Resilience of genetic diversity in forest trees over the Quaternary

Pascal Milesi[1,2]*†, Chedly Kastally[3,4]†, Benjamin Dauphin[5]†, Sandra Cervantes[6,7]†, Francesca Bagnoli[8], Katharina B. Budde[9,10], Stephen Cavers[11], Bruno Fady[12], Patricia Faivre-Rampant[13], Santiago C. González-Martínez[14], Delphine Grivet[15], Felix Gugerli[5], Véronique Jorge[16], Isabelle Lesur Kupin[14,17], Dario I. Ojeda[18], Sanna Olsson[15], Lars Opgenoorth[5,19], Sara Pinosio[8,20], Christophe Plomion[14], Christian Rellstab[5], Odile Rogier[16], Simone Scalabrin[21], Ivan Scotti[12], Giovanni G. Vendramin[8], Marjana Westergren[22], GenTree Consortium, Martin Lascoux[1,2]*‡, Tanja Pyhäjärvi[3,4]*‡

Affiliations:

[1]Department of Ecology and Genetics, Evolutionary Biology Centre, Uppsala University, 752 36 Uppsala, Sweden
[2]Science for Life Laboratory (SciLifeLab), 752 36 Uppsala, Sweden
[3]Department of Forest Sciences, University of Helsinki, 00014 Helsinki, Finland
[4]Viikki Plant Science Centre, University of Helsinki, 00014 Helsinki, Finland
[5]Biodiversity and Conservation Biology, Swiss Federal Research Institute WSL, 8903 Birmensdorf, Switzerland
[6]Department of Ecology and Genetics, University of Oulu, 90014 Oulu, Finland
[7]Biocenter Oulu, University of Oulu, 90014 Oulu, Finland
[8]Institute of Biosciences and Bioresources, National Research Council of Italy (IBBR-CNR), 50019 Sesto Fiorentino, Italy
[9]Department of Forest Genetics and Forest Tree Breeding, Georg-August-University Goettingen, 37077 Göttingen, Germany
[10]Department of Forest Genetic Resources, North Western German Forest Research Institute, 34346 Hann. Münden, Germany
[11]UK Centre for Ecology & Hydrology (UKCEH), EH26 0QB Bush Estate, United Kingdom
[12]INRAE, URFM, Ecology of Mediterranean Forests, 84914 Avignon, France
[13]University of Paris-Saclay, INRAE, EPGV, 91000 Evry, France
[14]University of Bordeaux, INRAE, BIOGECO, 33610 Cestas, France
[15]Institute of Forest Sciences (ICIFOR-INIA), CSIC, 28040 Madrid, Spain
[16]INRAE, ONF, BioForA, 45075 Orléans, France
[17]Helix Venture, 33700 Mérignac, France
[18]Department of Forest Biodiversity, Norwegian Institute of Bioeconomy Research (NIBIO), 1433 Aas, Norway
[19]Plant Ecology and Geobotany, Philipps-Universität Marburg, 35043 Marburg, Germany
[20]Institute of Applied Genomics (IGA), 33100 Udine, Italy
[21]IGA Technology Services S.r.l., 33100 Udine, Italy
[22]Slovenian Forestry Institute, 1000 Ljubljana, Slovenia

*Corresponding authors. E-mail(s): pascal.milesi@scilifelab.uu.se, tanja.pyhajarvi@helsinki.fi, martin.lascoux@ebc.uu.se

†These authors contributed equally to this work

‡These authors contributed equally to this work

Abstract: The effect of past environmental changes on the demography and genetic diversity of natural populations remains a contentious issue and has rarely been investigated across multiple, phylogenetically distant species. Here, we performed comparative population genomic analyses and demographic inferences for seven widely distributed and ecologically contrasted European forest tree species based on concerted sampling of 164 populations across their natural ranges. For all seven species, the effective population size, N_e, increased or remained stable over many glacial cycles and up to 15 million years in the most extreme cases. Hence, surprisingly, the drastic environmental changes associated with the Pleistocene glacial cycles have had little impact on the level of genetic diversity of dominant forest tree species, despite major shifts in their geographic ranges. Based on their trajectories of N_e over time, the seven tree species can be divided into three major groups, highlighting the importance of life history and range size in determining synchronous variation in genetic diversity, over time. Altogether, our results indicate that forest trees were able to retain their evolutionary potential over very long periods of time despite strong environmental changes.

One-Sentence Summary: Population genomic analyses reveal that forest trees retained their evolutionary potential over the Quaternary.

Main Text:

Extant northern temperate and boreal tree species have existed for millions of years and survived multiple glacial cycles. Micro- and macrofossil data indicate that these tree species have undergone huge range changes and large fluctuations in their census population sizes (N_c) over time[1]. This was largely corroborated by numerous phylogeographic studies[2]. Yet, today most tree species harbor large amounts of genetic diversity[3] and they have been shown to respond rapidly to recent environmental challenges such as the Little Ice Age[4] or the Last Glacial Maximum (LGM)[5,6]. While these rapid responses to new selection pressures are consistent with their current large population sizes, high realized outcrossing rates and efficient gene flow, they may seem paradoxical in view of the large census size changes suggested by the fossil records. Unfortunately, we still lack a comprehensive view of the impact of past demographic changes on N_e, the key evolutionary parameter defining the genetic diversity and efficacy of selection[7]. For example, did N_e fluctuate strongly through time or, on the contrary, was it retained and stable over repeated glacial cycles despite changes in N_c? Are changes in N_e primarily driven by climatic events or do they also reflect intrinsic biological characteristics such as life history or physiological properties? In the former case one would expect a high synchronicity in changes across multiple species independently of their biological properties while, in the latter case, one would expect species to form categories according to their patterns of diversity changes and shared biological properties[8,9].

In Europe, the LGM which occurred approximately 27,000 to 19,000 years ago, and ensuing Holocene recolonization have often been assumed to be the main drivers of the current distribution of genetic diversity, with southern populations being typically more diverged than those from the northern core-range[10,11]. Earlier analyses of the demographic histories of European forest tree species generally relied on organellar markers whose polymorphisms are informative on a shorter time span than nuclear markers since in monoecious species their effective population size is half of that for nuclear markers and in dioecious a quarter[12,13]. Quite naturally, outcomes were interpreted from the perspective of only the most recent glacial period (i.e., LGM)[11]. Further, inferences based on organellar markers that behave as a single locus and, in most cases, are maternally inherited and only disperse via seeds, have limited relevance for nuclear genetic diversity where most of the genetic variation lies. Whole-genome sequencing combined with coalescence-based demographic methods allowed inferring the demographic history of forest trees and its timescale well beyond the LGM, up to millions of years. However, most studies so far address single species or focus on inferring the timing of divergence and the extent of gene flow between populations or closely related species[14–18]. Congruence between population history and glacial oscillations has been observed in some species[14,19] but has remained far from being conclusive in others[20,21]. A general conclusion on the drivers of temporal changes in genetic diversity across species cannot, however, be drawn from the compilation of these studies, or even from the reanalysis of the data they present, due to the heterogeneity of sampling strategies, genomic sources of polymorphism, and numbers of loci.

Here, we carried out a comprehensive demographic inference of seven major European tree species, distributed from the boreal to the Mediterranean regions (Table 1), based on a common strategy both for sampling populations across Europe and for sequencing genomic regions. All seven species are wind-pollinated, three are conifers (*Picea abies*, *Pinus pinaster*, and *Pinus sylvestris*) and four are angiosperms (*Betula pendula*, *Fagus sylvatica*, *Populus nigra*, and *Quercus petraea*). We conducted targeted nuclear DNA sequencing (~10,000 species-specific probes that covered ~3 Mbp of largely orthologous sequences) on a total of 3,407 adult

trees collected from 19 to 26 locations per species (~25 individuals each) across large parts of their natural ranges (Figs. 1, S1, Supplementary Text, Tables 1, S1–S2)[22]. We first conducted a comprehensive survey of the distribution of current genetic diversity in all seven species and used state-of-the-art coalescent approaches to reconstruct changes in effective population sizes over multiple glacial cycles and test for synchronous changes across species.

