Investigating past range dynamics for a weed of cultivation, *Silene vulgaris*

Megan Elizabeth Sebasky
Andover, Massachusetts

Bachelor of Science, University of Richmond, 2010

A Thesis presented to the Graduate Faculty
of the University of Virginia in Candidacy for the Degree of
Master of Science

Department of Biology

University of Virginia
December 1, 2014

Abstract

Since the last glacial maximum (LGM) ~21 kya, when colder climates and ice sheets restricted most European taxa to southern regions, the warming climate has allowed certain species to colonize previously unsuitable regions. This phenomenon of post-glacial expansion from glacial refugia has been documented in studies of numerous plant and animal taxa. During contemporary biological invasions, species also experience dramatic range expansions but by very different mechanisms. For example, human mediated dispersal may allow species to expand into suitable, but previously unoccupied, sites. Weeds of cultivation may have spread globally following the expansion of agriculture and/or ruderal habitats associated with human-mediated disturbance. In this thesis, I tested whether the range expansion of *Silene vulgaris* across Europe fit the classical model of post-glacial expansion from southern refugia, or followed known routes of the expansion of human agricultural practices. I used Species Distribution Modeling (SDM) to predict patterns of post-glacial expansion and contrasted these with the patterns of human agricultural expansion. A population genetic analysis using microsatellite loci was then used to test which scenario was better supported by spatial patterns of genetic diversity and structure. Genetic diversity was highest in Southern Europe and declined with increasing latitude, and locations of ancestral demes from genetic cluster analysis were consistent with areas of predicted refugia. These results support post-glacial colonization while refuting the East to West agricultural spread as the main mode of expansion for *S. vulgaris*. We know that *Silene vulgaris* has recently colonized many regions (including North America and other continents) via human-mediated dispersal, but there is no evidence for a direct link between the Neolithic expansion of agriculture and current patterns of genetic diversity in Europe. Therefore, *S. vulgaris* likely participated in a long history of post-glacial expansion since the Last Glacial Maximum, but has since spread around the globe by other means.

Acknowledgements

I would like to thank my advisor Doug Taylor and previous and current Taylor lab members Stephen Keller, Deb Triant, Andrea Berardi, and Peter Fields for their help developing and executing this project, preparing presentations, and editing this document. My committee members, Ben Blackman and Laura Galloway, have also provided invaluable comments and support in the preparation and execution of my thesis. I would also like to thank my wonderful friends Amanda Hanninen, Danielle Racke, and Mike Hague for moral support, watching practice presentations, and helping edit multiple documents related to my thesis. And many thanks to my loving parents who supported me in so many ways during these two years.

Introduction

Evolutionary biologists have long been interested in how and why species ranges change over time. Large-scale climate change is known to be important for determining species distribution shifts (Davis & Shaw, 2001; McCarty, 2001; Walther et al., 2002). Historical climatic cycles associated with glacial ice ages caused the cyclical expansion and contraction of many species ranges. Evidence for the 'expansion-contraction' model (Provan & Bennett, 2008) has come from terrestrial flora and fauna in Europe and North America (reviewed in Hewitt, 2004; Schmitt, 2007). However, multiple environmental factors may structure phylogeographic patterns of genetic diversity. For example, plastic species that can survive in a wide range of climates may be better able to endure environmental changes *in situ*, and thus not adhere to the typical 'expansion-contraction' model.

Weeds of cultivation, or those widely dispersed unintentionally by human agriculture into disturbed areas, are a common example of a species that can inhabit a broad range of climates and whose expansion may not be linked to the expansion-contraction model. For a weed of cultivation, the human-mediated dispersal through agricultural expansion may be the dominant process shaping the spatial pattern of its diversity rather than the effects of post-glacial expansion since the last glacial maximum (LGM) 20,000 years ago (Balfourier et al., 2000). Previous phylogeographic studies of post-glacial expansion have mainly focused on animal and tree species. The few phylogeographic studies of widespread, weedy plant species tend to find different patterns of diversity than expected from models of post-glacial expansion most likely due to recent human-mediated dispersal (e.g. Tyler, 2002; Jiménez-Mejías et al., 2012). A recent study of two weeds of cultivation, *Lolium perenne* and *L. rigidum*, in Europe found genetic clusters correlated with historical agricultural routes (Balfourier et al., 2000), while another study found clear evidence for post-glacial expansion with multiple demes structured between different putative glacial refugia in *Arabidopsis thaliana* (Beck et al., 2008). Testing these alternative hypotheses for the expansion of the species would benefit from the use of species distribution modeling (SDM) to refine predictions about habitat-diversity relationships. In this study I use the combination of phylogeographic

analysis and SDM to examine whether *Silene vulgaris*, a weed of cultivation, has become widespread due to the common post-glacial expansion route or through recent agricultural expansion.

Recently, phylogeographic approaches have been used in concert with SDM to reconstruct past species range dynamics, especially for taxa where fossil records are not available. Both methods have their own assumptions and limitations, but when used together, are powerful for testing phylogeographic hypotheses (Waltari et al., 2007; Schorr et al., 2012, 2013; Waltari & Hickerson, 2012). Population genetic analyses allow the identification of glacial refugia areas with high genetic diversity where populations are hypothesized to be able to survive through the LGM (Hewitt, 2000; Petit et al., 2003). This signature of high genetic diversity in older, refugial populations with low diversity on the outskirts of the range would support the 'leading edge' model of range expansion in which only a subset of individuals (e.g., founders) at the expansion front establish new populations (Hewitt, 1996).

In Europe, post-glacial expansion primarily occurred by species spreading northward from southern glacial refugia on Europe's Mediterranean peninsulas (Hewitt, 2000). In contrast, the more recent agricultural spread by humans originated in the Middle East and spread westward into Europe (Ammerman & Cavalli-Sforza, 1971; Pinhasi et al., 2005). Therefore, the competing hypotheses for expansion can be tested by comparing estimated clines in genetic diversity with either latitude or longitude. A significant negative correlation between diversity and latitude would support northward post-glacial expansion from Southern refugia, while a longitudinal cline in diversity decreasing from East to West would support a model of expansion associated with the spread of agriculture.

Genetic structure analyses can also contribute to the identification of putative glacial refugia by locating ancestral source populations. Populations descended from the same glacial refugia tend to show common ancestry in genetic clustering analyses, with up to three demes in Europe whose descendents are spread latitudinally from refugia in the three southern peninsulas of Iberia, Italy, and the Balkans (Taberlet et al., 1998; Hewitt, 2004; Provan & Bennett, 2008). Similarly, demes can be identified for weeds of

cultivation that cluster based on the route of westward agricultural spread (Balfourier et al., 2000). Population genetic methods can be quite useful in determining past range dynamics, but historical processes can be difficult to disentangle from more recent events, especially in the case of contemporary admixture (Petit et al., 2003).

Species distribution models (SDMs) can also be used to hypothesize about the location of glacial refugia, augmenting findings based on population genetic data. SDMs use current species locations and environmental variables to build habitat suitability models. Built from current climate variables, these models can be projected onto climate data in different regions or at different time periods to infer species distributions in those alternative scenarios. For past range dynamics, current climate suitability models can be projected onto the reconstructed climate data for the LGM to determine putative locations of glacial refugia (Kozak et al., 2008). SDMs have many assumptions and limitations (e.g. see Araújo & New, 2007; Diniz-Filho et al., 2009) as well as high uncertainty when projecting in space and time (e.g. see Elith & Leathwick, 2009; Nogués-Bravo, 2009). However, combining SDMs with population genetic data can enable more robust assessments of historical range dynamics (Waltari et al., 2007; Schorr et al., 2012, 2013; Waltari & Hickerson, 2012).

SDMs based on current environmental variables in addition to climate can also be used to test the competing hypotheses of agricultural spread or post-glacial expansion. For a weed of cultivation that colonizes disturbed areas in various climates, land use and soil type may show higher importance than climate in an SDM. This result would suggest that the primary mode of expansion of the species was agricultural rather than post-glacial. Alternatively, if climate is more important in the SDM, it is more likely that the species tracked suitable climates. This result would also bolster confidence in predictions about past species distributions based on the reconstruction of past climates. Although genetic and SDM methods are both susceptible to limitations, using them together allows us to test the competing hypotheses for a weed of cultivation in multiple ways, allowing a more complete understanding of past range dynamics.

In this study I used the combined application of phylogeographic analysis and SDM to examine whether the weedy plant *Silene vulgaris* (Caryophyllaceae, “bladder

campion") has become widespread during an historical post-glacial expansion or through more recent agricultural expansion. Past genetic studies of *S. vulgaris* have found weak phylogeographic signatures of post-glacial expansion, with members of ancestral demes dispersed throughout Europe, making it difficult to make predictions of past range dynamics (Taylor & Keller, 2007; Keller & Taylor, 2010; Keller et al., 2014). Here I add to the data from Keller et al. (2014) by analyzing microsatellite data for additional samples with a greater representation of eastern European populations. This more robust sampling of populations across the European range allowed me to test whether the range expansion in *S. vulgaris* tended to follow post-glacial expansion northward from southern refugia since the LGM or followed the spread of agriculture, westward from the Middle East as humans created disturbed land and transported seeds.

