Low genetic differentiation across an entire continent: the Scots pine paradox

Jade Bruxaux[1#], Wei Zhao[1#], David Hall[1,2#], Alexandru Lucian Curtu[3], Piotr Androsiuk[4], Andreas D. Drouzas[5], Oliver Gailing[6], Heino Konrad[7], Alexis R. Sullivan[1], Vladimir Semerikov[8], Xiao-Ru Wang[1*]

[1]Department of Ecology and Environmental Science, Umeå Plant Science Center, Umeå University, Umeå, Sweden.

[2]Forestry Research Institute of Sweden (Skogforsk), Sävar, Sweden.

[3]Department of Silviculture, Transilvania University of Braşov, Braşov, Romania.

[4]Department of Plant Physiology, Genetics and Biotechnology, Faculty of Biology and Biotechnology, University of Warmia and Mazury in Olsztyn, Olsztyn, Poland.

[5]Laboratory of Systematic Botany and Phytogeography, School of Biology, Aristotle University of Thessaloniki, Thessaloniki, Greece.

[6]Department of Forest Genetics and Forest Tree Breeding, University of Göttingen, Göttingen, Germany.

[7]Austrian Research Centre for Forests (BFW), Department of Forest Biodiversity and Nature Conservation, Unit of Ecological Genetics, Vienna, Austria.

[8]Institute of Plant and Animal Ecology, Ural Division of Russian Academy of Sciences, Ekaterinburg, Russia.

[#]These authors contributed equally to the work.

[*]Correspondence to Xiao-Ru Wang, xiao-ru.wang@umu.se

Abstract

Scots pine (*Pinus sylvestris* L.) rapidly established dominance across Eurasia after the last ice age. Within this vast region, Scots pine shows remarkable ecological breadth and occurs in environments ranging from temperate rainforests to arid tundra margins. Such expansive distributions can be favored by different demographic and selective processes and the potential feedback between these forces. In this study, we conducted range-wide population genetic analyses and found a high and uniform genetic diversity across the entire range and surprisingly little impact of the last glacial maximum on historical population sizes. Genetic-environmental associations identified only a few SNPs that were significantly linked with environmental gradients. Our results suggest that extensive gene flow is responsible for the genetic patterns observed in Scots pine, and open exciting opportunities for further exploration into how genetic adaptation evolves under gene flow and changing environment.

Introduction

Life history traits, demographic events, and selection shape the amount and distribution of genetic diversity across a species range[1]. Outcrossing species are generally expected to have diverse and weak genetic structure compared to inbreeding species, but their greater adaptive potential could promote divergence in a heterogeneous fitness landscape[2,3]. Demographic changes, such as those induced by range contraction and expansion, can erode genetic diversity and increase differentiation via genetic drift[4-7]. Contingencies like these influence the contemporary genetic structure of a species[4-6] and, importantly, can constrain their ability to track, tolerate, and adapt to changing environmental gradients in the future. Understanding the legacy of demography, geography, and selection is particularly important for boreal forest trees, where repeated colonization and extinction during the Quaternary glaciations may determine their adaptive potential under intensifying climate change at high latitudes[8].

Scots pine (*Pinus sylvestris* L.) is the foundation species of diverse forested ecosystems across Eurasia. It has the widest distribution of any pine species, with a range spanning from the Iberian and Balkan Peninsulas north to the Barents Sea and from Scotland east to the Russian Pacific coast (Figure 1)[9]. It has an exceptionally wide niche, tolerating annual averaged temperatures ranging from -3°C to 16°C, and annual precipitations between 400 and 2,900 mm, with temperatures of the coldest month ranging from -17°C to 11°C, and precipitations during the driest month between 5 and 130 mm[10]. It can be found from the sea level in its most northern populations and up to 2,600 m above the sea level in the Caucasus[10]. It is a light demanding species able to grow on very nutrient poor sites[10]. The variable habitat types that Scots pine occupies do appear to have promoted distinct local adaptation as reflected by characteristic growth forms, ecotypes and clines in physiological traits[9,11]. A clear illustration of local adaptation is that northern populations set buds and develop frost hardiness significantly earlier and show higher cold tolerance than southern populations[12-14]. Clinal trait variation is a response to spatially varying selection along environmental gradients, which is expected to generate gene frequency clines at associated loci[15].

The current distribution of Scots pine has long been considered as being shaped by migration out of several scattered refugia that survived the last glacial maximum (LGM) in southern Europe, Turkey and the Russian Plain[16,17]. Mitochondrial polymorphisms suggested that Scots pine expanded into Fennoscandia from west-central Europe and the Russian Plain 10,000 – 7,000 years ago as they tracked receding ice sheets from ice age refugia[18,19]. However, the discovery of a tree dated at 11,700 before present (BP) in central Sweden, the

continuous presence of conifers during the interval 32-16,000 BP in central and eastern Europe, the high chloroplast diversity at intermediate latitudes for 22 European trees and shrubs, as well as a vegetation model reconstruction for *P. sylvestris* during the LGM question the importance of the species distribution contraction during that period and suggest the existence of northern refugia[20-23].

Theory predicts that rapid colonization from few isolated refugia would leave distinct genetic signature in the extant populations such as declining genetic diversity along migration routes due to drift, and genetic lineages that track their recent ancestry[24,25]. If so, we would expect that geographically separated populations of Scots pine, due to its long generation times, have been and still are under the influence of both genetic drift and local adaptation. However, if the species survived in many refugia scattered across the current distribution, we should observe very little decline of the diversity in the north and a reduced differentiation between populations. Testing these hypotheses requires comprehensive coverage of the species distribution and sufficient genetic information. However, genome-wide patterns of genetic diversity across the species range, as well as the impact of landscapes and environments on population differentiation, remain scarcely studied for Scots pine (but see[26,27]), hindering the recovery of its evolutionary history and of its potential to manage climate change.

We genotyped 2,321 Scots pine using genotyping-by-sequencing from more than 200 populations across its natural range. This is the largest collection of genotypes ever amassed for this species at this scale. Our primary goal was to understand whether the Quaternary glacial cycles have left Scots pine with distinct genetic structure and whether its past success sets the stage for inevitable demise. Specifically, we aim to 1) determine the level and structure of genetic diversity in Scots pine at regional and range-wide scales to understand the impact of geography and gene flow on the spatial genetic patterns, 2) establish the demographic history of Scots pine to understand the timeline of its diversification and accompanying population size changes; 3) evaluate the genetic and environmental associations to comprehend climatic drivers of population genetic differentiation and local adaptation. By providing better insights into how diversity correlates with geography, demography, and climate adaptation, we can better predict how the species may respond to changing environmental conditions and inform management plans for the future.

Results

Genotype dataset

We genotyped 2,895 individuals of Scots pine from 211 natural stands covering the entire species distribution range. After quality and relatedness filtering, we kept 2,321 individuals from 202 stands in our final dataset (Figure 1A, Supplementary Figure 1, Supplementary Table 1). We used the likelihood approach in ANGSD[28] to calculate genotype probabilities after mapping the sequenced reads against a concatenated reference based on the *Pinus taeda* genome v.1.01[29,30]. Samples averaged 2.42 million (M) mapped reads (0.33 – 49.56 M reads), 2.83 Mbp (million base pairs) with at least 5x depth (1.37 – 7.58 Mbp), and a read depth of 85.26x (17.38 – 953.61x). After quality filtering, we obtained a final genotype likelihood dataset which included 1,432,399 sites, of which 386,426 were SNPs. Filtering for minor allele frequency (MAF) > 0.05 reduced the dataset to 25,830 SNPs.

Genetic diversity and population structure

To ensure adequate sample sizes for location-based diversity estimations, we merged neighboring poorly-sampled stands, resulting in 123 larger populations for subsequent population genetic analyses (minimum six samples, on average 19). Three sample sites that were both small and isolated were left unmerged (Supplementary Table 1). We estimated the pairwise nucleotide diversity (π) among samples at each location and found little variation among them with an average value of 0.0062 differences per base pair (Table 1). Over the whole distribution, π values decreased slightly with both increased latitude and longitude (Supplementary Figure 2). In general, the 0-fold pairwise nucleotide diversity (π_0, where any mutation would be non-synonymous) was around half of the 4-fold pairwise nucleotide diversity (π_4, where any mutation would be synonymous), indicating either negative selection spanning the entire genome, or more probably an increase of the population size. The same tendency to lower values at high latitude and longitude was present for both π_0 and π_4 (Supplementary Figure 2). The average observed heterozygosity (H_o) was 0.258, and the expected heterozygosity (H_e) was 0.277 (Table 1, Supplementary Table 1), with a small decrease at high latitudes, and no variation along the longitudinal gradient (Supplementary Figure 2). As expected for a species that is highly outcrossing, we recovered a very low inbreeding coefficient (F_{IS}), in general close to 0 at all locations (on average 0.059, ranging from -0.088 to 0.2; Table 1, Supplementary Table 1). Slightly higher F_{IS} values were observed in northern Scandinavia (e.g., Alta, Vågå or Kirkesmoen in Norway) and a few

isolated populations (Black Sea or Massif Central in France), with a general tendency for marginally higher values at high latitudes (Supplementary Figure 2).

To gain an overview of the distribution of the genetic diversity present in our dataset, we performed a principal component analysis (PCA) on the full collection (Figure 1B, Supplementary Figure 3A). We observed a pattern congruent with the geographic distribution: the first axis (PC1) followed a rough South-North direction in west-central Europe, starting from Spain on the left and reaching Scandinavia on the right; the second axis (PC2) followed a more west-east direction for northern European, Russian, and Asian populations. The distribution was rather continuous, with only the Spanish populations being separated from the rest of the samples. The first two PC axes, however, explained less than 2% of the overall genetic variation, meaning that almost all the diversity is shared among all populations. The third and fourth axes, explaining less than 0.4% each, separated the Scottish and Turkish / Black Sea populations from the rest of the samples (Supplementary Figure 3B).

We then tried to identify how many distinct genetic groups could describe the current diversity by performing a clustering analysis using NGSadmix[31]. We tested between 1 and 20 groups but did not recover a clear optimal number of clusters (*K*): the Evanno method[32] indicated that two groups explained the genetic diversity best, while the logarithm method[33] indicated 20 groups. At *K*=2, we observed a clear gradient between the western (Spanish) and eastern (Chinese) populations, corresponding to the two most distant ones (Figure 2A). At *K*=3, the Scandinavian populations formed a new cluster (Figure 2B). At *K*=4, a component for western Europe populations appeared. The Russian populations west of the Ural Mountains differentiated from the rest of the Asian populations if we considered five clusters. The Turkish and Black Sea populations split from the western Europe populations at *K*=6 (Figure 2C). One Romanian population (from Piatra Craiului National Park) stood out from nine clusters, and the Scottish populations were differentiated when we considered 10 groups. At *K*=12, the Caucasus populations separate from Asia Minor and no additional pattern emerged at *K* >12 (Supplementary Figure 4).