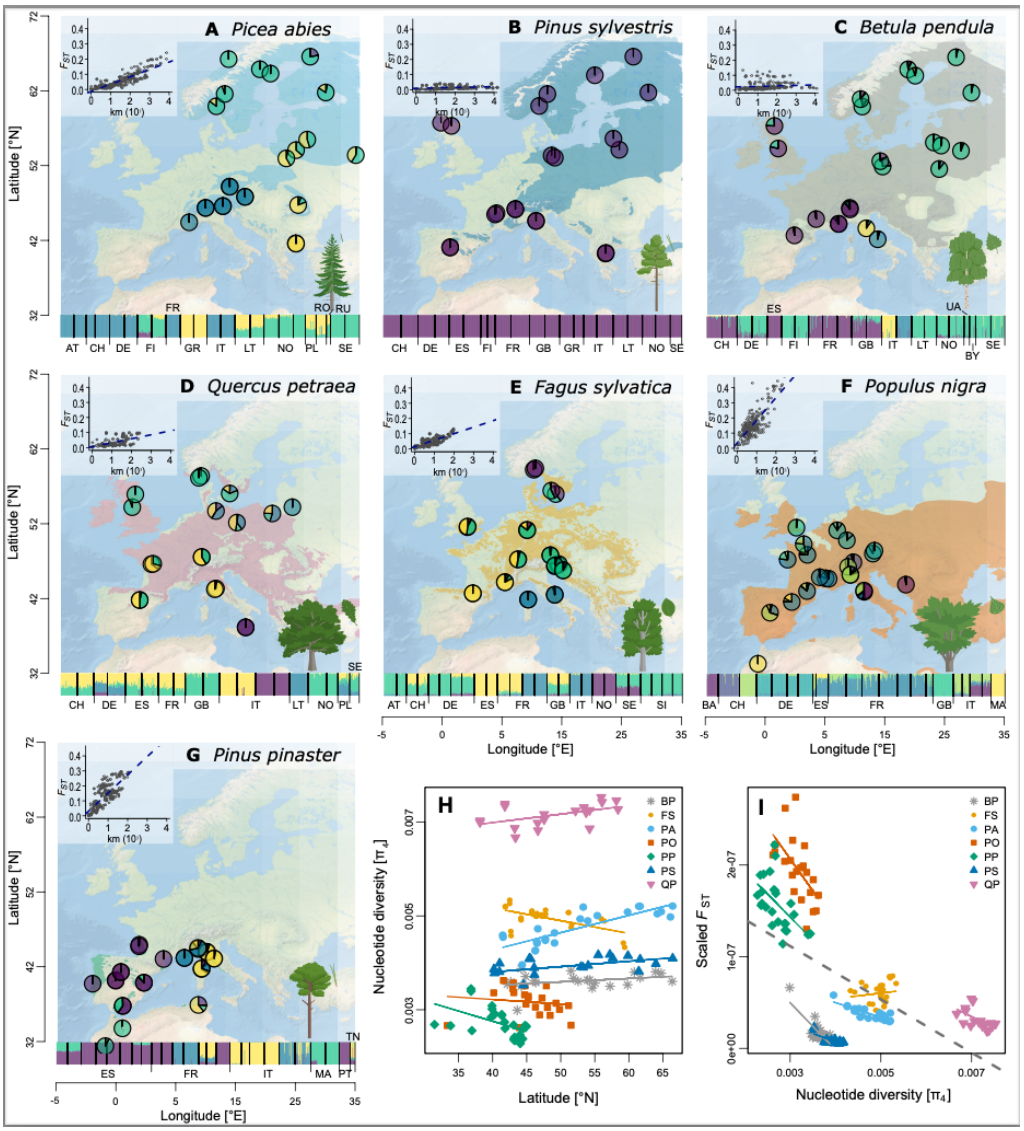

Fig. 1. Admixture, isolation by distance, genetic diversity and divergence patterns of the seven studied European tree species. (A–G) Pie charts represent population average admixture coefficients. Four genetic clusters are shown to visualize genetic structure, except for *Pinus sylvestris* (K=1; B) and *Populus nigra* (K=7; F) (see Figs. S3–S16 for other cluster numbers). An admixture bar plot of all individuals is shown at the bottom of each panel (country codes are explained in Table S1). Background maps represent species' ranges. Inset graphs show patterns of isolation by distance. **(H)** Nucleotide diversity at four-fold sites, π_4, per base pair as a function of latitude. **(I)** Population-specific scaled differentiation estimated as the average of the ratios of pairwise F_{ST} over pairwise distance for all population pairs as a function of π_4. Different symbols and colors represent the populations of the different species with their respective trend lines. Species codes are explained in Table S2. The few populations from Russia are not represented here and can be found in Figures S1 and S12.

Results

Patterns of genetic diversity do not reflect phylogeny or environmental preferences

Overall, current genetic diversity and structure in the seven species (Fig. 1) reflected neither phylogeny nor current environmental preferences. The patterns likely followed from a combination of biological and ecological characteristics as well as range-limit constraints (biotic or abiotic). With respect to genetic diversity and genetic differentiation among populations, the seven species can be divided into four sets: highly genetically structured and low diversity *P. pinaster* and *P. nigra*; moderately structured, and intermediate diversity *F. sylvatica* and *P. abies*; moderately structured, high diversity *Q. petraea*; and finally, panmictic and moderate diversity *P. sylvestris* and *B. pendula* (Fig. 1I).

Nucleotide diversity at four-fold degenerate, synonymous sites (π_4) ranged from 0.0027 to 0.0072 per bp across the seven species (Tables 1, S3), as is typical of outcrossing trees[23], and was remarkably similar among populations of a given species (Fig. 1H, Table S3). *Quercus petraea* had the highest genetic diversity and it increased towards the north. Also, boreal species (*P. abies*, *P. sylvestris* and *B. pendula*) exhibited slightly higher diversity at high latitudes, whereas genetic diversity tended to decrease northwards for the temperate species *P. nigra*, *F. sylvatica* and *P. pinaster* (Fig. 1H, Table S3). Thus, the geographic distribution of genetic diversity did not consistently follow the south–north latitudinal gradient that is often considered as a proxy for postglacial recolonization history[24].

Genetic differentiation between populations (F_{ST}) was low for most species, except for *P. pinaster* and *P. nigra* (Table 1). Isolation-by-distance at the range scale was significant for most species, likely reflecting the distance-dependency of wind-mediated pollen dispersal over mere population-level genetic drift (Fig. 1A–G). However, the level of divergence was not uniform across the species distributions. Generally, the most genetically divergent populations were found at southern latitudes (Fig. S2), similarly to what Petit et al.[11] found. This result is also supported by the spatial distribution of ancestry proportions and principal component analysis (Figs. 1A–G, S3-S23, Table S3). Additionally, populations at higher elevations were more differentiated from the rest of the range (Fig. S2, Table S4). Globally, genetic structure coincided with the main discontinuities in the species' distributions, but with considerable variation across species (Figs. 1A–G, S3–S16).

Main divergence events largely predate the last glacial maximum in all seven species

To compare divergence and demographic events across species, we analyzed different subsets of the data with, in each case, simple and consistent modeling choices and methodology. The origin and timing of the divergence between populations (Table 1) was studied using demographic models implemented in fastsimcoal2[25,26]. To estimate the timeframe of population separation we analyzed in each species two non-admixed populations representing the main southern and northern clusters (Table S5, Fig. S24). In all species, divergence models with migration had better support than models without (Table S5), and the estimated divergence times between major clusters largely predated the LGM, extending from 0.6 Mya (middle Quaternary) up to 17 Mya within the Miocene (Fig. S25, Tables S5–S7). Hence, for all species the formation of the main genetic groups was an outcome of demographic events having occurred over multiple glacial cycles, and, importantly, these groups were preserved through glacial cycles despite extensive gene flow. Consequently, the overall pattern of genetic differentiation reflects

topography and other persistent barriers to movement of populations, instead of recent divergence during the LGM. For example, the mountains of southern Europe (e.g. Pyrenees, Alps) could have driven the recurrent formation of 'sky islands', i.e. isolated high-elevation regions to which cold-adapted species repeatedly shifted during interglacial periods, resulting in higher divergence between southern high elevation populations than between populations at lower altitudes or populations at more northern latitudes[27].

General increase in N_e over multiple glacial cycles

To infer the timescale of changes in N_e in the seven species and over many glacial cycles, we used Stairway Plot 2[28], a composite likelihood method that infers changes in N_e over time from site frequency spectra (SFS). Because Stairway Plot 2 is model-flexible and inferences can be biased towards complex models, with more demographic events occurring (overfitting), we tested the robustness of the results with the more constrained fastsimcoal2 2-epoch model. Importantly, to account for the effect of sampling on demographic inferences[29], we conducted analyses at the species, population, and one-sample-per-population levels. With the last level, the analysis was focused on the dominating 'collecting' phase of the genealogy[30,31].

