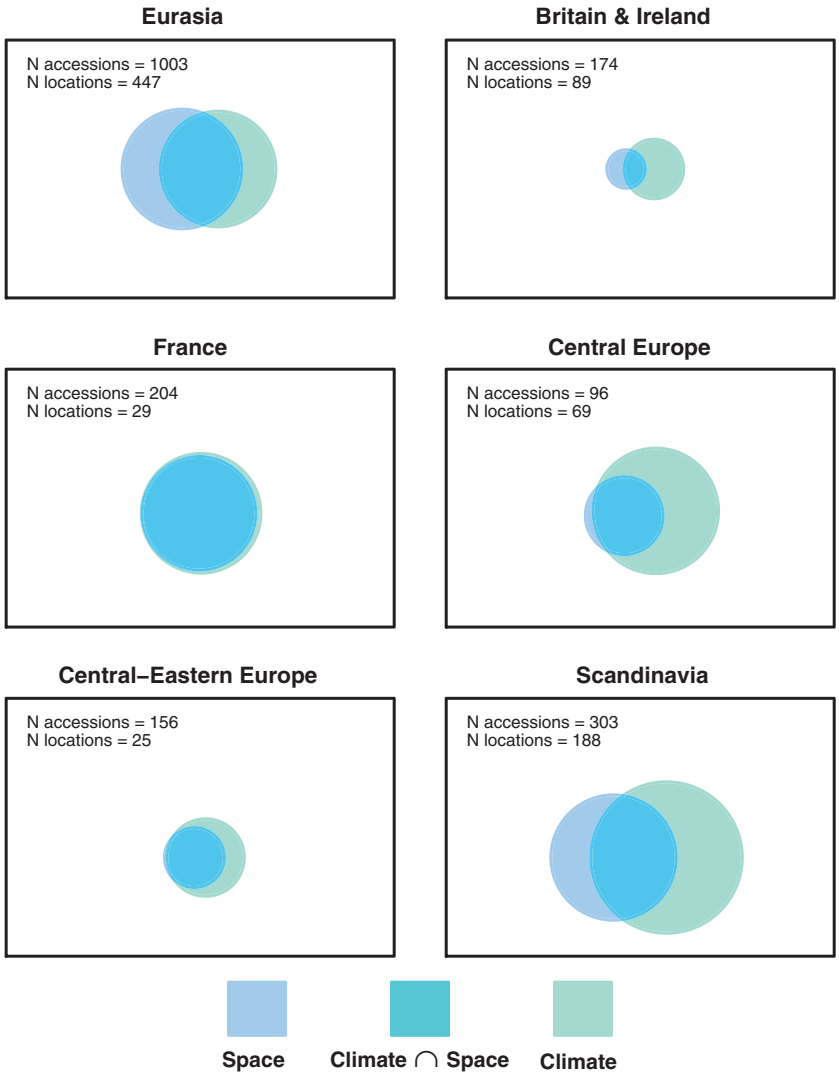

Fig. 3 Variance partitioning results for different subsets of accessions. R^2 (adjusted) of explanatory variables is represented by area, unexplained residual variation is shown in white. Even-month climate variables were removed in variance partitioning for regional subsets because regional subsets contained fewer observations.

for each category in any of 1000 permutations (two-tailed permutation tests for NS and S-SNPs, both $P < 0.002$, Fig. 4). Additionally, the proportion of NS- and S-SNP variation explained by climate independent of spatial structure was greater than the total SNP variation in all permuted SNP sets (two-tailed permutation test $P < 0.002$). Unlike variation in S- and NS-SNPs, the proportion of intergenic (IG) SNP variation explained by climate was not significantly different from null permutations (proportion explained by climate, two-tailed permutation test $P = 0.59$; proportion explained by climate independent of space, $P = 0.36$). Climate explained 0.07% more variation among S-SNPs compared with among NS-SNPs, although this difference was not significant (two-tailed permutation test between NS and S-SNPs, $P = 0.43$). Climate also explained 0.16% more variation among S-SNPs compared with among NS-SNPs after removing spatial structure and the difference was significant (two-tailed permutation test $P < 0.002$).

Outlier loci

Squared SNP scores on the first RDA axis (indicating loci associated with the first multivariate climatic axis) varied widely across the genome (Fig. 5A). The SNP with the highest squared score on the first RDA axis was in the coding region of an unknown gene next to the WRKY38 transcription factor (Table 2). Squared SNP scores on the first axis were different for partial RDA after removing the effects of spatial structure (compare panels A and B; Fig. 5). After removing spatial structure, the SNP with the highest correlation to the first RDA axis was in the intergenic region in the MAF2-5 (MADS-box affects flowering) cluster of four related transcription factors (Table 3).

Table 1 Climate variables and the per cent of single nucleotide polymorphism (SNP) variation they explain in redundancy analysis (RDA) (100*P_x). Only the top 15 climate variables are shown for each RDA

RDA on raw SNPs		Partial RDA after removing effects of spatial structure	
Climate variable	Percent of SNP variation explained	Climate variable	Percent of SNP variation explained
Max. April temp.	5.51	Mean monthly min. temp. grow. seas.	0.83
Min. February temp.	5.40	August prec.	0.83
Max. March temp.	5.39	Min. October temp.	0.82
Mean April temp.	5.36	Mean prec. grow. seas.	0.82
Min. March temp.	5.32	Prec. warmest quart.	0.82
Min. December temp.	5.32	Prec. wettest month	0.81
Min. January temp.	5.32	June prec.	0.80
Min. temp. coldest month	5.31	Prec. wettest quarter	0.79
Mean February temp.	5.24	Mean diurnal temp. range	0.79
Min. November temp.	5.22	August min. temp.	0.77
Mean temp. coldest quarter	5.20	September min. temp.	0.75
Mean January temp.	5.19	July min. temp.	0.74
Mean March temp.	5.16	May prec.	0.73
Mean November temp.	5.16	June min. temp.	0.73
Mean May temp.	5.14	Mean temp. wettest quarter	0.71