To estimate the genetic differentiation between populations, we calculated the fixation index (F_{ST}) between location pairs. The global average F_{ST} among the 123 populations was 0.048, ranging from 0.009, between the Alta and Kirkesmoen sites in Norway, the two northernmost locations in our sampling separated by 190km, to 0.127 between a population in the Gunibsky District close to the Black Sea and one in the Sakha region in the most eastern part of Russia. The Gunibsky District stand was also very different from the Glen Orchy location in Scotland (F_{ST} = 0.119). The third highest differentiation was

between the Sakha site and the Spanish Embalse del Ebro location (F_{ST} = 0.117). In general, the genetic differentiation correlated strongly with the geographic distance separating the populations (Figure 3A; Mantel test ρ: 0.586, p-value < 2.2×10^{-16}), showing a pattern of isolation by distance (IBD) [34]. By contrast, the correlation of the genetic differentiation with the environmental distance (Mahalanobis distance calculated within the redundancy analysis (RDA) [35]; see below) was weak (Figure 3B; Mantel test ρ: 0.059, p-value: 4.5×10^{-7}), indicating no clear pattern of isolation by environment (IBE) [5]. The correlation between the geographic and environmental distances was also weak (Figure 3C; Mantel test ρ: 0.104, p-value < 2.2×10^{-16}). This is not surprising considering the wide range of complex climate zones and landscapes covered in the study, which could distort a linear relationship between the two distances as ocean currents and topography can play a significant role in modifying local climates.

Because of the signal of IBD, we further analyzed population clusters using tess3r[36] that explicitly considers the geographic distance between populations. We could not identify a specific number of clusters that best explained the genetic diversity based on the cross-validation analysis (Supplementary Figure 5). However, the clustering results of tess3r (Supplementary Figure 6) were in general very similar to the ones obtained with NGSAdmix.

Since IBD was present but weak, we tried to localize potential barriers to gene flow which could disrupt IBD. We performed an effective migration surface analysis with the python package feems[37] and found several regions with substantially reduced gene flow: the Pyrenees, North Sea, Aegean Sea, Caucasus Mountains and Khingan Mountains (Figure 3D). The Ural Mountains did not seem to impact the gene flow as much as the other landmarks, although their presence was detectable.

Demographic history

To understand how Scots pine colonized its current distribution and how this could have impacted the current genetic diversity and differentiation, we inferred past demographic changes using two complementary approaches. The first, Stairway Plot 2[38], performs multi-epoch coalescent inference of population size changes through time. It is suitable for demographic analyses where no previous knowledge is available but can only analyze populations independently. The second, fastsimcoal2[39,40], implements multi-population demographic models that take divergence and migration among populations into consideration.

We performed a Stairway Plot analysis on each of the three most distant populations in the PCA: the Spanish, Chinese, and Scandinavian (from northern Norway) populations. We also considered the entire species as a single population, given the low global differentiation between locations. Because singletons are prone to sequencing errors that may bias demographic inferences, we performed the analysis with and without them. If we exclude singletons and assume a generation time of 50 years[41], we detected an initial phase of population expansion starting around 6 Mya (million years ago, Figure 4A, Supplementary Figure 7). The Spanish population was the first to expand, followed by the Chinese and Scandinavian populations, until they reached their contemporary sizes around 300 kya (thousands of years ago) for the Spanish population, 80 kya for the Chinese population, and 15 kya for the Scandinavian population. Both the Chinese and Scandinavian population sizes peaked at around 150 kya before reaching their contemporary sizes. The three populations reached a current effective population size (N_e) of about 500,000 individuals. For the entire species, we observed a sharp increase of the population size starting slightly before 1 Mya and then a rapid contraction at ca. 400 kya. The maximal population size was reached around 300 kya, with a decrease around 20 kya, corresponding to the LGM[42], followed by a growth reaching 1.1 million individuals around 3,000 years ago. Using a generation time of 20 years[43], the final N_e were again rather similar between the three populations but larger, around 1.2–1.3 million individuals (Supplementary Figure 7). The N_e of the entire species was estimated to be around 2.7 million individuals (Supplementary Figure 7). If we included singletons, the N_e estimates became larger in all populations, with an estimate of 1.4 or 3.6 million individuals for the whole species using a generation time of 50 or 20 years, respectively. The final size was reached in both cases around 300 kya, with a short period of contraction around 50 – 30 kya.

We then tested 20 demographic models with fastsimcoal2, of which the first two best-fitting models, based on the maximum value of their likelihood and their Akaike's weight of evidence[44], were three-populations models with a pattern of isolation-with-migration (Supplementary Figure 8, Supplementary Table 2). The expected site frequency spectrum (SFS) simulated with these two models reproduced the observed SFS well (Supplementary Figure 9), suggesting that the recovered demographic parameters were good estimates of the past population history. The estimated demographic parameters from the second best-fitting model (Model 3) overlapped better with those estimated from the Stairway Plot analysis than the first best-fitting model, in terms of N_e and timing of population size changes (Supplementary Table 3). This model estimated an ancestral population size N_e of 9.5×10^3

individuals (95% CI: 9.15×10^3–1.48×10^4), which diverged into the Spanish and the ancestral population of the Chinese and Scandinavian groups at 6.20 Mya (95% CI: 5.17–6.20 Mya; Figure 4B, Supplementary Table 3). Both populations expanded after the splitting. The divergence time between the Chinese and Scandinavian populations was estimated at 0.57 Mya (95% CI: 0.49–0.63 Mya; Figure 4B, Supplementary Table 3), followed by expansion. All three populations experienced an instantaneous growth at 60.75 kya (95% CI: 27.95–88.07 kya), where N_e reached 6×10^5, 3.76×10^5, and 9.20×10^5 for the Chinese, Spanish and Scandinavian groups, respectively (Figure 4B, Supplementary Table 3). The historical gene flow among the three groups was very small (with only 0–0.71 migrants per generation), while the current gene flow among them was extremely high (30–211 migrants per generation; Figure 4B, Supplementary Table 3), suggesting that climate oscillations during the Quaternary likely caused range shifts that facilitated contacts between the previously separated populations. This high rate of gene flow among populations is in accordance with the observed low genetic differentiation among the populations.

The best-fitting model (Model 7, Supplementary Table 3) uncovered a similar demographic history with estimates of N_e, divergence time, and gene flow among groups in the same orders of magnitude as those of the first model. However, the order of population split was different, as it inferred that the Chinese group split first from the ancestral population. This scenario is plausible, given that the sister species to *P. sylvestris*, *P. densiflora*, is present in Asia and partly overlaps with the eastern margin of *P. sylvestris*[45-47]. However, the current effective population sizes of the Chinese, Spanish and Scandinavian groups were estimated to be 6.60×10^5, 9.97×10^5, and 2.40×10^5, respectively (Supplementary Table 3), corresponding to a Spanish N_e twice as high as in the model 3, and a Scandinavian N_e divided by three. This would not be in agreement with the current census sizes of the two populations, even though they do not directly reflect the size of the effective population and the slightly higher nucleotide diversity for the Spanish populations could imply a marginally higher N_e. The lower connectivity of the Spanish populations to the rest of the distribution is also visible in both the PCA and feems results.

Genotype and environment association (GEA)

With its large distribution covering wide environmental gradients, we anticipate the presence of genetic adaptation in *P. sylvestris*. To test for associations between genotype and environment, we first applied gradient forest (GF)[48] that correlates allele frequency turnover with environmental gradients and estimates the importance of the variables in explaining the

overall genomic pattern. Based on GF ranking we selected 21 pseudoindependent environmental variables (Supplementary Table 4 & see M&M) to perform a redundancy analysis (RDA)[35] to estimate the relative importance of environmental and geographic factors to the observed population differentiation. Forward selection on the 21 environmental variables retained 10 (see Supplementary Table 4), which were used to produce the environmental distance matrix. We also performed a forward selection on the PCNM-axes (principal coordinates of neighborhood matrix, a set of orthogonal spatial structures computed from a distance matrix), retaining 21 of the 64 axes to generate the geographic distance matrix. The variance partitioning showed that the two matrices together explained 18.9% of the total genetic variance (adjusted R^2), of which 10% was confounded between the two (Figure 5A). Partial RDA on the environmental matrix conditional on the distance explained a total of 1.28% of the total variance, while distance alone explained 7.72%. Only the first RD-axis was identified as significant, and 16 SNPs along this axis appeared as significant outliers (p-value ≤ 0.001; Figure 5B). However, none deviated from the expected distribution of p-values (Figure 5C), meaning that these 16 SNPs are probably false positives. We found only low correlations of RD1 to any of the "environmental" factors: the top two correlations were with bio4 (standard deviation of the monthly mean temperatures) and elevation, with a correlation coefficient of -0.18 and 0.19, respectively.

In addition to RDA, we used a BayPass analysis to detect loci with divergent allele frequencies across populations and strong association with the 21 variables (covariates) from the GF analysis. We identified 103 outlier SNPs that showed strong differentiation (large X^tX-values) among populations (p-value(X^tX) ≤ 0.001, analogous to F_{ST} outliers; Figure 5D) and 41 SNPs associated with four environmental variables (standard deviation of the effect estimate not crossing zero, $|\beta|-\sigma(\beta) > 0$, which corresponds to posterior inclusion probability $\geq$ 0.55): aridity index (ai), proportion of silt particles in the fine soil fraction, volumetric fraction of coarse soil fragments (cfvo), and presence of mixed forest (consensus4; Figure 5E, Supplementary Table 5). Of the 41 SNPs, 11 were also highly differentiated (Supplementary Table 5).

Overall, the associations detected by RDA and BayPass between genetic differentiation and environment were weak across the range-wide collection of Scots pine populations, based on the analysis of 25.8K SNPs. The putative outliers identified by RDA were probably all false positive and showed no overlap with the SNPs identified by BayPass, where we have higher confidence in the correlation with environmental variables. However, the number of significant GEA loci was small.

Discussion

High and uniform genetic diversity across Eurasia

Scots pine is outstanding for its ability to rapidly expand and (re)establish its distribution across Eurasia in the post LGM period. Rapid expansion of a refugial population is expected to leave a spatial pattern of declining genetic diversity due to increasing drift and inbreeding on the leading edge of the colonization front. However, our findings in Scots pine do not align well with these expectations. Firstly, while there are observable patterns in the diversity (both π and H_o) along latitudes and longitudes, the magnitude of these differences is very small. Thus, the genetic diversity is only marginally reduced in northern Europe and in Asia compared to the rest of the distribution, meaning that the founder effect that could result from a recolonization from southern refugia is very faint. Secondly, the inbreeding coefficient (F_{IS}) was close to 0 in most populations, indicating that the expected increase of inbreeding in geographically marginal populations does not hold strong in Scots pine.

