Population genomics of temperate forest trees

Gabriele Nocchi

School of Biological and Behavioural Sciences,

Queen Mary University of London,

Mile End Road,

London E1 4NS

Supervisor: Professor Richard J. A. Buggs

A thesis submitted to

Queen Mary University of London

for the degree of Doctor of Philosophy

May 2022

Declaration

I, Gabriele Nocchi, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below, and my contribution indicated. Previously published material is also acknowledged below. I attest that I have exercised reasonable care to ensure that the work is original and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material. I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis. I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university. The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

I acknowledge collaboration as follows:

- In the oak project (Chapter 2) Nathan Brown sampled and phenotyped trees, Tim Coker extracted DNA, William Plumb extracted DNA, Jonathan Stocks extracted DNA, Sandra Denman selected the sites and organised the sampling, Richard Buggs obtained funding, and oversaw the project. I performed SNP calling, all analyses and wrote the chapter.
- In the Asian white birch project (Chapter 3), Nian Wang performed collection of leaf material, DNA extraction, arranged and funded sequencing, and performed SNP calling on both the China and Eurasian birch datasets. Richard Buggs oversaw the project. I performed all analyses and wrote the chapter.

Signature: Gabriele Nocchi **Date:** 25/05/2022

Publications

The research presented in this thesis has directly contributed to the following publications:

- Nocchi, G., Brown, N., Coker, T., Plumb, W., Stocks, J., Denman, S., & Buggs, R. (2022). Genomic structure and diversity of oak populations in British parklands. *Plants, People, Planet*, *4* (2), 167-181.
- Gathercole, L., Nocchi, G., Brown, N., Coker, T., Plumb, W., Stocks, J., Nichols, R., Denman, S., Buggs, R. (2021). Evidence for the widespread occurrence of bacteria implicated in acute oak decline from incidental genetic sampling. *Forests*, 12, 1683.

Funding

This thesis is part of a PhD programme run jointly between Queen Mary University of London and the Royal Botanic Gardens, Kew funded by the UK Government Department for Environment, Food and Rural Affairs (Defra), in association with Action Oak. Oak sequencing was funded by Defra under the Future Proofing Plant Health scheme. Asian White Birch sequencing was funded by the National Natural Science Foundation of China (31770230 and 31600295) to Nian Wang.

Acknowledgements

I thank Forest Research and its Technical Service Unit for sending us oak leaf samples and metadata. I thank Joan Cottrell (Forest Research) for sharing with us haplotype data of 163 British ancient oak woodlands. I thank Richard Nichols for his expert advice on statistical analyses. I thank Alberto Carmagnini (Queen Mary University of London, Laurent Frantz’s Lab) for his numerous constructive comments on bioinformatics analyses. I thank Stefan Götz, David Seide and the BioBam

Bioinformatics team for kindly letting us use their software OmicsBox for the functional annotation and functional enrichment analysis of the identified birch adaptive genes. I thank Nian Wang for allowing me to analyse the birch sequencing data that he collected and generated. I thank all woodland owners, managers and staff that allowed us and our collaborators to access and sample trees at their sites. This research utilised Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT: http://doi.org/10.5281/zenodo.438045.

Lastly, I would like to thank my grandmother, Natalina, and my father, Andrea, for supporting me and motivating me throughout the last four years.

Abstract

The current geographic distribution and genetic structure of plant species have been greatly affected by the Quaternary climatic oscillations. The alternating glacial and interglacial periods resulted in cycles of populations expansion and contraction that have provided opportunities for contacts and hybridization between advancing populations, as well as for differentiation due to long-lasting geographic isolation and/or adaptation to different environments.

Here, I first focused on the genetic differentiation, structure and the patterns of hybridization between the native British oak species *Q. robur* and *Q. petraea*, using whole genome sequence data of 386 oak individuals sampled across four managed British parklands environments. Additionally, I reconstructed these species demographic histories in view of paleoclimatic events and searched the *Q. robur* genome for signatures of recent positive selection. I then explored chloroplast DNA variation and compared it with data from other studies of oaks in ancient woodlands across Europe, to assess whether the parklands populations sampled derive from native and local seed stock.

Secondly, I explored the genetic structure and diversity of Asian white Birch (*B. platyphylla)* in China and its hybridization with the closely related silver birch, *B. pendula*, across Eurasia, based on whole genome sequencing data of 83 *B. platyphylla* individuals sampled across 74 natural Chinese populations and *B. pendula* whole sequencing data from a previous Eurasian study. I developed ecological niche models to define the present habitat and predict the future distribution of this species in China. Furthermore, I explored the genetic structure and evolution of *B. platyphylla* Chinese populations and performed genome-environment association (GEA) analyses to detect signals of local adaptation in China. I used the identified adaptive genetic loci to assess the degree of maladaptation of these populations to future climatic conditions and highlight regions of China where this species may be particularly threatened by climate change.

Contents

List of figures and tables

Chapter 1: General Introduction

Summary

The climatic fluctuations of the Quaternary period have had a profound impact on the distribution and evolution of temperate plant and animal species. In this introduction I first describe the Quaternary climate and the role of glacial refugia for the survival of plant species during harsh climatic episodes. I discuss the geography of Europe and the Sino-Japanese floristic region at the last glacial maximum and explore theories on the location of glacial refugia. Secondly, I discuss the genetic consequences of the Quaternary glacial cycles on plant and more specifically tree species, such as hybridization, introgression and divergence. Thirdly, I present findings about ongoing climate change, based on the report of the recent sixth intergovernmental panel on climate change. I discuss how plant species might be able to cope with the rapidly changing climate, through migration/habitat tracking and in situ via local adaptation, and I briefly comment the conservation strategies available to aid adaptation. Finally, I provide a broad introduction to the study systems used in this thesis, the British native oak species *Q. robur* and *Q. petraea*, and Asian White Birch (*B. platyphylla*). First, I describe the morphological and ecological characteristics of these species and their phylogeny. Then, I discuss relevant genomics research related to hybridization and divergence, demographic history, post-glacial migration and local environment adaptation. Furthermore, I provide a brief parenthesis on Acute Oak decline, an oak syndrome of particular concern in Britain during the last decade. I conclude with a description of the research questions that I aimed to address.

Europe and Asia during the Quaternary

The current geographic distribution and genetic structure of plant and animal species in Europe and Asia have been greatly influenced by the Quaternary glacial cycles, which started between 2.9 and

2.4 M years ago with the establishment of the Arctic ice cap (Raymo, 1994). These cycles consisted in cold long-lasting glacial periods characterized by southward advancing ice sheets followed by warmer and shorter interglacial periods (Hewitt, 1999; Hewitt, 2000). Carbon and oxygen isotopes, pollen and other types of biological signatures recovered from the bottom of seas, lakes and ice sheets have been used to reconstruct the paleoclimatic history of the planet (Hewitt, 2000; Raymo, 1994). From the onset of the Quaternary (approx. 2.4 Ma) until 0.9 M years ago, glacial-interglacial cycles lasted approximately 40,000 years, while thereafter they have increased in severity to longer 100,000 years cycles (Berger & Loutre, 2010; Hewitt, 2000). Historically glacial periods have been between seven and nine times longer than interglacials (Berger & Loutre, 2010; Hewitt, 1999).

Glacial refugia were critically important for the survival of plants and animals' species during the Quaternary glaciations: these were unglaciated habitats which allowed the growth and reproduction of species during adverse environmental conditions, providing less severe and ice-free environments (Birks & Willis, 2008). Refugia were also the sources of re-colonization during interglacial periods, when environmental conditions became more favourable (Birks & Willis, 2008). Three main types of glacial refugia have been described in the literature: nunataks or cryptic refugia, which are located at higher altitudes on mountain peaks that protrude above the icesheet; peripheral refugia, which are located at the border of a mountain system; and low-land refugia which are located outside mountains systems beyond the limits of the ice sheet (Holderegger & Thiel-Egenter, 2009; Provan & Bennet, 2008). When the icesheets retreated, species started the recolonization of previously inhospitable and glaciated areas from the refugia.

In Europe, studies based on fossil pollen and chloroplast DNA variation, suggest that temperate plant species survived the Quaternary glaciations mainly in glacial refugia located at lower latitudes, principally in three areas: the Iberian Peninsula, Italy and the Balkans (Birks & Willis, 2010; Hewitt, 1999; Petit et al., 2002a; Petit et al., 2002b; Provan & Bennett, 2008). The term last glacial maximum (LGM) refers to the coldest period of the last glaciation when ice sheets reached their maximum extent (18,000 – 25,000 years ago). Vast areas of northern Europe were covered by ice sheets at the LGM, which extended through Poland and Germany, with additional glaciers on mountains at lower latitudes, such as in the Alps (Clark et al., 2009; Comes and Kadereit, 1998). Permafrost extended from south of the ice sheet as far as modern-day southern Hungary (Clark et al., 2009). Following deglaciation, it is thought that plants species re-colonized Europe by migrating north-ward as the climate got warmer, mostly from these three major refugial areas (Hewitt, 1999; Petit et al., 2002a; Petit et al., 2002b), and reached the current distribution roughly 6,000 years ago (Brewer et al., 2002; Hewitt, 1996). However, this model of expansion-contraction of latitudinal range change characterized by southward retreat into refugia and subsequent north-ward post-glacial re-expansion does not entirely explain species demography during the Quaternary glacial cycles (Qiu et al., 2011). In addition to the three major southern European refugia, a strong argument has been built for the role of cryptic refugia (or nunataks), which are micro-refugia located at higher latitudes which provided pockets of favourable environmental conditions for species survival (Provan & Bennett, 2008). Therefore, the current species distribution in Europe seems a result of both major long-distance migration and dispersal from the southern refugia, as well as minor local recolonization from cryptic refugia at higher latitudes (Bhagwat & Willis, 2008; Provan & Bennett, 2008).

On the other side of the world, the “Sino-Japanese floristic region” (SJFR), which harbours the world’s most diverse temperate flora, has never been covered by extensive ice sheets such as Europe at the LGM (Shi et al., 1986; Liu, 1988), nonetheless the Quaternary climatic fluctuations still had a strong impact in shaping the distribution of plants and animals’ species in this region (Qiu et al., 2011). Moreover, changes in topography, altitudes, river drainage patterns and strong fluctuations in sea levels between China, Korea and Japan, provided opportunities for changes in species distribution, population fragmentation as well as hybridization (Qiu et al., 2011). During the Quaternary glaciations, ice sheet developed only locally or at high altitudes in the SJFR and permafrost extended as far south as Beijing (Clark et al., 2009), making southwestern and subtropical China important refugial areas for many temperate plant species throughout the Quaternary (Qiu et al., 2011). Therefore, the scenario in Eastern Asia, and particularly modern-day China, was different than that in Europe, and the current distribution of species appears to reflect multiple short postglacial re-colonization paths from local refugia, rather than a major long-distance northward migration such as that thought to have taken place in Europe (Bao et al., 2015; Chen & Lou, 2019; Qiu et al., 2011).

The theoretical genetic consequences of the Quaternary climatic oscillation on plant species

The theoretical genetic consequences of the Quaternary glaciations on plant species have been described by Hewitt (1996). He suggested that populations near the refugia should harbour higher genetic diversity as there is a tendency for loss of alleles and increased homozygosity due to bottlenecks occurring during population expansion. These populations might also be genetically differentiated from each other in order to suit different selective pressures imposed by different environments. Therefore, in Europe more diversity should be maintained in the stably present refugial populations in the south, which persisted through glacial cycles by moving up and down mountain systems hence experiencing fewer steep drops in population sizes (Hewitt, 1996; Petit et al., 2003). These patterns of genetic diversity have been shown that can result from both long-distance dispersal as well as continuous-range expansion (Hallatschek, 2007; Hewitt, 1996; Ibrahim et al., 1996). Once the climate warmed populations from different refugia could meet and mix as they recolonised vacant areas, in some cases forming hybrid zones. These have interesting properties, as reported by Hewitt (1996): first they can be characterized by some sort of reduction in fitness, which determines the width and extent of the hybrid zone itself; secondly, hybrid zones may remain confined to area of low dispersal until major climatic changes occur, therefore hybrid zones might help maintain the genetic diversity and integrity of populations and/or species (Hewitt, 1996). Interestingly, a study based on chloroplast DNA diversity of 22 woody species found higher genetic diversity at mid-latitudes in Europe, probably due to admixture of different lineages migrating from separate lower latitude refugia and converging in central Europe (Petit et al., 2003), but this could also be due to the presence of cryptic refugia at higher latitudes, which complicates the interpretation of the population structure of some species.

Speciation, differentiation and hybridization

The repeated cycles of contraction/expansion of the geographical ranges of species characteristic of the Quaternary have profoundly shaped the genetic structure of plants (Comes & Kadereit, 1998), and have led to opportunities for speciation by long-lasting geographic isolation and/or exposure to different selective pressures, as well as hybridization and introgression among several species and differentiated populations of the same species (Comes and Kadereit, 1998; Goczal et al., 2020; Hewitt, 1996; Hewitt, 1999; Hewitt, 2000).