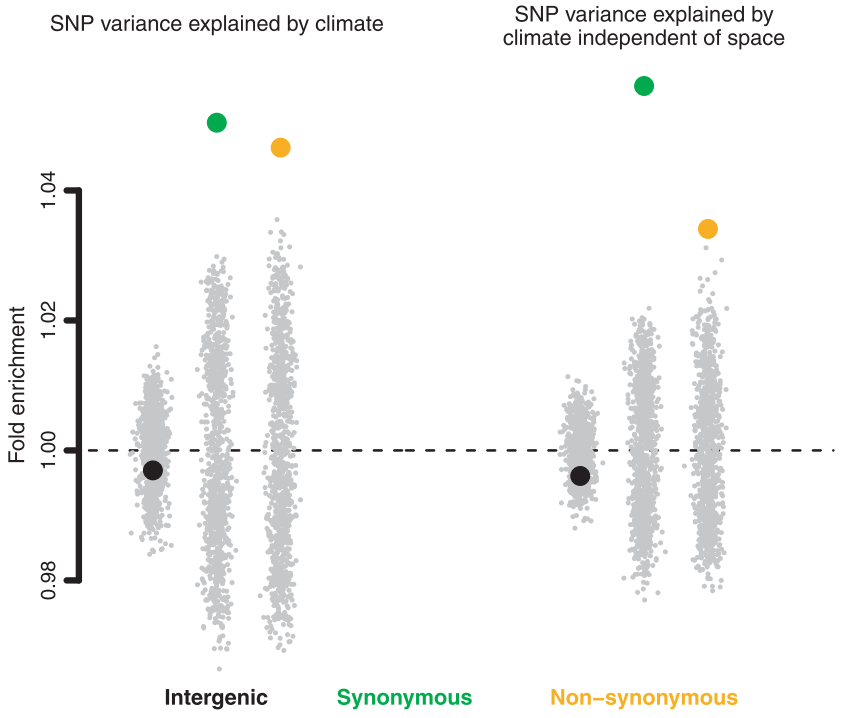

Fig. 4 Enrichment analysis of climatic structure in different classes of SNPs. The *y*-axis shows fold enrichment, which equals the portion of SNP variation explained divided by the mean portion explained in null permuted SNP sets. Grey dots represent 1000 null permutations of SNP categories. Large dots represent observed climatic structure in SNP sets.

Fifty-seven types of molecular function (GO terms) were significantly enriched (FDR < 0.05) in the tail of SNPs with the greatest squared scores on the first canonical axis (Table 4). Thirteen of the terms with the strongest enrichment were associated with stimulus responses, including a number of abiotic stress responses. After removing the effects of spatial structure, only 24 GO terms were significantly enriched (Table 4). Four of the significant terms were associated with positive regulation of cellular processes.

Discussion

Among all accessions, we found that climate (15.7%) and space (16.9%) explained roughly similar portions of genomic variation. A large portion of genomic variation

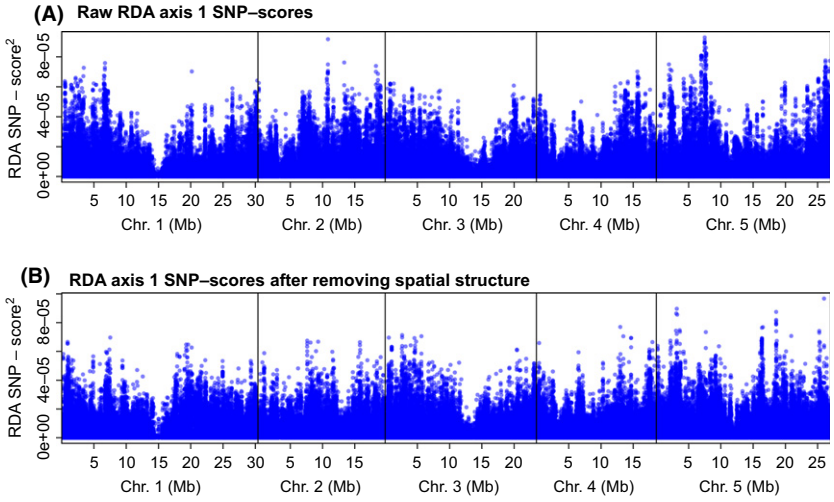

Fig. 5 Genome-wide plots of squared SNP loadings on the first multivariate climatic redundancy analysis (RDA) axes. (A) SNP loadings on the first RDA axis of raw SNP variation. (B) SNP loadings on the first RDA axis after removing the effects of spatial structure.

explained by space is likely due to population structure generating isolation by distance (Sharbel *et al.* 2000). However, latent spatially structured environmental variables, such as edaphic gradients, may be also represented by spatial variables (Manel *et al.* 2010). Climatic variation and isolation by distance among accessions may be confounded across the regions with the most sampling. In northern and western Europe, climate gradients are relatively shallow and strongly spatially autocorrelated. Accordingly, variance partitioning showed that a large portion of single nucleotide polymorphism (SNP) variation explained by climate was also spatially structured. Mountainous terrain creates sharp climatic gradients that could disrupt environment-dispersal correlations. However, sampling of *Arabidopsis thaliana* is sparse in mountainous regions. Further sampling of *A. thaliana* populations located in mountainous regions would increase our ability to disentangle the drivers of genomic variation (Beck *et al.* 2008).

Climate explained very different amounts of SNP variation across regions of the *A. thaliana* range. SNP variation in Scandinavia showed the greatest climatic structure (27% of SNP variation explained by climate). Scandinavian accessions tend to require vernalization for flowering, a phenotype associated with winter annual phenology and other physiological triats (McKay *et al.* 2003; Stinchcombe *et al.* 2004; Atwell *et al.* 2010). The relatively consistent life histories of Scandinavian accessions may result in consistent selective gradients for any given climatic variable and thus stronger SNP-climatic associations (Korves *et al.* 2007; Wilczek *et al.* 2009) compared with locations in Southern Europe where both spring and winter annuals are common (Picó 2012). Additionally, local adaptation may have been relatively stronger in Scandinavia because accessions farther south in Europe may be homogenized by anthropogenic dispersal with large-scale agriculture in central Europe (Mitchell-Olds & Schmitt 2006). Climate-genome correlations may also be weaker within other regional samples owing to sampling effects. Repeatedly sampling the same populations limits the variation in climate sampled and increases the share of total genetic variation found within populations. Scandinavia was the most extensively sampled region, although in most regions the strength of climate-SNP correlation was not clearly related to sampling differences.