Methods

Population samples and genotyping

I sampled 167 individuals from 73 populations across the native range of *S. vulgaris* in Europe, with one to ten individuals sampled per population. Samples were collected as seeds from maternal families or as leaf tissue dried on silica gel (Keller & Taylor, 2010). Genomic DNA was extracted from leaf tissue using Qiagen DNeasy plant mini kits. I genotyped ten of the 15 markers used in Keller et al. (2014) and derived from *S. latiolia* as described by Moccia et al., (2009) (Table 1). Microsatellite amplification and fragment analysis were performed as described in Keller et al. (2014). Genotyping and binning of the 167 new samples as well as 79 samples from Keller et al. (2014) were executed using GeneMarker 2.6.2 (Softgenetics). One marker, SL_eSSR17, was removed from the analysis due to peaks of varying sizes inconsistent with the known number of repeats. For all genetic analyses, the dataset was reduced to individuals with scores for at least six of the nine microsatellite loci to avoid issues with missing data, giving a total of 191 individuals in 76 populations (Fig. 1, Appendix S1).

Table 1: Microsatellite markers and associated genetic diversity metrics.

Locus	Indiv. scored	No. alleles	Eff. alleles	H_O	H_S	H_t	H'_t	G_{is}
SL_eSSR01	123	5	1.387	0.274	0.579	0.558	0.557	0.526
SL_eSSR03	137	12	1.42	0.323	0.584	0.816	0.819	0.447
SL_eSSR04	182	7	1.233	0.269	0.28	0.49	0.493	0.038
SL_eSSR05	172	7	1.414	0.414	0.421	0.567	0.569	0.017
SL_eSSR012	168	17	1.859	0.613	0.744	0.903	0.906	0.176
SL_eSSR016	184	9	1.455	0.408	0.478	0.632	0.634	0.145
SL_eSSR20	160	5	1.232	0.166	0.38	0.52	0.522	0.562
SL_eSSR22	174	6	1.118	0.106	0.203	0.226	0.227	0.48
SL_eSSR28	154	9	1.39	0.381	0.468	0.622	0.624	0.187
Overall		8.556	1.39	0.328	0.46	0.593	0.595	0.286

Observed heterozygosity (H_O); heterozygosity within populations (H_S); total heterozygosity (H_t); corrected total heterozygosity (H'_t); inbreeding coefficient (G_{is})

Genetic diversity

Due to many populations with small sample sizes, genetic diversity estimates were obtained at the individual level and compared to diversity estimates for populations of n>1. Multilocus heterozygosity was calculated for each individual using the standardized heterozygosity metric within the Rhh package (Alho et al., 2010), which is calculated as the proportion of heterozygous typed loci/mean heterozygosity of typed loci (Coltman et al., 1999). Observed (H_o) heterozygosity estimates for the 45 populations of n>1 were calculated in GenoDive (Meirmans & Van Tienderen, 2004). To assess the spatial distribution of genetic diversity, the two metrics were interpolated across the study area using the inverse distance weighting method through the SPATIAL ANALYST extension in ArcGIS 10.1 (ESRI, Redlands, CA, USA).

Phylogeographic structure

To assess patterns of population structure, I used Bayesian clustering to assign multilocus genotypes into clusters using the program STRUCTURE version 2.3 (Pritchard et al., 2000). I performed ten independent runs for each K (1-10) with the default program settings and 1,000,000 MCMC iterations after a burn-in period of

500,000 iterations. The optimal number of clusters (K) was determined based on the Evanno et al. (2005) method provided by STRUCTURE HARVESTER (Earl & vonHoldt, 2012). I used CLUMPP (Jakobsson & Rosenberg, 2007) to align the ancestry coefficients (Q-values) of the ten replicates. Results were visualized using *distruct* (Rosenberg, 2004) and pie charts of population-averaged ancestry coefficients were mapped using ArcGIS 10.1.

Species distribution modeling

To spatially assess putative refugia, I used SDM to predict where *S. vulgaris* occurred during the last glacial maximum based on its current environmental niche. Because the only environmental data available for the LGM is climatic, I first assessed whether the current niche of *S. vulgaris* depends more on climate than other factors that may influence its distribution. I created one SDM including climate as well as present-day environmental variables that are not available for the LGM. I created a second SDM including only climate data to project onto the LGM climate data.

Occurrence data

For both the current and LGM models, I trained the model based on D. R. Taylor's seed collection database and records from the Global Biodiversity Information Facility (GBIF) database. My initial dataset included 219 sampling locations and 14,238 post-1950 records available from GBIF with a spatial resolution of 10 km or less. However, inspection of the GBIF dataset revealed that reported occurrences in the United Kingdom were likely inaccurate because they covered nearly the entire region. I therefore excluded the GBIF occurrences in that region and supplemented with updated records obtained directly from the main source of GBIF data in that region, the Botanical Society of Britain and Ireland (BSBI). Overall, GBIF occurrence locations were severely biased towards Western Europe, with many points covering Western Europe and extremely few in Eastern Europe, where *S. vulgaris* is also widespread based on the Atlas Florae Europeae (Jalas & Suominen, 1986). To reduce spatial bias, I reduced the dataset to one point per 10 km grid cell using ArcGIS 10.1, thus matching the resolution of the environmental data as well as the resolution reported for the majority of GBIF records.

The resulting occurrence dataset for model training included 3,173 points. I further accounted for the clear spatial bias in the dataset using a bias grid (see below).

Present day SDM environmental data

For the present-day and LGM SDM, I obtained data for 19 bioclimatic variables from the WorldClim database (Hijmans et al., 2005). These variables are all derivatives of temperature and precipitation patterns known to be important in determining species distributions. I downloaded the dataset with a resolution of 2.5 arc minutes and resampled to 10-km resolution in ArcGIS 10.1. To account for the collinearity among the 19 bioclimatic variables for the present day SDM, I ran a principal components analysis (PROC PRINCOMP; SAS version 9.4, SAS Institute, 2012). The first two principal components accounted for 74.3% of the variation in the 19 bioclimatic variables (Appendix S2). For the present day model I also included available data on soil type, land use, and human influence, as these variables could be important for describing the niche of a widespread weed of cultivation. For soil type, I used multiple datasets from the European Soil Database (ESDB) version 2 at 10 km resolution: full soil code of the soil typological unit (STU) from the World Reference Base (WRB) for Soil Resources (WRB-FULL), dominant parent material of the STU (PAR-MAT-DOM), full soil code of the STU from the 1974 (modified CEC 1985) FAO-UNESCO Soil Legend (FAO85-FULL), and dominant land use (USE-DOM). I also included a more detailed land use dataset from the European Environment Agency, the Corine Land Cover 2006 database, version 16, downloaded at 250-meter resolution (Copyright © European Environment Agency). For another measure of human disturbance, I used the Last of the Wild version 2 Human Influence Index dataset at 1-km resolution (Wildlife Conservation Society - WCS & Center for International Earth Science Information Network - CIESIN - Columbia University, 2005). The Human Influence Index was created from data layers including population density, land use and infrastructure, and human access (coastlines, roads, railroads, rivers). All datasets were resampled to 10-km resolution. Because some datasets did not cover the entire study area, I extracted all data to the smallest extent of the input grids. I performed a second principal component analysis for these data to account for the inherent collinearity among the soil, land use, and human influence

datasets, (PROC PRINQUAL; SAS version 9.4, SAS Institute, 2012). The first two principal components explain 92.8% of the variation in the datasets (Appendix S3). The two bioclimatic and two land-type principal components were used as environmental variables for the SDM (Table 2).

LGM SDM environmental data

The 19 bioclimatic variables for current conditions were used to train the LGM MaxEnt model. To address collinearity issues, I ran a correlation analysis in ENMTools (Warren et al., 2010) and reduced the dataset to 7 bioclimatic variables (Table 2) with correlation coefficients of less than 0.7. From each pair of variables with R>0.7, I chose one variable to keep in the model based on variable importance and degree of extrapolation in the LGM in initial model runs. The degree of extrapolation was determined by viewing the most dissimilar variable (MoD) output maps provided by MaxEnt (see Elith et al., 2010). Climatic data for the LGM were obtained from the WorldClim database (Hijmans et al., 2005) at 2.5 arc minute-resolution for both the available datasets based on CCSM and MIROC general circulation models (GCMs) and resampled to 10-km resolution. Independent models projecting *S. vulgaris* distributions in the LGM were performed using both LGM climate datasets based on different GCMs for comparison, as neither GCM is known to be more accurate.

Table 2: Environmental variables used in the current and LGM SDMs.