All estimates of past and present N_e were in the range of tens or hundreds of thousands; these values are much smaller than any reasonable estimate of the species' current census sizes (N_c), which is in the scale of billions of individuals for most species[32,33]. While the observed ratio was less than common mean estimates of N_e/N_c, it fits with the trend of species with large N_c displaying below-average estimates of N_e[34].

All species, except *F. sylvatica*, showed an excess of rare variants in the SFS (Fig. 2) revealing a global signal of ancient population growth (from 0.6 Mya in *P. sylvestris* to 15 Mya in *Q. petraea*). This signal was consistent across sampling schemes and the two inference methods (Figs. 3A–B, S26) and is in line with earlier studies on, e.g., *P. abies* and *P. sylvestris*[17,35]. Strikingly, very few populations exhibited a signal of decreasing N_e through time, and those populations were often disconnected from the rest of the range and, thus, likely to have experienced stronger genetic drift (Fig. S27, Table S8). The magnitude of increase in N_e varied across species and was largest for *P. sylvestris* (from ~5,000 to 500,000) and weakest for *F. sylvatica*, for which the proportional increase was only two-fold (from 100,000 to 200,000).

Crucially, these patterns suggest that the overall genetic diversity of each species, or equivalently for neutral loci their N_e, has been maintained even during the range contractions caused by glacial advances. In other words, dominant forest trees with large ranges, large N_c, and efficient gene flow have retained genetic diversity over long periods of time, despite regional extinctions during ecologically unfavorable periods. This contrasts with previous studies that generally tended to consider the LGM as the major cause of genetic change in forest tree species in Europe.

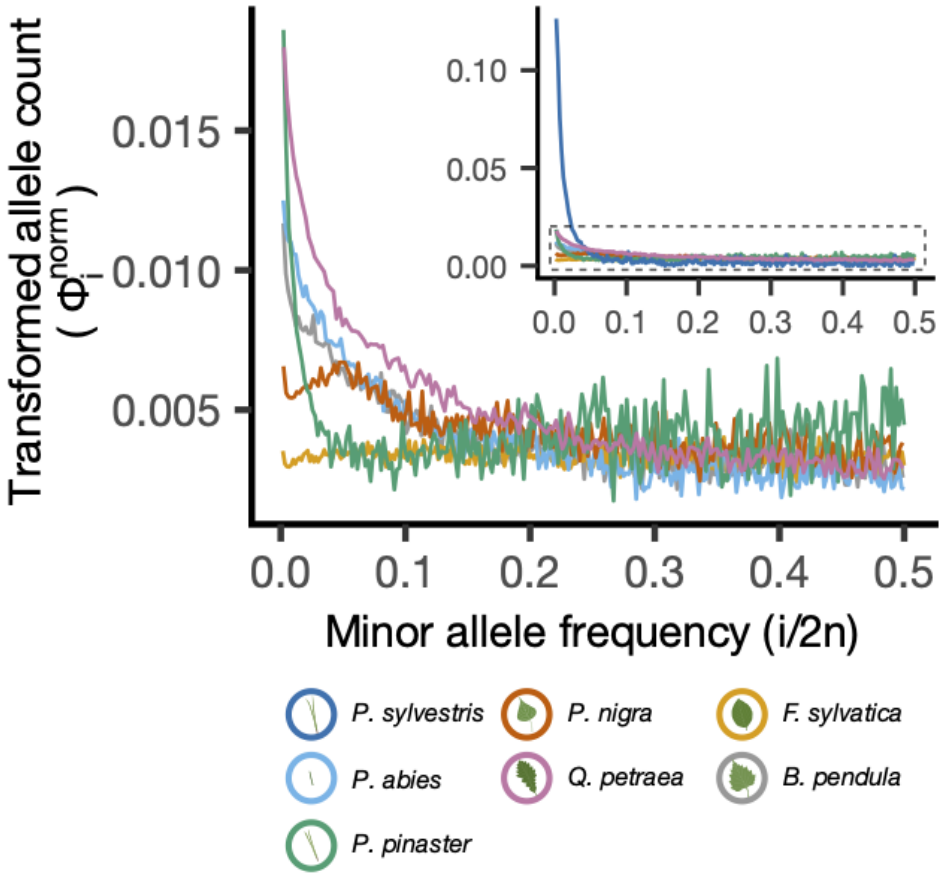

Fig. 2. Folded site frequency spectra (SFS) of all seven studied species (inset) and of the remaining six species after excluding *Pinus sylvestris* (with adjusted scale in the y axis). The SFS were transformed and normalized following[36] (see Materials & Methods) so that they are flat under the standard neutral model, facilitating the visualization of the SFS that depart from expectations of the standard neutral model[37,38].

Synchronicity and idiosyncrasy in changes in N_e

Across the seven species, the changes in N_e over time were not entirely species-specific (Fig. 3C, Fig. S28 and Table S9). Instead, three groups of species could be distinguished on the basis of their demographic trajectories (Fig. 3C). A first group included the three boreal species (*P. abies*, *P. sylvestris* and *B. pendula*) and the riparian *P. nigra*, a second group comprised the two major temperate broadleaves (*F. sylvatica and Q. petraea*), and the only Mediterranean conifer, *P. pinaster*, constituted a group of its own. Interestingly, this grouping differs from the one established above based on current nucleotide diversity and population genetic structure. Neither patterns of N_e of individual species nor synchronous changes align well with the known glacial and interglacial periods (Fig. 3). This contrasts with some earlier results of individual forest tree species (e.g.,[14,19,39]), but is in line with other comparative studies[8,9]. The overall lack of correlation between changes in N_e and climatic oscillations suggest that forest tree populations remained highly interconnected and acted as a single, large metapopulation, whose N_e was less affected by climatic fluctuations than N_c. Still, repeated synchronous phases of slight decrease in N_e can be detected despite the global increasing trends, likely being the signature of shared,

recurrent and possibly delayed influence of glacial cycles on these metapopulations (Fig. S28 and Tab. S9).

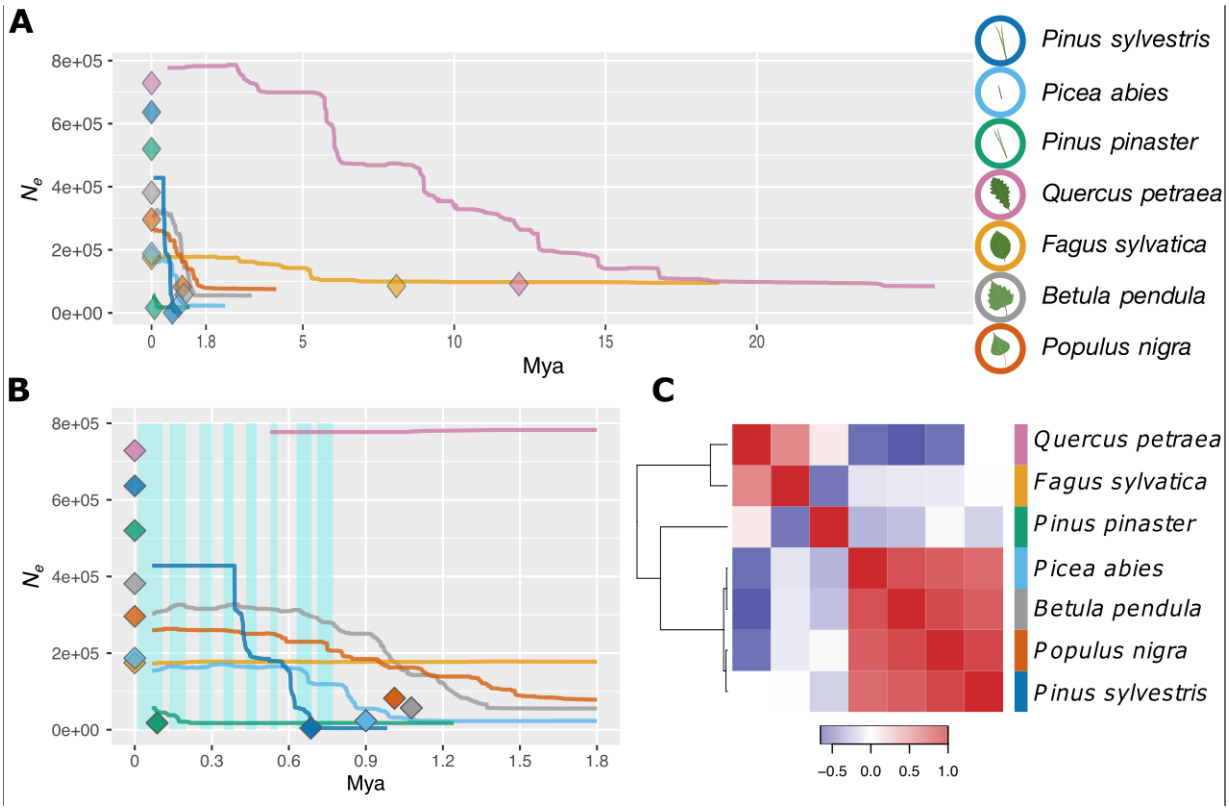

Fig. 3. Demographic change across time for the seven tree species. (**A**) Change in effective population size (N_e) through time (million years ago, Mya), inferred with Stairway Plot 2 (lines, one-sample-per-population dataset; see Table S8) or with fastsimcoal2 (diamonds, 2-epoch model and one-sample-per-population dataset; see Tables S6–S7). The median changes in N_e are reported from both methods. (**B**) A zoom-in of (A) focusing on the 0–1.8 Mya period. Blue shaded rectangles delineate glacial periods. Timing of the glacial periods was inferred from composite CO_2 records publicly available at ([link]). (**C**) Heatmap based on Kendall's correlation coefficients computed from changes in N_e through time between each pair of species. The order of species along the x-axis is the same as that along the y-axis. Blue and red colors represent negative and positive correlation values, respectively.