Significant enrichment of SNPs in coding regions, both nonsynonymous (NS) and synonymous (S), for climatic structure suggests an adaptive basis for a substantial portion of the observed climate-SNP correlations. NS-SNPs encode phenotypic variation in amino acid sequences, which may be more likely to be associated with fitness consequences than intergenic (IG) SNPs that likely have weaker linkage to protein polymorphism. Accordingly, IG-SNPs had weaker climatic structure than random SNPs. Both NS and S-SNPs were likely enriched for climate correlation owing to hitch-hiking (Maynard Smith & Haigh 1974) with nearby (potentially uncharacterized) NS polymorphisms. While NS-SNPs were strongly enriched for climate variation, they are not necessarily the polymorphism under selection. There are many polymorphisms that were uncharacterized by the 250 k SNP chip and these may underlie locally adapted genetic variation, but many of these should be in linkage disequilibrium with nearby SNPs. Climatic selection on sites linked with nearby NS sites could increase frequency of even deleterious polymorphisms, provided that they are outweighed by fitness benefit of linked sites.

Unexpectedly, S-SNPs had greater climatic structure after removing spatial variation than NS-SNPs. There

Table 2 Gene models located within 5 kb of the top 10 single nucleotide polymorphisms (SNPs) having the greatest squared score on the first redundancy analysis (RDA) axis (RDA on raw SNPs). To remove redundant markers, SNPs were removed from the list that were within 100 kb of SNPs higher on the list

Chr.	SNP position	SNP category	Locus	Start	Stop	RDA score^2	Description (if known)
5	7493047	Synonymous	AT5G22555	7489450	7490296	9.3×10^{-5}	
			AT5G22560	7491544	7493097	9.3×10^{-5}	Plant protein of unknown function (DUF247)
			AT5G22570	7495608	7496707	9.3×10^{-5}	WRKY DNA-binding protein 38
2	10846314	Non-synonymous	AT2G25480	10843449	10845343	9.2×10^{-5}	TPX2 (targeting protein for Xklp2) protein family
			AT2G25470	10838420	10841881	9.2×10^{-5}	Receptor like protein 21
			AT2G25490	10848018	10850275	9.2×10^{-5}	EIN3-binding F box protein 1
			AT2G25482	10845884	10846348	9.2×10^{-5}	Protein of unknown function (DUF784)
5	7701756	Intergenic	AT5G23000	7696234	7697712	8.6×10^{-5}	myb domain protein 37
			AT5G23010	7703173	7706769	8.6×10^{-5}	methylthioalkylmalate synthase 1
5	6964995	Synonymous	AT5G20580	6958790	6962592	8.1×10^{-5}	
			AT5G20590	6963517	6966006	8.1×10^{-5}	Trichome birefringence-like 5
			AT5G20600	6966345	6967943	8.1×10^{-5}	
			AT5G20610	6969184	6972794	8.1×10^{-5}	
5	26203511	Synonymous	AT5G65570	26203968	26206184	7.8×10^{-5}	Tetratricopeptide repeat (TPR)-like superfamily protein
			AT5G65580	26207654	26207962	7.8×10^{-5}	
			AT5G65560	26201012	26203759	7.8×10^{-5}	Pentatricopeptide repeat (PPR) superfamily protein
			AT5G65550	26198410	26199810	7.8×10^{-5}	UDP-Glycosyltransferase superfamily protein
5	26885612	Synonymous	AT5G67385	26884754	26887083	7.7×10^{-5}	Phototropic-responsive NPH3 family protein
			AT5G67380	26881156	26883383	7.7×10^{-5}	Casein kinase alpha 1
			AT5G67390	26887883	26888512	7.7×10^{-5}	
2	13361973	Synonymous	AT2G31320	13354046	13359578	7.6×10^{-5}	poly(ADP-ribose) polymerase 2
			AT2G31340	13361614	13364633	7.6×10^{-5}	Embryo defective 1381
			AT2G31345	13365496	13365708	7.6×10^{-5}	
			AT2G31335	13360985	13361167	7.6×10^{-5}	
1	6715711	Non-synonymous	AT1G19397	6711040	6711336	7.6×10^{-5}	
			AT1G19400	6712222	6713676	7.6×10^{-5}	Erythronate-4-phosphate dehydrogenase family protein
			AT1G19410	6714492	6716439	7.6×10^{-5}	FBD/Leucine Rich Repeat domains containing protein
5	1954643	Unknown	AT5G06390	1952939	1955047	7.5×10^{-5}	FASCICLIN-like arabinogalactan protein 17 precursor
			AT5G06380	1949632	1950072	7.5×10^{-5}	
			AT5G06400	1955959	1959051	7.5×10^{-5}	Pentatricopeptide repeat (PPR) superfamily protein
2	18279230	Intergenic	AT2G44200	18276302	18278240	7.5×10^{-5}	CBF1-interacting co-repressor CIR
			AT2G44195	18274806	18275539	7.4×10^{-5}	CBF1-interacting co-repressor CIR
			AT2G44220	18283803	18285690	7.4×10^{-5}	Protein of Unknown Function (DUF239)
			AT2G44190	18272346	18274332	7.4×10^{-5}	Family of unknown function (DUF566)
			AT2G44210	18280809	18282591	7.4×10^{-5}	Protein of Unknown Function (DUF239)
			AT2G44198	18276066	18276164	7.4×10^{-5}	

Coordinates and annotation are from TAIR10 (http://www.arabidopsis.org)

are at least two possible causes of the higher enrichment of S-SNPs. First, removing spatial structure is an imperfect method of controlling for population structure and may have been slightly biased towards removing spatially structured adaptive variation in linkage disequilibrium with NS-SNPs. Accounting for spatial

Table 3 Gene models located within 5 kb of the top 10 single nucleotide polymorphisms (SNPs) having the greatest squared score on the first partial redundancy analysis (RDA) axis after removing spatial structure effects. To remove redundant markers, SNPs were removed from the list that were within 100 kb of SNPs higher on the list