Current SDM	LGM (CCSM and MIROC) SDMs
Bioclim PC1	Max. temp. warmest month
Bioclim PC2	Temp. annual range
Landcover C1	Mean temp. wettest quarter
Landcover C2	Mean temp. coldest quarter
	Precip. driest month
	Precip. seasonality
	Precip. coldest quarter

SDM procedure

I utilized a machine learning method based on maximum entropy implemented in the program MaxEnt 3.3.3k (Phillips et al., 2006) to assess the current environmental niche of *S. vulgaris* and predict its distribution during the LGM. I chose MaxEnt because it was designed for niche modeling and provides options for handling spatial bias and overfitting. To further address sample bias, in addition to rarefying out points, I created a sample bias grid using SDMToolbox version 1.0b (Brown, 2014) using the Gaussian kernel density of sampling localities. The sampling bias distance to create the grid was chosen to minimize the influence of very high sampling density in parts of Western Europe and give a projected current distribution that corresponds well with observations from the Atlas Florae Europeae and S. R. Keller & D. R. Taylor (personal observations). An optimal sampling bias distance of 20 km was chosen for the Gaussian kernel density grid. Each model was averaged for 10 replicates using default MaxEnt settings and the addition of the bias grid. For the LGM models, multivariate environmental similarity surface (MESS) and MoD maps (Elith et al., 2010) were evaluated to assess the accuracy of LGM models based on the extent of extrapolation of climate variables. The plausibility of LGM models was also assessed using knowledge of the landscape at the time. From climate reconstructions, we know that the LGM ice sheet extended south down to about 52° N and permafrost covered most areas south to 47° N (Hewitt, 2004). The MIROC model predicted substantial suitable habitat up to 50° N, and therefore was removed from the analysis (Appendix S4). The CCSM model was used for analysis because the predictions were plausible based on the knowledge of the landscape.

Correlation between genetic diversity and spatial data

I performed a correlation analysis between standardized heterozygosity and four variables: latitude, longitude, and LGM climate suitability values using JMP 9 (SAS Institute, 2012) to understand spatial patterns of genetic diversity in light of the two competing hypotheses. Standardized heterozygosity values for individuals were used rather than population-level estimates due to a majority of populations with $n < 2$. Under the assumptions described above, a significant correlation between diversity and longitude would point toward agricultural expansion being an important determinant of

current genetic diversity. Similarly, a significant correlation between diversity and latitude or LGM suitability would support the post-glacial expansion hypothesis.

Results

Genetic diversity

The spatial distributions of individual standardized heterozygosity (Fig. 1) and population observed heterozygosity were similar. I therefore used the standardized heterozygosity of individuals for subsequent analyses. No single clear latitudinal or longitudinal pattern in the genetic diversity of *S. vulgaris* across Europe was apparent. In Western Europe, heterozygosity was highest in Spain and decreased northeastward to low levels from Italy to the UK and Ireland. In Eastern Europe, diversity was highest in Greece and Belarus, and decreased to the east of these countries (Fig. 1). These patterns manifested in four latitudinal groups based on similar diversity levels: 1) high diversity in Spain, 2) low diversity from Italy to Ireland, 3) high diversity from Greece to Estonia and northwestern Russia, and 4) low diversity from Lebanon and Turkey to southwestern Russia.

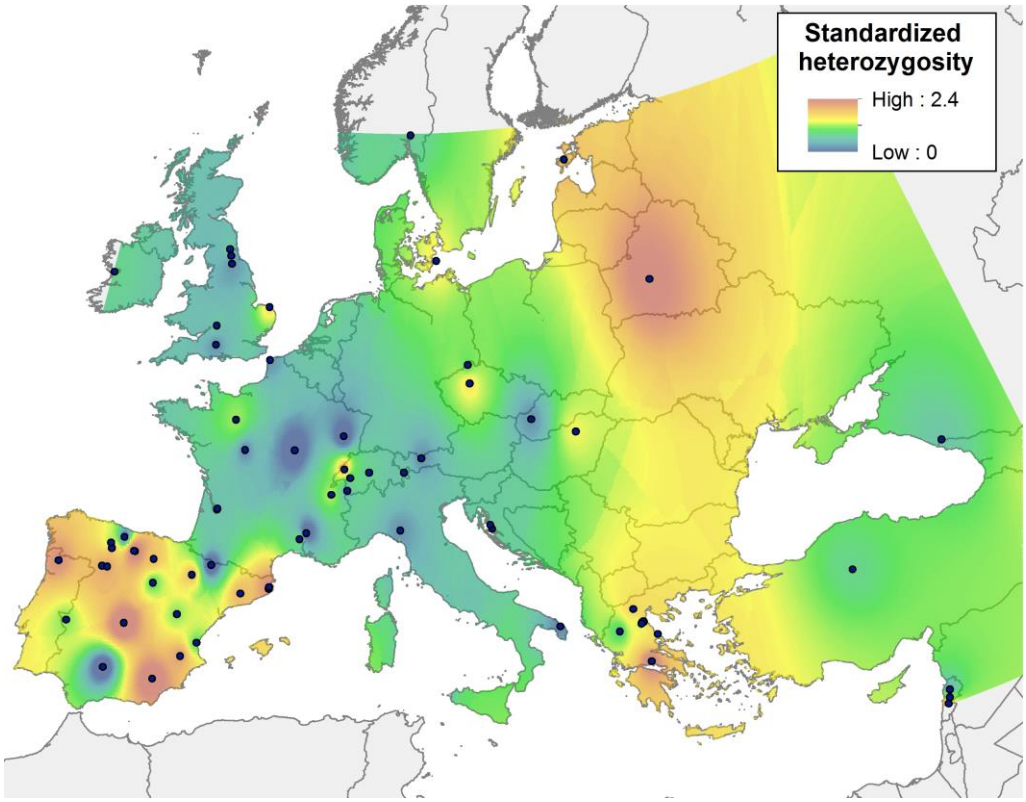


Figure 1 Standardized heterozygosity calculated for individuals and interpolated using the inverse distance weighting method (IDW) in ArcGIS 10.1. Warmer colors show higher genetic diversity, while cooler colors show lower genetic diversity. Dark blue dots show populations used in the analysis.

To assess the spatial distribution of genetic diversity, a Spearman's rank correlation was implemented because of the non-normality of the variables. Latitude was the only variable significantly correlated with heterozygosity (Table 3). Latitude had a significant negative correlation with heterozygosity, where individuals at higher (northern) latitudes have lower heterozygosity. There was no evidence of increased genetic diversity at the origins of agricultural expansion in the east, so the data are more consistent with post-glacial expansion from Southern refugia. These results persist when admixed individuals (when K=2) were removed from the analysis. In addition, when admixture is removed there was a significant ($p=0.034$) positive relationship between LGM suitability and heterozygosity.

Table 3: Correlation to assess the relationship between genetic diversity (standardized heterozygosity) and four variables: geographic location (latitude, longitude), and LGM climate suitability based on the CCSM general circulation model. Bold indicates significant *P*-values after Bonferroni correction.

Variable	Variable	Spearman's ρ	Prob>|ρ|
Latitude	Heterozygosity	-0.3478	**<.0001**
Longitude	Heterozygosity	-0.0919	0.2061
Suitability (CCSM)	Heterozygosity	0.0791	0.277
Suitability (CCSM)	Latitude	-0.4692	**<.0001**

Genetic structure

Bayesian clustering through the program STRUCTURE described an optimal model of K = 2 clusters based on the ΔK method (Evanno et al., 2005) (Appendix S5). In Western Europe, the two clusters corresponded with two geographic groups – one including only the Iberian Peninsula, and the other spanning from Italy to the UK and Ireland (Fig. 2a). In Eastern Europe, many populations were of mixed ancestry. A majority of the populations in the farthest eastern region clustered with the Iberian Peninsula populations. There was also moderate support for K = 3 clusters (Appendix S5). Many populations in eastern EU showed high posterior probability for this cluster, but it was also present in many western populations (Fig. 2b). Overall, the analysis roughly distinguishes the southwest (Iberia, green), the northwest (blue), and the eastern (yellow) populations (Fig. 2b).