Discussion

Current population genetic diversity is the result of long-term processes

Our study demonstrates that tree species have been able to retain their evolutionary potential through multiple glacial cycles. This is in agreement with recent studies showing the ability of tree species to rapidly respond to environmental challenges[4,6] and to swiftly colonize new areas as they become suitable[40]. This potential is likely a reflection of their unique biological features. Very large and genetically connected populations allowed tree species to retain genetic diversity through time, despite intermittent, substantial geographical range contractions. This diversity has taken shape and accumulated over very long periods of time, involving multiple glacial-

interglacial cycles. Similarly, an investigation of the Distribution of Fitness Effects (DFE) of new mutations for a subset of the same data showed differences in DFE between species but no differences among populations within species. This finding also suggests that differences in DFE parameters accrue over long time periods[41] and essentially reflect the collecting phase of the coalescent process of a many-demes model[30,31].

While the use of pollen and macrofossil records as well as low-resolution uni-parentally inherited organellar DNA markers in previous studies revealed post-LGM migration patterns[11], these approaches did not give a full picture of the dynamics of genetic diversity over time, simply because both data types cover a too short time period. Pollen records are informative on past plant population movements but chronologies of tree pollen records are limited[42], often uncertain especially at the species level[43], and heterogeneous across space[44]. Further, they comprise no information on genetic relationships. Long-lived organisms, such as forest trees, have fewer generations per glacial cycle than annual and herbaceous species, and therefore will, for the same number of generations, experience a larger number of glacial cycles. Hence, as was indeed observed, the current structure of genetic diversity reasonably reflects the impact of many glacial cycles (e.g.[45,46]). Importantly, our results explain how tree species that have survived repeated glacial cycles were able to retain genetic diversity and, hence, a capacity to respond to new environmental challenges. However, genetic stability across millions of years does not exclude drastic changes in the short term, e.g., in species distributions and local abundances, which can have major impacts on ecosystem and forest functions.

Since we focused on inferring these demographic events based on current genetic diversity, we disregarded individuals with a high degree of admixture. However, in most plant species, including trees, hybridization contributes significantly to genetic diversity[14,47,48]. This situation primarily holds for *Q. petraea*[49], which shows high nucleotide diversity (Fig. 1H) among the seven species we studied, but it may also be relevant for others that show hybridization at least in parts of their ranges (Table 1). For demographic inference, hybridization introduces signals of even older evolutionary events and leads to elevated estimates of N_e. For predicting population responses to climate change, more information on groups of closely related species will be essential, especially as introgression can be important in environmental adaptation[50].

Changes in N_e are not only driven by glacial cycles

The genetic diversity of all seven species was maintained across long periods of time: most of the time, N_e either increased or remained stable. However, we did not observe a single shared dynamic of changes in N_e across the seven species, which would be expected if glacial cycles were the main drivers. Instead, there were three clear groups based on N_e change over time (Fig. 3C). In this respect, our results are congruent with previous studies that also showed a tendency to low levels of overall synchronicity among species and even populations within species ([8,9] and references therein). Bai et al.[8] carried out a comparative demographic study in walnut species and showed that the timing and amplitude of changes in N_e differed among species. They concluded that the population histories of these walnut species were not driven by extrinsic environmental changes alone and that different species responded idiosyncratically to similar environmental challenges[8]. Similarly, Helmstetter et al.[9] reconstructed the trajectories of N_e of three palm species and four Annonaceae tree species from African rainforest. Here too, evolutionary responses were largely asynchronous and individualistic. Interestingly, the three

palm species had large N_e (around 500,000) that increased regularly through time, as observed for our seven temperate forest tree species. This was in contrast to the four Annonaceae species, whose N_e were significantly smaller (around 50,000) and fluctuated strongly over the same time period. In the present study, the peculiarity of *P. pinaster* likely mirrors its more southern distribution as well as its highly fragmented range compared to the other six species while the synchronicity between *Q. petraea* and *F. sylvatica* likely finds its origin in similarities in both biology (e.g., generation time, ecological niche, dispersal) and geographical distribution. The recovered demographic histories of *Q. petraea* and *F. sylvatica* go back much further in time (i.e., ~32 and ~17 Mya, respectively) than for the other studied species. The depth of the *Q. petraea* genealogy and its relatively large N_e are likely a consequence of continuous hybridization with other abundant and closely related white oak species[51], combined with long generation times. *Fagus sylvatica* may have had small, secondary refugia outside the core refugial areas, maintaining local reservoirs of genetic diversity across glacial cycles[44,52]. The synchronicity of the three boreal species also reflects some similarities in both biology and geographical distribution, although they also show marked differences in population genetic estimates. One explanation for the grouping of *P. nigra* with the boreal species could be that as a riparian species it was less affected by global climatic patterns and could have survived glacial times in microrefugia close to rivers, also in colder regions. Finally, it is worth pointing out that the two pine species, *P. pinaster* and *P. sylvestris*, were very distinct from each other, despite similar mutation rates, rates of adaptive evolution and generation times, reflecting their different ecological characteristics and geographic ranges[5,53].

Conclusion

The present study highlights the existence of commonalities in the N_e trajectories of seven major European forest tree species. Firstly, the N_e of all seven species we analyzed showed a mostly monotonic growth over a large part of the Quaternary. Hence, the genetic diversity of large tree metapopulations has been strikingly resilient to the drastic environmental changes during which the species experienced regional extinctions and extensive shifts in species distributions. Secondly, the trajectories of N_e through time were correlated within groups of species sharing ecological and biogeographical properties, but not across all species. This supports the idea that changes in N_e are not solely driven by climatic events but also reflect species' shared biological characteristics such as life history or physiological properties. For instance, two species, such as *B. pendula* and *P. abies*, can have rather different extant population genetic structure across their similar distribution ranges, and yet have highly similar N_e trajectories.

Finally, it has been suggested that understanding individual species' responses to past climatic oscillations is critical to predict their ability to cope with climate change[9]. Our results, in contrast, suggest that species' idiosyncratic responses, while highly desirable to understand for other purposes, may not be required for predicting their evolutionary response to ongoing rapid climate change. However, recovering the evolutionary response of a wider spectrum of species would be needed to establish a reliable typology of demographic histories and their effects on genetic diversity.

Methods

Sampling

We sampled seven tree species of considerable economic and/or ecological importance in Europe: *B. pendula*, *F. sylvatica*, *P. nigra*, *P. abies*, *P. pinaster*, *P. sylvestris* and *Q. petraea*. For each species, we sampled a minimum of 20 populations (Tables S1–S2) from across the species' natural distribution ranges (Fig. S1). Sampling was carried out within the framework of the EU Horizon 2020 project GenTree. The majority of the sampled populations are the same as reported in Opgenoorth et al.[22], with additional samples reported in Table S2. Sampling principles and details are described in Opgenoorth et al.[22]. We dried the samples with silica gel and collated them in a single lab per species, where we extracted DNA (see Table S10 for details). We eluted the DNA in water, quality-checked it by UV spectrophotometry, and treated it with RNAse. We then sent all DNA extracts to IGA-TS (Udine, Italy) for targeted sequencing.