Chr.	SNP position	SNP category	Locus	Start	Stop	RDA score^2	Description (if known)
5	25986868	Unknown	AT5G65060	25987527	25991065	9.7×10^{-5}	K-box region and MADS-box transcription factor family protein
			AT5G65050	25982415	25986114	9.7×10^{-5}	AGAMOUS-like 31
5	3165299	Intergenic	AT5G10120	3169732	3171147	9.0×10^{-5}	Ethylene insensitive three family protein
			AT5G10110	3166660	3167938	9.0×10^{-5}	
5	18566945	Unknown	AT5G45760	18561121	18563005	8.7×10^{-5}	Transducin/WD40 repeat-like superfamily protein
			AT5G45770	18563568	18564845	8.7×10^{-5}	Receptor like protein 55
			AT5G45780	18566946	18569625	8.7×10^{-5}	Leucine-rich repeat protein kinase family protein
			AT5G45775	18565281	18566496	8.7×10^{-5}	Ribosomal L5P family protein
4	12986185	Synonymous	AT4G25420	12990982	12992409	7.7×10^{-5}	2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein
			AT4G25410	12985772	12987149	7.7×10^{-5}	Basic helix-loop-helix (bHLH) DNA-binding superfamily protein
			AT4G25400	12981295	12982335	7.7×10^{-5}	Basic helix-loop-helix (bHLH) DNA-binding superfamily protein
5	16511395	Unknown	AT5G41260	16503997	16506970	7.7×10^{-5}	Protein kinase protein with tetratricopeptide repeat domain
			AT5G41300	16515004	16516102	7.7×10^{-5}	Receptor-like protein kinase-related family protein
			AT5G41280	16509532	16510729	7.7×10^{-5}	Receptor-like protein kinase-related family protein
			AT5G41270	16507797	16508813	7.7×10^{-5}	
			AT5G41290	16512326	16513500	7.7×10^{-5}	Receptor-like protein kinase-related family protein
5	16354866	Synonymous	AT5G40820	16343860	16353847	7.6×10^{-5}	Ataxia telangiectasia-mutated and RAD3-related
			AT5G40830	16354611	16355855	7.6×10^{-5}	S-adenosyl-L-methionine-dependent methyltransferases superfamily protein
			AT5G40840	16359611	16363722	7.6×10^{-5}	Rad21/Rec8-like family protein
5	3615859	Intergenic	AT5G11340	3619226	3621068	7.5×10^{-5}	Acyl-CoA N-acyltransferases (NAT) superfamily protein
			AT5G11330	3617342	3618861	7.5×10^{-5}	FAD/NAD(P)-binding oxidoreductase family protein
			AT5G11320	3611429	3613361	7.5×10^{-5}	Flavin-binding monooxygenase family protein
5	3046145	Intergenic	AT5G09800	3043123	3044352	7.4×10^{-5}	ARM repeat superfamily protein
			AT5G09805	3047218	3047517	7.4×10^{-5}	Inflorescence deficient in abscission (IDA)-like 3
5	25246808	Unknown	AT5G62910	25250830	25252015	7.4×10^{-5}	RING/U-box superfamily protein
			AT5G62900	25248872	25249725	7.4×10^{-5}	
			AT5G62890	25243723	25247075	7.4×10^{-5}	Xanthine/uracil permease family protein
5	7663161	Synonymous	AT5G22910	7660927	7663829	7.3×10^{-5}	Cation/H+ exchanger 9
			AT5G22900	7657224	7659868	7.3×10^{-5}	Cation/H+ exchanger 3
			AT5G22920	7665143	7667031	7.3×10^{-5}	CHY-type/CTCHY-type/RING-type Zinc finger protein

Coordinates and annotation are from TAIR10 (http://www.arabidopsis.org).

structure in genetic data is meant to remove effects of population structure. However, if the climatic gradients most important in local adaptation and associated with NS polymorphisms are strongly spatially structured, then removing spatial effects may remove a substantial portion of patterns of local adaptation. Second, environmental gradients might select for changes in *cis* regulatory sequences that are linked with S-SNPs while

Table 4 Gene ontology (GO) terms significantly enriched in the tail of single nucleotide polymorphisms (SNPs) with greatest squared SNP scores on the first canonical redundancy analysis (RDA) axes. Enrichment column gives the proportion of genes in the tail gene set belonging to the GO term divided by the proportion of genes in the genome belonging to the GO term. Only GO terms with false discovery rate (FDR) < 0.05 are shown

Analysis	Term	Enrichment	FDR
RDA on raw SNPs	Cell part	1.25	<0.0001
	Cell	1.25	<0.0001
	Intracellular part	1.34	<0.0001
	Intracellular membrane-bounded organelle	1.39	<0.0001
	Intracellular organelle	1.37	<0.0001
	Membrane-bounded organelle	1.38	<0.0001
	Organelle	1.37	<0.0001
	Intracellular	1.32	<0.0001
	Cytoplasmic part	1.31	0.0003
	Cytoplasm	1.29	0.0003
	Response to stimulus	1.48	0.0004
	Catalytic activity	1.25	0.0007
	Response to organic substance	1.80	0.0022
	Response to chemical stimulus	1.62	0.0022
	Response to endogenous stimulus	1.92	0.0022
	Cellular process	1.21	0.0024
	Binding	1.21	0.0029
	Regulation of biological process	1.43	0.0032
	Biological regulation	1.38	0.0051
	Membrane	1.32	0.0055
	Lipid storage	10.01	0.0074
	Response to hormone stimulus	1.84	0.0081
	Cold acclimation	9.01	0.0092
	Cellular response to chemical stimulus	2.28	0.0092
	Response to stress	1.50	0.0092
	Macromolecule modification	1.57	0.0120
	Mitochondrion	1.61	0.0120
	Membrane part	1.58	0.0120
	Cellular response to organic substance	2.28	0.0130
	Cellular response to stimulus	1.85	0.0140
	Nucleus	1.40	0.0150
	Lipid binding	2.63	0.0160
	Transferase activity	1.39	0.0160
	Kinase activity	1.57	0.0160
	Reproduction	1.68	0.0190
	Regulation of cellular process	1.37	0.0190
	Localization	1.50	0.0190
	Regulation of transcription, DNA-dependent	1.74	0.0190
	Transcription, DNA-dependent	1.72	0.0190
	RNA biosynthetic process	1.72	0.0190
	Regulation of RNA metabolic process	1.73	0.0200
	Intrinsic to membrane	1.71	0.0210
	Multi-organism process	1.83	0.0230
	Lipid localization	2.80	0.0260
	RNA metabolic process	1.53	0.0300
	DNA binding	1.40	0.0310
	Plastid	1.33	0.0310
	Protein modification process	1.53	0.0370
	Cellular metabolic process	1.19	0.0370
	Defense response	1.76	0.0400
	Nutrient reservoir activity	4.63	0.0400
	Post-translational protein modification	1.56	0.0410
	Post-embryonic development	1.62	0.0410
	Response to jasmonic acid stimulus	2.71	0.0410
	Cellular response to endogenous stimulus	2.23	0.0410
	Metabolic process	1.16	0.0460
	Response to auxin stimulus	2.15	0.0470
Partial RDA after removing spatial effects	Cell part	1.24	<0.0001
	Cell	1.24	<0.0001
	Intracellular	1.29	<0.0001
	Intracellular part	1.29	<0.0001
	Cytoplasm	1.32	<0.0001
	Intracellular organelle	1.29	<0.0001
	Organelle	1.29	<0.0001
	Cytoplasmic part	1.32	<0.0001
	Catalytic activity	1.26	<0.0001