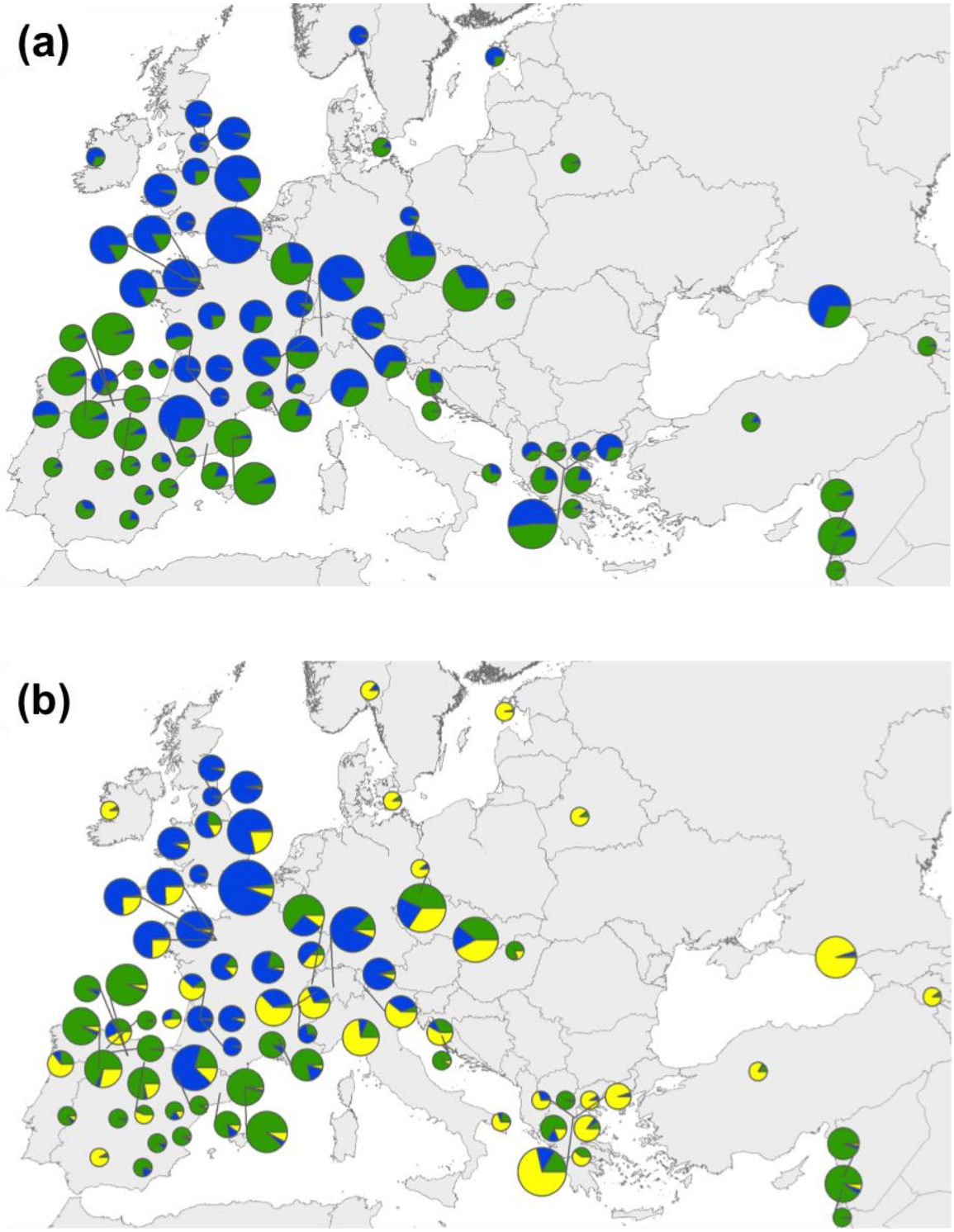


Figure 2 Ancestry assignment from STRUCTURE models. (a) Map showing pie charts of population-averaged ancestry assignment for K=2. Size of circles indicates sample size of each population. (b) Same as (a) for K=3.

SDM for current conditions

The final MaxEnt model was an adequate fit based on test AUC (0.705) and the known current distribution of the species. The AUC value metric ranges from 0 to 1 while a value of 0.5 indicates that the model is not performing better than random, and an AUC above 0.7 indicates 'fair' model performance (Swets, 1988; Araújo et al., 2005). However, the maximum achievable AUC is lower for a widespread species (Phillips et al., 2006), and widespread species typically have higher commission error. Using a bias grid of Gaussian kernel density at 20 km resulted in a prediction of current distribution consistent with field observations and the Atlas Florae Europeae (Jalas & Suominen, 1986) (Appendix S6). All variable importance metrics showed both climatic principal components having an overwhelmingly large effect when compared to the land-usage variables on the distribution of *S. vulgaris* (Table 4). A similar result was found using a model with all of the original variables before principal components analyses (data not shown).

Table 4 Percent contribution and permutation importance values for each environmental variable used in the MaxEnt model for the prediction of current distribution. Each environmental variable is a principal component axis summarizing multiple datasets. Bioclim PC1 and PC2 are the first two principal components of the 19 bioclimatic variables. Landcover C1 and C2 are the two components representing soil type, land cover, and human influence metrics.

Variable	Percent contribution	Permutation importance
Bioclim PC2	70.2	60.5
Bioclim PC1	27.5	32.1
Landcover C1	1.4	6
Landcover C2	1	1.4

Predicting LGM distribution

The MaxEnt model based on only climatic variables also showed an adequate fit based on test AUC (0.772). The prediction of current distribution (Appendix S7) is also consistent with field observations and the Atlas Florae Europeae. In fact, the fit was better for this model than the model that included both climate and land-use variables. That climate variables are of such importance makes it reasonable to predict distributions during the LGM, for which only climate data are reconstructed. The LGM prediction from the CCSM model shows moderately suitable habitat in most of the regions of Europe not covered by Eurasian ice sheets (Fig. 3) with higher suitability in southern Europe along the coasts. The model predicts high values for suitable habitat on all three European peninsulas.

A common concern with the SDM approach is the extent that findings can be affected by extrapolating variables outside the species range. Variable extrapolation was not a concern for this model because variables were only extrapolated for northern regions that were known to be uninhabitable during the LGM. Specifically, The MESS map (Appendix S8) showed that the variables in the model were only extrapolated from their training range in northern Europe, predominantly in the location of the Eurasian ice sheet, above 52° N. Even with the extrapolation, the MaxEnt models correctly predict very low suitability in this area, except farther north near Finland and Russia where there was the largest degree of extrapolation.

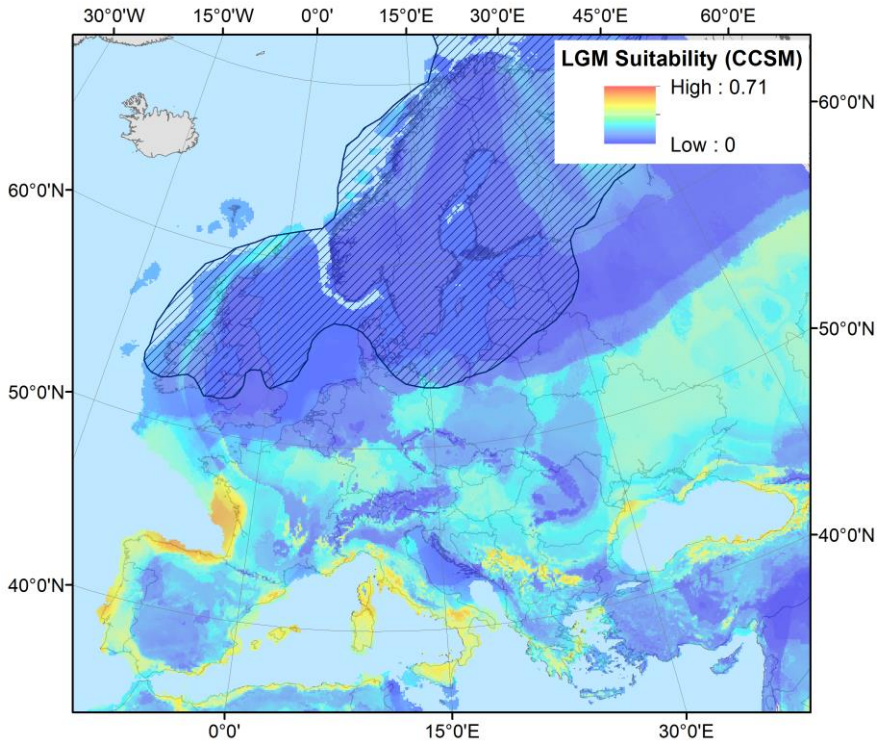


Figure 3 Predicted habitat suitability in the LGM from MaxEnt model for the CCSM climate prediction. The hashed blue area shows a generalized extent of the ice sheets during the LGM (Svendsen et al., 2004).

Discussion

I used population genetic analysis of microsatellite diversity and occurrence data for *S. vulgaris* throughout Europe in order to reconstruct past range dynamics and test two competing hypotheses for range expansion: post-glacial expansion from the South versus agricultural expansion from the East. The results support the post-glacial expansion from southern refugia hypothesis and are not consistent with predictions of agricultural spread from the East. These findings are in contrast with the only other study that has contrasted the same hypotheses for two weeds of cultivation, *Lolium perenne* and *L. rigidum* (Balfourier et al., 2000). Their finding was based on detecting genetic structure among populations along longitudinal trade routes; however this study did not estimate clines in genetic diversity that would point to centers of origin and range expansion. My findings are similar to those found in a recent study of *A. thaliana*, (Beck et al., 2008). Evidence of post-glacial expansion is still apparent in the genomes of both *A.*

thaliana and *S. vulgaris* even though admixture caused by human-mediated dispersal is expected to obscure phylogeographic patterns (Petit et al., 2003).

Genetic diversity trends support post-glacial expansion

Genetic diversity of *S. vulgaris* was higher in southern regions than northern regions. High genetic diversity in Spain and Greece supports the existence of glacial refugia in those areas, consistent with typical refugia on the southern peninsulas. The Italian peninsula also commonly served as a refugium (Taberlet et al., 1998; Hewitt, 1999). Although the spatial patterns of genetic diversity data suggest the existence of refugia on only two of the three southern peninsulas, my sampling on the Italian peninsula was sparse, including only one individual from one population in southern Italy. Further sampling could capture individuals with higher genetic diversity in this area. Another unexpected result was high genetic diversity in northeast Europe, originating in Belarus, though sampling in this region was also sparse. Further sampling to obtain more individuals per population and more populations in regions with low sampling could help decipher patterns in genetic diversity. Based on the current dataset, the overall trend of higher genetic diversity in southern regions is more consistent with post-glacial than agricultural expansion.