Sequencing and SNP calling

We focused on a limited part of the genome (3 Mbp) using targeted sequencing. The targeted regions consisted of orthologous genes involved in putative functions of interest, species-specific candidate genes, and randomly selected genes (Table S11). We first established a list of 2,639 genes involved in functions of interest (e.g. response to stress, immune response, circadian clock, detection of abiotic stimulus) in *Arabidopsis thaliana* using a term search in Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Table S12). For each species independently, we then identified putative orthologs to those 2,639 genes using a reciprocal best hit approach based on protein sequences (Blastp, BLAST v.2.5.0+). Finally, we used Orthofinder v.1.1.4[54] on the complete set of putative orthologs to define orthogroups across the seven species. The *Quercus robur* reference genome was used to identify orthologs for *Q. petraea*, a closely related species with which it often hybridizes. The *Populus trichocarpa* reference genome was used for *P. nigra*. We selected 811 'best orthogroups' that included at least one gene for at least six of the seven tree species (Table S13). For some species, we also included a variable number of other orthogroups including genes that were found only in a reduced set of species (e.g., 59 additional genes across the three conifer species; Table S14). For each species, we then completed the ortholog list with genes of interest and randomly selected genes to reach up to 6 Mbp of sequence to serve as a template for probe definition. Starting from the 6 Mbp of sequence mentioned above, Roche designed a set of uniquely mapping probes based on either a reference genome or coding sequence (CDS) data coordinates (Table S11), relying on its custom probe design pipeline (454 Life Sciences, a Roche company, Branford, CT, USA) including the SSAHA algorithm[55]. We then restricted candidate probes to cover 3 Mbp of sequence, prioritizing probes covering best-ortholog genes (Table S15).

To estimate the quality of genomic DNA, we quantified random samples from each 96-well plate using a Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, CA, USA) and a NanoDrop 1000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). We quantified all 4,754 samples using the GloMax Explorer System (Promega Corporation, Madison, WI, USA) and prepared libraries for target enrichment using the SeqCap EZ – HyperPlus kit (Roche Sequencing Solutions, Pleasanton, CA, USA) with 100 ng/µl of input DNA, following the manufacturer's instructions. For *P. abies* we conducted a second round of library preparation using 200 ng/µL of input DNA. We evaluated library size using the Bioanalyzer High Sensitivity DNA assay (Agilent Technologies, Santa Clara, CA, USA) and quantified libraries using a Qubit 2.0 Fluorometer. We sequenced the libraries on a HiSeq 2500 (125 cycles per read) for *P. abies* and *B. pendula* and on a NovaSeq 6000 (Illumina, San Diego, CA, USA; 150 cycles per read) for

the remaining species, in both cases working in paired-end mode. We used Illumina bcl2fastq v.2.20 for base calling and demultiplexing, and we used ERNE v.1.4.6[56] and Cutadapt[57] for quality and adapter trimming, both with default parameters.

For mapping we used BWA mem v.0.7.17[58] and samtools v.1.7[59] against the available reference genome of the same or closely related species, adding the mitochondrial and chloroplast genomes when those were missing from the reference (Tables S16–S17, see [link] for exact commands). In brief, we removed reads mapping on the organellar genomes, multiple mapping reads using samtools, marked duplicates with Picard and removed them. We maintained only a relevant portion of the genome with sufficient depth (> 5n, where n = sample size) for the next step to speed up the SNP calling stage, which can be computationally demanding for large, fragmented genomes.

We performed SNP calling using the software package GATK v.4.0.10.0[56]. Briefly, we ran *HaplotypeCaller* in GVCF mode to call potential variant sites at the single-sample level, then used *GenomicsDBImport* and *GenotypeGVCFs* to perform joint genotyping on the entire cohort of samples. Command lines are available at [link].

QC and SNP filtering

We conducted initial quality control of the SNP data using principal component analysis (PCA) and ADMIXTURE. We removed non-desired samples, e.g. misidentified species or obvious hybrids. We also removed samples with excessive amounts of missing data (60% for all species except *B. pendula* where threshold was 20%) or extreme values of heterozygosity (6% of the calls heterozygous) for *P. nigra* and *B. pendula*). Jupyter notebooks of initial quality control are available at [link].

We retained bi-allelic variants and followed GATK recommendations to exclude poorly supported SNPs with scores QD < 0.25, QUAL < 20, SOR > 3.0, MQ < 30, MQRankSum < -12.5 and/or ReadPosRankSum < -8.0. Filtering code is available at [link]. We identified putative false SNPs derived from paralogs mapping to a single location in the reference genome. We based the identification on heterozygote excess (H > 0.6) and deviation from the expected read ratio (D < -20 or D > 20) using the HDplot method [link][60].

We further utilized the location information of each SNP to delineate regions with an especially large proportion of paralog-derived SNPs

$$\frac{\text{number of paralog SNPs}}{\text{number of SNPs}} > 10\% \text{ within 250 bp}$$

and excluded all additional polymorphic positions included in those regions ([link]).

In addition to the variant-level filtering described above, we applied genotype-level filtering. We reported genotype calls with depth (DP) < 8 or genotype quality (GQ) < 20 as missing data to ensure high-quality genotypes[61], and we filtered out SNPs with > 50% missing calls to produce the v.5.3.2 VCF files.

To estimate the size of the genome with sufficient quality and depth for each species ('available genome', Table S17), we applied the same filtering procedure to the monomorphic positions, removing positions with DP < 8 and GQ < 20. Additionally, we excluded the same areas enriched with paralogs. The exact limit of those areas was defined at mid-distance between the last paralog and the next retained SNP.

Site-based annotation (4-fold degenerate, 2-3-fold degenerate and 0-fold degenerate sites) of detected SNPs was completed using the python script NewAnnotateRef.py available at [link].
SNPs were classified as 'intergenic', 'intron', 'stop', 'up' and 'down' using ANNOVAR[62] (Tables S18 and S19).

SFS scaling

Inference of demographic history can be done using site frequency spectra (SFS)[28,63]. The SFS must be estimated on a sample of a given size for all sites, however, in real datasets sample size varies among sites due to missing data. We therefore used the SNP set v.5.3.2 (see Supplementary Methods), removed SNPs with > 50% of missing data in any population and down-sampled the SFS for each population or subset used for demographic inference (see demographic analyses section for details) to half the initial sample size. The source code for resampling the SFS is available at [link]. The SFSs produced were then used in downstream analyses (see below).

Population genetic structure and isolation-by-distance

To characterize the main genetic clusters among populations, we used SNP from dataset v.6.3.2 with only putatively neutral SNPs (i.e., 0-fold degenerate sites or located in introns or intergenic regions), pruned for SNPs in high linkage-disequilibrium and excluding singletons (plink v.1.9.), to compute ancestry components (Q score) for each individual using ADMIXTURE v.1.3 (*50*). We performed an unsupervised analysis of individual ancestry proportions based on maximum likelihood and implemented 20 replicates for each K value (1–12) to assess cross-validation errors. Then, we averaged the Q scores of individuals for each population and visualized the geographic distribution of the genetic groups identified in ADMIXTURE using the 'raster' R package v.3.1.5[64].

We calculated genetic distances with dataset v.6.3.1 based on pairwise F_{ST} values[65], implemented with the *stamppFst* function of the 'StAMPP' R package v.1.6.3[66]. We then performed a hierarchical analysis of molecular variance (AMOVA), using the *poppr.amova* function of the 'poppr' R package v.2.9.3[67], to assess the partitioning of genetic variation: (i) between populations, (ii) between individuals within populations, and (iii) within individuals. We tested levels of significance using the *randtest* function of 'poppr'.

We quantified the pattern of isolation-by-distance (IBD) with dataset v.6.3.1 by regressing the genetic distance between pairs of populations (F_{ST}[65]) over the natural logarithm of the geographic distance between populations, following Rousset's[68] approach for a two-dimensional stepping-stone model.

$$\frac{\widehat{F_{ST_i}}}{\left(1-\widehat{F_{ST_i}}\right)} = \beta \ln(x_i) + \alpha + \varepsilon_i$$

For each pair of populations *i*, we estimated F_{ST} using vcftools (v.0.1.13)[69]. x_i is the geodesic distance separating the pair of populations in km ([70] 'geosphere' R package v.1.5-10), β is the slope of the regression, α is the intercept, and ε is the error term. The slope term is inversely proportional to the product of population density D by the second moment of dispersal distance, i.e. $\beta = 1/4\pi D\sigma^2$. This estimate of $4\pi D\sigma^2$ (from 1/b) can be interpreted as neighborhood size (Ns), with individuals within a neighborhood mating randomly.

Principal component analysis

We first conducted principal component analysis (PCA) using EIGENSOFT with default parameters (v.7.2.0, [link]), for each species independently, to obtain a general picture of the main population structure using all populations included in the dataset v5.3 while keeping only 4-fold, intronic and intergenic SNPs pruned for high linkage disequilibrium. The whole procedure was repeated using dataset v.6.3.1, but removing the most divergent populations to visualize more subtle population structure.