Table 4 *Continued*

Analysis	Term	Enrichment	FDR
	Intracellular membrane-bounded organelle	1.28	<0.0001
	Membrane-bounded organelle	1.27	<0.0001
	Cytosol	1.88	0.0008
	Membrane	1.31	0.0027
	Organelle part	1.40	0.0055
	Intracellular organelle part	1.40	0.0055
	Biological regulation	1.38	0.0071
	Cellular process	1.18	0.0190
	Regulation of biological quality	1.95	0.0190
	Nucleus	1.35	0.0290
	Plasma membrane	1.43	0.0320
	Positive regulation of metabolic process	3.92	0.0330
	Positive regulation of biosynthetic process	4.21	0.0330
	Positive regulation of cellular biosynthetic process	4.21	0.0330
	Positive regulation of cellular metabolic process	3.92	0.0330

strong, global purifying selection on amino acid sequences might reduce environmental variation of NS-SNPs. Jones *et al.* (2012) found that most environmentally divergent sequence polymorphisms for stickleback occurred as S-SNPs and at intergenic sites, which they attributed to selection on regulatory regions. However, Jones *et al.* (2012) did not use the null permutation approach that we adopted from Hancock *et al.* (2011). The conclusions of Jones *et al.* (2012) contrast with those of Hancock *et al.* (2011), who found stronger univariate climate enrichment of NS compared with S-SNPs. These explanations for our finding are highly speculative and require further investigation.

The first spatial eigenvector (PCNM) explained the greatest portion of SNP variation and separated western Europe from the rest from Eurasia. Our finding is consistent with reports of strong east–west population structure across Eurasia (Sharbel *et al.* 2000; Nordborg *et al.* 2005; Schmid *et al.* 2006; Beck *et al.* 2008; Horton *et al.* 2012). The smallest scale PCNM eigenvectors explained relatively little genomic variation, consistent with the monotonic isolation by distance in Europe (Platt *et al.* 2010). However, Schmid *et al.* (2006) found a hump-shaped isolation by distance pattern in central Asia. The methods we employed could model hump-shaped isolation by distance, but the monotonic range-wide pattern we observed was probably dominated by heavily sampled Europe.

Early spring and winter temperatures explained the greatest portion of SNP variation among all accessions. Recent experimental evidence suggests that low temperature represents a significant selective gradient (Ågren & Schemske 2012). Accession freezing tolerance is correlated with local cold extremes (Hannah *et al.* 2006). Hoffmann (2002) found that mean April temperature best explained the northern range limit of *A. thaliana,* which was nearly congruent with the 0.1 °C isotherm, whereas April maximum temperature explained the most SNP variation in our analysis. Winter minimum temperature was the climate variable explaining the most genetic variation of the confamilial *Arabis alpina* (Manel *et al.* 2010). Additionally, Fournier-Level *et al.* (2011) found that SNP alleles associated with locally increased survival appeared to be particularly limited by temperature variables. However, much of early spring temperature variability occurs along a continental-coastal axis that may be an axis of population structure in *A. thaliana* (Nordborg *et al.* 2005).

We attempted to control for population structure by removing spatial structure via partial regression (Urban 2011) and found that minimum growing season temperatures and summer precipitation explained the most genomic variation. Variation in climate conditions during growing periods may represent critical selective gradients because annual plants do not avoid stress via dormancy at this time. We conclude that our predicted growing season minimum temperatures and precipitation are likely selective gradients driving local adaptation in *A. thaliana.* We used our results to generate hypotheses about selective pressures and the role of local adaptation and population structure in *A. thaliana*. *In situ* common garden experiments could assess the role of the April temperature and conditions during predicted growing seasons in fitness variation and local adaptation.

Three-quarters of SNP variation among all accessions remained unexplained. Variation unexplained by RDA may be attributed to many factors: (i) variation in mechanisms of adaptation (e.g. convergence), (ii) variation in life history, (iii) environmental variation occurring at scales smaller than our data, (iv) nonlinear climate-genetic relationships, (v) unmeasured selective gradients with little spatial structure, (vi) balancing selection, mutation, drift and other processes that maintain local diversity, (vii) population structure unexplained by spatial relationships and (viii) recent human-assisted dispersal that has altered spatial genomic variation from environmental selective regimes. Biplots of RDA canonical axes revealed that there were nonlinear correlations

between SNPs and climate in some cases (e.g. Fig. S12, Supporting Information), although we did not include nonlinear terms because of the high number of climate variables. Additional tools that model complex multivariate relationships will be vital for addressing this important issue in the future.

Our identification of outlier loci associated with RDA axes complements recent genome-wide association studies for climate in *A. thaliana*. Hancock *et al.* (2011) demonstrated enrichment of nonsynonymous SNPs for associations with climate variables and enrichment of certain functional processes in climate associated SNPs. Fournier-Level *et al.* (2011) found that SNP alleles associated with fitness variation across environments tended to have nonrandom distributions relative to climate. Our outlier analysis adds to these studies by identifying loci correlated with linear combinations of climatic gradients, whereas Hancock *et al.* (2011) studied single climatic variable-SNP correlations. Additionally, the loci we identified were representative of multi-loci associations with multi-variate climate gradients, which may represent local adaptation across many loci of small effect or substantial hitch-hiking (Hill & Robertson 1966).