Allozyme and chloroplast DNA studies have looked at the genetic diversity between and within plant species and have inferred inter and intra specific divergence times. Furthermore, the post-glacial colonization routes of several species were reconstructed, coupling molecular data with fossil pollen maps (Comes & Kadereit, 1998). The Pleistocene-Holocene boundary appears to be a cluster point for the intraspecific divergence of many temperate tree species, as well as the speciation of several short-lived herbaceous plant species (Comes & Kadereit, 1998). Speciation is the process by which a new species is created and occurs when a population separates from other members of its species and becomes reproductively isolated. This can be caused by different selective pressures imposed by different environments; therefore, the modes of speciation have been traditionally classified on a geographical context (Butlin et al., 2008). There are four main types of speciation: allopatric, peripatric, parapatric, and sympatric, as described in Coyne & Orr (2004). Allopatric speciation occurs when two populations are separated by a physical barrier, such as mountains, rivers, or lakes, which act as barriers to gene-flow causing the two populations to develop differently, due to genetic drift and possibly different natural selection forces. Peripatric speciation is similar to allopatric and is characterized by the presence of a physical barrier to gene-flow, however the main difference is that in peripatric speciation a significantly smaller group breaks-off from the ancestral population and forms a new species. Parapatric speciation occurs when a species occupies a large heterogenous geographic area, such as two adjacent environmental niches. Even in the absence of a physical barrier, members of an area tend to mate more frequently with those that occupy the same niche, however an hybrid zone usually arises between the two diverging populations. Subsequently, the hybrid zone acts as a barrier between the diverging populations, until it gets eliminated due to selective disadvantage, which completes the speciation process. Sympatric speciation occurs when there are no physical barriers to gene flow and all members of the ancestral population are in proximity to each other and a new species develops spontaneously from some of its members.

On the other hand, introgression refers to the incorporation, via hybridization and repeated backcrossing, of alleles from one species or population into the gene pool of another (Anderson, 1949; Harrison & Larson, 2014). Hybridization can be both beneficial and detrimental for a species. It can bring adverse consequences, and in the most severe cases lead to extinction, by reducing a species reproductive effectiveness, its competitive status and its interactions with pathogens and other animals, particularly in rare species with small population sizes (Levin et al., 1996). On the other hand, hybridization can also be favourable through adaptive introgression, by increasing genetic diversity and can be a key driver to speciation (Becker et al., 2013; Hamilton & Miller, 2015; Brennan et al., 2014; Mallet, 2015; Thomas, 2015). Hamilton & Miller (2015) suggested that introgression may increase a species ability to respond to rapid climate changes or changing environment, by improving the adaptive potential that would be given by mutations alone, particularly in cases of limited standing genetic variation and between populations exhibiting strong ecotypic differences. However, in some cases hybridization results in higher fitness which can lead to the extinction of one or both parental species/populations through a process called genetic assimilation, or in cases where one of the segregated populations has a significantly smaller population size and outbreeding is common, this population may be replaced by hybrids through a process known as genetic swamping (Hamilton & Miller, 2015; Levin et al., 1996). Another similar process is demographic swamping, which occurs when hybridization results in reduction of fitness (outbreeding depression), but hybridization is common and exceeds populations growth rate, then one of the parental populations may decline below replacement rate (Todesco et al., 2016). Hybridization can therefore maintain diversity in cases where the hybrid zone remains stable, alternatively diversity could be decreased through the breakdown of reproductive barriers and the merging of distinct evolutionary lineages (Todesco et al., 2016). Interestingly, Beatty et al. (2010) assessed hybridization at the limits of the species geographic distribution range in the genus *Pyrola* in Canada and showed that unidirectional hybridization may lead to the extinction of peripheral populations of *P. minor* through genetic assimilation via hybridization with the more abundant *P. grandiflora*, which in turn would limit the species ability to respond to climate change through habitat tracking (Beatty et al., 2010).

Climate change and trees adaptation

Human activities have resulted in a significant increase in well-mixed greenhouse gases concentrations since around 1750s, which has warmed the atmosphere, ocean, cryosphere and biosphere (Masson-Delmotte et al., 2021). Each of the last four decades has been warmer than any decade that preceded it since 1850, and currently the global surface temperature is 0.99 °C higher than in 1850-1900 (Masson-Delmotte et al., 2021). Globally precipitation over land has increased since 1950, the upper ocean temperature has warmed, arctic ice has reached its lowest level since the 1850, and sea levels raised of 0.20 m in the last century (Masson-Delmotte et al., 2021). Evidence of these changes is provided by the greater frequency and intensity of extreme climatic events such as heatwaves, heavy precipitation, droughts, tropical cyclones, accompanied by the decreased frequency of cold extreme events (Masson-Delmotte et al., 2021). The sixth report of the intergovernmental panel on climate change (Masson-Delmotte et al., 2021) warned that global land temperature will continue to increase until at least mid-century, under all emission and socioeconomic pathways tested, thus a rise of 1.5 – 2 °C might be exceeded by 2100 unless CO_2 and greenhouse gas emission are drastically reduced. Under the worst emission scenario, global surface temperature might rise by 3.3 to 5.7 °C by 2100 (Masson-Delmotte et al., 2021). Precipitation is projected to increase over high latitudes in all polar, northern European and northern North American regions, most Asian regions and two regions of South America, but decrease in southern Europe and in parts of the tropics and sub-tropics (Masson-Delmotte et al., 2021).

In response to the projected climate changes, Eurasian temperate trees are expected to shift their geographic ranges northward and to higher elevations, or to adapt locally. Although adaptation to changing climate have been documented, it is not known whether this will be sufficient to prevent species extinctions and the extent to which populations will adapt will depend on many factors, such as phenotypic variation, standing genetic variation, strength of selection, fecundity, interspecific competition, and biotic interactions (Parmesan, 2006; Aitken et al., 2008). Aitken et al. (2008) hypothesized that there are three possible outcomes for forest tree populations in a rapidly changing climate: persistence through migration to track suitable habitats; persistence *in situ* through local adaptation; and extinction. There is compelling evidence of the capacity of forest trees to adapt rapidly to new environments, such as the development of genetic and phenological clines during post-glacial migration (Parmesan, 2006; Davis and Shaw, 2001). In addition, fossil pollen maps and molecular data, suggests capacity of strong geographic range shifts in forest trees during the Quaternary glacial and interglacial periods (Aitken et al., 2008; Davis & Shaw, 2001). Fossil pollen maps and chloroplast DNA analyses have been used to infer both the latitudinal and altitudinal migrations of trees out of the glacial refugia and showed that species' geographic ranges have shifted in correlation with the Quaternary climatic cycles (Davis & Shaw, 2001; Petit et al., 2004a). Species range expansion studies suggest an average migration rate of less than 100 m per year for boreal and temperate forest trees, considering both expansion from the northern edge of the species distribution range and from possible cryptic refugia located at higher latitudes (Aitken et al., 2008); rare long-distance dispersal event may increase this estimate however these are hard to quantify and were not included in the models (Aitken et al., 2008). As it stands, this speed of dispersal does not meet the dispersal requirement necessary for habitat tracking under the projected climate change scenarios (Aitken et al., 2008). In addition, the ability to track and migrate to suitable environments may be hindered by the presence of competing neighbouring species, which reinforce the importance of trees to locally adapt to some extent (Case & Taper, 2000).

Lynch & Lande (1993) examined a model of continuously changing conditions and found that a species can survive by maintaining a steady rate of adaptation as long as the required change does not exceed a threshold, which is determined by a combination of standing genetic variation, reproductive success, strength of selection, interspecific competition, random environmental disturbances, and population size. In addition, a follow-up study found that the risk of extinction due to climate change are much higher when the population size is small, due to the combined effects of genetic drift and environmental stochasticity (Burger & Lynch, 1995); while when population sizes are large and the species are characterized by high fecundity, as it is often the case with forest trees, populations are more likely to adapt and only suffer adaptational lag for a few generations (Aitken et al., 2008; Burger & Lynch, 1995; Lynch & Lande, 1993). The rate of adaptation is increased in species with short generation times and capacity of long-distance seed dispersal (Aitken et al., 2008). This model (Lynch & Lande, 1993) is very simplistic and assumes that fitness is determined by a single trait, while in nature fitness and local adaptation are determined by several traits, which may be more or less correlated, therefore the rate of adaptation might be slower than that estimated. In addition, gene-flow can have either a beneficial or detrimental effect on a species ability to adapt. Commonly, in nature, a species density is higher at the centre of its distribution and lower in the periphery (Aitken et al., 2008). Gene-flow from the centre of the distribution towards the periphery can inhibit the adaptation of peripheral populations (Garcia-Ramos & Kirkpatrick, 1997; Butlin et al., 2003). On the other side, gene-flow from the centre towards the periphery can aid adaptation when the effects of genetic drift are strong, and therefore gene-flow can increase genetic variation; however, in forest trees population sizes are generally large and the effects of drift are negligible (Aitken et al., 2008). Peripheral populations which fall at higher latitudes or elevation may actually benefit from gene-flow from the centre of the distribution located at lower latitude, as this may introduce allele pre-adapted to warmer conditions (Davis & Shaw, 2001). The opposite scenario is envisioned for peripheral populations located at lower latitudes compared to the centre of distribution, and gene-flow coming from the centre would be detrimental and aid maladaptation (Aitken et al., 2008). Interspecific gene-flow can also have similar effects (Aitken et al., 2008; Leroy et al., 2019a; Rieseberg et al., 2003).

During expansion, founder populations first rely on phenotypic plasticity, which is the ability to shift phenotype in response to environmental changes (Nicotra et al., 2010), to persist until their population size increases and sufficient genetic variation accumulates to support adaptation (Aitken et al., 2008; Petit et al., 2003). To avoid adaptational lag and mitigate the effects of the current climate change, numerous conservation strategies have been suggested. In the context of plant adaptation, two main approaches have been proposed: assisted migration and assisted gene-flow (Borrel et al., 2019; Prieto-Benitez et al., 2021). Assisted migration refers to the translocation of seeds of maladapted or threatened individuals/populations to a more suitable environment (Prieto-Benitez et al., 2021; Ricciardi & Simberloff, 2009; Hallfors et al., 2014), while assisted gene-flow is the movement of genetic material/gametes from a donor pre-adapted population to a population threatened by changing conditions (Aitken & Whitlock, 2013; Whiteley et al., 2015).

Genomics studies on local adaption of *A. thaliana*, the first plant species genome ever sequenced, identified trends across abiotic gradients such as temperature, elevation and precipitation (Agren et al., 2017; Exposito-Alonso et al., 2017; Johnson et al., 2021; Montesinos-Navarro et al., 2011), paving the way for extending this type of research to non-model species.

The strongest patterns of adaptation in temperate forest trees have been initially found in the synchronization of the cycles of growth and dormancy with local seasonal temperatures (Aitken et al., 2008), and several studies have found geographic clines for height growth, bud phenology, cold resilience and drought resistance in a variety of species (Campbell, 1979; St. Clair et al., 2005; Rehfeldt, 1995; Mimura & Aitken, 2007). In recent years, with the advances and reduced cost of DNA sequencing, the number of landscape genomics studies in forest trees has increased and genomic signals of local adaptions for many climatic variables have been detected in several species, including the study systems in this thesis, oaks and birches (Jordan et al., 2007; Lovell et al., 2021; Pina-Martins et al., 2008; Rellstab et al., 2016; Borrel et al., 2019; Martins et al., 2018).

Oaks

Oaks are wind-pollinated temperate broadleaved trees widely distributed in the northern hemisphere, found from boreal to tropical latitudes (Kremer et al., 2012). Oaks are of huge economic, ecological and cultural value and have provided valuable resources to humans such as wood for fire, acorns for livestock and timber for construction since early history (Kremer et al., 2012; Eaton et al., 2016).

Oaks are part of the *Fagaceae*, which is a large angiosperm plant family including over 900 species classified in eight-ten genera; the three prevalent genera are *Quercus* (Oak), *Castanea* (Chestnut) and *Fagus* (Beech) (Kremer et al., 2012). The genus *Quercus* has been the focus of many phylogenetics studies with the first infrageneric classification in ten groups based on leaf and reproductive characteristics dated 1838 (Loudon, 1838; Denk et al., 2017). Before the use of molecular data in phylogenetics oak classifications tended to be discordant due to the different weights assigned to the morphological features considered and undetected homoplasy, which is the development of similar traits across different species as a consequence of convergent evolution or incomplete lineage sorting (Denk et al., 2017). Loudon classification (Loudon, 1838) established the subdivision of European oaks (*cerris*, *ilex* and *robur*), which although slightly modified, has remained largely unchanged in all later phylogenies. More recently, the classification proposed for the genus *Quercus* by Denk et al. (2017), based on both molecular and morphological evidence, has identified two major clades: a larger clade, referred to as the New World oaks, which includes group *Protobalanus* (golden cup oak), group *Lobatae* (red oaks), group *Ponticae*, group *Virentes* and group *Quercus* (white oaks); and a smaller clade, predominantly Eurasian, which includes group *Ilex*, group *Cerris* and group *Cyclobalanopsis* (cycle-cup oaks). This classification is similar to that of Loudon (1838), however some of the original sections were merged into sections *Lobatae* and *Ponticae* (Denk et al., 2017). An important discriminating characteristic used in this classification (Denk et al., 2017) is pollen sculpturing and ultrastructure, which was found to be highly conserved within groups and was often overlooked in previous studies (Denk et al., 2017). The latest time-calibrated oak phylogeny, based on both fossil data and restriction-site associated DNA sequencing (RAD-seq), was inferred by Hipp et al. (2019). This shows that oaks have first differentiated among continents, and subsequently have diverged ecologically within geographic regions (Hipp et al., 2019).

Quercus robur* and *Quercus petraea

Quercus robur (English or pedunculate oak) and *Quercus petraea* (sessile oak) are part of the white oaks group and are the most common and widespread species in Europe, found from Central Spain to the Urals and from Scandinavia to the South of Italy (Barreneche et al., 1998; Eaton et al., 2016). These species are native to Britain and constitute a great study system to investigate the delineation of species. As stated by Kleinschmit (1993), European oak species are not strict biological species: different species of the same taxonomic group, such as *Q. robur* and *Q. petraea*, hybridize naturally giving rise to intermediate forms exhibiting high levels of variation (Eaton et al, 2016; Kremer et al., 2012; Kleinschmit, 1993).