The significant negative correlation between heterozygosity and latitude (i.e. lower diversity at northern latitudes) is consistent with the typical post-glacial expansion. Suitability in the LGM model was also significantly negatively correlated with latitude. Taken together, the reconstruction of the climate during the LGM and current patterns of genetic diversity are consistent with expectations of a post-glacial expansion from southern European refugia. Further, when admixed individuals were removed from the analysis, there was a significant positive correlation between LGM suitability and heterozygosity. By contrast, there was not a significant correlation between heterozygosity and longitude, contrary to predictions of *S. vulgaris* migrating westward during agricultural expansion. In fact, the direction of the non-significant relationship between these variables was negative. There was also low habitat suitability in the Middle East in the current and LGM models, further supporting that the center of origin was not located in this area. These findings exemplify the benefit of combining genetic,

spatial, and suitability data to understand range dynamics and discriminate between competing hypotheses for the historical spread of weedy plants.

Genetic structure supports post-glacial expansion

In western Europe, west of Germany and Italy, the genetic clusters identified by STRUCTURE form two demes structured between two putative refugia – the Iberian and Italian peninsulas, consistent with expectations for groups of descendants from different southern glacial refugia based on previous studies (Taberlet et al., 1998; Hewitt, 2004; Provan & Bennett, 2008). My results suggest one refugia on the Iberian Peninsula with the Pyrenees mountains acting as a barrier to dispersal, a pattern seen in in other species as well (Taberlet et al., 1998; Hewitt, 2000; Schmitt, 2007). The origin of the western European deme is unclear. The LGM reconstructions suggest a possible refugium in coastal France or Italy, however, there was no clear area of higher genetic diversity within this second cluster of populations to aid in identifying a refugial location within this deme.

The population genetics of *S. vulgaris* in Eastern Europe are not straightforward. Many individuals in Eastern Europe, especially those in Lebanon, were genetically similar to individuals on the opposite edge of the species range on the Iberian Peninsula, and a second cluster originating in France/Italy separates these two regions. This pattern could have resulted from recent human-mediated dispersal, perhaps via trade routes along the North African coast (see Balfourier et al., 2000). However, a similar genetic clustering pattern was observed in a recent study of the European wild boar (Vilaça et al., 2014), which would not likely follow these trade routes. Vilaça et al. (2014) propose a scenario during the last interglacial period, where Iberian and eastern European populations could have traveled northward and become panmictic, with populations in France and Italy remaining isolated. This seems unlikely for *S. vulgaris* because the Pyrenees seem to be a barrier to dispersal that would have remained as such during past interglacial periods. However, Iberian and Lebanon populations may have originated from panmictic populations in northern Africa.

Another possible explanation for the unexpected similarity between eastern and western populations is that there are three refugial groups originating in the three southern peninsulas, but STRUCTURE was unable to separate a third cluster based on the data. The STRUCTURE analysis showed moderate support for K=3, which supports the differentiation between Iberia and eastern populations, except for those in Lebanon. With more data from eastern Europe, a third cluster could become more clear with eastern populations representing a third eastern refugial group. The difficulty in separating the third cluster could be due to recent admixture between populations descended from different refugial groups (Petit et al., 2003). Admixture is likely as *S. vulgaris* continues to travel great distances aided by human dispersal, becoming invasive in the United States and other countries. There was widespread occurrence of admixed individuals ($0.25<Q<0.75$) in this study, but the results did not differ when removing these individuals from the analysis.

SDMs support post-glacial expansion and identify potential refugia

The CCSM model predicted suitable habitat on the entire coastline of the Iberian Peninsula, western coast of France, Italy and surrounding islands, southern Greece, north of Greece around Serbia, coastal areas around the Black Sea, and the northern coast of Africa. These areas could all have potentially served as glacial refugia for *S. vulgaris*, making it possible for refugia to exist on all three southern peninsulas as seen in many other species (Taberlet et al., 1998; Hewitt, 2004; Schmitt, 2007). While there are multiple potential refugial areas within the central and eastern Europe regions based on SDMs, the genetic data do not allow a determination the specific number and locations of which of these refugia were likely to be occupied. However, the data show locations where *S. vulgaris* could exist based on SDM habitat suitability within the broad regions predicted by the genetic data. Further genetic sampling using more markers could clarify the results and make it possible to match predicted refugia locations with areas of suitable habitat.

The initial model of current *S. vulgaris* distribution showed that climate was by far the primary explanatory factor while other land-based attributes contributed little predictive value. Therefore it is likely that *S. vulgaris* tracked suitable climates since the

LGM, supporting the post-glacial expansion hypothesis. Moreover, *S. vulgaris'* high dependence on climate strengthens the findings of the LGM suitability models. These models can only use climate variables, but this is not a limitation in this study because climate is the most important factor in determining *S. vulgaris* distribution when compared with other expected important abiotic environmental variables. Many studies using paleoclimate modeling do not include estimates of how important climate is for understanding species distributions in comparison to other environmental variables potentially important for the species. It is important to check the importance of climate before projecting LGM distributions, as the LGM prediction would most likely be incorrect if climate is not important in determining the species' distribution.

Conclusion

Post-glacial expansion from glacial refugia since the LGM has been supported for a variety of taxa on many different parts of the planet. Many of these studies use genetic analyses and are beginning to use SDM as a powerful complement. However, one type of species that may not adhere to typical climate-tracking trends is a widespread weed capable of growing in a variety of climates. Weeds of cultivation may have spread primarily by agriculture rather than by post-glacial expansion, as was recently found for two *Lolium* species (Balfourier et al., 2000) but not for *Arabidopsis thaliana* (Beck et al., 2008). My results for the widespread weed *S. vulgaris* support the hypothesis of post-glacial expansion from southern refugia in Europe. This finding was stronger than those of past studies on *S. vulgaris* due to a larger sample size, addition of SDM analyses, and direct testing of two competing hypotheses. As with past phylogeographic studies of weedy plants, my results did not show the exact expected pattern of post-glacial expansion seen in other species, but still adequately support many aspects of this route, especially when compared with the alternative agriculture expansion. My results did not allow for strong predictions of exact refugial locations due to the low resolution of spatial trends in genetic diversity and structure, the inference of which would benefit from further sampling in Italy and Eastern Europe. Future studies using partial genome sequencing would be useful if additional analyses with the current microsatellites remain unclear. Finally, I found that including data from environmental variables other than climate and the known distribution of ice sheets and permafrost can significantly enhance one's confidence in the distribution of species during the LGM, and should be considered in future SDM studies.

References

Alho J.S., Välimäki K., & Merilä J. (2010) Rhh: an R extension for estimating multilocus heterozygosity and heterozygosity–heterozygosity correlation. *Molecular Ecology Resources*, **10**, 720–722.

Ammerman A.J. & Cavalli-Sforza L.L. (1971) Measuring the Rate of Spread of Early Farming in Europe. *Man*, **6**, 674.

Araújo M.B. & New M. (2007) Ensemble forecasting of species distributions. *Trends in ecology & evolution*, **22**, 42–7.

Araújo M.B., Pearson R.G., Thuiller W., & Erhard M. (2005) Validation of species–climate impact models under climate change. *Global Change Biology*, **11**, 1504–1513.

Balfourier F., Imbert C., & Charmet G. (2000) Evidence for phylogeographic structure in *Lolium* species related to the spread of agriculture in Europe. A cpDNA study. *Theoretical and Applied Genetics*, **101**, 131–138.

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

Brown J.L. (2014) SDMtoolbox: a python-based GIS toolkit for landscape genetic, biogeographic and species distribution model analyses. *Methods in Ecology and Evolution*, n/a–n/a.

Coltman D.W., Pilkington J.G., Smith J.A., & Pemberton J.M. (1999) Parasite-Mediated Selection against Inbred Soay Sheep in a Free-Living, Island Population. *Evolution*, **53**, 1259–1267.

Davis M.B. & Shaw R.G. (2001) Range shifts and adaptive responses to Quaternary climate change. *Science (New York, N.Y.)*, **292**, 673–9.

Diniz-Filho J.A.F., Mauricio Bini L., Fernando Rangel T., Loyola R.D., Hof C., Nogués-Bravo D., & Araújo M.B. (2009) Partitioning and mapping uncertainties in ensembles of forecasts of species turnover under climate change. *Ecography*, **32**, 897–906.

Earl D.A. & vonHoldt B.M. (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, **4**, 359–361.

Elith J., Kearney M., & Phillips S. (2010) The art of modelling range-shifting species. *Methods in Ecology and Evolution*, **1**, 330–342.

Elith J. & Leathwick J.R. (2009) Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 677–697.

Evanno G., Regnaut S., & Goudet J. (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology*, **14**, 2611–2620.