Changes in gene expression may underlie much of local adaptation (Hodgins-Davis & Townsend 2009; Des Marais & Juenger 2010; Juenger *et al.* 2010; Des Marais *et al.* 2012). Genes involved in the transcription regulation were significantly enriched in the tail of SNPs having strong RDA associations to climate (Table 4). Transcription factors were at the top of outlier analyses (RDA and partial RDA). A SNP near WRKY38 had the strongest association to the first multivariate RDA axis. WRKY38 is known to play a role in defence against pathogens in *A. thaliana* (Kim *et al.* 2008) and cold and drought response in barley (Marè *et al.* 2004). After removing spatial structure effects via partial RDA, a SNP in the MAF2-5 (MADS-box affects flowering) cluster of transcription factors had the strongest association to the first multi-variate climate axis. MAF2-5 is similar to the floral regulator FLC and has highly polymorphic sequence and transcription that affect flowering time (Caicedo *et al.* 2009), a trait that is associated with response to abiotic stress (McKay *et al.* 2003, Korves *et al.* 2007).

Enrichment of GO terms for gene function in the tail of SNP RDA scores may indicate that RDA modelled patterns of local adaptation. Many of the terms with the most significant enrichment were responses to environmental stimuli and stress; these genes may be under divergent selection along the environmental gradients modelled by RDA. However, after controlling for spatial structure, SNPs in the tail of strong association to climate were relatively weakly enriched for GO terms. Controlling for spatial structure may have removed a large portion of genomic variation associated with local adaptation to spatially autocorrelated selective gradients.

Redundancy analysis (RDA) and related multivariate methods can be powerful tools for ecological genomics, although they have only recently been used in this context (Manel *et al.* 2010; Sork *et al.* 2010; Lee & Mitchell-Olds 2011; Salathé & Schmid-Hempel 2011). We have demonstrated how RDA, variance partitioning and PCNM can be used to determine correlations between various factors affecting genomic variation. Eigenanalyses such as RDA allow one to simplify the system by decomposing SNP, climate and spatial variables into orthogonal axes. These tools have allowed us to address a largely overlooked issue: the importance of different climate and spatial variables in explaining total genomic variation (c.f. Manel *et al.* 2010; Sork *et al.* 2010; Lee & Mitchell-Olds 2011 for fewer loci and Montesinos-Navarro *et al.* 2011; Urban 2011 for phenotype-environment correlations).

Large spatial gradients and winter, spring and growing season temperatures explained the greatest portion of SNP variation in *A. thaliana*. These patterns are likely due to both population structure and local adaptation to climate. Enrichment of climatic structure with SNPs that coded phenotypic variation for amino acid substitutions suggests fitness consequences and local adaptation are partly the source of observed correlations.

Acknowledgements

We foremost thank Magnus Nordborg, Joy Bergelson, Justin Borevitz and associates for making available the SNP data upon which our analyses depended. We thank Ginnie Morrison, Sam Taylor, Kate Behrman, Tania Pena, Betsy Kreakie, Colin Addis, Amanda Kenney, Liz Milano, Jacob Soule and Eli Meyer for their comments on this manuscript. Wei-Jia Xu of iPlant and the Texas Advanced Computing Center (TACC) provided assistance and computing. Craig Dupree of Center for Computational Biology and Bioinformatics at the University of Texas at Austin also provided computing assistance. This research was supported by NSF EF 1064901 to THK, 2010 programme funding to TEJ (DEB-0618347), JHR (DEB-0618294), JKM (DEB-0618302, DEB-1022196) and IOS 0922457 to THK and TEJ. Support from the California and Colorado Agricultural Experiment Stations is acknowledged.

Author contributions

JRL conceived and performed analyses and prepared the manuscript; DLD performed experiments and helped prepare the manuscript; JKM, JHR and TEJ helped frame the questions and analyses and prepare the manuscript; and THK helped with framing the question, directing statistical analysis and preparing the manuscript.

References

Ågren J, Schemske DW (2012) Reciprocal transplants demonstrate strong adaptive differentiation of the model organism *Arabidopsis thaliana* in its native range. *The New Phytologist*, 4, 1112–22.

Anastasio AE, Platt A, Horton M *et al.* (2011) Source verification of mis-identified Arabidopsis thaliana accessions. *The Plant Journal*, **67**, 554–566.

Atwell S, Huang YS, Vilhjalmsson BJ *et al.* (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, **465**, 627–631.

Beck JB, Schmuths H, Schaal BA (2008) *Native range genetic variation in Arabidopsis thaliana is strongly geographically structured and reflects Pleistocene glacial dynamics*. *Molecular Ecology*, **17**, 902–915.

Bergelson J, Roux F (2010) Towards identifying genes underlying ecologically relevant traits in Arabidopsis thaliana. *Nature Reviews Genetics*, **11**, 867–879.

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

Caicedo AL, Richards C, Ehrenreich IM, Purugganan MD (2009) Complex Rearrangements Lead to Novel Chimeric Gene Fusion Polymorphisms at the Arabidopsis thaliana MAF2-5 Flowering Time Gene Cluster. *Molecular Biology and Evolution*, **26**, 699–711.

Christman MA, Richards JH, Mckay JK, Stahl EA, Juenger TE, Donovan LA (2008) Genetic variation in *Arabidopsis thaliana* for night-time leaf conductance. *Plant, Cell & Environment*, **31**, 1170–1178.

Corbesier L, Coupland G (2005) Photoperiodic flowering of Arabidopsis: integrating genetic and physiological approaches to characterization of the floral stimulus. *Plant, Cell & Environment*, **28**, 54–66.

Des Marais DL, Juenger TE (2010) Pleiotropy, plasticity, and the evolution of plant abiotic stress tolerance. *Annals of the New York Academy of Sciences*, **1206**, 56–79.

Des Marais DL, McKay JK, Richards JH, Sen S, Wayne T, Juenger TE (2012) Physiological genomics of response to soil drying in diverse Arabidopsis accessions. *The Plant Cell*, **24**, 893–914.

Du Z, Zhou X, Ling Y, Zhang Z, Su Z (2010) agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Research*, **38**, W64–70.

Endler J (1986) *Natural Selection in the Wild*. Princeton University Press, New Jersey.

Fournier-Level A, Korte A, Cooper MD, Nordborg M, Schmitt J, Wilczek AM (2011) A Map of Local Adaptation in *Arabidopsis thaliana*. *Science*, **334**, 86–89.

Hancock AM, Witonsky DB, Gordon AS *et al.* (2008) Adaptations to climate in candidate genes for common metabolic disorders. *PLoS Genetics*, **4**, e32.

Hancock AM, Brachi B, Faure N *et al.* (2011) Adaptation to climate across the *Arabidopsis thaliana* genome. *Science*, **334**, 83–86.