Individuals from these two species live long, from 100s to over 1000 years in some cases, and usually reach height of 30 m and trunk diameter up to 1 m, however there are reports of larger individuals (Jones, 1959). The major morphological differences between *Q. robur* and *Q. petraea* are usually in the leaves, acorns, terminal buds, bark and trunk, given in order of importance as reported by Jones (1959). Leaves in *Q. robur* are narrow at the base, wider well above the middle and are irregularly lobed with very short petioles while in *Q. petraea* leaves are more ovate and regularly lobed with longer petioles. Acorns in *Q. robur* are usually large and oblong with a long stalk, while in *Q. petraea* acorns are smaller and rounder and have a very short or absent stalk. The terminal buds in *Q. robur* are usually small and obtuse while in *Q. petraea* are usually large and acute. The bark has a rectangular blocks architecture in both species but in *Q. petraea* this is thinner and tends to exfoliate. The trunk of *Q. robur* tends to disappear in the crown and boughs are irregularly branched while in *Q. petraea* the trunk usually persists through the crown and boughs tend to be straighter (Jones, 1959). In addition, leaf hair is another important discriminating characteristic in distinguishing the two species: *Q. robur* is hairless on the lamina while *Q. petraea* has stellate hair (Kremer et al., 2002; Rellstab et al., 2016b). *Q. robur* and *Q. petraea* have wide ecological ranges and are often found in sympatry however the two species exhibit different preferences: *Q. robur* prefers to grow on natural, moist and alkaline soils and is more tolerant to flooding and waterlogs while *Q. petraea* tends to prefer lighter, well-drained and acidic soils and is more drought tolerant (Eaton et al., 2016; Saintagne et al., 2004; Barreneche et al., 1998).

Oak's genomics research

Oak genomics research kicked off around two decades ago with the construction of a genetic linkage map based on random amplification of polymorphic DNA (RAPD), sequence characterized amplified regions (SCAR), microsatellite, minisatellite, isozymes and 5S rDNA markers in 1998 (Barreneche et al., 1998). This was the first linkage map constructed for the genus *Quercus* and the first linkage map in the *Fagaceae* plant family (Barreneche et al., 1998). Barreneche et al. (1998) genotyped and analysed the mendelian segregation of a set of markers in a progeny of 94 individuals, including markers derived from previous investigations and newly generated RAPD fragments (Barreneche et al., 1998). The markers were subdivided in four groups depending on the segregation patterns observed in the progeny and were used to re-construct two separate parental maps, each represented by 12 linkage groups. The linkage map constructed in this study was essential for much research that followed: molecular markers and linkage maps are the pre-requisite for comparative genomics studies, quantitative trait loci (QTL) identification and provide an important backbone for genome assemblies (Saintagne et al., 2003).

A comparative mapping between *Quercus* and *Castanea* by Barreneche et al. (2004) relied on the newly constructed oak linkage map (Barreneche et al., 1998) to assess molecular markers transferability in the *Fagaceae*. This study was motivated by research which revealed synteny and QTL conservation between many cereals and between some forest trees (Barreneche et al., 2004); furthermore, the construction of linkage maps for *Quercus robur* and *Castanea sativa* reported similar estimated genome size and identical number of linkage groups between the two genera (Barreneche et al., 1998; Casaoli et al, 2001). Barreneche et al. (2004) mapped microsatellite repeats to the available *Q. robur* and *C. sativa* linkage maps to compare their respective locations. Two species specific sets of microsatellite markers were constructed; the set constructed in *Q. robur* was tested for amplification in *C. sativa* and vice versa. The markers which exhibited cross-amplification were genotyped in two progenies (one per genera) of 94 individuals each and those that showed mendelian segregation were mapped to the available linkage maps (Barreneche et al., 1998; Cassaoli et al., 2001). Nineteen markers were mapped in both species matching nine linkage groups, and interestingly the order of the markers was also conserved (Barreneche et al, 2004). The nucleotide sequences of the matched markers were compared revealing high sequence identity at the microsatellite flanking region and significant conservation of the repeat region between *Q. robur* and

C. sativa, thus suggesting orthology of these markers and confirming the potential of microsatellites in comparative analysis in the *Fagaceae* family (Barreneche et al, 2004).

Species differentiation between *Q. robur* and *Q. petraea* has also been investigated with the aid of microsatellites and RAPD based linkage maps. Initial efforts, such as the investigation by Zanetto et al. (1994), failed to identify any species-specific marker. As previously mentioned, *Q. robur* and *Q. petraea* are interfertile therefore their genetic boundary is thin, and it is difficult to detect variation when comparing a relatively small number of markers; the study by Bodénès et al. (1997) confirmed this hypothesis by assessing the molecular diversity between these two species through the screening of 2800 RAPD fragments in natural populations reporting very low nucleotide divergence (~0.5% overall). Overall, initial differentiation studies based on small numbers of loci (RAPDs, SCARs, Isozymes, SSRs and AFLPs) showed extremely low levels of differentiation between English and sessile oak (Barreneche et al., 1996; Bodenes et al., 1997; Coart et al., 2002; Mariette et al., 2002; Saintagne et al., 2004; Zanetto et al., 1994).

A study that successfully detected species-specific regions in *Q. robur* and *Q. petraea* was the investigation by Saintagne et al. (2004) which aimed to uncover QTLs involved in 15 leaf morphological differences. QTLs for 13 of the 15 investigated traits were identified using an expanded version of the previously developed linkage map (Barreneche et al, 1998) and five showed significant variation between the two species. The QTLs identified for these traits were found distributed in clusters located on multiple linkage groups. Similar studies relied on the oak linkage map (Barreneche et al., 1998) to identify QTLs controlling oak adaptive traits, such as water use efficiency and tolerance to water logging (Brendel et al., 2007; Parelle et al., 2007). Furthermore, Guichoux et al. (2011) successfully developed a set of 20 microsatellites markers which allowed discrimination between *Q. robur* and *Q. petraea* and facilitated the identification of hybrids.

After the first wave of RAPD and microsatellite-based linkage maps and related QTL research, considerable efforts were devoted to developing a denser oak linkage map based on expressed sequence tags (EST) (Bodénès et al., 2012). Expressed sequence tags are short and randomly selected sequences derived from cDNA libraries reverse transcribed from mRNA; they are used for a variety of tasks such as gene discovery, gene structure identification, single nucleotide polymorphism (SNP) profiling and are also a cheap and fast alternative to generate DNA markers such as simple sequence repeats (SSRs) (Nagaraj, Gasser and Ranganathan, 2006; Bodénès et al., 2012). Bodénès et al. (2012) mined EST collections available for *Q. robur* and *Q. petraea* to identify EST-SSRs markers and 255 were assigned to chromosomal locations using a selective binning approach (Bodénès et al., 2012). These markers constitute an advantage over genomic SSRs: EST-SSRs are reverse transcribed from coding sequences therefore represent expressed DNA and can be used to investigate gene variation directly, which is ideal in comparative genomics, QTL and genetic diversity studies (Bodénès et al., 2012). ESTs were also used to construct the oak unigene set, which will be pivotal for the study of the oak transcriptome, the identification of improved molecular markers as well as the mapping and annotation of the oak genome (Ueno et al., 2010).

Ueno et al. (2010) constructed the first oak unigene set by de novo assembly of EST sequences derived from Sanger sequencing and pyrosequencing. Following annotation, the oak transcriptome showed high levels of homology with other species (such as *Vitis vinifera*) and candidate genes for important traits such as bud phenology, cuticle formation and cell wall formation were identified with similarity searches (Ueno et al., 2010). The oak unigene set was further expanded by the work of Lesur et al. (2015) which also investigated the regulation of genes involved in bud dormancy release, an important adaptive trait strongly dependant on temperature and therefore very likely to be affected by the ongoing climate change (Lesur et al., 2015).

The final contribution towards the construction of a dense oak linkage map came with Lepoittevin et al. (2015) who developed a single nucleotide polymorphism (SNP) assay for *Q. robur* and *Q. petraea* (Bodénès et al., 2016; Lepoittevin et al., 2015). Lepoittevin et al. (2015) identified a set of 7,913 SNPs in six parental trees (three *Q. robur*, three *Q. petraea*) derived from mining available EST sequence data and the recently developed oak unigene set (Lesur et al., 2015). The SNPs were genotyped in intra and interspecific progenies and in three natural populations from the southwest of France; ~80% were recovered in the progenies, ~55% in the natural population and high levels of shared SNPs (91%) were recorded between *Q. robur* and *Q. petraea*.

The SNPs characterized by Lepoittevin et al. (2015) were pivotal for the construction of the densest oak linkage map by Bodénès et al. in 2016, which would be later used to guide the oak genome assembly (Plomion et al., 2016; Plomion et al., 2018). For the construction of the SNPs-based linkage map, Bodénès et al. (2016) crossed six parental trees, three of species *Q. robur* and three of species *Q. petraea* to generate four families: an intraspecific family for *Q. robur* and *Q. petraea* respectively and two interspecific families. The 7,913 SNPs identified in the study by Lepoittevin et al. (2015) were genotyped in the four families and parental maps were re-constructed by analysing the segregation of SNPs. The resulting eight parental maps were joined in a composite map of 4,261 SNPs covering 742 cM of the 12 oak chromosomes: 1 SNP marker per 0.2 cM, considerably denser than the earlier EST-SSRs based map (Bodénès et al., 2012).

The dense SNP-based oak linkage map (Bodénès et al., 2016) and the improved oak unigene set (Lesur et al., 2015) constituted essential resources for the oak genome project, which was carried by the French consortium led by Christopher Plomion and officially started in 2012 (Plomion et al., 2016). The French group sequenced (whole shotgun sequencing) the genome of a 100 years-old *Q. robur* tree at the INRA Pierroton forestry research station in France, combining different technologies (Illumina, Roche 454 and ABI-capillary sequencing). The assembly was completed in 2018 with the construction of a highly contiguous haploid sequence composed of 1,409 scaffolds (N50 of 1,343 kb). 871 of these scaffolds, covering 96% (716.6 Mb) of the estimated physical size of the oak genome (736 Mb/1C - 740 Mb/1C, estimated with k-mer analysis and flow-cytometry respectively) and including 90% of the 25,808 predicted protein coding genes, were mapped to the 12 oak chromosomes in the dense SNP-based oak linkage map (Bodénès et al. in 2016; Plomion et al, 2018). Further genome analysis revealed that 52% of the *Q. robur* genome is composed of transposable elements (TE), most of whom are class I retrotransposons. The genetic diversity was assessed at both SNP and nucleotide level in 20 trees reporting high levels of population diversity and mutations in oaks when compared to other plants and animals (Plomion et al, 2018).

Plomion et al. (2018) also reconstructed the paleo-history of oak in the *Rosid* clade, revealing that five fissions and 14 fusions events led to the modern 12 chromosomes configuration from an initial 21 chromosomes architecture. Another interesting finding of this research was the identification of 524 orthologous groups which seem to have expanded in oaks. Following annotation, the identified expanded groups showed to be associated with genes involved in the immune response, such as R genes (Plomion et al, 2018). By comparing these findings with the data available for other long-living trees, such as eucalyptus, a similar expansion of disease-resistant genes was observed suggesting that the expansion of these gene families may contribute to the long survival of some species (Plomion et al, 2018). The sequencing and public release of the *Q. robur* genome was followed by the draft assemblies of two other *Quercus* species: *Q. lobata* (valley oak) and *Q. suber* (cork oak) (Sork et al., 2016; Ramos et al., 2018). Overall, the availability of reference genome sequences for species of the genus *Quercus* coupled with the advances and decreased cost of sequencing, have enabled to better understand the function of oaks in the environment and has opened and have facilitated numerous research opportunities in topics such as the identification of genes that control adaptive and important traits, the identification of genes that modulate the interaction with other organisms and the study of the speciation and hybridization in this very diverse and widely distributed genus.

***Q. robur* and *Q. petraea* genomics research in recent years**

The last decade has also seen an increase in the studies aimed at uncovering the differentiation between *Q. robur* and *Q. petraea* using dense SNPs markers sets. Guichoux et al. (2012) assessed the interspecific differentiation between *Q. robur* and *Q. petraea* in south-west of France by genotyping 856 individuals scattered across six forests and identified between 13 and 75 outlier SNPs. This study showed that introgression occurs asymmetrically, predominantly from the resident species, *Q. robur* (early successional), to the invading species, *Q. petraea* (late successional). This finding has been validated further by other studies (Jurksiene et al., 2020; Lang et al., 2018; Lepais et al., 2013; Leroy et al., 2019a).

Lang et al. (2018) recently characterized over 10,000 SNPs in *Q. robur* and *Q. petraea* and used them to assess the diversity and the patterns of introgression between these two species in central and western Europe, focusing on genic regions. This study highlighted the low-mean differentiation (F_{st}) between these two species, and the relatively high levels (for the plant kingdom) of genetic diversity within both species, which is consistent with their longevity, reproductive success and long history of hybridization and introgression. Genetic diversity was found higher in *Q. petraea* rather than *Q. robur,* despite the study included significantly less sessile oaks (Lang et al., 2018). This pattern seems to be explained by their different modes of colonizing new stands, with Q. *petraea* being the invading species as previously discussed (Guichoux et al., 2012).

Lesur et al. (2018) developed a targeted captured-based next generation assay to characterize about 190,000 SNPs in 303 *Q. robur* and *Q. petraea* individuals from a mixed oak strand in the southwest of France. This investigation aimed to delineate species and hybrids and calculate relatedness and inbreeding within this stand. The SNP assay developed revealed high transferability across the white oaks group and constitutes an important resource for comparative studies among related white oaks species and populations, and also helps reducing the cost for large population genetic studies allowing to genotype a smaller set of SNPs over carefully selected target regions and therefore include more individuals.