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

Hewitt G.M. (1996) Some genetic consequences of ice ages, and their role in divergence and speciation. *Biological Journal of the Linnean Society*, **58**, 247–276.

Hewitt G.M. (1999) Post-glacial re-colonization of European biota. *Biological Journal of the Linnean Society*, **68**, 87–112.

Hewitt G.M. (2004) Genetic consequences of climatic oscillations in the Quaternary. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **359**, 183–95; discussion 195.

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. *International Journal of Climatology*, **25**, 1965–1978.

Jalas, J., and J. Suominen. 1986. Atlas Florae Europaeae vol. 3 Cambridge Univ. Press, Cambridge, U.K.

Jakobsson M. & Rosenberg N.A. (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, **23**, 1801–1806.

Jiménez-Mejías P., Luceño M., Lye K.A., Brochmann C., & Gussarova G. (2012) Genetically diverse but with surprisingly little geographical structure: the complex history of the widespread herb *Carex nigra* (Cyperaceae). *Journal of Biogeography*, **39**, 2279–2291.

Keller S.R., Fields P.D., Berardi a. E., & Taylor D.R. (2014) Recent admixture generates heterozygosity-fitness correlations during the range expansion of an invading species. *Journal of Evolutionary Biology*, **27**, 616–627.

Keller S.R. & Taylor D.R. (2010) Genomic admixture increases fitness during a biological invasion. *Journal of evolutionary biology*, **23**, 1720–31.

Kozak K.H., Graham C.H., & Wiens J.J. (2008) Integrating GIS-based environmental data into evolutionary biology. *Trends in Ecology & Evolution*, **23**, 141–148.

McCarty J.P. (2001) Ecological Consequences of Recent Climate Change. *Conservation Biology*, **15**, 320–331.

Meirmans P.G. & Van Tienderen P.H. (2004) genotype and genodive: two programs for the analysis of genetic diversity of asexual organisms. *Molecular Ecology Notes*, **4**, 792–794.

Moccia M., Oger-Desfeux C., Marais G.A., & Widmer A. (2009) A White Campion (*Silene latifolia*) floral expressed sequence tag (EST) library: annotation, EST-SSR characterization, transferability, and utility for comparative mapping. *BMC Genomics*, **10**, 243.

Nogués-Bravo D. (2009) Predicting the past distribution of species climatic niches. *Global Ecology and Biogeography*, **18**, 521–531.

Petit R.J., Aguinagalde I., de Beaulieu J.-L., Bittkau C., Brewer S., Cheddadi R., Ennos R., Fineschi S., Grivet D., Lascoux M., Mohanty A., Müller-Starck G., Demesure-Musch B., Palmé A., Martín J.P., Rendell S., & Vendramin G.G. (2003) Glacial refugia: hotspots but not melting pots of genetic diversity. *Science (New York, N.Y.)*, **300**, 1563–5.

Phillips S.J., Anderson R.P., & Schapire R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Pinhasi R., Fort J., & Ammerman A.J. (2005) Tracing the Origin and Spread of Agriculture in Europe. *PLoS Biology*, **3**, e410.

Prentice H.C., Malm J.U., & Hathaway L. (2008) Chloroplast DNA variation in the European herb *Silene dioica* (red campion): postglacial migration and interspecific introgression. *Plant Systematics and Evolution*, **272**, 23–37.

Pritchard J.K., Stephens M., & Donnelly P. (2000) Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, **155**, 945–959.

Provan J. & Bennett K.D. (2008) Phylogeographic insights into cryptic glacial refugia. *Trends in ecology & evolution*, **23**, 564–71.

Rosenberg N.A. (2004) distruct: a program for the graphical display of population structure. *Molecular Ecology Notes*, **4**, 137–138.

Schmitt T. (2007) Molecular biogeography of Europe: Pleistocene cycles and postglacial trends. *Frontiers in zoology*, **4**, 11.

Schorr G., Holstein N., Pearman P.B., Guisan A., & Kadereit J.W. (2012) Integrating species distribution models (SDMs) and phylogeography for two species of Alpine *Primula*. *Ecology and Evolution*, **2**, 1260–1277.

Schorr G., Pearman P.B., Guisan A., & Kadereit J.W. (2013) Combining palaeodistribution modelling and phylogeographical approaches for identifying glacial refugia in Alpine *Primula*. *Journal of Biogeography*, n/a–n/a.

Svendsen J.I., Alexanderson H., Astakhov V.I., Demidov I., Dowdeswell J.A., Funder S., Gataullin V., Henriksen M., Hjort C., Houmark-Nielsen M., Hubberten H.W., Ingólfsson Ó., Jakobsson M., Kjær K.H., Larsen E., Lokrantz H., Lunkka J.P., Lyså A., Mangerud J., Matiouchkov A., Murray A., Möller P., Niessen F., Nikolskaya O., Polyak L., Saarnisto M., Siegert C., Siegert M.J., Spielhagen R.F., & Stein R. (2004) Late Quaternary ice sheet history of northern Eurasia. *Quaternary Science Reviews*, **23**, 1229–1271.

Swets J.A. (1988) Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.

Taberlet P., Fumagalli L., Wust-Saucy a G., & Cosson J.F. (1998) Comparative phylogeography and postglacial colonization routes in Europe. *Molecular ecology*, **7**, 453–64.

Taylor D.R. & Keller S.R. (2007) Historical range expansion determines the phylogenetic diversity introduced during contemporary species invasion. *Evolution; international journal of organic evolution*, **61**, 334–45.

Tyler T. (2002) Geographical distribution of allozyme variation in relation to post-glacial history in *Carex digitata*, a widespread European woodland sedge. *Journal of Biogeography*, **29**, 919–930.

Vilaça S.T., Biosa D., Zachos F., Iacolina L., Kirschning J., Alves P.C., Paule L., Gortazar C., Mamuris Z., Jędrzejewska B., Borowik T., Sidorovich V.E., Kusak J., Costa S., Schley L., Hartl G.B., Apollonio M., Bertorelle G., & Scandura M. (2014) Mitochondrial phylogeography of the European wild boar: the effect of climate on genetic diversity and spatial lineage sorting across Europe. *Journal of Biogeography*, **41**, 987–998.

Waltari E. & Hickerson M.J. (2012) Late Pleistocene species distribution modelling of North Atlantic intertidal invertebrates. *Journal of Biogeography*, **40**, 249–260.

Waltari E., Hijmans R.J., Peterson A.T., Nyári Á.S., Perkins S.L., & Guralnick R.P. (2007) Locating Pleistocene Refugia: Comparing Phylogeographic and Ecological Niche Model Predictions. *PLoS ONE*, **2**, e563.

Walther G., Post E., Convey P., Menzel A., Parmesan C., Beebee T.J.C., Fromentin J., I O.H., & Bairlein F. (2002) Ecological responses to climate change. *Nature*, **416**, 389–395.

Warren D.L., Glor R.E., & Turelli M. (2010) ENMTools: a toolbox for comparative studies of environmental niche models. *Ecography*, **33**, 607–611.

Wildlife Conservation Society - WCS & Center for International Earth Science Information Network - CIESIN - Columbia University (2005) Available at: http://dx.doi.org/10.7927/H4BP00QC.

Appendix S1: Information on populations used for genetic analysis

Population code	No. indiv.	Country	Latitude	Longitude	Collector
ALC	1	Spain	38.869	-1.098	Remedios Alarc—n
ALN1	4	France	48.414	0.089	S. Keller and J. Keller
ALN4	4	France	48.397	0.093	S. Keller and J. Keller
ALR	2	Spain	42.660	-4.312	T. Giraud
ANK	1	Turkey	40.080	33.462	Mecit Vural
BAD2	1	Germany	50.952	14.247	C. Barr, S. Keller, D. Sowell
BIS	5	Spain	41.954	2.992	T. Giraud
BKNP	3	UK	51.928	-1.971	D. Taylor, B. Penna, M. Neiman, D. Sowell
BOL	3	France	44.275	4.774	C. Barr and D. Sowell
BRV	1	Spain	42.501	-3.280	T. Giraud
BUV	5	Spain	42.659	-4.385	T. Giraud
CAL	2	Spain	41.647	1.522	T. Giraud
CAN	1	Spain	37.796	-2.284	Remedios Alarc—n
CAT1	3	UK	54.381	-1.639	S. Keller and J. Keller
CAT2	2	UK	54.381	-1.630	S. Keller and J. Keller
CON	5	Switzerland	46.847	6.713	A. Berardi, P. Fields, D. Taylor, M. Neiman
CRE1	1	France	44.834	-0.282	C. Barr and D. Sowell
CRE2	1	France	44.830	-0.287	C. Barr and D. Sowell
CRE3	2	France	44.848	-0.287	C. Barr and D. Sowell
CRE4	2	France	44.857	-0.281	C. Barr and D. Sowell
CRE5	2	France	44.821	-0.284	C. Barr and D. Sowell
CRO	6	UK	52.934	1.290	S. Keller and J. Keller
DEUX	9	France	50.885	1.704	D. Taylor
DK3	1	Denmark	54.996	12.447	D. Taylor
DPI	1	Greece	38.470	22.503	S. Ribstein
EGR	1	Hungary	47.895	20.383	D. Taylor
EJEA	1	Spain	42.126	-1.137	T. Giraud
ELB	1	Lebanon	33.700	35.700	Dave West
EPI	2	France	48.147	6.582	D. Taylor
FAN	1	Ireland	53.116	-9.285	S. Keller and J. Keller
GAS	6	Switzerland	46.758	8.135	S. Keller and J. Keller
GUA1	3	Switzerland	46.778	10.154	Andrea Berardi
JOZ	1	Spain	39.419	-7.075	Javier Tard'o
KDS	3	Lebanon	34.233	35.983	Dave West
KRA	5	Russia	43.632	40.288	J. Andreeva via Helena Storchova.
LAR1	1	Spain	42.799	-5.690	C. Barr and D. Sowell
LBL	2	Spain	42.600	-5.571	Carmen Acedo
LTR	1	Spain	37.886	-4.777	D. Taylor
MET	2	Greece	39.848	21.175	H. Frierson
MIN	1	Belarus	53.219	26.683	M. Dzhus
MOL	1	Spain	40.495	-1.595	Remedios Alarc—n, Javier Tard'o
MTOP1	2	Greece	40.073	22.462	H. Frierson
MTOP2	7	Greece	40.083	22.373	H. Frierson
MTOP3	1	Greece	0.000	0.000	H. Frierson
MTOP4	1	Greece	40.018	22.316	H. Frierson