Hannah MA, Wiese D, Freund S, Fiehn O, Heyer AG, Hincha DK (2006) Natural genetic variation of freezing tolerance in Arabidopsis. *Plant Physiology*, **142**, 98–112.

Helmuth B, Kingsolver J, Carrington E (2005) Biophysics, physiological ecology, and climate change: does mechanism matter? Annual Review of Physiology, **67**, 177–201.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Hill WG, Robertson A (1966) The effect of linkage on limits to artificial selection. *Genetical Research*, **8**, 269–294.

Hodgins-Davis A, Townsend JP (2009) Evolving gene expression: from G to E to G × E. *Trends in Ecology & Evolution*, **24**, 649–658.

Hoffmann MH (2002) Biogeography of Arabidopsis thaliana (L.) Heynh. (Brassicaceae). *Journal of Biogeography*, **29**, 125–134.

Horton MW, Hancock AM, Huang YS *et al.* (2012) Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nature Genetics*, **44**, 212–216.

Huston MA, Wolverton S (2009) The global distribution of net primary production: resolving the paradox. *Ecological Monographs*, **79**, 343–377.

Johanson U, West J, Lister C, Michaels S, Amasino R, Dean C (2000) Molecular analysis of FRIGIDA, a major determinant of natural variation in Arabidopsis flowering time. *Science*, **290**, 344–347.

Johnson JD, Ferrell WK (1983) Stomatal response to vapour pressure deficit and the effect of plant water stress. *Plant, Cell & Environment*, **6**, 451–456.

Jones FC, Grabherr MG, Chan YF *et al.* (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, **484**, 55–61.

Juenger TE, Sen S, Bray E *et al.* (2010) Exploring genetic and expression differences between physiologically extreme ecotypes: comparative genomic hybridization and gene expression studies of Kas-1 and Tsu-1 accessions of *Arabidopsis thaliana*. *Plant, Cell & Environment*, **33**, 1268–1284.

Kalnay E, Kanamitsu M, Kistler R *et al.* (1996) The NCEP/NCAR 40-year reanalysis project. *Bulletein of the American Meteorological Society*, **77**, 437–471.

Kim S, Plagnol V, Hu TT *et al.* (2007) Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genetics*, **39**, 1151–1155.

Kim K-C, Lai Z, Fan B, Chen Z (2008) Arabidopsis WRKY38 and WRKY62 Transcription Factors Interact with Histone Deacetylase 19 in Basal Defense. *The Plant Cell*, **20**, 2357–2371 (Online).

Koornneef M, Alonso-Blanco C, Peeters AJM, Soppe W (1998) Genetic control of flowering time in Arabidopsis. *Annual Review of Plant Physiology and Plant Molecular Biology*, **49**, 345–370.

Korves TM, Schmid KJ, Caicedo AL *et al.* (2007) Fitness effects associated with the major flowering time gene FRIGIDA in *Arabidopsis thaliana* in the field. *The American Naturalist*, **169**, E141–E157.

Lee C-R, Mitchell-Olds T (2011) Quantifying effects of environmental and geographical factors on patterns of genetic differentiation. *Molecular Ecology*, **20**, 4631–4642.

Legendre P, Legendre L (1998) *Numerical Ecology*, 2nd edn. Elsevier, New York.

Lempe J, Balasubramanian S, Sureshkumar S, Singh A, Schmid M, Weigel D (2005) Diversity of flowering responses in wild *Arabidopsis thaliana* strains. *PLoS Genetics*, **1**, e6.

Li B, Suzuki JI, Hara T (1998) Latitudinal variation in plant size and relative growth rate in *Arabidopsis thaliana*. *Oecologia*, **115**, 293–301.

Manel S, Poncet BN, Legendre P, Gugerli F, Holderegger R (2010) Common factors drive adaptive genetic variation at different spatial scales in *Arabis alpina*. *Molecular Ecology*, **19**, 3824–3835.

Marè C, Mazzucotelli E, Crosatti C, Francia E, Stanca AM, Cattivelli L (2004) WRKY38: a new transcription factor involved in cold- and drought-response in barley. *Plant Molecular Biology*, **55**, 399–416.

Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. *Genetical Research*, **23**, 23–35.

McKay JK, Richards JH, Mitchell-Olds T (2003) Genetics of drought adaptation in *Arabidopsis thaliana*: I. Pleiotropy contributes to genetic correlations among ecological traits. *Molecular Ecology*, **12**, 1137–1151.

McKay JK, Richards JH, Nemali KS *et al.* (2008) Genetics of drought adaptation in *Arabidopsis thaliana* II. QTL analysis of a new mapping population, Kas-1 × Tsu-1. *Evolution*, **62**, 3014–3026.

Metcalf CJE, Mitchell-Olds T (2009) Life history in a model system: opening the black box with Arabidopsis thaliana. *Ecology Letters*, **12**, 593–600.

Michaels SD, Amasino RM (1999) FLOWERING LOCUS C Encodes a novel MADS domain protein that acts as a repressor of flowering. *The Plant Cell*, **11**, 949–956 (Online).

Mitchell-Olds T (2001) *Arabidopsis thaliana* and its wild relatives: a model system for ecology and evolution. *Trends in Ecology & Evolution*, **16**, 693–700.

Mitchell-Olds T, Schmitt J (2006) Genetic mechanisms and evolutionary significance of natural variation in Arabidopsis. *Nature*, **441**, 947–952.

Montesinos A, Tonsor SJ, Alonso-Blanco C, Picó FX (2009) Demographic and genetic patterns of variation among populations of *Arabidopsis thaliana* from Contrasting Native environments. *PLoS One*, **4**, e7213.

Montesinos-Navarro A, Wig J, Pico FX, Tonsor SJ (2011) *Arabidopsis thaliana populations show clinal variation in a climatic gradient associated with altitude*. *New Phytologist*, **189**, 282–294.

Murray FW (1967) On the computation of saturation vapor pressure. *Journal of Applied Meteorology*, **6**, 203–204.

New M, Lister D, Hulme M, Makin I (2002) A high-resolution data set of surface climate over global land areas. *Climate Research*, **21**, 1–25.

Nordborg M, Hu TT, Ishino Y *et al.* (2005) The Pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biology*, **3**, e196.