In the context of the ongoing climate change, Rellstab et al. (2016) genotyped more than 3,500 SNPs in 71 populations including *Q. robur, Q. petraea* and *Q. pubescens* to identify loci associated with environmental factors and then calculated the risk of non-adaptedness (RONA) to future conditions for these populations in Switzerland. This study revealed wide variability among species in regard to the environmentally associated SNPs and suggested that some of allele frequency changes at environmentally associated loci required to match future climate are substantial and unlikely to be achieved through standing genetic variation alone, given the long generation time of oaks.

More recently Leroy et al. (2019b) identified a huge set of over 30 million SNPs in populations of four oak species (*Q, robur, Q. petraea, Q. pubescens* and *Q. pyrenaica*) using a pool-sequencing approach, and used it to assess their differentiation, and their demographic and divergence history. This study suggested that extensive gene-flow among these four species throughout the last 20,000 years, coupled with pre and post zygotic selection (Lepais et al., 2013), have cleared most species-specific genetic structures except those at barrier loci creating a very diverse genetic landscape. Nevertheless, some genomic regions exhibit extremely high-levels of interspecific differentiation, which may be due to strong selection counteracting the effects of the interspecific gene-flow.

More recently Reutimann et al. (2020) developed a very condensed SNP marker set (58 SNPs) which allows to reliably differentiate between *Q. robur, Q. petraea* and *Q. pubescens* and to assess their degree of admixture in Swiss populations. This study found that the level of admixture is much higher between *Q. petraea* and *Q. pubescens* rather than between the other pairs.

Oak threats and diseases in Britain: acute oak decline (AOD)

Oaks interact with a huge number of species during their long survival, mainly bacteria, fungi and insects but also amphibians, reptiles, birds and mammals; in turn these communities interact with each other forming complex forest ecosystems (Plomion and Fievet, 2013; Tarkka et al., 2013). Some of these interactions are favourable to oaks, such as the formation of ectomycorrhizas (EMs) between the roots and some fungi, which facilitates the absorption of nutrients (Tarkka et al., 2013; Richard et al., 2005); other interactions have negative effects and are usually in the form of infections by bacteria or parasites, which are responsible for the deteriorating oak health observed in recent years.

Episodes of deteriorating oak health have been reported in Britain since the early 1900 and have been labelled with the term dieback or decline (Denman and Webber, 2009; Denman et al., 2014). The definition of decline diseases states that these are complex tree syndromes caused by a combination of biotic and abiotic factors which compromise the host health (Denman et al., 2014; Brown et al., 2016). An acute decline is characterized by the sudden episodic appearance of symptoms over a period of 5-10 years with high tree mortality until the disease slowly loses intensity and may even disappear; on the other side a chronic decline is characterized by low mortality and slow symptoms development, which may persist for decades and never disappear (Denman and Webber, 2009). Chronic Oak Decline is often caused by root related problems such as fungal infections while the acute form is usually characterized by mildew/defoliation which results in the progressive thinning of the canopy (Denman and Webber, 2009; Brown et al., 2016). The current British AOD episode, is characterized by stem damage: fluid filled lesions develop between bark plates (Denman and Webber, 2009). The oak parklands sampled in this thesis have been monitored for over a decade for this complex multifactorial syndrome. A disease with similar symptoms to the current British AOD syndrome is Sudden Oak Death (SOD), which has killed many native American oaks in California in the 1990s. This disease was characterized by stem bleeding and leaf necrosis and its causal agent was identified as *Phytophthora ramorum* (Denman et al., 2009). In Britain the only infection by *Phytophthora ramorum* reported were on beech and red oak; European oak species such as *Q. robur* and *Q. petraea* have shown higher tolerance to this pathogen than American oaks (Denman et al., 2009).

The current AOD episode was first reported in 1980s in East England; since then, the disease has spread to the midlands and further into Wales, with the most affected area being the southeast (Brown et al., 2016; Brady et al., 2017). A detailed description of the symptoms of the current AOD epidemic was published in 2014 (Denman et al., 2014), and the main characteristics of AOD described in this publication are reported in the next paragraph.

The main external evidence of AOD are stem bleeds; these are patches that develop on the tree trunk outer bark, usually at 1-5 m above the ground. The bleeds are usually found at distances of 5-20 cm from each other and are denser and more numerous in severely affected trees; these are active in the spring and autumn months when the fluid in the cavities is dark and translucent. The fluid penetrates in the inner bark and through several ring of sapwood causing evident internal necrosis in the form of dark patches; however, lesions show a gradual loss of intensity from the outer bark going into the sapwood and usually do not reach the heartwood. The bark of infected trees is characterized by cracks of 3-22 cm in length, which develop following the decay of the underlying tissue. Signs of infection by *Agrilus biguttatus* are usually present underneath the bark of infected trees and these can't be detected by external observation until the larvae have successfully colonized the tree and D-shaped exit holes are left on the bark. *A. biguttatus* larval galleries tend to avoid the necrotic patches and extend up into the sapwood. Some infected trees also show signs of healing with the formation of callus-like tissue around the cracks, which builds up until it occludes the lesions, pushing and loosening the outer bark plates causing some to fall, suggesting the presence of some sort of host resistance to AOD (Brown et al, 2016; Denman et al., 2014). In this study (Denman et al., 2014) no association between crown condition and AOD could be identified.

Although the cause of AOD mortality is very likely to be due to a combination of biotic and abiotic factors, significant efforts were devoted to understanding the polymicrobial cause of tissue necrosis and stem bleeds of affected British oak trees in recent years by Forest Research (Denman et al., 2016). First isolation of bacteria from stem lesions of symptomatic trees revealed that many of the most occurring microbes were either unknown or un-named species belonging to the families of *Enterobacteriaceae* and *Pseudomonadaceae* (Brady et al., 2017). 13 new species and two new genera were characterized with the most commonly isolated species being: *Brenneria goodwinii*, *Gibbsiella quercinecans*, *Lonsdalea quercina ssp. britannica*, *Rahnella victoriana* and *R. variigena* (Denman et al., 2012; Denman et al., 2016; Gathercole et al., 2021). On the other side attempts to isolate fungal pathogens from the bleeds have failed to identify the same species consistently, therefore fungi were excluded as possible causal agents of the current AOD outbreak (Brown et al, 2016).

A major advance in the understanding of the biotic components of AOD came in 2018 (Denman et al, 2018) when the microbiomes of healthy and AOD affected trees from several sites were analysed and compared to validate earlier findings and hypothesis using a systemic approach.

First bacteria were isolated from the trunks and cultured to reveal that the most commonly isolated bacteria from AOD lesions are of genera *Gibbsiella*, *Brenneria* and *Rahnella*. Interestingly, a significant co-occurrence between *G. quercinecans* and *B. goodwinii* was observed in affected trees and none of these two species was isolated from the sites with no AOD history. Following bacterial isolation, the trunk microbiomes of healthy and affected trees were sequenced. The most abundant genera in healthy trees found with metagenomics analysis were *Periglandula*, *Burkholderia*, *Streptomyces*, *Bacillus* and *Auriemonas*; on the other side *B. goodwinii*, *G. quercinecans* and *R. Victoriana* were detected in all the diseased trees microbiomes with *Brenneria* being the most abundant genus.

The genomes of *B. goodwinii*, *G. quercinecans* and *R. Victoriana* were sequenced revealing abundance in genes involved in carbohydrate metabolism, transport, virulence and disease thus confirming the necrotic capabilities of these bacteria. These findings were further confirmed by functional metagenomics and metatranscriptomics analysis of the symptomatic group microbiomes. Furthermore, the infection was reproduced by inoculating the three imputed causative bacteria into live oak logs, which confirmed the necrotic capabilities of two of them: *B. goodwinii* and *G. quercinecans*.

To conclude, the findings of this study (Denman et al., 2017) suggest that the decline characteristics typical of British oak trees affected by AOD are a consequence of vascular degradation initiated by *B. goodwinii* and *G. quercinecans* tissue necrosis, enhanced and spread by *A. biguttatus* galleries. This blocks carbon allocation, which accumulates in the roots negatively affecting water transport and availability (Denman et al., 2017). However, much of the effort has been devoted to the study of the disease, but little is known about the role of the host in AOD. Susceptibility to this syndrome may have a heritable genetic component, and if so, genomics could help to identify oak trees with increased tolerance to AOD that could be of great importance for forest planting and breeding programs.

Birches

Birches (genus *Betula*) are broad-leaved deciduous trees and shrubs belonging to the *Betulaceae* plant family, which includes five other genera: *Alnus* (Alder), *Carpinus* (Hornbeam), *Corylus* (Hazel), *Ostrya* (Hop Hornbeam) and *Ostryopsis* (Ashburner & McAllister, 2013; Wang et al., 2016). Birches are widely distributed in the northern hemisphere and are found over a large latitudinal range, from the sub-tropics to the arctic (Wang et al., 2016). Birches populate a wide variety of environments; however, they are particularly widespread in forests of temperate and boreal regions. Some *Betula* species have a large distribution while others have a very limited and restricted range, with some classified as endangered in the IUCN Red List (Ashburner & McAllister, 2013; Shaw et al., 2014; Wang et al., 2016). The taxonomy of this genus is complex, and several discordant classifications have been proposed in the literature, with the number of species varying between approximately 40 and 60 (Shaw et al., 2014). A potential cause for the discordance in the molecularly and morphologically based classifications of birch species could be the complexity added by the high frequency of interspecific hybridization and subsequent introgression within this genus (Dehond & Campbell, 1989; Ding et al., 2021; Hu et al., 2019; Karlsdottir et al., 2009; Schenk et al., 2008; Wang et al., 2016), in fact, some birch species appear to have a hybrid origin (Nagamitsu et al., 2006). The frequency of hybridization is augmented by the artificial introduction of cultivars outside of their distribution range, which provides additional opportunities for interspecific contact (Schenk et al., 2008). Introgression seems bidirectional (Williams & Arnold, 2001) but asymmetrical (Palme et al., 2004) among several species in the genus. Hybrids generally exhibit a morphology intermediate between the parental species but are not always morphologically distinct as a group (Schenk et al., 2008; Thórsson et al. 2001). Ploidy level was found to be of great importance in distinguishing some morphologically similar birch species, such as *B. pendula* (diploid) and *B. pubescens* (tetraploid) (Ashburner & McAllister, 2013; Wang et al., 2016). Ploidy level within the genus exhibit great variability, ranging from diploid (2n = 28) to dodecaploid (12n = 168) (Wang et al., 2016).

The first classification of birches (Regel, 1865) divided the genus in two major taxonomic groups, subgenus *Alnaster* and subgenus *Eubetula,* and further divided these two subgenera into seven sections. This classification was accepted until Winkler (1904) suggested to lower the ranks from subgenera to sections, and from sections to subsections, but proposed a very similar classification overall. In more recent years, De Jong (1993) suggested to divide the genus in five subgenera and this classification has been widely accepted until recently, when both molecular and morphologically based classifications agreed to split the genus in four subgenera and eight sections (Ashburner & McAllister, 2013; Schenk et al., 2008; Skvortsov, 2002). However, the most recent inferred phylogeny for the genus (Wang et al., 2021), and the first based on genome-wide loci using restriction site-associated DNA (RAD) loci, proposes to split the genus in two main taxonomic groups or subgenera, which interestingly differ in seed wings morphology and distribution, with one group with narrow seed wings and limited distribution and another with prominent seed wings and wide geographic distribution.

Fossil pollen data suggest that birches in eastern Asia had four major refugial area at the last glacial maximum: the Changbai mountains in north-eastern China, the mountainous area of central China, the southwestern Tibetan plateau and the Altay mountains and adjacent area (Cao et al., 2015). In addition, there is an argument cold-tolerant tree species persisted further north in Eastern Asia during the Quaternary glaciations, and several species, such as birches, seem to also have persisted in a fifth refugia located in north-eastern Siberia (Tsuda et al., 2017).

Betula platyphylla

Asian white birch (*B. platyphylla*), included in the subgenus *Betula* and section *Betula* in the latest phylogenetic classification of the *Betula* genus (Wang et al., 2021), is a temperate wind-pollinated tree of great ecological and commercial value widely distributed in the northern hemisphere, particularly in Eastern Asia including Russia, mainland China, Korea and Japan (Chen & Lou, 2019; Chen et al., 2021). Asian white birch is a pioneer species and can colonize open grasslands with light soil, can grow on roadsides and on the verge of forests (Ashburner & McAllister, 2013; Chen & Lou, 2019; Chen et al., 2021), and can take over after environmental disturbances such as forest fire (Chen & Lou, 2019). Its pollen can travel for thousands of kilometres transported by the wind (Sofiev et al., 2006). This species is monoecious, and individuals can grow rapidly and reach height up to approximately 20 m (Chen et al., 2021). Morphologically *B. platyphylla* is characterized by bright white bark, an oval or pyramidal crown and thin side branches. In China, it occupies primarily the mountainous areas that extend from the north-east in the Hinggan range, Changbai mountains and adjacent areas, to the southwest in the Qinghai–Tibetan Plateau (QTB) and is found at an elevation from 20 m up to over 4,000 m (Chen & Lou, 2019).