Another outstanding result is the low genetic differentiation in Scots pine, with a global F_{ST} of 0.048 across its continental distribution, which means that only 5% of the genetic diversity observed in the species is due to differences among populations. This is in agreement with previous results that estimated a F_{ST} of 0.02 on average, at a smaller geographic scale[26,43,49-51]. Despite this low value, we observed a pattern of isolation by distance, with distant populations being more differentiated than close ones[34], suggesting that spatial distance is a stronger barrier to gene flow than any landforms. Nonetheless, the feems analysis detected barriers to gene flow precisely where we can expect a decrease in pollen exchange: over large bodies of water, over mountain chains, or over wide spaces of steppe or broad-leaved woods. The Pyrenees, the Black and Caspian Seas and the Caucasus and Khingan Mountains are for example detectable barriers to gene flow, together with, to a lesser extent, the Eastern European plain. The relative isolation of the Spanish and the Caucasian populations identified by feems confirms previous results obtained using other genetic markers[18,52-55]. Those barriers, even if detectable, are overall faint, and the species can be considered as close to panmictic as reasonably expected. We were for example able to detect the existence of gene flow between Spain and the rest of western Europe thanks to a more even sampling of western Europe, including previously missing regions in France in particular, as shown by the NGSadmix analysis. The French samples are intermediate between the Spanish and central European populations, showing that the Pyrenees are a porous barrier and that the Spanish refugia contributed to the genetic diversity further north.

We confirm the genetic proximity of the Scottish populations to the mainland ones[49,56], and the presence of genetically distinct populations in Turkey and the Carpathian mountains detected at higher *K* values[18,50,57]. The uniqueness of these populations, and in general the differentiation between European populations, are more visible at the maternally inherited mitochondrial DNA markers due to the limited dispersal ability of seeds[16]. However, gene flow mediated by pollen is sufficiently strong to overcome seed dispersal limitations and to blur the ancestral information in the nuclear genomes of the extant populations.

Overall, Scots pine is exceptional with its large distribution but very low genetic differentiation among populations. This is in stark contrast to Norway spruce that has a distribution largely overlapping the western part of the Scots pine and a similar recolonization history after LGM but shows a strong population structure[58,59]. Both Norway spruce and Scots pine are wind pollinated with wind dispersed seeds and a marked tendency towards outcrossing[60-62]. One reason for their different genetic patterns could be the heavier spruce pollen compared to pine[63], which reduces the gene flow distance and enhances population genetic differentiation. Another reason could be their different demographic histories during and post LGM.

Resilience over the LGM

One unexpected result from our demographic analyses was the relatively stable effective population size since 60 kya and little fluctuation during the LGM. In the two demographic analyses performed here, we obtained similar results regarding the time of the major events and the final population sizes. The impact of the LGM on the species demography seems overall to be limited and therefore not detected by the two methods at the population scale. Only the Stairway Plot analysis at the entire species level showed a decrease in population size around 20–40 kya. This can be due to either a glacial period that is too short a time compared to the evolutionary timescale of the species to have any strong impact, or to the existence of multiple glacial refugia which could have conserved the entire species genetic variation. The demography of Scots pine recovered in this study supports strong changes in population size that trace back to much older history in Pleistocene or Pliocene[64,65]. This pattern is also observed in eight other wind pollinated tree species including three pine species in Eurasia[51,66].

The resilience of Scots pine over the last ice age is not entirely surprising for a cold-resistant species[13,14,67]. It faces strong competition from deciduous species when the climate becomes milder[68,69], making it maybe more sensitive to interglacial than to glacial periods. It

is thus possible that patches of trees persisted and were able to grow on surface debris accumulating on ice[70,71], and that the presence of trees was therefore not limited to ice-free areas. This would explain the macrofossil evidence of Scots pines at high latitudes very rapidly after the LGM[20,71,72], the absence of a strong founder effect, the mild impact of the LGM on the species demography, and the very low differentiation between populations in general. The Quaternary could therefore have been much more forested than initially thought.

Signals of genetic-environmental association (GEA)

Despite strong phenotypic and physiological divergence among Scots pine populations[9,11,14], we found only weak signals of genetic adaptation to environmental gradients. The overall pattern of isolation by environment was weak (ρ=0.059, p-value= $4.5e^{-7}$), RDA identified only 1% of the genetic differentiation among populations that could be explained by environment alone, and only 41 out of 25.8K SNPs were strongly associated with environmental variables, of which 11 SNPs also appeared as F_{ST} outliers. This may seem rather unexpected at first glance given the wide ecological gradients covered by the populations but could result from several factors. Firstly, adaptation to local environment, which is moderately to highly heritable in Scots pine[73], is not a simple trait but a combination of complex traits that together provide a fitness advantage for populations in their respective environments. Identifying the genetic basis of local adaptation in wild populations is challenging and requires dense sampling of the genome and unrealistic sample sizes[74]. In this study, we surveyed only around 0.01% of the genome, thus the possibility of detecting outliers of significant effects is slim, but very few outlier loci have been detected in a similar study despite the sequencing of 60,000 open reading frames by exon capture[26]. Secondly, the traits under selection are likely driven by very small changes in allele frequencies distributed across many loci, which makes their detection difficult with the current methods[14,75,76], or by more complex processes undetectable with SNP analyses, such as methylation, differences in gene expression or retrotransposons activation[77,78]. Thirdly, because of the recent and rapid recolonization of Scots pine from the South to the North or East to West, the impact of neutral demographic events could overshadow the signal of selection, thus blur the signals of GEA. This is particularly difficult to disentangle when colonization routes covary with climate gradients along latitudes or longitudes[6,79]. Finally, considering the short recolonization history after the LGM, it is unlikely that a few hundred generations of nucleotide substitutions under selection would produce a convincing signal in a species that exhibits phenotypic plasticity.

Nonetheless, among the four significant environmental variables that drove the few allele frequency changes in this study, the most important one was the aridity index. This is consistent with the ecology of Scots pine which is known to suffer from drought, resulting in a reduction in tree height in the southern part of the distribution[80], shorter and less numerous shoots[81,82] and an increase of mortality after severe drought events[83].

To conclude, our range-wide study on Scots pine revealed a few seemly paradoxical results. Scots pine is an outstanding species characterized by a high genetic diversity but low differentiation at neutral markers between populations across its Eurasian distribution. Its population size remained relatively stable since 60 kya, and the last glacial period did not appear to have had a noticeable impact on the species demography. These genetic patterns are to a large extent influenced by extensive gene flow, reducing the impact of drift and population differentiation, and maintaining a large effective population size. Thus, the success of Scots pine in colonizing new areas after the LGM was likely due to a combination of factors, including its ability to disperse over long distances, its resilience and adaptability. However, we detect very few loci that can be associated with environmental gradients despite previous evidence of local adaptation. Improving the resolution of GEA and our understanding of how natural selection shapes the genetic variation that underlies adaptive traits will require a more refined systems biology approach that considers the specific genetic, environmental, and phenotypic contexts of adaptation.

Methods

Sampling and GBS library preparation

We collected 2895 samples from 211 natural forests covering the entire geographic distribution of *P. sylvestris*. The samples were either needles, buds or bulk seeds of local stands. We selected native forests to the best of our knowledge, and the sampled trees in each stand were at least 50 meters apart. We extracted DNA from each individual using an EZNA SP Plant DNA Kit (Omega Biotek). GBS library preparation followed the protocol of Pan, et al.[84]. Briefly, 200 ng of DNA of each individual was digested with *Pst*I-HF (New England Biolabs) and ligated to sequencing adapters with individual barcodes. Digestion and ligation products from 300 samples (including intra-library and between-library replicates) were pooled into each library, and then purified and PCR amplified. The purified PCR products were separated on a E-gel EX 2% agarose gel (Thermo Fisher Scientific) to recover fragments in the size range of 350 – 450 bp. Paired-end sequencing (2×150 bp) was performed on Illumina HiSeq X Ten (Novogene, UK).

GBS data filtering

The sequence quality of each library was assessed with FastQC v. 0.11.9[85]. Reads with > 40% bases having a Phred-scaled quality <20 were discarded by fastp v. 0.23.1[86]. Adapter sequences and low-quality bases from the tail of each read were trimmed. Samples in each library were demultiplexed with the process_radtags module of Stacks v.2.0[87] allowing only one mismatch with the barcode. The first 6 bases of the reads, being the enzyme cutting site, were removed with the fastx_trimmer module of FASTX Toolkit v. 0.0.13[88]. Reads shorter than 36 bp were discarded. Clean reads were mapped on the *P. taeda* genome v.1.01[29,30] using the Burrows-Wheeler Aligner MEM algorithm v. 0.7.17 with default parameters[89]. Variants were called using SAMtools and BCFtools with default parameters[90,91]. Several filtering steps were performed to minimize genotyping errors: SNPs located in repetitive regions (reference to *P. taeda* genome v1.01) and with a mapping quality (MQ) <40 were removed; genotypes with a Phred-scaled genotype quality (GQ) <20 or read depth (DP) <5 were masked as missing; loci with a missing rate of >40%, minor allele frequency (MAF) <5%, heterozygosity >50% or not biallelic were also removed. Based on this first SNP dataset, we removed individuals with >60% missing data and highly related individuals in each population following the procedure of Hall, et al.[92]. After these different quality controls, we kept 2,321 individuals from 202 populations (Figure 1, Supplementary Table 1).

Genotype likelihoods

We estimated the genotype likelihoods for the 2,321 individuals with ANGSD v. 0.935[28]. Because the *P. taeda* genome v.1.01 is highly fragmented, we first concatenated all the regions with at least one mapped read into 20 pseudo scaffolds, separating each region by 300 N. We remapped the reads of each sample against this new reference. We set a maximal depth of 99.5% of the highest depth measured (5022.8x), a minimal depth of 5, a Phred-scaled base quality of 20 and a mapping quality of 40 for an individual site to be considered. We removed the sites with >40% missing data and more than two alleles. We estimated the genotype likelihoods with the GATK method, inferred the major and minor alleles and estimated the allele frequencies. Sites with an observed heterozygosity higher than 50% were discarded.

To remove the batch or library effect, we examined PCA grouping patterns for subsets of samples by country or region using PCAngsd v. 1.01[93] to identify the SNPs driving the library effect and removed them. Finally, we re-ran ANGSD with the sites kept, and produced one dataset that kept both the selected SNPs and all the invariant sites for nucleotide diversity and demographic analyses, and one SNP dataset with a minimal MAF of 0.05 which was used for all the other analyses. We transformed the genotype likelihoods to a vcf file for subsequent analyses where genotype likelihoods could not be applied.

Genetic diversity and population structure

To describe the genetic diversity within and among populations, we first combined the very small populations with their closest neighbors to have at least six individuals per population (Supplementary Table 1). This formed 123 large populations. Three small distant populations in central Russia were excluded. We estimated for each population the per-site observed (H_o) and expected heterozygosity (H_e) and inbreeding coefficient (F_{IS}) with ANGSD on the MAF 0.05 dataset. To estimate the pairwise nucleotide diversity (π)[94], we generated the folded site frequency spectrum (SFS) from the full dataset (including invariant sites) with ANGSD and realSFS, and estimated the nucleotide diversity with the realSFS saf2theta command and the thetaStat program in ANGSD. Nucleotide diversity at 0-fold (π_0) and 4-fold (π_4) sites (genes nucleotides whose mutations are never or always synonymous, respectively) were defined by annotation of the coordinates of those sites on the original reference.