MVRM	1	Greece	40.635	22.031	H. Frierson
NBS	4	Lebanon	33.933	35.867	Dave West
OSE1	2	Spain	43.129	-5.027	C. Barr and D. Sowell
OSL	1	Norway	59.914	10.739	D. Taylor
PAG	1	Croatia	44.441	15.053	Mirjana Vrbek
PAM	2	Greece	39.496	23.041	H. Frierson
PCO1	6	Spain	42.622	-0.194	C. Barr and D. Sowell
PIE	6	Slovakia	48.589	17.834	Lorne Wolfe
PIS	1	France	45.831	6.022	D. Taylor
POT	1	Spain	39.789	-4.168	Remedios Alarc—n
PRG4	7	Czech Rep.	50.186	14.252	C. Barr, S. Keller, and D. Sowell
PUM	1	Spain	39.477	-0.377	Remedios Alarc—n, Pilar Garcia
PZN	2	Croatia	44.615	14.965	Mirjana Vrbek
REC	3	Spain	41.563	-3.077	T. Giraud
RG	4	Italy	44.508	9.952	A. Berardi and P. Fields
RUP	4	Spain	42.021	2.993	T. Giraud
SAA	1	Estonia	58.429	22.406	Vilma Kuusk
SCN1	4	Switzerland	46.516	7.054	C. Barr, S. Keller, J. Keller, D. Sowell
SCT	1	Italy	40.353	18.174	Silvano Marchiori
SEE5	3	Austria	47.330	11.182	C. Barr, S. Keller, and D. Sowell
STH1	1	UK	51.179	-1.831	D. Taylor, B. Penna, M. Neiman, D. Sowell
STK1	2	UK	54.937	-1.918	Stephen Keller
TAB	4	Spain	41.836	-5.888	T. Giraud
TRU	2	France	47.280	0.839	D. Taylor (1999); S. Keller (2004)
UZE	2	France	44.011	4.419	T. Giraud
VAL	3	France	46.023	6.918	C. Barr and D. Sowell
VDM3	2	Portugal	41.633	-8.169	C. Barr and D. Sowell
VEZ1	3	France	47.464	3.740	S. Keller and J. Keller
VIL	4	Spain	41.848	-5.615	T. Giraud
WLW	1	UK	54.679	-1.781	S. Keller and J. Keller
YER	1	Armenia	40.160	44.509	Anush Nersesyan

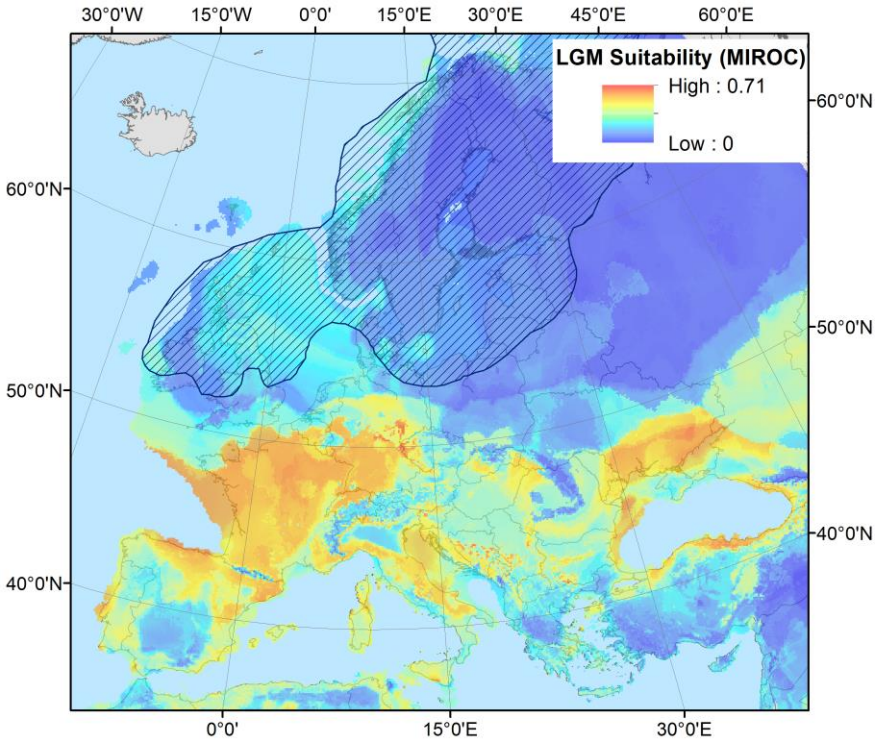
Appendix S2: Principal component loadings for the 19 Bioclim variables.

Bioclim variable name	Var. code	Prin1	Prin2
Annual mean temperature	BIO1	0.295651	0.183803
Mean diurnal range	BIO2	0.264726	-0.079484
Isothermality	BIO3	0.226693	0.221137
Temperature seasonality	BIO4	-0.030241	-0.343061
Max temp warmest month	BIO5	0.327054	0.017918
Min temp coldest month	BIO6	0.218027	0.286745
Temperature annual range	BIO7	0.076682	-0.321912
Mean temp wettest quarter	BIO8	0.002915	-0.156092
Mean temp driest quarter	BIO9	0.282637	0.193623
Mean temp warmest quarter	BIO10	0.319052	0.061528
Mean temp coldest quarter	BIO11	0.249632	0.263219
Annual precipitation	BIO12	-0.208449	0.302098
Precipitation wettest month	BIO13	-0.128839	0.293623
Precipitation driest month	BIO14	-0.276173	0.186871
Precipitation seasonality	BIO15	0.243416	-0.01037
Precipitation wettest quarter	BIO16	-0.138235	0.29895
Precipitation driest quarter	BIO17	-0.267248	0.209454
Precipitation warmest quarter	BIO18	-0.314683	0.049543
Precipitation coldest quarter	BIO19	-0.026353	0.352148

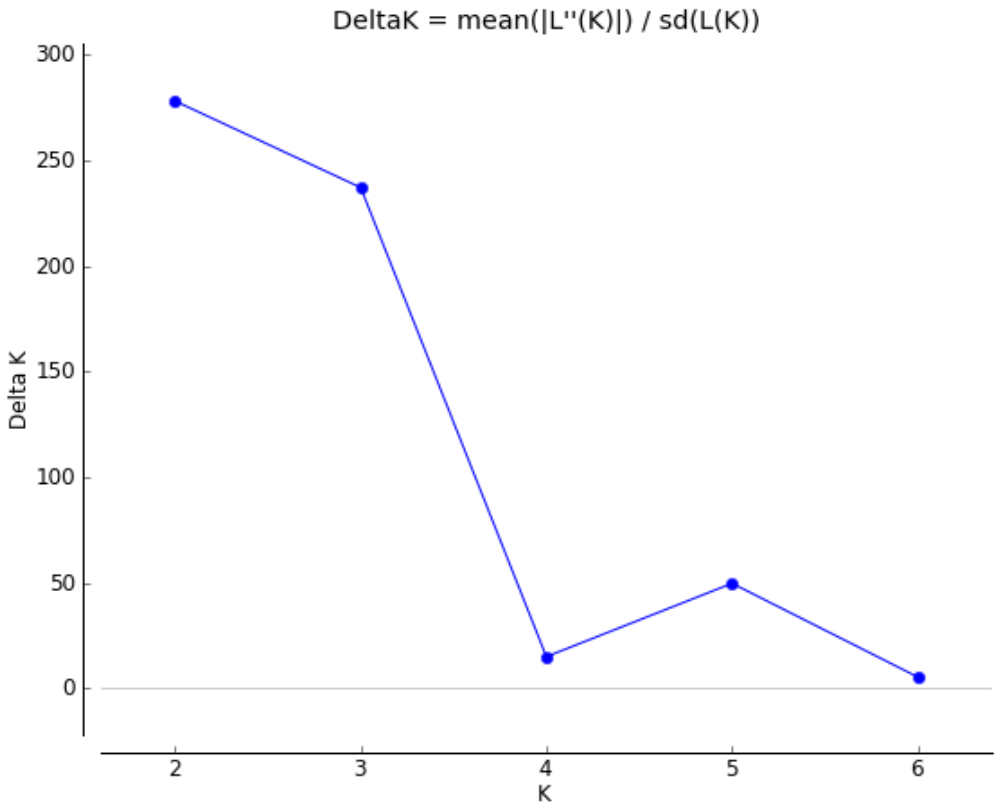
Appendix S3: Principal component loadings for land-based variables.