***Betula platyphylla* genomics research**

In recent years there have been numerous studies on the hybridization and introgression patterns across Eurasian birch species, including *B. platyphylla.* Tsuda et al. (2017) looked at the hybridization and divergence between silver birch, *B. pendula*, and *B. platyphylla,* sampling individuals from 47 populations scattered from the United Kingdom to Japan and detected a hybrid zone separating these species in central Asia in Siberia, between the Yenisei River and Lake Baikal (Tsuda et al., 2017). Admixture rate suggested a higher contribution of *B. pendula* to hybrids and the two species divergence was estimated to have occurred pre-LGM, approximately 36,000 years ago, with the subsequent admixture dated more recently, approximately 1,600 years ago. This study suggested that *B. platyphylla* persisted in Beringia and *B. pendula* in western Siberia and Russia at the LGM and then they came into contact during postglacial re-expansion.

B. platyphylla has also been found to hybridize with more distantly related birch species, such as *B. albosinensis* (red birch), even though this is much less frequent than with *B. pendula* (Hu et al., 2019). Limited admixture has also been found between *B. platyphylla* and *B. microphylla*, and *B. platyphylla* and *B. tianshanica* (Ding et al., 2021).

The *B. pendula* (Salojarvi et al., 2017) and *B. platyphylla* genomes (Chen et al., 2021) have been sequenced and assembled recently and their genomes showed high similarity (Chen et al., 2021). The *B. platyphylla* genome has 14 chromosomes and it has been found to contain a large percentage (43%) of transposable elements (TEs), primarily class I TEs, and 31,253 protein coding genes. The *Betula* genome was found not to have undergone any recent whole-genome duplication event and a phylogenetic tree based on gene families re-estimated the divergence time between *B. pendula* and *B. platyphylla* at 2.6 million years ago (Chen et al., 2021). The complete chloroplast DNA sequence of *B. platyphylla* has also been sequence and assembled recently (Wang et al., 2018).

Recently, a study of *B. platyphylla* populations in China based on ten microsatellites loci, suggested that this species is divided in five genetic cluster corresponding to geographic locations, and these were inferred to have diverged during the Pleistocene climatic oscillations (Chen & Lou, 2019). Signals of admixture were detected among these clusters and nucleotide diversity was found to increase with latitude, therefore contradicting the Europe-like southern refugia model (Harrison et al., 2001). However, this nucleotide diversity pattern may also be due to other factors, such as the different topography with the populations in northern China occupying a flatter region with fewer geographical barriers compared to the populations in the southern QTB, where mountains and deep valleys restrict pollen movement and therefore gene-flow. Long-time isolation, restricted gene flow, and genetic drift may have resulted in strong genetic differentiation and low genetic diversity in the south, while in the north admixture with other populations coming from the refugia identified in Beringia (Tsuda et al., 2017) may have contributed to increase the genetic diversity in the *B. platyphylla* populations in northern China. This, together with an environmental niche model projected to the LGM and mid-Holocene (Chen & Lou, 2019), seems to suggest that at LGM *B. platyphylla* still had a large latitudinal range and persisted separately both in the north and south in eastern Asia in four or five distribution centres: Beringia (Tsuda et al., 2017), the Altay mountains in the northwest China, the mountainous area of central China, the Changbai mountains and greater Khingan range in the north-east China (Wang et al., 2019), and the southern refugia in north-eastern (the Qilian Mountains) and south-eastern (the Hengduan Mountains) edge of Qinghai–Tibetan Plateau (Chen & Lou, 2019). Therefore, the current *B. platyphylla* Chinese distribution seems to reflect recolonization from multiple local refugia (Chen & Lou, 2019).

Research questions

In this thesis I analysed two separate forest trees whole genome sequencing datasets.

The first included whole genome sequencing data of 386 oak individuals, including *Q. robur*, *Q. petraea* and their hybrids, sampled across four managed British parklands. These parklands have been under long-term monitoring for the symptoms of AOD (Denman et al., 2014), which is widespread across all four sites. This dataset was initially assembled with the aim to uncover whether susceptibility and resistance to AOD has a genetic component. However due to the high complexity of this syndrome and its associated phenotypes, the sample size of this dataset was deemed insufficient for such an ambitious association analysis, as further discussed in Chapter 4. Therefore, the focus of this work was switched to the analysis of the patterns of hybridization and differentiation between English and sessile oak, which are known to hybridize frequently and therefore constitute a great model to study species delineation (Curtu, Gailing & Finkeldey, 2007; Jensen et al., 2009; Kremer et al., 2012; Leroy et al., 2019a; Leroy et al., 2019b; Petit et al., 2004b). Furthermore, I assessed the population structure of the sampled oaks as well as the patterns of relatedness within each managed parkland environment, which can be correlated to different planting regimes. Moreover, I inferred the demographic and divergence history for both species and looked for signals of strong recent natural selection in *Q. robur*. Finally, I took advantage of the huge amount of genetic DNA data generated to assemble *de novo* the whole chloroplast genome sequences of most of the oaks sampled. I assessed chloroplast DNA variation across the parklands sampled and compared it with data generated by previous work on oaks on natural woodlands across Britain and Europe, which characterized the main European oak chloroplast haplotypes and re-constructed the post-glacial colonization routes of oaks from the southern refugia based on chloroplast DNA and fossil pollen maps (Petit et al., 2002a; Petit et al., 2002b; Cottrell et al., 2002), to estimate whether the oaks in the parklands sampled derive from local and/or native seed stock.

The second dataset included whole genome sequencing data for 83 Asian white birch individuals (*B. platyphylla*) across 74 natural population scattered throughout the known distribution of this species in China. The ultimate aim of this study was to detect genetic signals of local adaptation in this ecologically and economically important Asian species and define its current and future geographic distribution in China. *B. platyphylla* provides a good model system to study adaptation as its distribution span a large latitudinal range and occupies a variety of environments, from low to very high elevation, and characterized by different levels of precipitation and temperature (Ashburner & McAllister, 2013; Chen & Lou, 2019; Chen et al., 2021). In addition, the recently published reference genomes of both *B. pendula* (Salojarvi et al., 2017) and *B. platyphylla* (Chen et al., 2021) constitute great resources which enhance the potential studies on this species. I made use of the results generated to assess the possible degree of maladaptation to future climate (2080 – 2100) of the sampled populations and highlight areas in China where this species may be particularly threatened by climate change, particularly rising temperature, in order to inform conservation strategies. In addition, I assessed the population structure of this species across Eurasia and its hybridization with the closely related silver birch, *B. pendula*, by including data from another study (Salojarvi et al., 2017). Finally, I assessed the demographic history of Chinese *B. platyphylla* populations in the context of the Quaternary climatic oscillations and post-LGM migration.

Chapter 2: Genomic structure and diversity of oak populations in British parklands

The research presented in this chapter has been published and I am the lead author:

Nocchi, G., Brown, N., Coker, T., Plumb, W., Stocks, J., Denman, S., & Buggs, R. (2021). Genomic structure and diversity of oak populations in British parklands. *Plants, People, Planet*, *4*(2), 167-181.

This chapter presents the above publication in an extended form, including additional unpublished analyses.

Abstract

The two predominant oak species in Britain are *Quercus robur* (English or pedunculate oak) and *Q. petraea* (sessile oak), and large populations of these species are found in British parklands: managed wood pastures up to 1000 years old. We sequenced the whole genomes of 386 oak trees from four British parkland sites and found over 50 million nuclear single nucleotide polymorphisms (SNPs), allowing us to identify 360 *Q. robur*, ten *Q. petraea* and 16 hybrid individuals using clustering methods. We assessed the population structure and demographics histories of these species as well as their patterns of admixture and differentiation. Comparing *Q. robur* and *Q. petraea* trees from Attingham Park, we found that the nuclear genomes of the two species are largely undifferentiated but identified 81 coding regions exhibiting strong interspecific differentiation. The nuclear genomes of the *Q. robur* individuals showed no clear differentiation among the four parkland sites. In addition, scans for selective sweeps in *Q. robur* highlighted regions containing genes with putative involvement in stress tolerance, one of which was moderately differentiated from *Q. petraea*. Reconstructions of past effective population sizes suggested a long population size decline in both *Q. robur* and *Q. petraea* over the Pleistocene, but population growth after the last glacial maximum. Lastly, we assembled the whole chloroplast genomes of 287 *Q. robur*, 8 *Q. petraea* and 14 hybrid trees. In a phylogenetic network, these fell into five major haplotypes, which were shared among species but differed in frequency among parkland sites. We matched our chloroplast genome haplotypes to restriction enzyme fragment haplotypes identified in older studies that had surveyed ancient woodlands in Britain and much of Europe. This suggested that the parkland populations in our study derive from local seed sources.

Introduction

Oaks (genus *Quercus*) are some of the most common and widely distributed forest trees in the northern hemisphere, found throughout North America, Europe and Asia comprising between 300 and 500 species (Denk et al., 2017; Leroy et al., 2019b). Differentiating oak species morphologically has been termed a “botanical nightmare” due to ambiguous and extremely variable phenotypes, which is partly due to extensive hybridization and backcrossing (Denk et al., 2017; Kleinschmit, 1993; Kremer et al., 2012; Lepais et al., 2009; Leroy et al., 2019b).

A recent phylogenomic study showed that the oak phylogeny is highly reticulate, with phylogenetic incongruence widespread throughout the genome (Hipp et al., 2019). The Eurasian white oaks form a clade in the phylogeny of Hipp et al. (2019) that is nested within clades of largely American oak species. White oaks are thought to have crossed the North Atlantic land bridge in the Oligocene and split into European and Asian clades in the Miocene, which then rapidly diversified (Hipp et al., 2019). Evidence from chloroplast variation and fossil pollen suggests that at the last glacial maximum (~20 Kya), European white oaks were constrained to refugia in southern Europe and the vast majority of Britain native oak trees stem from a lineage that migrated from a western Iberia glacial refugium (Cottrell et al., 2002; Petit et al., 2002a; Petit et al., 2002b). In total, six chloroplast lineages have been identified in European oaks, corresponding to glacial refugia in the Balkans (lineage A), in western Spain (lineage B), in the south of Italy (lineage C), in eastern Spain (lineage D), in the eastern Balkans (lineage E) and north-east of the Black Sea (Crimea) (lineage F) (Petit et al., 2002a; Petit et al., 2002b). In Europe today, the two most common oak species are *Q. robur* (English oak or pedunculate oak) and *Q. petraea* (sessile oak), which are found from central Spain to the Urals and from Scandinavia to the south of Italy (Eaton et al., 2016). Their lineages seem to have diverged in the late Miocene/early Pliocene, and they are found in sister clades that each also contain a small number of species distributed in southern Europe and/or North Africa (Hipp et al., 2019).

Quercus robur and *Q. petraea* are sympatric across most of their geographic range but exhibit different ecological preferences (Eaton et al., 2016; Saintagne et al., 2004). *Q. robur* reaches more northerly and easterly ranges in Europe, prefers moist and alkaline soils with high nutrient availability and it is more tolerant to periodic flooding and water logging (Eaton et al., 2016; Saintagne et al., 2004). On the other hand, *Q. petraea* prefers more acidic soils and it is more resistant to drought (Eaton et al., 2016; Saintagne et al., 2004). *Q. robur* is a pioneer species that colonizes open ecosystems while *Q. petraea* appears to be more of a successional species that expands in area already occupied by *Q. robur* but can also behave as a pioneer itself (Eaton et al., 2016; Levy et al., 1992; Truffaut et al, 2017; Petit et al., 2004b). Hybridization between *Q. robur* and *Q. petraea* is common and introgression of adaptive alleles appears to have played an important role in *Q. petraea* postglacial northward migration and expansion into cooler and wetter areas (Lepais et al., 2009; Leroy et al., 2019a; Petit et al., 2004b). Millennia of hybridization and introgression among European oaks are responsible for a high degree of phenotypic variability within species, however key morphological and ecological differences between *Q. robur* and *Q. petraea* are still present in nature (Beatty et al., 2016; Curtu, Gailing & Finkeldey, 2007; Jones, 1959; Kremer et al., 2002; Lesur et al., 2018). Controlled pollination experiments suggest that pre- and postzygotic barriers to gene flow in European white oaks maintain the nuclear genetic diversity between species and these barriers are often asymmetric (Lepais et al., 2013; Leroy et al., 2017; Truffaut et al., 2017), for example pollen of *Q. petraea* pollinates *Q. robur* more frequently than vice-versa, under both natural and experimental conditions (Lepais et al., 2013; Truffaut et al., 2017). Furthermore, genomic patterns of admixture suggest that introgression occurs predominantly from *Q. robur* to *Q. petraea* (Guichoux et al., 2012; Leroy et al., 2019a; Petit et al., 2004b).

Recent genetic studies based on approximate Bayesian computation (ABC) models have shown that massive secondary contacts between European white oak species have occurred recently after a long period of isolation, probably at the end of the last glacial period or at the start of the current interglacial period (Leroy et al., 2017; Leroy et al., 2019b). These contacts are thought to have homogenized the chloroplast genome and the majority of the nuclear genome of European oak species, with the exception of some barrier regions accumulated during the long periods of isolation that preceded secondary contacts, which are responsible for maintaining species integrity (Lepais et al., 2013; Leroy et al., 2017; Leroy et al., 2019b). Past studies based on small numbers of loci (RAPDs, SCARs, Isozymes, SSRs and AFLPs) have shown that the genetic differentiation between *Q. robur* and *Q. petraea* is subtle and their genomes are characterized by high homogeneity (Barreneche et al., 1996; Bodenes et al., 1997; Coart et al., 2002; Mariette et al., 2002; Scotti-Saintagne et al., 2004; Zanetto et al., 1994). More recent efforts, based on genome-wide SNP markers, have successfully identified some regions significantly differentiated between these two closely related oak species, but confirmed that the majority of their genomes appear permeable to interspecific gene flow (Guichoux et al., 2012; Lang et al., 2018; Leroy et al., 2019b; Lesur et al., 2018; Reutimann, Gugerli & Rellstab, 2020).