Scaling F_{ST} and computing population-specific F_{ST}

To investigate the change in population differentiation along latitude, longitude and elevation, we computed the average population-specific F_{ST} scaled by geographic distance. For each focal population, we computed the average pairwise F_{ST} divided by the average distance (km) between each pair of populations in which the focal population was involved.

Measuring genetic diversity

We estimated the pairwise nucleotide diversity, π, for 4-fold (π_4) and 0-fold sites (π_0) as well as for silent sites (π_s, comprising intergenic, intronic and 4-fold sites). Based on the allele frequencies obtained from the SFS after down-sampling per species and per population based on SNP set v.5.3.2 (Table S20), we estimated the expected heterozygosity for each polymorphic site and summed the resulting values over all segregating sites. To obtain a π value per site, we divided this sum by the total number of sites, including monomorphic sites:

$$\pi = \frac{n}{L(n-1)} \sum_{i=1}^{S} \left(1 - \sum_{j=1}^{n} p_{ji}^2 \right)$$

where *n* is the sample size, i.e. the number of allele copies, *S* is the number of segregating sites, p_{ji} is the allele frequency at a polymorphic site *i* for the jth allele, and *L* is the total number of sites (polymorphic and monomorphic) in a class. Nucleotide diversity was estimated from and averaged over 1,000 resampled replicates using [link] and [link].

Demographic modeling

To infer past changes in the effective population size (N_e), we applied two approaches based on the SFS: model-based estimates using fastsimcoal2[63] and model-free estimates with Stairway Plot 2 v.2.1.1[28]. For both approaches, we used folded and rescaled SFS (see above) based on 4-fold, intergenic and intronic sites (SNP set v5.3.2. without 0-fold sites). For the estimates using

Stairway Plot 2, we applied different hierarchical levels and subsamples. First, we made stairway plots by pooling all the samples together by species, to maximize the power to detect relatively recent demographic events by using large sample sizes[71]. Considering that the SFS resulted from pooling all the samples together, different populations and especially population genetic clusters were not always equally represented, due to unequal missing data across populations and unequal representation (in terms of populations and individuals) of different clusters. To investigate the impact of hierarchical population clustering on the stairway plots, we reran Stairway Plot 2 at different hierarchical levels: (i) one sample per population (to account for unequal representation of populations) and (ii) separately for each population.

For the four broad-leaved species, we used the mutation rate of 7.77×10^{-9} per site per generation[72] estimated for *Prunus*, which is close to estimates of the mutation rate per generation of 7×10^{-9} for *A. thaliana*[73] and *Silene latifolia*[74]. To scale estimates of timing from generations to years, we assumed a generation time of 60 years for *Q. petraea* and *F. sylvatica*, and 15 years for *P. nigra* and *B. pendula*. For the three conifers, we used a mutation rate of 2.7×10^{-8} per site per generation estimated based on somatic mutations of *Picea sitchensis*[75]. Assuming a 25-year generation time, this value is similar to estimates obtained from fossil calibration[17,76]. We ran all stairway plot estimates using 67% of the sites for training and 200 resamplings from the SFS. We calculated breakpoints following the suggestions in the stairway plot manual, i.e. at n/4, n/2, n*¾ and n-2 with n indicating the sample size.

To confirm the results obtained with Stairway Plot 2, we used fastsimcoal2[25,63] to infer the past demographic history of each species. We used the same folded SFSs as for Stairway Plot 2. We explored three single population models (Fig. S24): an equilibrium model (standard neutral model, SNM), a model with one demographic change (epoch-2) and a model with two demographic changes (epoch-3). For all parameters (N_e at each step and each time of events, specifically: NCUR = most recent N_e; NANC = ancestral N_e; NBOT = N_e after the first demographic event in the 3-epoch model; TBOT = time of the first demographic event; TENDBOT: time of the second demographic event in the 3-epoch model), we set a log prior with a range of 10 to 10^7. We first inferred the best model by running 100 independent runs for each of the three models using fastsimcoal2 with 10^6 simulations (-n 1,000,000), a minimum of 10 conditional maximization algorithm (ECM) cycles for likelihood computations (-l 10), 40 ECM cycles for parameter estimations (-L 40), and a minimum of one allele count for parameter estimation (--minSFScount 1; default value). For each model, we then identified the run with the best likelihood score. We computed the AIC score for each run, and we compared the AIC values of different models. Finally, following Excoffier et al.[63], we computed confidence intervals for each parameter, selecting the parameter values of the best run for both 2-epoch and 3-epoch models to simulate 100 SFS using fastsimcoal2. We then ran, for each of those 100 SFSs, 100 independent runs of fastsimcoal2 using the same settings as before, obtaining 100 sets of parameters for each model, which we used to compute confidence intervals for each model parameter. We used this same approach for each species, using the SFS computed over all samples and the SFS computed using only one haplotype per population.

To further test if more realistic models would better fit the observed data and impact our inference of past demographic changes, we tested four models of demographic changes including events of divergence. For these models, we subset, for each species, samples from two locations, a southern and a northern one (we used admixture results to make sure that the samples came from two distinct gene pools; Figs. S3–S16). From these sets, we explored the same three

models, with no divergence events and pooling all samples together regardless of origin, and four models of divergence differing regarding whether migration was possible after the divergence and including or not a demographic change before the divergence event. We used fastsimcoal2 to assess models, using the folded SFS of all samples pooled together for reference models, and the folded joint SFS of the population pairs for models of divergence. We conducted the simulations and inference with fastsimcoal2 by running 100 independent runs with 100,000 simulations each (-n 100,000), the same number of ECM cycles as used previously (-l 10 -L 40) and in folded mode (-m). We explored all parameters (including all or some of the following for the divergence models: NCUR = N_e before the split but after the demographic event; NANC = ancestral N_e; NPOP1 = N_e after the split for population 1; NPOP2 = N_e after the split for population 2; TDIV = time of the split; TSEP = time of the demographic event; N1M21 = effective migration rate from population 1 to population 2; N2M12 = effective migration rate from population 2 to population 1; and the same parameters as above for the panmictic models) with a logunif prior between 10 and 10^9, except for the parameters related to the number of effective migrants (the product between the migration rate and the effective population size, $N_e \times$ m), which used logunif priors set between 10^{-4} and 10^4. Finally, we compared the seven models using the likelihood and AIC score of the single best run out of 100, to identify the best model for each species and its associated parameters.

Synchronicity analysis

We then assessed whether the dynamics of changes in N_e over time were species-specific or synchronous across species. To do so, we first computed Kendall's correlation coefficients between the output of Stairway Plot 2 for each pair of species and then investigated their covariance across the seven species using a heatmap (heatmap3 R package, v.1.1.9 [link] with default parameters, Fig. 3C). This global approach allowed us to quantify the synchronous pattern in change in N_e over time between some species and test whether it was primarily driven by the global increase in N_e over the period studied. This approach captures the main trend but it will not allow the identification of specific periods of high synchronicity between several species. In particular, it will miss decreases in N_e that could be expected to be associated with glacial periods. To specifically test whether periods over which the species experienced a decrease in N_e showed higher synchronicity than expected given the actual change in N_e over time for the various species, we used a randomization approach. This analysis was conducted independently for the two groups of species showing the highest synchronicity, *F. sylvatica* and *Q. petraea* on the one hand, and *P. abies*, *B. pendula*, *P. sylvestris* and *P. nigra*, on the other hand.

More specifically, the Stairway Plot 2 output consists of a succession of intervals, defined by a given N_e estimate and two different time points. To simplify this, for each species independently, each interval was represented by a pair of values: the unique N_e value characteristic of the interval and the midpoint of the two time points. We excluded the most recent time point of the output of Stairway Plot 2 as it does not correspond to a proper step. As the time points at which Stairway Plot 2 provides estimates of N_e are species-specific, we inferred N_e at every time point in the joint dataset by considering for each species independently the value of the closest N_e estimated by Stairway Plot 2. To smooth random fluctuations between time points and to mitigate the effects associated with small deviations to our estimates of generation time and mutation rate, we averaged the changes in N_e between two time points ($DN_e = N_{e(t)} - N_{e(t+1)}$) over

250 consecutive time points, using sliding windows (mDN_e). Finally, we used a randomization approach to detect periods of time where the pattern of synchronicity in N_e change is stronger than what would be expected given species-specific change in N_e over time; *i.e.* periods over which a larger number of species experienced a decrease in N_e than expected if changes in N_e were independently distributed across time in each species. For each species independently, we first randomized the vector of DN_e values and averaged the randomized values over 250 consecutive time points using sliding windows (mDN_e), as we did for the observed data. From the randomized time series, we recorded the maximum number of species experiencing a decrease simultaneously, as well as the longest span over which synchronicity was conserved (*i.e.* the maximum number of consecutive positive DN_e). We repeated the whole procedure 10,000 times and compared the observed values with the 95% percentile of the distribution of the maximum values obtained through the 10,000 simulations (Table S9 and Fig. S28).