To characterize the spatial organization of genetic diversity, we first performed a PCA on the MAF filtered dataset with PCAngsd. We then examined the population structure using NGSadmix v. 32[31] with a potential number of clusters *K* ranging from 1 to 20, 50 replicates

for each analysis, and a maximum of 200,000 iterations. The likelihood for each *K* was obtained with the R package pophelper v. 2.3.0, and the results were combined by CLUMPAK[95].

To understand the role of geography on genetic differentiation, we performed a Mantel test on genetic distance between populations (as normalized population differentiation F_{ST}) *vs*. geodistances using the R package vegan v. 2.5-6[96] in 9999 permutations. F_{ST} values between populations were obtained from ANGSD and the geodesic values were from the geosphere package v. 1.5-10 in R[97]. Similarly, we performed Mantel tests on genetic *vs* environmental distance, and geodistance *vs* environmental distance. Environmental distances between populations were obtained from RDA, see next section.

To account for the impact of isolation by distance (IBD) on population differentiation, we examined population structure using tess3r which considers IBD in defining genetic clusters[98]. We ran the tess3 analysis for *K* ranging from 1 to 20, with 20 replicates for each value of *K* and a maximum of 200 iterations. We estimated the root mean-squared errors for cross-validation.

Finally, to identify spatial gene flow barriers across the distribution range, we estimated the effective migration surfaces with the python package feems[37]. We transformed the vcf into a bed file with PLINK v. 1.9b5[99,100] and used a triangular grid covering the species distribution at a resolution 7 created with the package dggridR v. 3.0.0 in R[101], corresponding to triangles of 1,557 km^2. We performed a leave-one-out cross-validation for lambda values (tuning parameter) varying from 10^{-6} to 10^{2}, with 20 values tested.

Demographic history

Following the PCA and admixture results, we estimated the demography of the three most divergent populations in Scots pine (i.e. China, 73 individuals; Spain, 70 individuals; and northern Norway, 168 individuals coming from the Beiarn, Hemne and Molde populations) using two complementary approaches, Stairway Plot v. 2[38,102] and fastsimcoal2 v. 2.6[39,40]. We produced folded SFS by ANGSD on the full dataset (including invariant sites) for these three groups, which included 1,431,145; 1,431,110 and 1,429,721 sites for China, Spain and Norway, respectively. We also produced 2D-SFS which included 1,430,535 sites for the China-Spain pair, 1,429,276 sites for the Norway-Spain comparison, and 1,429,401 sites for the Norway-China pair.

Stairway Plot v2 is a flexible multi-epoch coalescent approach that infers population size changes through time without assuming any a priori demographic model. We ran

Stairway Plot for each of the three populations separately and for the entire species (2,321 individuals) under the default parameters and tested four different numbers of random breakpoints using 200 replicates. We used an average generation time of 20[43] and 50 years[41,103] and a mutation rate of 7×10^{-10} mutations per site and year[103]. Because Stairway Plot analyses can be sensitive to rare variants which could potentially be due to sequencing/genotyping errors[38], we considered the entire allele frequency spectrum with or without singletons.

Fastsimcoal2 considers population divergence with migration. We tested 20 scenarios, all of which began with the ancestral population splitting into three different populations but differed in terms of: (i) whether the split events occurred simultaneously or not; (ii) whether divergence occurred with migration or not; (iii) levels of gene flow between two populations, (iv) the mode of population size change after the split, and (v) whether admixture was involved (Supplementary Figure 7). Considering the low confidence of rare alleles, we ran fastsimcoal2 based on SFS with "--nosingleton". Each model was run 50 independent times, with 100,000 coalescent simulations as well as 40 expectation-conditional maximization cycles for the calculation of the global maximum-likelihood. The best-fitting model was selected based on the maximum value of likelihood over the 50 independent runs of each model and their Akaike's weight of evidence[44]. The goodness of fit of the best model was tested by comparing the observed SFS with the expected SFS, which was obtained by 100,000 coalescent simulations under the maximum-likelihood estimates of population parameters. Parameter confidence intervals (95% CI) of the best model were obtained by running 100 parametric bootstraps, with 50 independent runs in each bootstrap. We used a mutation rate of 7×10^{-10} mutations per site and year and a generation time of 50 years to convert model parameters to absolute values.

Genotype and environment association

To identify important environmental variables responsible for the overall genomic pattern, we first performed a gradient forest analysis (GF)[48]. We then used two different methods, RDA[35] and BayPass[104], to test whether the sampled genetic variation correlated with biotic and abiotic variables. A GF analysis develops non-parametric functions that describe allele frequency turnovers along environmental gradients. We extracted 74 variables (Supplementary Table 6) describing the environment of each of the large populations (123) and used a GF analysis to rank them by importance. The variables with a correlation coefficient $\geq |0.75|$ (Spearman's R) to the top-ranked variable were removed. The remaining

21 uncorrelated variables were then used as input for the RDA and BayPass analyses. The GF models were validated by 10 different runs using shuffled environmental data.

RDA is a method that tries to find the best combination of environmental variables (among a constrained list) that can explain population allele frequencies, assuming that the environmental variables are independent, and that the allelic distribution is dependent on those variables. We built two explanatory matrices to test their influence on allele frequencies across populations. The first matrix was based on the distance between populations through principal coordinates of neighbourhood matrix (PCNM) that converts distances to a rectangular representation. The second matrix was based on the environmental variables extracted from the GF analysis. Each of these matrices was first analysed separately through a forward selection step where the allele frequencies were set as response variables with the 'ordistep' function in the vegan-package. In these steps we want to maximize the variance explained by a set of predictors. For each step, the predictor that has the greatest explanatory power of the observed variance was added and the model was evaluated. We ran RDA on the distance and environment models and a partial RDA-model where the environmental factors are conditional on the geodistance between populations. This allowed us to estimate the environmental factors that drive allele frequency change while accounting for the spatial relationship.

BayPass builds on previous methods established in BayEnv2[105] and provides more robust estimates of population covariance matrixes. It includes a calibration method for the outlier (X^tX) statistic and uses the linkage disequilibrium information to better pinpoint loci in significant associations with environmental co-variables. We first used the core model to establish the covariance of allele frequencies among populations that is a result of their shared history. We ran the model with 200,000 iterations and a thinning rate of 20 (-nval 10,000 and -thin 20) with the remaining parameters set to default values (burnin 5000, max number of pilot runs 20 and pilot run length 500). The core model was run three times to ensure stable results. We then created a pseudo-observed dataset (POD) based on the posterior means of the a_π and b_π parameters of the beta prior distribution assumed for SNP reference allele frequency across populations, π_i, produced by the core model.

$$\pi_i | a_\pi, b_\pi \sim \beta(a_\pi; b_\pi)$$

This is then used to simulate a reference set of X^tX differentiation values[105] that can be used to calibrate observed X^tX values. We compared the observed and POD distributions of X^tX-values but due to the very different types of distributions we relied on the observed log-normal distribution of X^tX-values to establish significance threshold of outliers (*p*-value

<0.001, see Figure 5D). We then ran the auxiliary covariate model with spatial dependency on the markers using an Ising prior, $b_{is} = 1$, with the same MCMC sampling parameters as in the core model above, for each of the GF selected environmental variables independently, considering both outliers detected by the X^tX statistics and other SNPs not differentiated between populations. Finally, we identified the environmental variables presenting an association with at least one SNP and ran the previous model with all these variables simultaneously for the final results.

References

1 Ellegren, H. & Galtier, N. Determinants of genetic diversity. *Nature Reviews Genetics* **17**, 422-433 (2016).

2 Savolainen, O., Lascoux, M. & Merila, J. Ecological genomics of local adaptation. *Nature Reviews Genetics* **14**, 807-820 (2013).

3 Hamrick, J. L. & Godt, M. J. W. Effects of life history traits on genetic diversity in plant species. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **351**, 1291-1298 (1996).

4 Orsini, L., Vanoverbeke, J., Swillen, I., Mergeay, J. & De Meester, L. Drivers of population genetic differentiation in the wild: isolation by dispersal limitation, isolation by adaptation and isolation by colonization. *Molecular Ecology* **22**, 5983-5999 (2013).

5 Wang, I. J. & Bradburd, G. S. Isolation by environment. *Molecular Ecology* **23**, 5649-5662 (2014).

6 Zhao, W. *et al.* Effects of landscapes and range expansion on population structure and local adaptation. *New Phytologist* **228**, 330-343 (2020).

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

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

9 Carlisle, A. & Brown, A. *Pinus sylvestris* L. *Journal of Ecology* **56**, 269-307 (1968).

10 Houston Durrant, T., De Rigo, D. & Caudullo, G. *Pinus sylvestris* in Europe: distribution, habitat, usage and threats. in *European Atlas of Forest Tree Species* (eds Jesús San-Miguel-Ayanz *et al.*) 132-133 (Publications Office of the European Union, 2016).

11 Rehfeldt, G. E. *et al.* Intraspecific responses to climate in *Pinus sylvestris*. *Global Change Biology* **8**, 912-929 (2002).

12 Andersson, B. & Fedorkov, A. Longitudinal differences in Scots pine frost hardiness. *Silvae Genetica* **53**, 76-80 (2004).

13 Hurme, P., Repo, T., Savolainen, O. & Pääkkönen, T. Climatic adaptation of bud set and frost hardiness in Scots pine (*Pinus sylvestris*). *Canadian Journal of Forest Research* **27**, 716-723 (1997).

14 Hall, D. *et al*. Divergent patterns between phenotypic and genetic variation in Scots pine. *Plant Communications* **2**, 100139 (2021).

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

16 Tóth, E. G., Köbölkuti, Z. A., Pedryc, A. & Höhn, M. Evolutionary history and phylogeography of Scots pine (*Pinus sylvestris* L.) in Europe based on molecular markers. *Journal of Forestry Research* **28**, 637-651 (2017).

17 Willis, K. J., Rudner, E. & Sümegi, P. The full-glacial forests of central and southeastern Europe. *Quaternary Research* **53**, 203-213 (2000).

18 Naydenov, K., Senneville, S., Beaulieu, J., Tremblay, F. & Bousquet, J. Glacial vicariance in Eurasia: mitochondrial DNA evidence from Scots pine for a complex heritage involving genetically distinct refugia at mid-northern latitudes and in Asia Minor. *BMC Evolutionary Biology* **7**, 233 (2007).

19 Pyhäjärvi, T., Salmela, M. J. & Savolainen, O. Colonization routes of *Pinus sylvestris* inferred from distribution of mitochondrial DNA variation. *Tree Genetics & Genomes* **4**, 247-254 (2008).