Oksanen J, Blanchet FG, Kindt R *et al.* (2011). vegan: Community Ecology Package. R package version 1.17-6. [WWW document] URL http://CRAN.R-project.org/package=vegan

Peres-Neto PR, Legendre P, Dray S, Borcard D (2006) Variation partitioning of species data matrices: estimation and comparison of fractions. *Ecology*, **87**, 2614–2625.

Picó FX (2012) Demographic fate of *Arabidopsis thaliana* cohorts of autumn- and spring-germinated plants along an altitudinal gradient. *Journal of Ecology*, **100**, 1009–1018.

Platt A, Horton M, Huang YS *et al.* (2010) The Scale of population structure in *Arabidopsis thaliana*. *PLoS Genetics*, **6**, e1000843.

Salathé RM, Schmid-Hempel P (2011) The Genotypic structure of a multi-host bumblebee parasite suggests a role for ecological niche overlap. *PLoS One*, **6**, e22054.

Schmid K, Törjék O, Meyer R, Schmuths H, Hoffmann M, Altmann T (2006) Evidence for a large-scale population structure of *Arabidopsis thaliana* from genome-wide single nucleotide polymorphism markers. *TAG Theoretical and Applied Genetics*, **112**, 1104–1114.

Sharbel TF, Haubold B, Mitchell-Olds T (2000) Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. *Molecular Ecology*, **9**, 2109–2118.

Sork VL, Davis FW, Westfall R *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. *Molecular Ecology*, **19**, 3806–3823.

Stenseth NC, Mysterud A (2002) Climate, changing phenology, and other life history traits: nonlinearity and match–mismatch to the environment. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 13379–13381.

Stinchcombe JR, Weinig C, Ungerer M *et al.* (2004) A latitudinal cline in flowering time in *Arabidopsis thaliana* modulated by the flowering time gene FRIGIDA. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 4712–4717.

Tonsor SJ, Alonso-Blanco C, Koornneef M (2005) Gene function beyond the single trait: natural variation, gene effects, and evolutionary ecology in *Arabidopsis thaliana*. *Plant, Cell & Environment*, **28**, 2–20.

Urban MC (2011) The evolution of species interactions across natural landscapes. *Ecology Letters*, **14**, 723–732.

Walter H, Lieth H (1960) *Klimadiagramm-Weltatlas*. Gustav-Fischer Verlag, Jena.

Whittaker RH, Levin SA, Root RB (1973) Niche, Habitat, and Ecotope. *The American Naturalist*, **107**, 321–338.

Wilczek AM, Roe JL, Knapp MC *et al.* (2009) Effects of genetic perturbation on seasonal life history plasticity. *Science*, **323**, 930–934.

van den Wollenberg A (1977) Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika*, **42**, 207–219.

Zomer RJ, Bossio DA, Trabucco A, Yuanjie L, Gupta DC, Singh VP (2007) *Trees and Water: Smallholder Agroforestry on Irrigated Lands in Northern India*. International Water Management Institute, Colombo, Sri Lanka, pp. 45. (*IWMI Research Report 122*).

Zomer RJ, Trabucco A, Bossio DA, Verchot LV (2008) Climate change mitigation: a spatial analysis of global land suitability for clean development mechanism afforestation and reforestation. *Agriculture, Ecosystems & Environment*, **126**, 67–80.

J.R.L. is interested in the understanding the drivers of spatial biodiversity patterns, in both the genetic and ecological community levels. D.L.D. studies the molecular genetic basis of physiological adaptations to the environment and the mechanisms that drive the process of evolution. J.K.M. and T.E.J. study the ecology, evolution and genetics of local adaptation in natural plant populations. J.H.R. studies plant physiological ecology and stress physiology. T.H.K. is interested in modeling microecological mechanisms related to individual traits and physical processes to predict macroecological outcomes such as population

persistence, community organization, ecosystem function, biogeographic patterns and climate change impacts.

Data accessibility

Climate data (including growing season data), SNP data, SNP scores along RDA axes (shown in Fig. 5) and flowering time category predictions for studied accessions are available: DRYAD entry doi:10.5061/dryad.2gp18.

Supporting information

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

Table S1 All climate variables for which data were obtained.

Table S2 Flowering time experiments used as training data in SVM model.

Table S3 SNPs used as predictor variables in SVM models of flowering time (TAIR 10).

Table S4 Proportion of total SNP variation explained by climate and spatial variables.

Table S5 Bioclim abbreviations from the WorldClim data set.

Table S6 Climate variables and the percent of SNP variation among early-flowering accessions they explain in RDA ($100*P_x$).

Table S7 Climate variables and the percent of SNP variation among late-flowering accessions they explain in RDA ($100*P_x$).

Fig. S1 Correlation matrix between values of climatic variables at the 389 unique collection locations in Eurasia.

Fig. S2 Flowering times of accessions from 13 experiments used to train a genetic SVM model of early vs. late-flowering phenotype.

Fig. S3 The first two principal components of flowering time in the absence of vernalization.

Fig. S4 Histogram of the distribution of accessions along the first principal component of flowering time variation shown in Fig. S3.

Fig. S5 Standardized flowering time for 27 accessions that were used to validate previous flowering time predictions. The first plant to flower was considered day 0.

Fig. S6 Portion of SNP variation explained (P_x) by PCNM eigenvectors (only those with positive eigenvalues are shown).

Fig. S7 Portion of SNP variation explained by PCNM eigenvectors (P_x) vs. Moran's I for each eigenvector.

Fig. S8 The first two RDA axes for all accessions combined. Climate variables with the strongest correlation to each quadrant are shown.

Fig. S9 The first two RDA axes for all accessions combined. Spatial structure variables were first removed in partial RDA.

Fig. S10 The first two RDA axes for early-flowering accessions. Climate variables with the strongest correlation to each quadrant are shown.

Fig. S11 The first two RDA axes for early-flowering accessions after removing spatial structure.

Fig. S12 The first two RDA axes for late-flowering accessions. Climate variables with the strongest correlation to each quadrant are shown.

Fig. S13 The first two RDA axes for late-flowering accessions. Spatial structure variables were first removed with partial RDA.

Fig. S14 Venn diagrams of variance partitioning results for early and late-flowering accessions.

Fig. S15 Comparison of the SNP variation explained by climate variables (P_x) in early vs. late-flowering accessions.

Fig. S16 Distribution of flowering time groups across the Eurasian sample.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.