A common and culturally important habitat for oaks in Britain is parkland. Parklands are wood pastures that vary in age from 100 to 1000 years, often containing high levels of deer or other stock such as sheep (Rackham, 1990). Parklands are largely anthropogenic habitats and trees within them can be planted, naturally regenerated, or retained from previous landscapes (Rackham, 1990). Previous studies on the genetic structure of oak populations throughout Britain concentrated on ancient woodland sites, to try to minimise anthropogenic influences and maximise the chances of sampling locally native trees (Cottrell et al., 2002). The chloroplast DNA haplotype structure of ancient woodlands in Britain has thus been substantially documented (Cottrell et al., 2002; Cottrell et al., 2004; Lowe et al., 2004). However, little is known about the genetic structure and background of oaks in parklands. Such information would help in the future management of these populations, and also help us understand the basis of health issues faced by parkland oaks, especially acute oak decline (AOD) (Denman et al., 2018). This disease has affected British native oak species for the past three decades and it is particularly widespread in south-eastern England (Brown et al., 2016). It is characterised by stem bleeds that get denser and more numerous as the disease progresses and can ultimately lead to tree death (Brown et al., 2016; Denman et al., 2018). Research has shown that AOD is a complex multifactorial and polymicrobial tree disease, but its primary causes are unclear, and the role of host genetics has not been investigated (Denman et al., 2018).

In this study we sequenced the whole genome of 386 oak trees from four British parkland sites, including *Q. robur*, *Q. petraea* and their hybrids. We detected over 50 million SNPs and made use of this huge genetic variation to: characterize the structure, diversity and demographic histories of *Q. robur* and *Q. petraea* parkland populations, find loci differentiated between the species, and to detect signatures of recent positive selection in *Q. robur*. Furthermore, we looked for evidence of parent-offspring or sibling relationships within populations. We also assembled whole chloroplast genomes and assigned these to haplotypes identified by restriction enzyme fragment sizes in previous studies of oaks in ancient woodlands across Europe, enabling us to test the origins of these parkland oak trees.

Materials and Methods

Sampling and DNA extraction

Leaf material was collected by the Forest Research Technical Service Unit from 386 oak trees in autumn 2017. Collections were from four British parkland sites: 82 trees in Attingham Park (Atcham, Shropshire, England, 52.688965° N, -2.667944° W), 80 in Hatchlands Park (East Clandon, Surrey, England, 51.257072° N, -0.472332° W), 124 in Langdale Wood (Malvern, Worcestershire, England, 52.085446° N, -2.307670° W) and 100 in Sheen Wood (Richmond, London, England, 51.456229° N, -0.269298° W). These parklands are being monitored for the symptoms of AOD (Denman et al., 2018). Based on morphology, the collectors tentatively identified 376 the trees as *Q. robur* and 10 as *Q. petraea*. They did not attempt to identify hybrids as these are known to be morphologically cryptic (Curtu, Gailing & Finkeldey, 2007). The collectors classified the sampled trees as either: young (approx. diameter at breast height “DBH" < 15 cm, Raimbault stages 1 - 3), semi-mature (approx. 15

cm < DBH < 40 cm, Raimbault stages 4 -5), mature (approx. 40 cm < DBH < 70 cm Raimbault stages 7 -8) or over mature (approx. DBH > 70 cm, Raimbault stages 9 - 10). Whole genomic DNA was extracted from leaf tissue at RBG Kew using Qiagen DNeasy protocol and was sent to Novogene (Hong Kong) for library preparation and whole-genome shotgun sequencing.

Genotyping

Shotgun libraries with fragment sizes of 350 bp were prepared with NEBNext DNA Library Prep Kit and were sequenced with 150 bp paired-end Illumina NovaSeq 6000 technology at 22x depth of coverage by Novogene. Trimmomatic v0.36 (Bolger, Lohse and Usadel, 2014) was used to remove sequencing adapters and for trimming, scanning all reads in windows of four bases, and cutting at the leftmost position if the average Phred base quality score dropped below 20 (99% base accuracy). Reads shorter than 70 bases after trimming were discarded. Processed reads were mapped to the haploid chromosome-level version of the *Q. robur* reference genome (Plomion et al., 2018) with BWA-MEM v0.7.15 (Li & Durbin, 2009) run with default settings. After alignment to reference, PCR duplicates were removed using Samtools v1.9 (Li et al., 2009). Variant calling was performed on all samples simultaneously with the software Haplotype Caller in joint genotyping mode, available in GATK v4.0.8.1 (DePristo et al., 2011). SNPs were extracted and filtered to exclude loci with either: quality by depth less than two, Fisher strand test greater than 60, root mean square mapping quality less than 50, mapping quality rank sum test less than -2 or read position rank sum test less than -2. We refer to the resulting SNP set as the genome-wide set.

Assignment of individuals to species

The genome-wide SNP set was quality filtered with vcftools v0.1.16 (Danecek et al., 2011) removing loci with missing genotypes and with either: individual mean depth less than 15, minor allele count less than three, minor allele frequency less than 0.005, individual depth less than five or mapping quality less than 20. This set was further reduced by excluding SNPs located in the transposable elements identified by Plomion et al. (2018), with bedtools v2.28.0 (Quinlan & Hall, 2010), and multiallelic sites, with bcftools v1.8. Finally, SNPs were pruned by linkage disequilibrium ($r^2 > 0.4$) using the indep-pairphase function of PLINK v2.0 (Chang et al., 2015) with window size of 50 markers and step of 5. If two SNPs in a window are linked, this function keeps the SNPs with higher minor allele frequency and if there is a tie the first SNP is kept. We refer to the resulting SNP set as the reduced set.

We used fastSTRUCTURE v.1.0 (Raj, Stephens, & Pritchard, 2014) to infer admixture levels and to assign individuals to species. fastSTRUCTURE was run with a simple prior, 5-fold cross-validation and number of ancestral populations (K) from one to ten. To select the model that best explain structure in the data we used the cross-validation profile to identify the number of K for which the prediction error is minimized, and we compared it with the output of the python function “chooseK”, available in fastSTRUCTURE, which computes two values for K: one that maximises the log-marginal likelihood lower bound (LLBO) of the dataset ($K^{*}\varepsilon$) and aims to identify strong structure, and another one which reports the model components that have a cumulative ancestry contribution of at least 99% ($K_{\emptyset}c$) aimed at capturing additional weak underlying structure (Raj, Stephens, & Pritchard, 2014). Individual trees were assigned to three categories according to the value of the admixture coefficient (q) computed with fastSTRUCTURE with K=2: pure *Q. robur* ($q \geq 0.9$), hybrid trees ($q > 0.1$ & $q < 0.9$) or pure *Q. petraea* ($q \leq 0.1$), as described in Truffaut et al. (2017). The species that each fastSTRUCTURE cluster represented was determined on the basis of our collectors’ morphological classification of the sampled individuals. Principal component analysis (PCA) was performed on the reduced SNP set using Plink v2.0 (Chang et al., 2015).

Interspecific differentiation

Genotypes of 10 *Q. petraea* and 10 unrelated *Q. robur* individuals from Attingham park (Atcham, Shropshire, England, 52.688965° N, -2.667944° W) were extracted from the reduced SNP set and were filtered to exclude loci with minor allele frequencies (MAF) below 0.05 using bcftools v1.8. Weir and Cockherham F_{st} (Weir & Cockerham, 1984) was calculated between species in non-overlapping 10 kb sliding windows and by SNP site. We identified outlier SNP loci (1% of most extreme F_{st} values) by generating smoothed quantiles from the empirical distribution of the F_{st}-heterozygosity relationship using the R package “fsthet” (Flanagan & Jones, 2017). This method is similar to that described by Beaumont & Nichols (1996) however it avoids assumptions about the underlying distribution and population structure parameters. This is particularly useful when the number of populations is small (< 10), and migration rate is low because the F_{st}-heterozygosity relationship does not usually fit the expected distribution and confidence intervals generated with the null infinite island model in such cases (Flanagan & Jones, 2017). The method implemented in the package “fsthet” is not a statistical test and always identifies a specific number of outliers from the empirical distribution, depending on the number of loci tested and the chosen confidence interval (Flanagan & Jones, 2017). To mitigate this limitation and minimize the number of false positives we conservatively selected outlier loci showing strong differentiation, with an F_{st} at least two standard deviations from the whole genome mean F_{st}. We then used a nonoverlapping 10 kb window approach to estimate the number of outliers per window across the genome, similarly to the method employed in Leroy et al. (2019b). We selected the top 1% outlier enriched windows and identified the genes within or flanking (within 5 kb) these regions with bedtools (Quinlan & Hall, 2010), based on the gene models reported by Plomion et al. (2018). To test the species discriminatory power of the identified genes we ran a second fastSTRUCTURE (Raj, Stephens & Pritchard, 2014) taxonomic assignment at K=2 of all individuals based only on SNP loci located within these genic regions.

Population size history and species divergence

We inferred *Q. robur* and *Q. petraea* population size and separation history using a multiple sequentially Markovian coalescent approach implemented in the software MSMC (Schiffels & Durbin, 2014). MSMC is an extension of the pairwise sequential Markovian coalescent (PSMC) model (Li & Durbin, 2011), which analyses two homologous sequences from a diploid individual, therefore limiting the resolution of population size estimation of more recent history (Schiffels & Durbin, 2014). MSMC employs a simplification which extends PSMC to multiple sequences (Schiffels & Durbin, 2014). Using more haplotypes lead to a more recent estimate of the time from the first coalescence event, while when using only two sequences from a diploid individual the most recent common ancestor is usually estimated further back in time (Schiffels & Durbin, 2014). For population size history inferences, we ran MSMC on one, two and four high coverage (mean depth > 20x) individuals of each species separately, to achieve higher resolution for both distant and recent history, respectively. For species separation history we used four individuals, two *Q. robur* and two *Q. petraea* trees. Biallelic genotypes were extracted from the genome-wide set, excluding loci with missing data, and MSMC was run over the 12 chromosomes with default time segment patterning. Mappability mask files of repeat elements (Plomion et al., 2018) were provided in bed format to exclude non-unique regions of the genome from the analysis. Statistical phasing was performed using Beagle 5.1 with default settings (Browning & Browning, 2007). MSMC outputs times and population sizes scaled by mutation rate per base pair per generation. To convert the output to real time and sizes it is necessary to divide estimates by the mutation rate and further multiply these by the generation time. As the mutation rate of *Q. robur* and *Q. petraea* was unknown, we used the substitution rate 7.5×10^{-9} of *A. thaliana*, retrieved from Buschiazzo, Ritland, Bohlmann, & Ritland (2012), as it was done for Fraxinus excelsior in Sollars et al. (2017). We used a generation time of 50 years (Leroy et al., 2019b).

Population structure and positive selection in Quercus robur

Genotypes of 360 *Q. robur* individuals identified with fastSTRUCTURE were extracted from the genome-wide set and were used to estimate linkage disequilibrium decay along the oak genome using two methods: the r^2 function available in Plink v 2.0 and the tool PopLDdecay, including only SNPs with MAF > 0.05 (Zhang, Dong, Xu, He, & Yang, 2018). We calculated *Q. robur* nucleotide diversity π (Nei & Li, 1979) along the genome including repeat regions and restricted to genic and intergenic regions, using the vcftools function window-pi in non-overlapping windows of 5 kb.

To study *Q. robur* population structure SNPs were extracted from the linkage disequilibrium pruned reduced SNP set. PCA and fastSTRUCTURE with number of ancestral clusters K from one to ten and 5-fold cross-validation were performed on the pruned *Q. robur* set excluding loci with minor allele frequencies below 0.05.

Marker based realized genomic relatedness was computed across sites including all sampled individuals and restricted to each site, according to the formula by VanRaden (2008), implemented in the kin function of the R package synbreed (Wimmer, Albrecht, Auinger, & Schon, 2012), using the pruned *Q. robur* SNP set (MAF > 0.05). Relatedness was removed by excluding an individual from each pairwise comparison with realized relatedness (VanRaden, 2008) greater than 0.05 using the r package plinkQC (Meyer, 2020). The relatedness filter implemented in plinkQC aims to minimize the number of individuals removed to eliminate any relatedness above a chosen threshold (Meyers, 2020). PCA was re-computed in Plink v2.0 for the unrelated *Q. robur* individuals. We identified potential parent-offspring and siblings relationships based on our relatedness estimates and the age class of trees derived from the assessment of DBH and crown condition (Raimbault, 1995). Pairs of trees with relatedness estimates above 0.45 (~0.5 expected for first degree relatives), were classified as putative siblings if they were in the same age class, or as parent-offspring if they were in different age categories.