20 Kullman, L. Boreal tree taxa in the central Scandes during the Late-Glacial: implications for Late-Quaternary forest history. *Journal of Biogeography* **29**, 1117-1124 (2002).

21 Willis, K. J. & van Andel, T. H. Trees or no trees? The environments of central and eastern Europe during the Last Glaciation. *Quaternary Science Reviews* **23**, 2369-2387 (2004).

22 Petit, R. J. *et al.* Glacial refugia: hotspots but not melting pots of genetic diversity. *Science* **300**, 1563-1565 (2003).

23 Cheddadi, R. *et al.* Imprints of glacial refugia in the modern genetic diversity of *Pinus sylvestris*. *Global Ecology and Biogeography* **15**, 271-282 (2006).

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

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

26 Tyrmi, J. S. *et al.* Genomics of clinal local adaptation in *Pinus sylvestris* under continuous environmental and spatial genetic setting. *G3: Genes, Genomes, Genetics* **10**, 2683-2696 (2020).

27 Sannikov, S. N. & Petrova, I. V. Phylogenogeography and genotaxonomy of *Pinus sylvestris* L. populations. *Russian Journal of Ecology* **43**, 273-280 (2012).

28 Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* **15**, 356 (2014).

29 Neale, D. B. *et al*. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biology* **15**, R59 (2014).

30 Zimin, A. *et al*. Sequencing and assembly of the 22-Gb Loblolly pine genome. *Genetics* **196**, 875-890 (2014).

31 Skotte, L., Korneliussen, T. S. & Albrechtsen, A. Estimating individual admixture proportions from next generation sequencing data. *Genetics* **195**, 693-702 (2013).

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

33 Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959 (2000).

34 Wright, S. Isolation by distance. *Genetics* **28**, 114 (1943).

35 Forester, B. R., Lasky, J. R., Wagner, H. H. & Urban, D. L. Comparing methods for detecting multilocus adaptation with multivariate genotype–environment associations. *Molecular Ecology* **27**, 2215-2233 (2018).

36 Caye, K., Jay, F., Michel, O. & François, O. Fast inference of individual admixture coefficients using geographic data. *The Annals of Applied Statistics* **12**, 586-608 (2018).

37 Marcus, J., Ha, W., Barber, R. F. & Novembre, J. Fast and flexible estimation of effective migration surfaces. *eLife* **10**, e61927 (2021).

38 Liu, X. & Fu, Y.-X. Stairway Plot 2: demographic history inference with folded SNP frequency spectra. *Genome Biology* **21**, 280 (2020).

39 Excoffier, L. *et al.* fastsimcoal2: demographic inference under complex evolutionary scenarios. *Bioinformatics* **37**, 4882–4885 (2021).

40 Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C. & Foll, M. Robust demographic inference from genomic and SNP data. *PLoS Genetics* **9**, e1003905 (2013).

41 Tóth, E. G., Bede-Fazekas, Á., Vendramin, G. G., Bagnoli, F. & Höhn, M. Mid-Pleistocene and Holocene demographic fluctuation of Scots pine (*Pinus sylvestris* L.) in the Carpathian Mountains and the Pannonian Basin: Signs of historical expansions and contractions. *Quaternary International* **504**, 202-213 (2019).

42 Patton, H., Hubbard, A., Andreassen, K., Winsborrow, M. & Stroeven, A. P. The build-up, configuration, and dynamical sensitivity of the Eurasian ice-sheet complex to Late Weichselian climatic and oceanic forcing. *Quaternary Science Reviews* **153**, 97-121 (2016).

43 Pyhäjärvi, T., Kujala, S. T. & Savolainen, O. 275 years of forestry meets genomics in *Pinus sylvestris*. *Evolutionary Applications* **13**, 11-30 (2020).

44 Akaike, H. Factor analysis and AIC. *Psychometrika* **52**, 317–332 (1987).

45 Jin, W. T. *et al.* Phylogenomic and ecological analyses reveal the spatiotemporal evolution of global pines. *Proceedings of the National Academy of Sciences* **118** (2021).

46 Saladin, B. *et al.* Fossils matter: improved estimates of divergence times in *Pinus* reveal older diversification. *BMC Evolutionary Biology* **17**, 95 (2017).

47 Zeb, U. *et al.* Comparative plastid genomics of *Pinus* species: Insights into sequence variations and phylogenetic relationships. *Journal of Systematics and Evolution* **58**, 118-132 (2020).

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

49 Wachowiak, W., Perry, A., Zaborowska, J., Gonzalez-Martinez, S. C. & Cavers, S. Admixture and selection patterns across the European distribution of Scots pine, *Pinus sylvestris* (Pinaceae). *Botanical Journal of the Linnean Society* **200**, 416–432 (2022).

50 Tóth, E. G., Vendramin, G. G., Bagnoli, F., Cseke, K. & Höhn, M. High genetic diversity and distinct origin of recently fragmented Scots pine (*Pinus sylvestris* L.) populations along the Carpathians and the Pannonian Basin. *Tree Genetics & Genomes* **13**, 47 (2017).

51 Milesi, P. *et al.* Synchronous effective population size changes and genetic stability of forest trees through glacial cycles. Preprint at https://www.biorxiv.org/content/biorxiv/early/2023/01/06/2023.01.05.522822.full.pdf (2023).

52 Soranzo, N., Alia, R., Provan, J. & Powell, W. Patterns of variation at a mitochondrial sequence-tagged-site locus provides new insights into the postglacial history of European *Pinus sylvestris* populations. *Molecular Ecology* **9**, 1205-1211 (2000).

53 Wachowiak, W. *et al.* Phylogeography of Scots pine in Europe and Asia based on mtDNA polymorphisms. *Journal of Systematics and Evolution* **61**, 315-327 (2022).

54 Semerikov, N. V. *et al.* Cytoplasmic DNA variation does not support a recent contribution of *Pinus sylvestris* L. from the Caucasus to the main range. *Tree Genetics & Genomes* **16**, 59 (2020).

55 Dering, M. *et al.* The evolutionary heritage and ecological uniqueness of Scots pine in the Caucasus ecoregion is at risk of climate changes. *Scientific Reports* **11**, 22845 (2021).

56 Wachowiak, W., Salmela, M. J., Ennos, R. A., Iason, G. & Cavers, S. High genetic diversity at the extreme range edge: nucleotide variation at nuclear loci in Scots pine (*Pinus sylvestris* L.) in Scotland. *Heredity* **106**, 775-787 (2011).

57 Dering, M. *et al.* Tertiary remnants and Holocene colonizers: Genetic structure and phylogeography of Scots pine reveal higher genetic diversity in young boreal than in relict Mediterranean populations and a dual colonization of Fennoscandia. *Diversity and Distributions* **23**, 540-555 (2017).

58 Milesi, P. *et al.* Assessing the potential for assisted gene flow using past introduction of Norway spruce in southern Sweden: Local adaptation and genetic basis of quantitative traits in trees. *Evolutionary Applications* **12**, 1946-1959 (2019).

59 Chen, J. *et al.* Genomic data provide new insights on the demographic history and the extent of recent material transfers in Norway spruce. *Evolutionary Applications* **12**, 1539-1551 (2019).

60 Burczyk, J., DiFazio, S. P. & Adams, W. T. Gene flow in forest trees: how far do genes really travel? *Forest Genetics* **11**, 179 (2004).

61 Robledo-Arnuncio, J. J. & Gil, L. Patterns of pollen dispersal in a small population of *Pinus sylvestris* L. revealed by total-exclusion paternity analysis. *Heredity* **94**, 13-22 (2005).

62 Piotti, A., Leonardi, S., Piovani, P., Scalfi, M. & Menozzi, P. Spruce colonization at treeline: where do those seeds come from? *Heredity* **103**, 136-145 (2009).

63 Di-Giovanni, F. & Kevan, P. G. Factors affecting pollen dynamics and its importance to pollen contamination: a review. *Canadian Journal of Forest Research* **21**, 1155-1170 (1991).

64 Pyhäjärvi, T. *et al.* Demographic history has influenced nucleotide diversity in European *Pinus sylvestris* populations. *Genetics* **177**, 1713-1724 (2007).

65 Kujala, S. T. & Savolainen, O. Sequence variation patterns along a latitudinal cline in Scots pine (*Pinus sylvestris*): signs of clinal adaptation? *Tree Genetics & Genomes* **8**, 1451-1467 (2012).

66 Xia, H. *et al.* Combining mitochondrial and nuclear genome analyses to dissect the effects of colonization, environment, and geography on population structure in *Pinus tabuliformis*. *Evolutionary Applications* **11**, 1931-1945 (2018).

67 Savolainen, O., Bokma, F., García-Gil, R., Komulainen, P. & Repo, T. Genetic variation in cessation of growth and frost hardiness and consequences for adaptation of *Pinus sylvestris* to climatic changes. *Forest Ecology and Management* **197**, 79-89 (2004).

68 Boisvert-Marsh, L. & Blois, S. d. Unravelling potential northward migration pathways for tree species under climate change. *Journal of Biogeography* **48**, 1088-1100 (2021).

69 Galiano, L., Martínez-Vilalta, J. & Lloret, F. Drought-induced multifactor decline of Scots pine in the Pyrenees and potential vegetation change by the expansion of co-occurring oak species. *Ecosystems* **13**, 978-991 (2010).

70 Fickert, T., Friend, D., Grüninger, F., Molnia, B. & Richter, M. Did debris-covered glaciers serve as Pleistocene refugia for plants? A new hypothesis derived from observations of recent plant growth on glacier surfaces. *Arctic, Antarctic, and Alpine Research* **39**, 245-257 (2007).

71 Zale, R. *et al.* Growth of plants on the Late Weichselian ice-sheet during Greenland interstadial-1? *Quaternary Science Reviews* **185**, 222-229 (2018).

72 Parducci, L. *et al.* Glacial survival of boreal trees in northern Scandinavia. *Science* **335**, 1083-1086 (2012).

73 Kujala, S. T. *et al.* Genetic heterogeneity underlying variation in a locally adaptive clinal trait in *Pinus sylvestris* revealed by a Bayesian multipopulation analysis. *Heredity* **118**, 413-423 (2017).

74 Gienapp, P. Opinion: Is gene mapping in wild populations useful for understanding and predicting adaptation to global change? *Global Change Biology* **26**, 2737-2749 (2020).

75 Nadeau, S., Meirmans, P. G., Aitken, S. N., Ritland, K. & Isabel, N. The challenge of separating signatures of local adaptation from those of isolation by distance and colonization history: The case of two white pines. *Ecology and Evolution* **6**, 8649-8664 (2016).

76 Barton, N., Hermisson, J. & Nordborg, M. Population genetics: Why structure matters. *Elife* **8**, e45380 (2019).

77 Alakärppä, E. *et al.* Natural variation of DNA methylation and gene expression may determine local adaptations of Scots pine populations. *Journal of experimental botany* **69**, 5293-5305 (2018).