References and Notes:

1. Birks, H. J. B. & W, T. in *European Atlas of Forest Tree Species* (eds San-Miguel-Ayanz, J., de Rigo, D., Caudullo, G., Houston Durrant, T. & Mauri, A.) e010c45+ (Publication Office of the European Union, Luxembourg, 2016).
2. Lascoux, M., Palmé, A. E., Cheddadi, R. & Latta, R. G. Impact of Ice Ages on the genetic structure of trees and shrubs. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **359**, 197-207 (2004).
3. Kremer, A. How well can existing forests withstand climate change. *EUFORGEN Climate Change and Forest Genetic Diversity: implications for sustainable forest management in Europe, Paris, France, 15-16 March 2006.* 3-17 (2007).
4. Saleh, D. et al. Genome-wide evolutionary response of European oaks during the Anthropocene. *Evolution Letters* **6**, 4-20 (2022).
5. Alberto, F. J. et al. Potential for evolutionary responses to climate change - evidence from tree populations. *Global Change Biology* **19**, 1645-1661 (2013).
6. Li, L. et al. Teasing apart the joint effect of demography and natural selection in the birth of a contact zone. *New Phytologist* **236**, 1976-1987 (2022).
7. Waples, R. S. What is Ne, anyway. *Journal of Heredity* **113**, 371-379 (2022).
8. Bai, W. et al. Demographically idiosyncratic responses to climate change and rapid Pleistocene diversification of the walnut genus *Juglans* (Juglandaceae) revealed by whole genome sequences. *New Phytologist* **217**, 1726-1736 (2018).
9. Helmstetter, A. J., Béthune, K., Kamdem, N. G., Sonké, B. & Couvreur, T. L. P. Individualistic evolutionary responses of Central African rain forest plants to Pleistocene climatic fluctuations. *Proceedings of the National Academy of Sciences* **117**, 32509-32518 (2020).
10. Taberlet, P., Fumagalli, L., Wust-Saucy, A. & Cosson, J. Comparative phylogeography and postglacial colonization routes in Europe. *Molecular Ecology* **7**, 453-464 (1998).
11. Petit, R. J. et al. Glacial refugia: Hotspots but not melting pots of genetic diversity. *Science* **300**, 1563-1565 (2003).
12. Ballard, J. W. O. & Whitlock, M. C. The incomplete natural history of mitochondria. *Molecular ecology* **13**, 729-744 (2004).
13. Doyle, J. J. Defining coalescent genes: theory meets practice in organelle phylogenomics. *Systematic biology* **71**, 476-489 (2022).
14. Liu, S. et al. Demographic history and natural selection shape patterns of deleterious mutation load and barriers to introgression across Populus genome. *Molecular Biology and Evolution* **39**, msac008 (2022).
15. Olsson, S. et al. Evolutionary history of the mediterranean Pinus halepensis-brutia species complex using gene-resequencing and transcriptomic approaches. *Plant Molecular Biology* **106**, 367-380 (2021).
16. Mayol, M. et al. A multiscale approach to detect selection in nonmodel tree species: Widespread adaptation despite population decline in *Taxus baccata* L. *Evolutionary Applications* **13**, 143-160 (2020).
17. 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).
18. Hou, Z. & Li, A. Population genomics reveals demographic history and genomic differentiation of *Populus davidiana* and *Populus tremula*. *Frontiers in Plant Science* **11**, 553736 (2020).

19. Salojärvi, J. et al. Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch. *Nature Genetics* **49**, 904-912 (2017).
20. Cai, M., Wen, Y., Uchiyama, K., Onuma, Y. & Tsumura, Y. Population genetic diversity and structure of ancient tree populations of *Cryptomeria japonica* var. *sinensis* based on RAD-seq data. *Forests* **11**, 1192 (2020).
21. Capblancq, T. et al. Whole□exome sequencing reveals a long□term decline in effective population size of red spruce (*Picea rubens*). *Evolutionary Applications* **13**, 2190-2205 (2020).
22. Opgenoorth, L. et al. The GenTree Platform: growth traits and tree-level environmental data in 12 European forest tree species. *GigaScience* **10**, giab010 (2021).
23. Chen, J., Glémin, S. & Lascoux, M. Genetic diversity and the efficacy of purifying selection across plant and animal species. *Molecular Biology and Evolution* **34**, 1417-1428 (2017).
24. Eckert, C. G., Samis, K. E. & Lougheed, S. C. Genetic variation across species' geographical ranges: the central–marginal hypothesis and beyond. *Molecular Ecology* **17**, 1170-1188 (2008).
25. Excoffier, L. & Foll, M. Fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* **27**, 1332-1334 (2011).
26. Excoffier, L. et al. fastsimcoal2: demographic inference under complex evolutionary scenarios. *Bioinformatics* **37**, 4882-4885 (2021).
27. Hirao, A. S. et al. Genetic diversity within populations of an arctic-alpine species declines with decreasing latitude across the Northern Hemisphere. *Journal of Biogeography* **44**, 2740-2751 (2017).
28. Liu, X. & Fu, Y.-X. Stairway Plot 2: demographic history inference with folded SNP frequency spectra. *Genome Biology* **21**, (2020).
29. Chikhi, L., Sousa, V. C., Luisi, P., Goossens, B. & Beaumont, M. A. The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. *Genetics* **186**, 983-995 (2010).
30. Wakeley, J. & Aliacar, N. Gene genealogies in a metapopulation. *Genetics* **159**, 893-905 (2001).
31. Wakeley, J. Nonequilibrium migration in human history. *Genetics* **153**, 1863-1871 (1999).
32. Petit, R. J. & Hampe, A. Some evolutionary consequences of being a tree. *Annu. Rev. Ecol. Evol. Syst.* **37**, 187-214 (2006).
33. Pyhäjärvi, T., Kujala, S. T. & Savolainen, O. 275 years of forestry meets genomics in *Pinus sylvestris*. *Evolutionary Applications* **13**, 11-30 (2020).
34. Palstra, F. P. & Fraser, D. J. Effective/census population size ratio estimation: a compendium and appraisal. *Ecology and Evolution* **2**, 2357-2365 (2012).
35. 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).
36. Lapierre, M., Lambert, A. & Achaz, G. Accuracy of demographic inferences from the site frequency spectrum: The case of the Yoruba population. *Genetics* **206**, 439-449 (2017).
37. Nawa, N. & Tajima, F. Simple method for analyzing the pattern of DNA polymorphism and its application to SNP data of human. *Genes and Genetic Systems* **83**, 353-360 (2008).
38. Achaz, G. Frequency spectrum neutrality tests: one for all and all for one. *Genetics* **183**, 249-258 (2009).

39. Jaramillo-Correa, J. P. et al. Evolutionary rate and genetic load in an emblematic Mediterranean tree following an ancient and prolonged population collapse. *Molecular Ecology* **29**, 4797-4811 (2020).
40. Dial, R. J., Maher, C. T., Hewitt, R. E. & Sullivan, P. F. Sufficient conditions for rapid range expansion of a boreal conifer. *Nature* **608**, 546-551 (2022).
41. James, J. et al. Between but Not Within-Species Variation in the Distribution of Fitness Effects. *Molecular Biology and Evolution* **40**, msad228 (2023).
42. Birks, H. J. B. Contributions of Quaternary botany to modern ecology and biogeography. *Plant Ecology & Diversity* **12**, 189-385 (2019).
43. Petit, R. J. et al. Identification of refugia and post-glacial colonisation routes of European white oaks based on chloroplast DNA and fossil pollen evidence. *Forest ecology and management* **156**, 49-74 (2002).
44. Magri, D. et al. A new scenario for the Quaternary history of European beech populations: palaeobotanical evidence and genetic consequences. *New Phytologist* **171**, 199-221 (2006).
45. Shalev, T. J. et al. The western redcedar genome reveals low genetic diversity in a self-compatible conifer. *Genome Research* **32**, 1952-1964 (2022).
46. Gugerli, F. et al. A range□wide postglacial history of Swiss stone pine based on molecular markers and palaeoecological evidence. *Journal of Biogeography* **50**, 1049-1062 (2023).
47. Whitham, T. G., Morrow, P. A. & Potts, B. M. Conservation of hybrid plants. *Science* **254**, 5033 (1991).
48. Fu, R. et al. Genome-wide analyses of introgression between two sympatric Asian oak species. *Nature Ecology & Evolution* **6**, 924-935 (2022).
49. Petit, R. J., Bodénès, C., Ducousso, A., Roussel, G. & Kremer, A. Hybridization as a mechanism of invasion in oaks. *New Phytologist* **161**, 151-164 (2004).
50. Suarez-Gonzalez, A., Lexer, C. & Cronk, Q. C. B. Adaptive introgression: A plant perspective. *Biology Letters* **14**, (2018).
51. Leroy, T. et al. Extensive recent secondary contacts between four European white oak species. *New Phytologist* **214**, 865-878 (2017).
52. Magri, D. Patterns of post□glacial spread and the extent of glacial refugia of European beech (*Fagus sylvatica*). *Journal of Biogeography* **35**, 450-463 (2008).
53. Grivet, D. et al. High rate of adaptive evolution in two widespread European pines. *Molecular Ecology* **26**, 6857-6870 (2017).
54. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* **16**, 1-14 (2015).
55. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Research* **11**, 1725-1729 (2001).
56. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297-1303 (2010).
57. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10-12 (2011).
58. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997* (2013).
59. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
60. McKinney, G. J., Waples, R. K., Seeb, L. W. & Seeb, J. E. Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. *Molecular Ecology Resources* **17**, 656-669 (2017).