We used SweeD v3.2.1 (Pavlidis, Živković, Stamatakis, & Alachiotis, 2013) and OmegaPlus (Alachiotis, Stamatakis, & Pavlidis, 2012) to perform selective sweep scans of the *Q. robur* genome, including repeat regions and all allele frequencies but excluding multiallelic loci and loci with missing calls. SweeD employs a composite likelihood ratio (CLR) test to detect hard sweeps based on the site-frequency spectrum (SFS) of SNPs in whole-genome data (Pavlidis, Živković, Stamatakis, & Alachiotis, 2013). OmegaPlus searches for specific linkage-disequilibrium patterns characteristics of recent selective sweeps (Alachiotis, Stamatakis, & Pavlidis, 2012) and output the ω-statistic (Kim & Nielsen, 2004). SweeD and OmegaPlus were run with a grid parameter that resulted in a measurement of CLR and ω-statistic every 5,000 bp and each chromosome was scanned separately. In OmegaPlus the minimum and maximum size of the sub-region around a position which was included in the calculation of the ω-statistic were fixed to 500 and 100,000 base pairs, respectively. Ancestral allele states for the SweeD analysis were inferred using two outgroups, *Fagus sylvatica L.* (Mishra et al., 2018) and *Castanea mollissima* (Xing et al., 2019). Raw whole genomic sequencing reads for both outgroups were downloaded from the European Nucleotide Archive (ENA) (PPRJEB24056, RJNA527178) and were mapped to the *Q. robur* reference genome using Bowtie2 in local alignment mode. The alignments generated were sorted and processed with Samtools v1.9 (Li et al., 2009) to remove PCR duplicates. Outgroups genotypes, including homozygous reference calls, were inferred with bcftools v1.8 and the vcftools v0.1.16 utility "fill-aa" was used to record the ancestral state for the *Q. robur* SNPs. The ancestral allele was determined only for SNP loci which appeared homozygous for the same allele in both outgroups. If a site was either not covered by the outgroup reads, was heterozygous in one or both outgroups or the outgroups were homozygous for alleles not found in oaks, then the ancestral state for that locus was not determined. We identified the common outliers between the SweeD and OmegaPlus runs, representing the top 1% CLR and ω-statistic values per chromosome. We used bedtools to identify the gene models (Plomion et al., 2018) within or flanking the common outlier windows detected.

Chloroplast network

The chloroplast DNA sequences of the 386 individual trees were assembled *de novo* using Novoplasty 3.7.2 (Dierckxsens, Mardulyn & Smits, 2017). The software was run in chloroplast mode with k-mer length of 39, the minimum length of overlap necessary to join adjacent reads in the assembly. We used the *Zea mays* chloroplast gene for the large subunit of ribulose bisphosphate carboxylase (RUBP) as seed sequence for the assembly and the *Quercus lobata* isolate SW786 complete chloroplast sequence (Sork et al., 2016) as reference. The assembly was performed *de novo* by Novoplasty with the reference used to aid resolution of difficult regions, such as inverted repeats. Novoplasty was set with read length of 150 and insert-size of 350 and the software was allowed to automatically finetune these values.

The fully assembled and circularized chloroplast sequences were linearized and shifted to the same origin, which was set at the seed RUBP sequence, using the Perl package fasta-tools (https://github.com/b-brankovics/fasta_tools). The shifted complete chloroplast sequences were aligned using MAFFT v7 (Katoh & Standley, 2013) and the alignment was exported in Phylip format. PopArt v1.7 (Leigh & Bryant, 2015) was used to generate a median joining network from the alignment file as described by Bandelt, Forster, & Rohl (1999).

Chloroplast haplotypes identification

Representative chloroplast DNA sequences for each of the clusters identified in the median joining network were analysed by reproducing *in silico* the polymerase chain reaction-restriction fragment length polymorphism (PCR-RFLP) method used to characterize the main oak chloroplasts lineages in Europe, including over 40 haplotypes, in Petit et al. (2002b). The haplotypes identified in our chloroplast network were matched to known oak chloroplasts haplotypes according to the information provided by four non-coding chloroplast DNA fragments extracted and digested *in silico* with custom scripts. The four chloroplast DNA fragments and restriction enzyme pairs were: trnD-trnT (DT) with TaqI, psaA-trnS (AS) with HinfI, psbC-trnD (CD) with TaqI and trnT-trnF (TF) with AluI (Petit, Demesure, & Dumolin, 1998; Petit et al., 2002b). The length variants obtained with restriction digestion simulation of the chloroplast DNA fragments were compared and matched with those reported for previously characterized oak haplotypes (Appendix B in Petit et al., 2002b). The trnD-trnT and trnT-trnF fragments were further characterized for the presence of a point mutation, using AluI and CfoI restriction enzymes respectively, as described in Petit et al. (2002b).

We compared the distribution of the chloroplast haplotypes identified with that of 163 ancient woodlands distributed across England and Wales, surveyed as part of a previous study (Cottrell et al., 2002). To assess whether the parklands surveyed matched the local ancient woodlands dominant haplotype, we performed an ordinary kriging linear regression (R, “gstat” package) of Cottrell et al. (2002) data to define haplotypes dominance regions in England and Wales, similarly to Lowe et al. (2004). Kriging linear regression is a spatial prediction method that combines regression, in this case based on haplotype count data at know locations, with spatial interpolation, in order to predict haplotype frequencies across geographic space (Pebesma, 2006).

Results

SNP discovery

We sequenced the whole genome of 386 oak trees from British parklands and identified 57,573,404 SNPs by mapping to the *Q. robur* reference genome (Plomion et al., 2018), prior to applying more stringent quality filters. This set of 57M SNPs includes only those sites that were polymorphic among the sampled individuals, an additional 1,737,384 sites were monomorphic in our data but differed from the reference genome. Mean individual depth of coverage of SNP loci was 21.5x, varying from 12.7x to 33.4x across samples and individual call rate varied from 1 to 5%, with mean of 4.2%. The Ts/Tv ratio was normally distributed around mean 2.58 with small variation. In total 4,665,342 SNPs (~8%) were in genic sequences, according to the gene models reported by Plomion et al. (2018).

Population structure and interspecific differentiation

We analysed the genome-wide variation of 386 oak trees with model-based inference and principal component analysis (PCA) based on the reduced SNP set, which included 2,768,547 unlinked SNPs ($r^2 < 0.4$).

fastSTRUCTURE function “chooseK” indicated K=2 as the number of ancestral clusters that maximises the LLBO (K*ε) of the data (Figure S2.1A, Supporting Information) and best explains additional weak underlying structure (Køc) (Raj et al., 2014). Furthermore, even though the fastSTRUCTURE 5-fold cross-validation analysis shows that the prediction error is minimized at K=9 (Figure S2.1B, Supporting Information), the cross-validation profile is not fully resolutive as the prediction error does not vary greatly with increasing model complexity (Figure S2.1B, Supporting Information). In particular, the prediction error at K=9 is within one standard error from that of K=2, therefore we chose the latter more parsimonious model to represent our data, as the use of prediction error for model selection with fastSTRUCTURE is prone to the overestimation of the number of populations and the function “chooseK” is recommended (Raj et al.,2014). In total, 10 individuals were assigned to *Q. petraea*, 360 individuals to *Q. robur* and the remaining 16 individuals were classified as hybrids, on the basis of the admixture coefficient computed with fastSTRUCTURE at K=2 (Figure 2.1). The vast majority of the sampled trees were *Q. robur*, according to our collectors’ morphological classification, therefore it was straightforward to determine which species each fastSTRUCTURE cluster represented. The fastSTRUCTURE species assignment matched the collectors’ classification based on leaf morphology for pure individuals, except for three *Q. petraea* individuals previously assigned to *Q. robur*. Three of the 16 admixed individuals identified were originally classified morphologically as *Q. petraea*, while the remaining 13 had been assigned to *Q. robur.* The collectors’ classification of the identified admixed individuals reflected the dominant contribution to their ancestry.

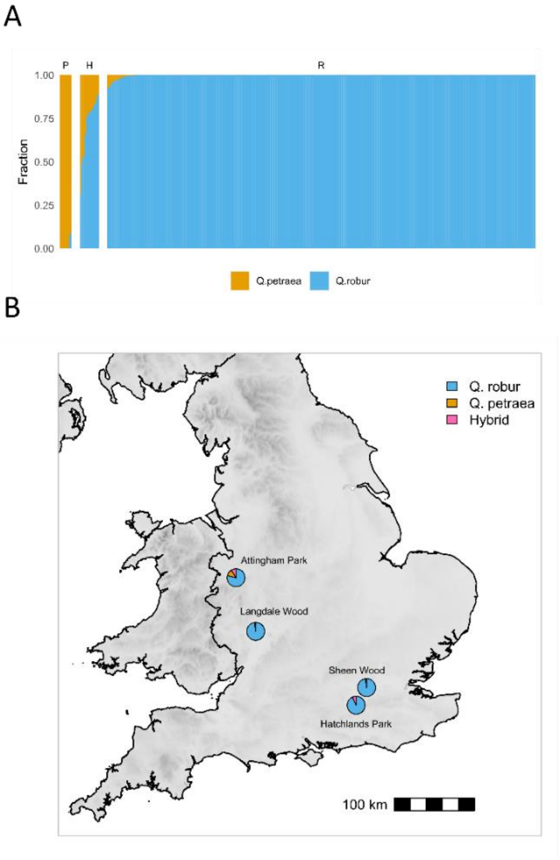


Figure 2.1. Oak species distribution across parklands. A) Bar plot of fastSTRUCTURE results of 386 oak individuals based on 2,768,547 unlinked SNPs ($r^2 < 0.4$) at K = 2. Taxonomic assignment based on the value of the admixture coefficient computed with fastSTRUCTURE: *Q. robur* (R) ($q \geq 0.9$), hybrid trees (H) ($q > 0.1$ & $q < 0.9$) or pure *Q. petraea* (P) ($q \leq 0.1$). Clusters assigned to species according to leaf morphology. B) Map of species distribution across sampling sites.

In PCA, the first principal component explained 38% of the total variance and clearly separated *Q. robur* and *Q. petraea* individuals while the admixed individuals identified with fastSTRUCTURE came between the main species clusters (Figure 2.2A). The second principal component accounted for 10% of the variance and separated the hybrid individuals from the rest, while keeping the two species closer. The explained variance levelled off at the third component which pulled apart six individual samples (Figure 2.2B-C). The six Sheen Wood individuals separated by the third principal component were determined to be closely related (relatedness > 0.125), as assessed later in the chapter.

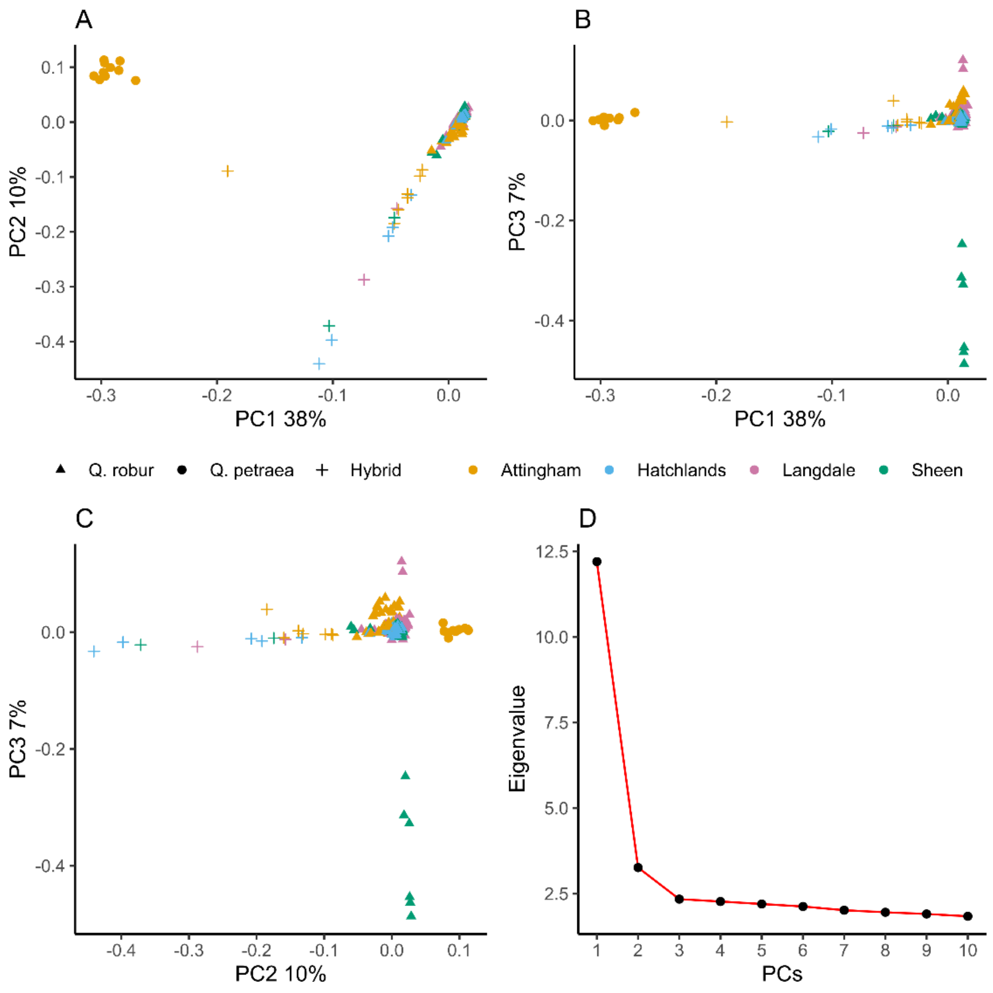


Figure 2.2. PCA of 386 individual oak trees of two species, based on 2,768,547 unlinked SNPs ($r^2 <$ 0.4). Taxonomic labels represent fastSTRUCTURE classification at K = 2. A) PC1 against PC2. B) PC1 against PC3. C) PC2 against PC3. D) Eigenvalues of the computed principal components.