78 Voronova, A., Belevich, V., Jansons, A. & Rungis, D. Stress-induced transcriptional activation of retrotransposon-like sequences in the Scots pine (*Pinus sylvestris* L.) genome. *Tree Genetics & Genomes* **10**, 937-951 (2014).

79 Capblancq, T., Lachmuth, S., Fitzpatrick, M. C. & Keller, S. R. From common gardens to candidate genes: exploring local adaptation to climate in red spruce. *New Phytologist* **237**, 1590-1605 (2023).

80 Hallingbäck, H. R. *et al*. Managing uncertainty in Scots pine range-wide adaptation under climate change. *Frontiers in Ecology and Evolution* **9**, 807 (2021).

81 Bachofen, C., Perret-Gentil, A., Wohlgemuth, T., Vollenweider, P. & Moser, B. Phenotypic plasticity versus ecotypic differentiation under recurrent summer drought in two drought-tolerant pine species. *Journal of Ecology* **109**, 3861-3876 (2021).

82 Taeger, S., Fussi, B., Konnert, M. & Menzel, A. Large-scale genetic structure and drought-induced effects on European Scots pine (*Pinus sylvestris* L.) seedlings. *European Journal of Forest Research* **132**, 481-496 (2013).

83 Haberstroh, S. *et al.* Central European 2018 hot drought shifts scots pine forest to its tipping point. *Plant Biology* **24**, 1186-1197 (2022).

84 Pan, J. *et al*. Optimization of the genotyping-by-sequencing strategy for population genomic analysis in conifers. *Molecular Ecology Resources* **15**, 711-722 (2015).

85 Andrews, S. FastQC: A quality control tool for high throughput sequence data v. 0.11.9, https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).

86 Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884-i890 (2018).

87 Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A. & Cresko, W. A. Stacks: an analysis tool set for population genomics. *Molecular Ecology* **22**, 3124-3140 (2013).

88 Gordon, A. & Hannon, G. J. Fastx-toolkit: FASTQ/A short-reads preprocessing tools v. 0.0.13, http://hannonlab.cshl.edu/fastx_toolkit (2010).

89 Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).

90 Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).

91 Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

92 Hall, D., Zhao, W., Wennström, U., Gull, B. A. & Wang, X.-R. Parentage and relatedness reconstruction in *Pinus sylvestris* using genotyping-by-sequencing. *Heredity* **124**, 633–646 (2020).

93 Meisner, J. & Albrechtsen, A. Inferring population structure and admixture proportions in low-depth NGS data. *Genetics* **210**, 719-731 (2018).

94 Nei, M. & Li, W.-H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences* **76**, 5269-5273 (1979).

95 Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A. & Mayrose, I. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Molecular Ecology Resources* **15**, 1179-1191 (2015).

96 Oksanen, J. *et al.* vegan: community ecology package v. 2.5-6, https://CRAN.R-project.org/package=vegan (2019).

97 Hijmans, R. J. geosphere: spherical trigonometry v. 1.5-10, https://CRAN.R-project.org/package=geosphere (2019).

98 Caye, K., Deist, T. M., Martins, H., Michel, O. & François, O. TESS3: fast inference of spatial population structure and genome scans for selection. *Molecular Ecology Resources* **16**, 540-548 (2016).

99 Chang, C. C. *et al*. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, s13742-13015-10047-13748 (2015).

100 Purcell, S. & Chang, C. PLINK v. 1.9b5, www.cog-genomics.org/plink/1.9/ (2017).

101 Barnes, R. & Sahr, K. dggridR: discrete global grids for R v. 3.0.0, https://github.com/r-barnes/dggridR/ (2021).

102 Liu, X. & Fu, Y.-X. Exploring population size changes using SNP frequency spectra. *Nature Genetics* **47**, 555-559 (2015).

103 Willyard, A., Syring, J., Gernandt, D. S., Liston, A. & Cronn, R. Fossil calibration of molecular divergence infers a moderate mutation rate and recent radiations for *Pinus*. *Molecular Biology and Evolution* **24**, 90-101 (2007).

104 Gautier, M. Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics* **201**, 1555-1579 (2015).

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

106 Caudullo, G., Welk, E. & San-Miguel-Ayanz, J. Chorological maps for the main European woody species. *Data in Brief* **12**, 662-666 (2017).

Acknowledgments

We are grateful to the collectors who sent us samples from the field: Philippe Castellar, Sébastien Cecchini (Office National des Forêts), Jean-Paul Collin, Véronique Guérin-Faublée, Caroline Guignier (Monteco), Cédric Hache (Lycée agricole du Haut-Languedoc), Jean-Philippe Hell (Office National des Forêts), Thierry Helminger (Musée national d'histoire naturelle de Luxembourg), Bertrand Jarri (Mayenne Nature Environnement), Isabelle Kerzaon (Université Claude Bernard Lyon 1), Laurent Lathuillière (Office National des Forêts), Marc Oesterle, Benoît Renaux (Conservatoire botanique national du Massif Central), Claude Roulet, Florent Sabatier, Hugues Savay-Guerraz, Hubert Schmuck (Office National des Forêts), Gilles Thébaud (Université Clermont-Auvergne), Hervé Tournier (Parc Naturel Régional du Vercors), Mircea Verghelet and Ionut Bordea (Piatra Craiului National Park), Natalia Demidova (NRIF), Seppo Ruotsalainen (LUKE), Arne Steffenrem and Øyvind Meland Edvardsen (NIBIO). The SAGE data were used with permission of The Center for Sustainability and the Global Environment, Nelson Institute for Environmental Studies, University of Wisconsin-Madison. The computation was enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) through High-Performance Computing Center North (HPC2N), partially funded by the Swedish Research Council through grant agreement no. 2018-05973. This study was supported by Formas (2018-00842, 2021-02155), Carl Tryggers Stiftelse and T4F program Sweden.

Author contributions

XRW designed the study and participated in genotyping; JB, WZ and DH conducted genotyping and data analyses; ALC, PA, ADD, OG and HK collected samples from natural populations from their respective regions; AS and VS provided DNA samples; JB, WZ, DH and XRW wrote the manuscript draft, with input from all authors.

Competing interests

The authors declare no competing interests.

Materials & Correspondence

The raw read data have been deposited in the NCBI database under the accession number PRJNA976641 (https://www.ncbi.nlm.nih.gov/bioproject/). The correspondence between the file names and the sampling location is available as a supplementary file.

Figure legends

Figure 1: Sampling and genetic distribution in Scots pine. Panel A: Geographic distribution of the species with sampling sites. Green triangles represent sampling sites. Three populations indicated by grey triangles were used only in PCA and admixture analyses. Species distribution map from Caudullo, Welk and San-Miguel-Ayanz[103]. Panel B: Principal component analysis (PCA) with individuals colored by origin. The same figure colored by country is available in Supplementary Figure 2.

Figure 2: Genetic clustering in Scots pine. Panel A: Admixture inference showing the proportion of each cluster component for K = 2 - 10. Each color represents one genetic cluster. The populations are ordered by country following the PCA results: from Spain to Scandinavia then to China. In each country, the populations are ordered by longitude except for Scandinavia where they are ordered by latitude. For clustering patterns at K >10, see Supplementary Figure 3. Panel B and C: Cluster distribution on a map for K=3 and K=6. The colors follow the result in panel A. To reduce the number of pies and increase the readability, the populations represented here are the merged populations (Supplementary Figure 1).

Figure 3: Spatial genetic patterns. Panel A: isolation by distance, with Mantel test on geodistance *vs* genetic distance between populations. The genetic distance is presented by normalized F_{ST} between populations. Panel B: isolation by environment with Mantel test on environmental distance *vs* genetic distance. Panel C: Relationship between geodistance and environmental distance. Panel D: Estimation of the effective migration surface between pairs of populations across the distribution range of Scots pine. Circles correspond to the sampled populations, with radius proportional to the sampling size. Brown areas correspond to a reduction of the gene flow compared to what would be expected given the distance separating populations, and blue areas correspond to an increase of the gene flow.

Figure 4: Population history of Scots pine. Panel A: Stairway Plot analysis of the three main lineages and of the entire species. The plain lines represent the estimated median effective population sizes (N_e), and the shaded areas correspond to the 95% confidence interval. Panel B: fastsimcoal2 analysis. Each block represents a current or ancestral population with their estimated N_e. Arrows indicate the direction of gene flow with the estimated migration rate labeled above or below the arrow. The timing of the splitting into two groups and their size change are indicated in million years ago (MYA) or thousand years ago (KYA), respectively. The estimates were based on a mutation rate of 7×10^{-10} mutations /site /year and a generation time of 50 years.

Figure 5: Genotype environment association analyses. Panel A: Venn diagram showing the partition of the observed total variance into geographic and environmental contributions, with the two factors jointly explain 10%, while 8% and 1% independently. Panel B: RDA showing 16 SNPs (●) with a *p*-value of 0.001 or lower. Main environmental variables correlating to RDA1 were bio4 and elev with a correlation of -0.18 and 0.19, respectively. Panel C, no *p*-values of the 16 SNPs from the RDA deviated from the expected distribution, suggesting that they are likely false positives. Panel D: Comparison of simulated outlier values based on the pseudo-observed data (POD) with observed values. POD estimates are based on the posterior mean and standard deviation of the two parameters of the Beta prior distribution assumed for the (across populations) frequencies of the SNP reference allele of the core model, see material and methods. Inset QQ-plot shows observed vs expected *p*-values given the log-normal distribution of observed X^tX outlier estimates (X^tX values > 311.93). Blue colored dots (● 103) in the inset figures are significant outlier SNPs at $p < 0.001$. Panel E: Loci with significant effect size of four climatic/soil variables from the BayPass analysis. Four variables were found to be associated with two or more loci. Significance was determined when the absolute value of allele effect size (β) minus its standard deviation is larger than zero, $|\beta|-\sigma(\beta) > 0$, which for this data corresponds to a posterior inclusion probability (PIP) of 0.55 or larger. Because the variables were scaled to a unit's variance, an allele effect size of -0.06 of ai means that the arid index is 0.06 standard deviations lower when the alternative allele is present. X^tX outliers are loci that have differentiated allele frequencies across populations.