Variable name	Prin1	Prin2
Human influence index	0.03927	-0.552748
Full soil code from WRB	0.434563	0.024799
Dominant land use	0.454067	0.026748
Dominant parent material	-0.454068	-0.026778
Land use	-0.04486	0.551791
Full soil code (1974 FAO-UNESCO)	0.43436	0.026196

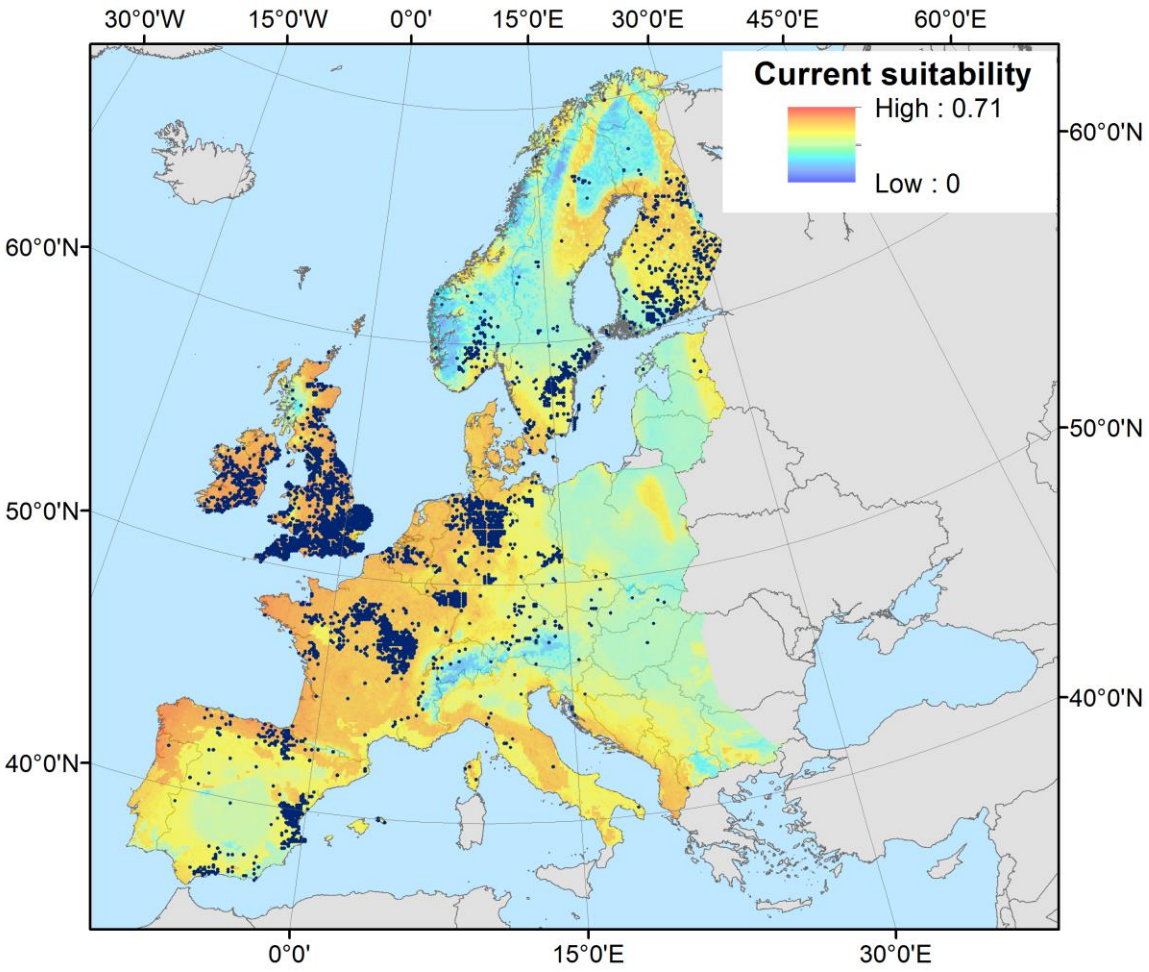
Appendix S4: LGM species distribution model for the MIROC climate scenario. Extent of the Eurasian ice sheet is sketched based on Svendsen et al., 2004.



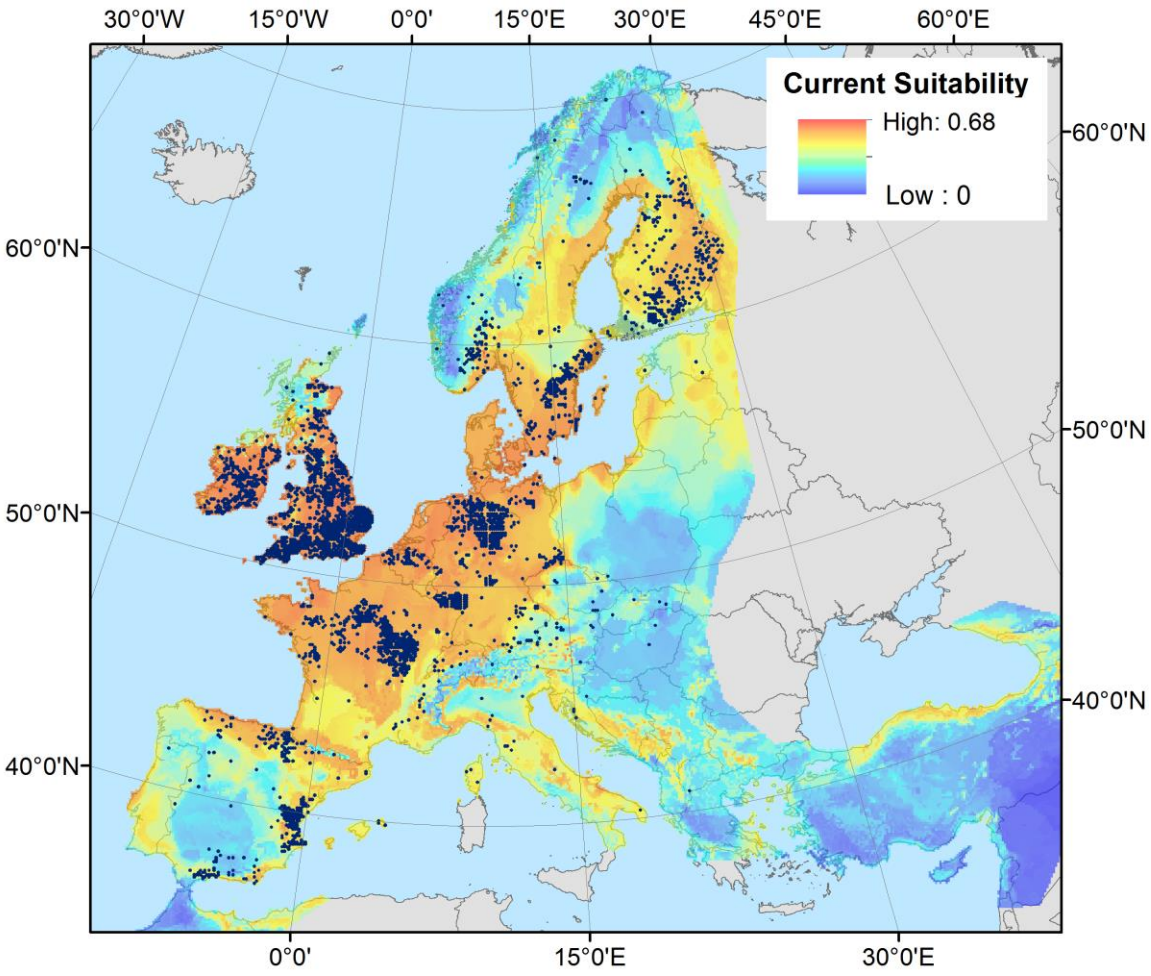
Appendix S5: Delta K likelihood for the STRUCTURE analysis.



Appendix S6: Current climate suitability based on MaxEnt model using land and climate variables. Points shown and used in the model are a combination of edited data from GBIF, BSBI, and lab collections.



Appendix S7: Current suitability from the MaxEnt model based on only climate variables. Points shown and used in the model are a combination of edited data from GBIF, BSBI, and lab collections.



Appendix S8: MESS maps for LGM Maxent models based on only climate variables. CCSM model shown above; MIROC below. Red areas represent those where one or more environmental variables are outside of their training range.