61. Carson, A. R. et al. Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. *BMC Bioinformatics* **15**, 1-15 (2014).
62. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **38**, e164-e164 (2010).
63. 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).
64. Alexander, D. H. & Lange, K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* **12**, 1-6 (2011).
65. Weir, B. S. & Cockerham, C. C. Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **38**, 1358-1370 (1984).
66. Pembleton, L. W., Cogan, N. O. I. & Forster, J. W. St AMPP: An R package for calculation of genetic differentiation and structure of mixed□ploidy level populations. *Molecular Ecology Resources* **13**, 946-952 (2013).
67. Kamvar, Z. N., Tabima, J. F. & Grünwald, N. J. Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* **2**, e281 (2014).
68. Rousset, F. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics* **145**, 1219-1228 (1997).
69. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011).
70. Hijmans R. J. Spherical Trigonometry, R package geosphere version 1.5-10. (2019).
71. Keinan, A. & Clark, A. G. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **336**, 740-743 (2012).
72. Xie, Z. et al. Mutation rate analysis via parent–progeny sequencing of the perennial peach. I. A low rate in woody perennials and a higher mutagenicity in hybrids. *Proceedings of the Royal Society B: Biological Sciences* **283**, 20161016 (2016).
73. Ossowski, S. et al. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327**, 92-94 (2010).
74. Krasovec, M., Chester, M., Ridout, K. & Filatov, D. A. The mutation rate and the age of the sex chromosomes in *Silene latifolia*. *Current Biology* **28**, 1832-1838. e4 (2018).
75. Hanlon, V. C. T., Otto, S. P. & Aitken, S. N. Somatic mutations substantially increase the per-generation mutation rate in the conifer *Picea sitchensis*. *Evolution Letters* **3**, 348-358 (2019).
76. 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).

Acknowledgments: The computations were enabled by resources in project SNIC 2017/7-328, SNIC 2020/15-107 and SNIC 2021/5-540 provided by the Swedish National Infrastructure for Computing (SNIC) at UPPMAX, partially funded by the Swedish Research Council through grant agreement no. 2018-05973. CSC – IT Center for Science, Finland, Genotoul Bioinformatics platform Toulouse Occitanie and the Genetic Diversity Centre at ETH Zurich, Switzerland, are acknowledged for generous computational and storage resources, and GBFOR, INRAE, Forest Genetics and Biomass Facility and Natural Resources Institute Finland (LUKE) for sample collection.

Funding:

European Union's Horizon 2020 Research and Innovation Programme grant agreement no. 676876

Academy of Finland grant nr. 287431 (TP)

Swiss Secretariat for Education, Research and Innovation (SERI) contract no. 6.0032 (FG)

Spanish Ministry of Agriculture, Fisheries and Food (MAPAMA) grant no. AEG 17-048 (DG)

European Agricultural Fund for Rural Development (EAFRD) (DG)

Swedish Research Council for Sustainable Development (FORMAS) grant nos. 2016-00780, 2020-01456 (ML)

Spanish Ministry of Economy and Competitiveness (MINECO) contract n. PTA2015-10836-I (SO)

LIFEGENMON (LIFE13 ENV/SI/000148) (MW)

The Slovenian Research Agency, research core funding no. P4-0107 (MW)

Biocenter Oulu

Author contributions:

Conceptualization: StC, SCGM, DG, ML, PM, TP, GGV

Methodology: KB, PFR, SCGM, CK, ML, PM, SP, TP, CR, SS

Software: KB, BD, CK, SP, TP, SS, IS

Validation: SaC, VJ, PM, TP, SS

Formal Analysis: FB, KB, SaC, BD, PFR, SCGM, DIO, VJ, CK, ILK, PM, LO, SP, CP, TP, CR, OR, IS

Investigation: FB, SCGM, DG, FG, VJ, ML, ILK, PM, SO, LO, CP, TP, CR, IS, MW

Resources: BD, PFR, SCGM, DG, FG, ML, SO, LO, CP, CR, IS,

Data curation: BD, SaC, VJ, CK, PM, LO, TP

Visualization: BD, CK, PM, IS, MW

Funding acquisition: StC, BF, SCGM, DG, FG, ML, SO, LO, TP, GGV, MW

Project administration: StC, BF, SCGM, FG, ML, LO, TP, GGV

Supervision: FB, SCGM, FG, VJ, ML, LO, CP, TP, CR, GGV, MW

Writing – original draft: KB, SaC, BD, SCGM, VJ, CK, ML, PM, LO, SP, TP, CR, OR, IS

Writing – review & editing: FB, KB, BD, BF, PFR, SCGM, FG, VJ, CK, ML, ILK, PM, LO, CR, SS, IS, GGV, MW

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: Short reads are available under NCBI BioProjects: PRJNA602465, PRJNA602466, PRJNA602467, PRJNA602468, PRJNA602470, PRJNA602471, PRJNA602473. Vcf, ped and map -files area available at https://doi.org/10.57745/DV2X0M. Code is available at: https://doi.org/10.5281/zenodo.7943876.

Table 1. Biological characteristics and genetic summary statistics for seven European tree species. Age information was retrieved from the European atlas of tree species https://forest.jrc.ec.europa.eu/en/european-atlas/ and the European forest genetic resources program (EUFORGEN).

Species	**Biological characteristics**	**Hybridization with**	**Min. flowering age (years)**	**Max. age known (years)**	F_{ST}	π_4 **(per bp)**	**Ns (individuals)**
Betula pendula (Silver birch)	deciduous temperate to boreal wind pollination wind seed dispersal large and continuous range	*B. platyphylla* *B. pubescens*	10–25	90–150	0.03*	0.0036	235
Fagus sylvatica (European beech)	deciduous temperate wind pollination animal seed dispersal large and continuous range	*F. orientalis*	40–50	150–300	0.05*	0.0050	65
Populus nigra (Black poplar)	deciduous Mediterranean to temperate wind pollination water and wind seed dispersal vegetative and sexual reproduction intermediate and discontinuous range (riparian)	*P. nigra* 'Italica' *P. deltoides* *P. trichocarpa* *P. maximowiczii* *Populus* sp. hybrid cultivars	4–10	100–400	0.16*	0.0032	8
Quercus petraea (Sessile oak)	deciduous temperate wind pollination animal seed dispersal large and continuous range	*Q. robur* *Q. pubescens* *Q. pyrenaica*	40–100	>1000	0.04*	0.0072	115
Picea abies	conifer temperate to boreal wind pollination	*P. obovata*	20–40	200–300	0.06*	0.0048	30

(Norway spruce)	wind seed dispersal large and discontinuous range						
Pinus pinaster (Maritime pine)	conifer Mediterranean wind pollination wind seed dispersal limited and discontinuous range		6–20	120–250	0.13*	0.0027	17
Pinus sylvestris (Scots pine)	conifer temperate to boreal wind pollination wind seed dispersal large and continuous range	*P. mugo* *P. uliginosa*	15–30	400–750	0.01*	0.0039	392

Notes: mean pairwise genetic differentiation (F_{ST}), * denotes significantly differentiated populations based on AMOVA ($P < 0.05$); pairwise nucleotide diversity per site at four-fold sites (π_4); neighborhood size (Ns)