Figures

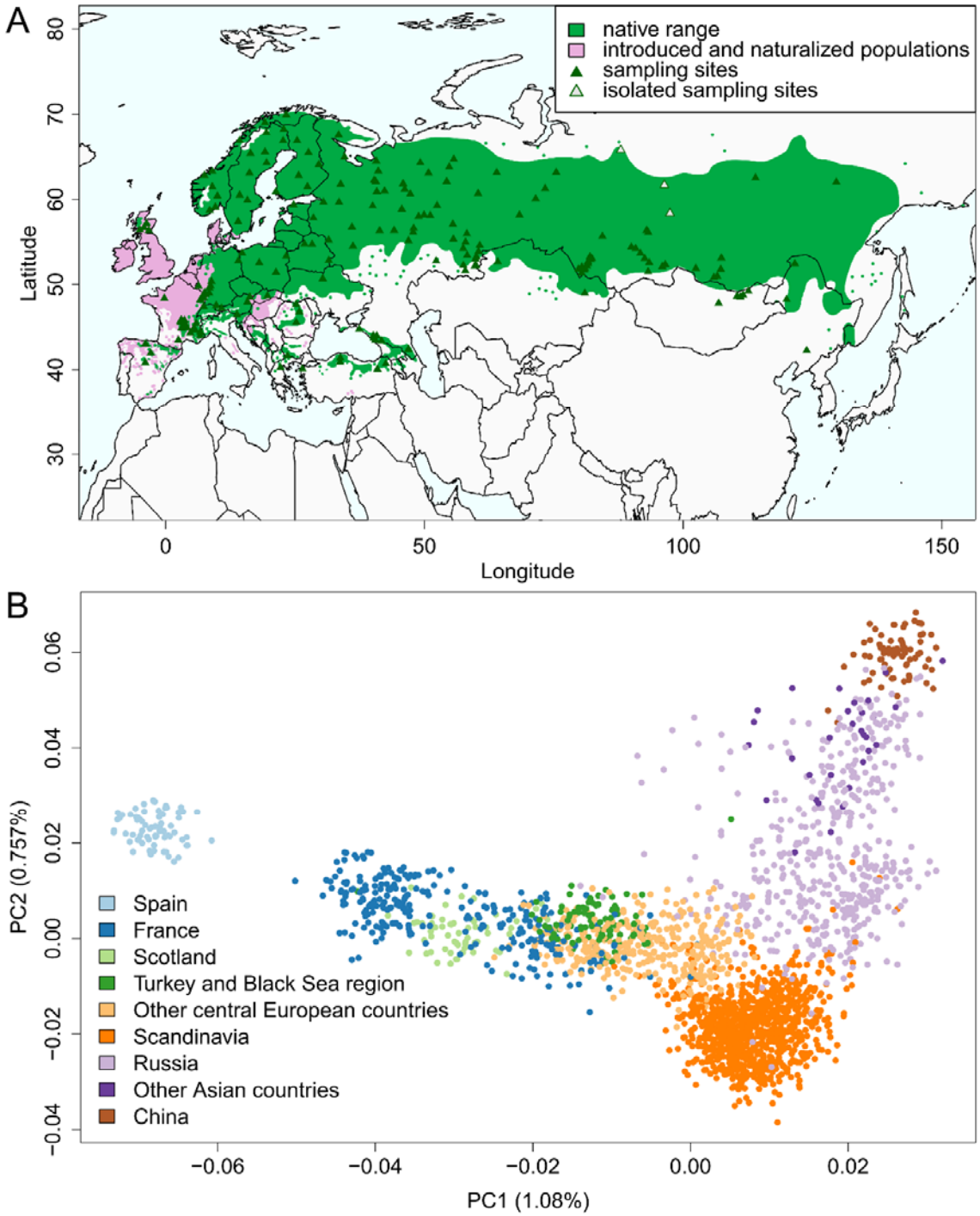


Figure 1: Sampling and genetic distribution in Scots pine. Panel A: Geographic distribution of the species with sampling sites. Green triangles represent sampling sites. Three populations indicated by grey triangles were used only in PCA and admixture analyses. Species distribution map from Caudullo, Welk and San-Miguel-Ayanz[106]. Panel B: Principal

component analysis (PCA) with individuals colored by origin. The same figure colored by country is available in Supplementary Figure 2.

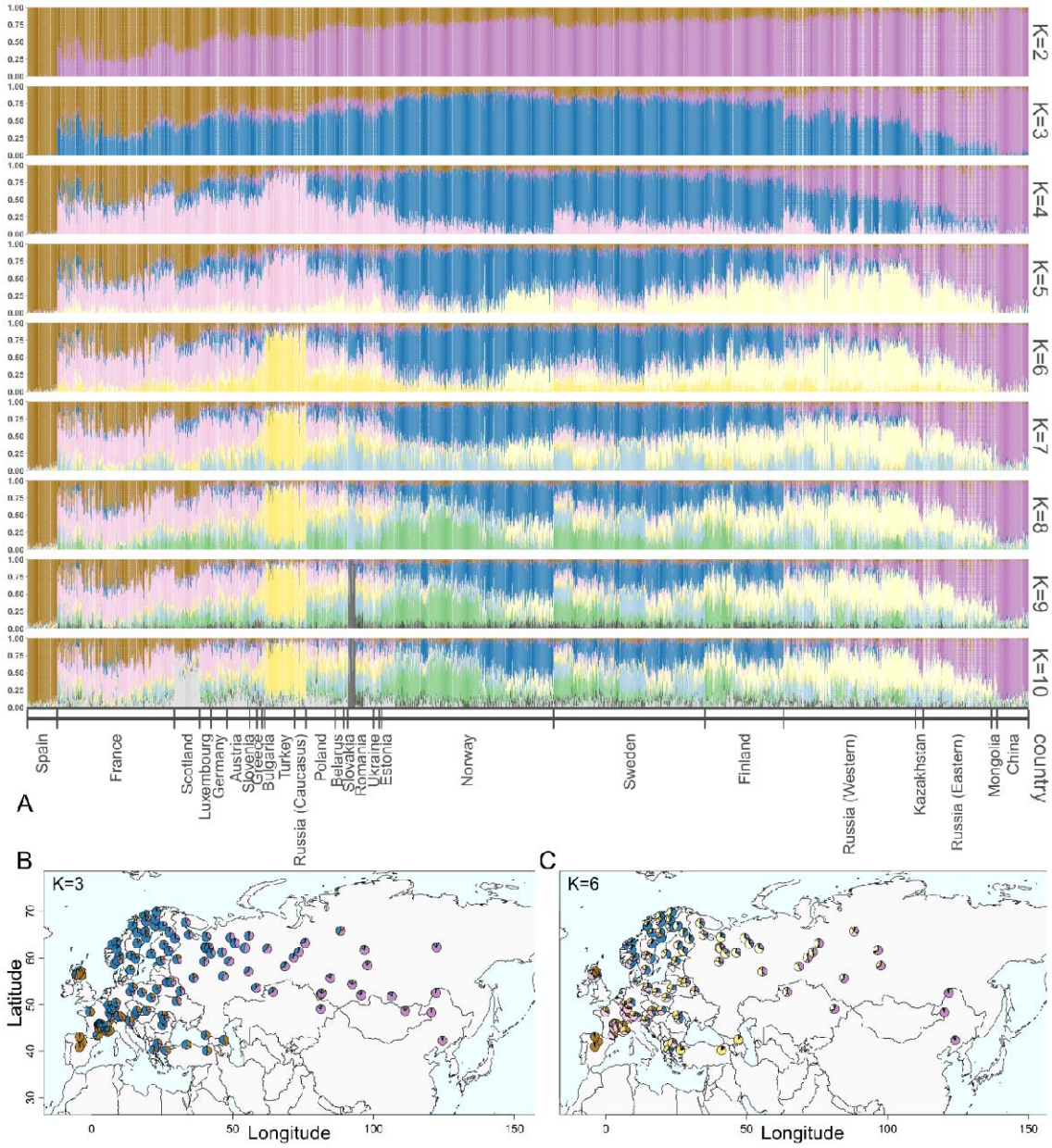


Figure 2: Genetic clustering in Scots pine. Panel A: Admixture inference showing the proportion of each cluster component for K = 2 - 10. Each color represents one genetic cluster. The populations are ordered by country following the PCA results: from Spain to Scandinavia then to China. In each country, the populations are ordered by longitude except for Scandinavia where they are ordered by latitude. For clustering patterns at K >10, see Supplementary Figure 3. Panel B and C: Cluster distribution on a map for K=3 and K=6. The colors follow the result in panel A. To reduce the number of pies and increase the readability, the populations represented here are the merged populations (Supplementary Figure 1).

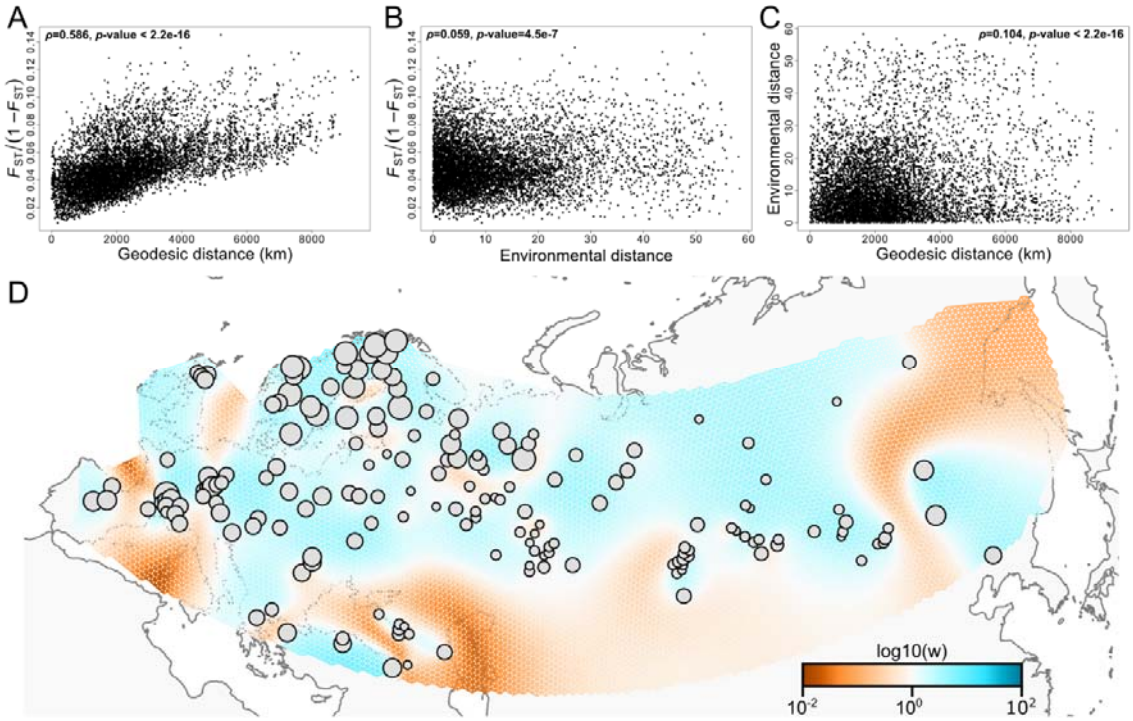
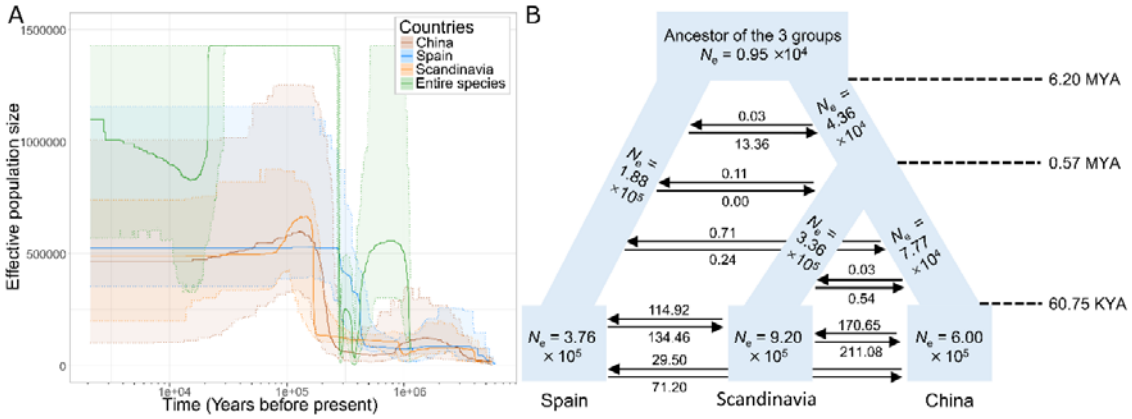


Figure 3: Spatial genetic patterns. Panel A: isolation by distance, with Mantel test on geodistance *vs* genetic distance between populations. The genetic distance is presented by normalized F_{ST} between populations. Panel B: isolation by environment with Mantel test on environmental distance *vs* genetic distance. Panel C: Relationship between geodistance and environmental distance. Panel D: Estimation of the effective migration surface between pairs of populations across the distribution range of Scots pine. Circles correspond to the sampled populations, with radius proportional to the sampling size. Brown areas correspond to a reduction of the gene flow compared to what would be expected given the distance separating populations, and blue areas correspond to an increase of the gene flow.



954

955 **Figure 4:** Population history of Scots pine. Panel A: Stairway Plot analysis of the three main
956 lineages and of the entire species. The plain lines represent the estimated median effective
957 population sizes (N_e), and the shaded areas correspond to the 95% confidence interval. Panel
958 B: fastsimcoal2 analysis. Each block represents a current or ancestral population with their
959 estimated N_e. Arrows indicate the direction of gene flow with the estimated migration rate
960 labeled above or below the arrow. The timing of the splitting into two groups and their size
961 change are indicated in million years ago (MYA) or thousand years ago (KYA), respectively.
962 The estimates were based on a mutation rate of 7×10^{-10} mutations /site /year and a
963 generation time of 50 years.

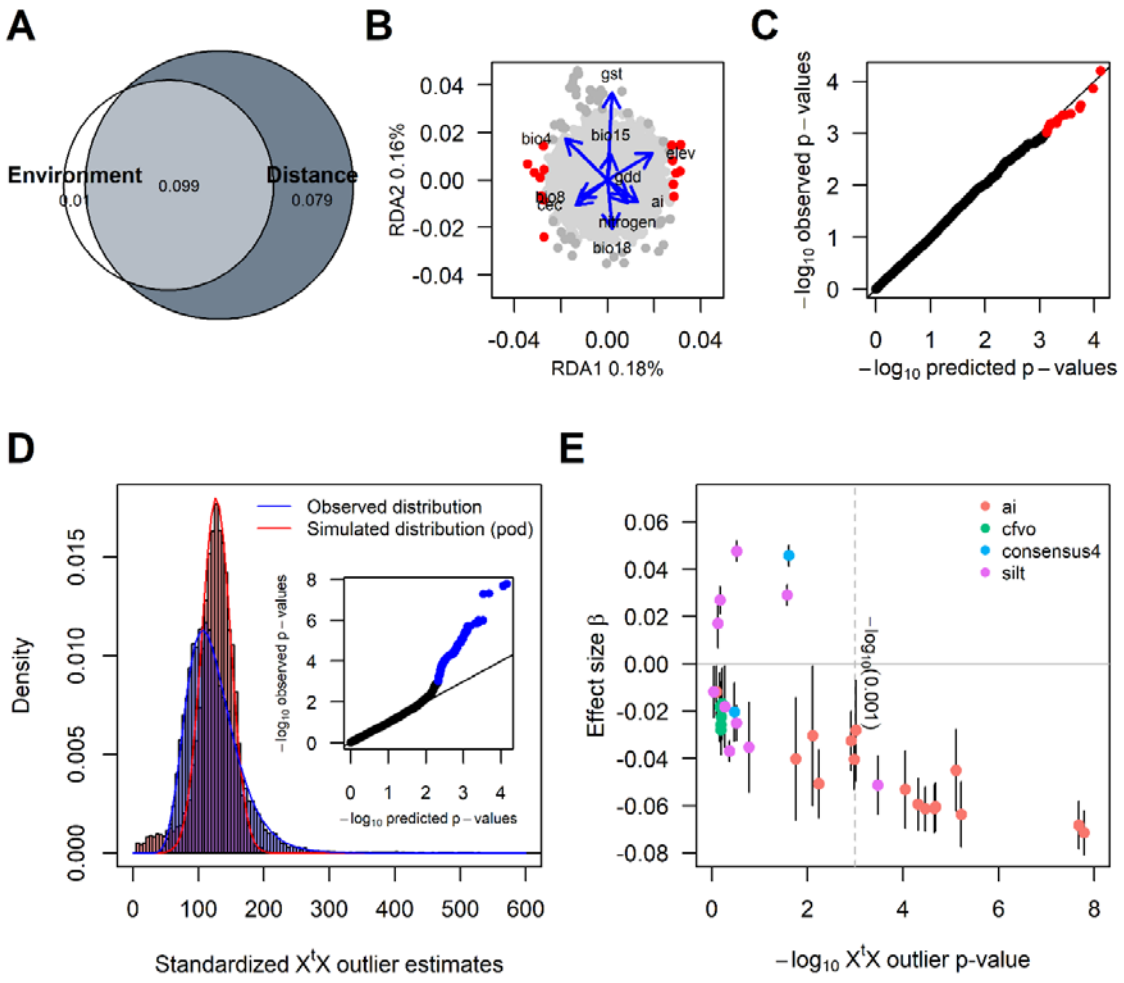


Figure 5: Genotype environment association analyses. Panel A: Venn diagram showing the partition of the observed total variance into geographic and environmental contributions, with the two factors jointly explain 10%, while 8% and 1% independently. Panel B: RDA showing 16 SNPs (●) with a *p*-value of 0.001 or lower. Main environmental variables correlating to RDA1 were bio4 and elev with a correlation of -0.18 and 0.19, respectively. Panel C, no *p*-values of the 16 SNPs from the RDA deviated from the expected distribution, suggesting that they are likely false positives. Panel D: Comparison of simulated outlier values based on the pseudo-observed data (POD) with observed values. POD estimates are based on the posterior mean and standard deviation of the two parameters of the Beta prior distribution assumed for the (across populations) frequencies of the SNP reference allele of the core model, see material and methods. Inset QQ-plot shows observed vs expected *p*-values given the log-normal distribution of observed X^tX outlier estimates (X^tX values > 311.93). Blue colored dots (● 103) in the inset figures are significant outlier SNPs at $p < 0.001$. Panel E: Loci with significant effect size of four climatic/soil variables from the BayPass analysis. Four variables

were found to be associated with two or more loci. Significance was determined when the absolute value of allele effect size (β) minus its standard deviation is larger than zero, $|\beta|-\sigma(\beta) > 0$, which for this data corresponds to a posterior inclusion probability (PIP) of 0.55 or larger. Because the variables were scaled to a unit's variance, an allele effect size of -0.06 of ai means that the arid index is 0.06 standard deviations lower when the alternative allele is present. X^tX outliers are loci that have differentiated allele frequencies across populations.

Table 1. Summary of the genetic diversity in Scots pine for each country/region or for the whole distribution. The number of populations (N_{pop}) and the number of individuals (N_{ind}) for each country or region are given in the two first columns. The overall pairwise nucleotide diversity (π) and diversity at the 0-fold (π_0) and 4-fold (π_4) sites were calculated on the full dataset including invariant sites. The inbreeding coefficient F_{IS}, the observed and expected heterozygosities (H_o and H_e) were based on the dataset filtered for MAF 0.05. Six individuals from three very distant Russian populations were not included in these analyses. The values presented are averages of the individual populations results, weighted for the number of individuals per population. Results by population are available in Supplementary Table 1.

Region / country	N_{pop}	N_{ind}	π	π_4	π_0	H_o	H_e	F_{IS}
Whole distribution		2315	0.00622	0.00890	0.00490	0.258	0.277	0.057
Spain	3	70	0.00631	0.00900	0.00504	0.265	0.273	0.027
Scotland	4	59	0.00638	0.00903	0.00504	0.284	0.292	0.020
Western and Central Europe		526	0.00650	0.00928	0.00520	0.282	0.290	0.016
France	25	271	0.00655	0.00941	0.00530	0.284	0.293	0.018
Luxembourg	1	26	0.00640	0.00913	0.00505	0.258	0.268	0.033
Germany	3	37	0.00663	0.00919	0.00519	0.278	0.287	0.019
Austria	4	53	0.00643	0.00906	0.00507	0.285	0.292	0.017
Slovenia	1	17	0.00641	0.00919	0.00502	0.278	0.280	0.007
Slovakia	1	9	0.00634	0.00900	0.00509	0.310	0.300	-0.036
Poland	4	68	0.00641	0.00912	0.00513	0.266	0.277	0.026
Belarus	2	20	0.00621	0.00921	0.00495	0.299	0.296	-0.018
Estonia	1	6	0.00633	0.00911	0.00506	0.346	0.325	-0.072
Bulgaria	1	6	0.00632	0.00912	0.00494	0.353	0.329	-0.076
Ukraine	1	13	0.00669	0.00949	0.00534	0.270	0.288	0.045
Southeastern Europe		140	0.00644	0.00913	0.00513	0.267	0.275	0.024
Greece	1	12	0.00709	0.00982	0.00564	0.324	0.304	-0.056
Romania	3	60	0.00641	0.00916	0.00518	0.261	0.274	0.039
Turkey	3	68	0.00635	0.00898	0.00500	0.263	0.271	0.026
Eastern Europe and Asia		589	0.00610	0.00878	0.00481	0.268	0.283	0.045
Russia	34	485	0.00613	0.00885	0.00483	0.266	0.285	0.051
Kazakhstan	2	18	0.00612	0.00892	0.00487	0.317	0.304	-0.045
Mongolia	1	13	0.00616	0.00892	0.00494	0.264	0.297	0.085
China	3	73	0.00591	0.00828	0.00467	0.265	0.269	0.016

Northern Europe		931	0.00610	0.00871	0.00473	0.235	0.265	0.099
Norway	9	398	0.00607	0.00862	0.00466	0.236	0.265	0.100
Sweden	10	351	0.00612	0.00880	0.00480	0.236	0.264	0.093
Finland	6	182	0.00610	0.00871	0.00472	0.233	0.267	0.109