Between ten unrelated *Q. robur* and the ten *Q. petraea* individual trees from Attingham park loci with minor allele frequencies over 0.05 (914,242 SNPs, ~24% genic and ~76% intergenic) gave a mean interspecific F_{st} of 0.158 (median and standard deviation were 0.111 and 0.156, respectively) (Figure 2.3, Figure S2.2, Supporting Information). Those SNPs located in the gene models predicted by Plomion et al. (2018) (223,319 SNPs), gave a mean F_{st} value of 0.155 (median and standard deviation were 0.111 and 0.153, respectively) whereas intergenic SNPs gave a mean F_{st} value of 0.159 (median and standard deviation were 0.111 and 0.157, respectively). The genome wide F_{st} distribution shows that the majority of loci exhibit small differentiation (Figure S2.2, Supporting

Information). A minority of loci clearly displayed very strong differentiation and 610 SNPs were fixed for different alleles in the two species (F_{st} = 1). Within-species F_{st} calculation between two populations (Attingham and Hatchlands) each composed of 10 unrelated *Q. robur* individuals gave a much lower mean F_{st} value of 0.024 (median and standard deviation were 0 and 0.044, respectively).

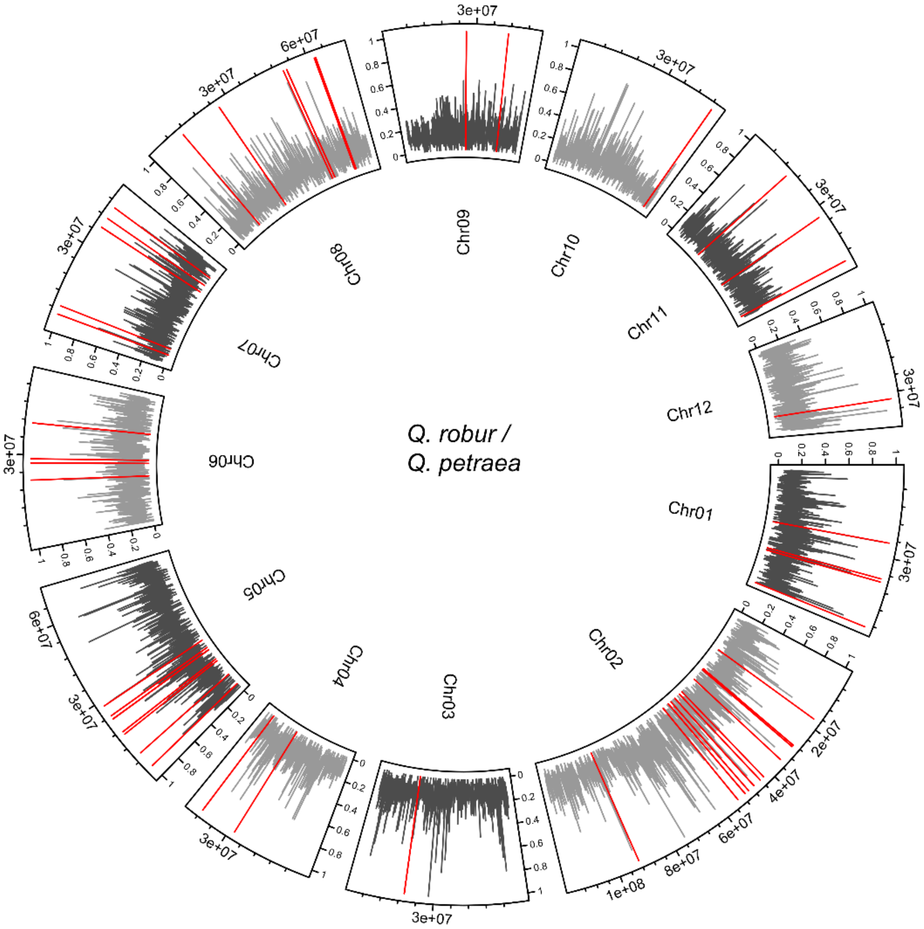


Figure 2.3. Genome-wide Fst between 10 *Q. robur* and 10 *Q. petraea* from Attingham Park**,** based on 914,242 SNPs (MAF > 0.05), computed in non-overlapping 10 kb sliding windows. In red the 81 coding regions with strong interspecific differentiation identified in this study.

Between *Q. robur* and *Q. petraea* we identified 8019 outlier SNPs, representing most extreme values of the F_{st}-heterozygosity distribution (Figure S2.3, Supporting Information). We searched the *Q. robur* gene models (Plomion et al., 2018) within or flanking the top 1% of outlier-enriched 10 kb windows (69 windows with 5 or more outliers) and found 81 genes (Table S2.1, Supporting Information). These were distributed across the entire genome and present on all chromosomes, with a hotspot visible near the middle of Chromosome 2 (Figure 2.3). Using only these 81 genic regions, which included 3,567 SNPs, provided the same assignment of individuals to species at K=2 in fastSTRUCTURE as the run based on 2,768,547 genome-wide SNPs, except that three individuals previously classified as *Q. robur* with signals of introgression from *Q. petraea* were classified as hybrids with a strong *Q. robur* component.

Species population size and separation history

Multiple sequentially Markovian coalescent (MSMC) plots for both species suggest a long-term history of population size decline followed by a steep increase in more recent history (Figure 2.4). Using 2, 4 or 8 nuclear haplotypes gave similar results for time periods where their estimates overlap (Figure 2.4A-C).

Estimation of cross-coalescence rates with MSMC suggests that *Q. robur* and *Q. petraea* populations started separating 8 Mya and reached full separation in the last 0.5/1 Mya (Figure 2.4D).

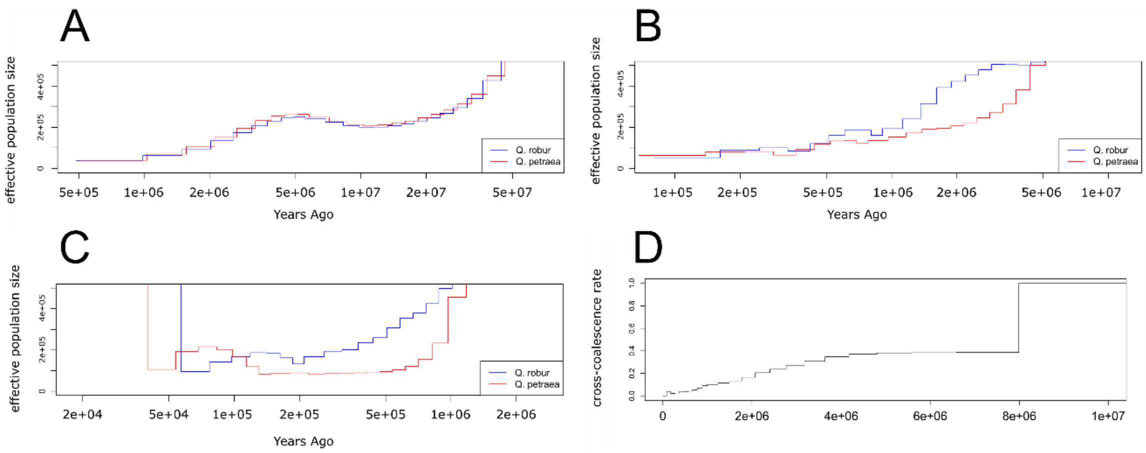


Figure 2.4. Effective population size and separation history of *Q. robur* and *Q. petraea* estimated with the MSMC method. Generation time: 50 years. Mutation-rate: 7.5e^{-9}. A) Estimates based on 2

haplotypes for each species. B) Estimates based on 4 haplotypes for each species. C) Estimates based on 8 haplotypes for each species. D) Relative cross-coalescence rate (CCR) ratio per year. The CCR ratio is a measure of divergence and represents the ratio of between-populations over within-populations coalescence rate. Values close to 0 indicate that populations have diverged, while values close to 1 indicate that populations have not yet diverged as between and within populations coalescence rates are equal.

Quercus robur *genetic diversity and positive selection*

The PCA restricted to the *Q. robur* individuals, based on 839,911 unlinked SNPs (MAF > 0.05) showed that there is little difference in the variance explained by each principal component and there is no strong clustering visible in the plots, with slight population-specific pattern visible for Attingham Park, Langdale Wood and Sheen Wood (Figure S2.4, Supporting Information). Realized genomic relatedness (VanRaden, 2008) among the 360 *Q. robur* trees was normally distributed around 0 (Figure S2.5, Supporting Information) and relatively low with 123 pairwise estimates above 0.125 (expected for third-degree relatives) among 64,620 pairwise estimates. There were noticeable differences in within-site relatedness among sites (Figure S2.6, Supporting Information). Trees within Hatchlands Park had the lowest level of relatedness (mean = 0.0016) with only 7 pairwise estimates above 0.125, involving 14 trees out of the 75 *Q. robur* sampled at this site (18 %). Sheen Wood similarly had relatively low relatedness levels (mean = 0.0019) with 23 estimates above 0.125, involving 21 trees out of 98 (21%). In contrast, Langdale Wood (mean = 0.003) and particularly Attingham Park (mean = 0.0065) reported substantially higher relatedness levels, with 48 and 45 pairwise estimates above 0.125 respectively, involving 53 trees (43%) at Langdale Wood and 40 trees (61%) at Attingham Park. Based on our relatedness estimates and the approximate age of trees based on diameter at breast height, we identified two potential parent-offspring relationships and two full-sibs' pairs at Attingham Park, three full-sibs' pairs at Sheen Wood, five full-sibs' pairs at Langdale Wood and a single potential full-sib relationship at Hatchlands Park. In a global test of relatedness among sites, we found only two among-site pairs with relatedness above 0.05. These were between Hatchlands Park and Sheen Wood.

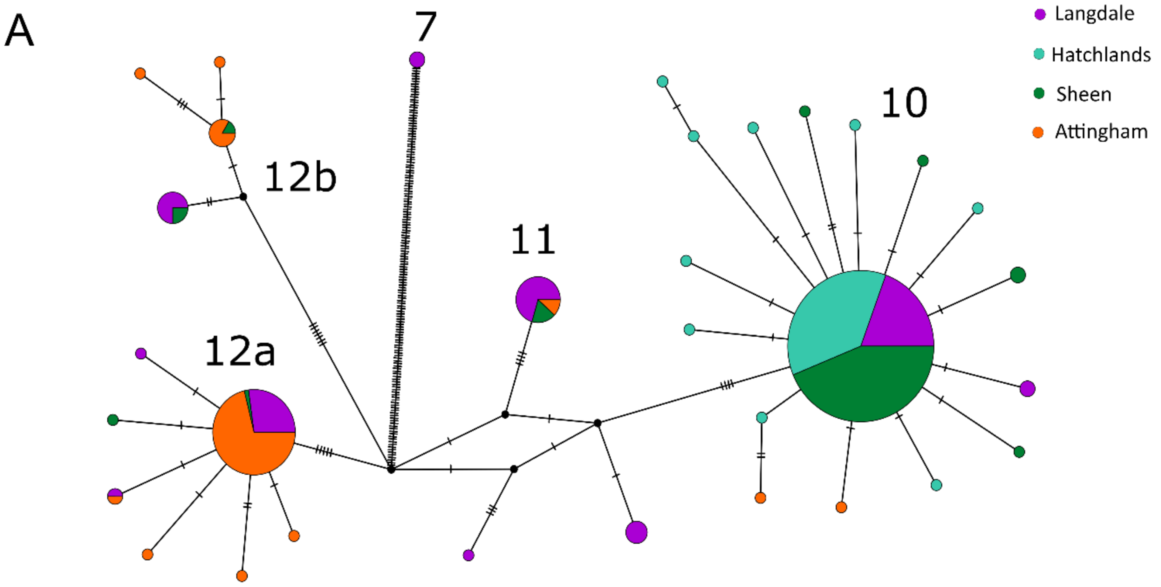
Examining genomic variability of *Q. robur* among all four populations with PCA, after excluding 99 individuals to remove any relatedness relationship above 0.05 between samples, PCA eigenvalues levelled off and the slight populations patterns observed prior to kinship filtering in Figure S2.4 (Supporting Information) vanished, reflecting the lack of any strong geographically correlated structure in the nuclear genomes of *Q. robur* individuals between the four parklands sampled (Figure S2.7, Supporting Information). Running fastSTRUCTURE on the 360 *Q. robur* individuals (K from one to ten) also confirmed this lack of structure within *Q. robur*: the two metrics (Køc and $K^*ε$) reported by fastSTRUCTURE function “chooseK” for the choice of ideal model complexity concordantly suggested a single ancestral population (Figure S2.1C, Supporting Information). The cross-validation profile further shows that the prediction error does not vary much with increasing model complexity and remains within one standard error from K=1 (Figure S2.1D, Supporting Information).

Linkage disequilibrium (LD) was found to decay quickly in the 360 *Q. robur* individuals, with average r^2 below 0.2 at 500 bases apart (Figure S2.8, Supporting Information). Over 90% of LD-blocks size estimates are within 5,000 bp length (Figure S2.8, Supporting Information). In windows of 5,000 bp, *Q. robur* pairwise nucleotide diversity showed genome-wide diversity π of ~0.007, with little variation between chromosomes and parklands sites (Figure S2.9-S2.10, Supporting Information). Nucleotide diversity was higher in the intergenic regions (π = 0.0066) than the coding regions (π = 0.0024).

Selective sweep scans every 5,000 bp along the 12 *Q. robur* chromosomes using SweeD (Pavlidis et al, 2013) and OmegaPlus (Alachiotis et al., 2012) each identified 1,311 outlier regions (1% of most extreme values). There were 10 common outlier regions found by both methods (Figure S2.11, Supporting Information), located on chromosomes 1 (2 regions), 2, 3 (2 regions), 9, 11 (3 regions) and 12. The gene models (Plomion et al., 2018) within or flanking the identified sweep regions included key developmental and stress response regulators (Table S2.2, Supporting Information), according to the available annotation (Plomion et al., 2018).

Chloroplast haplotypes phylogeny and identification

We *de novo* assembled and circularized the full chloroplast DNA sequence of 309 oak samples, including 287 *Q. robur*, 8 *Q. petraea* and 14 hybrids. Novoplasty was unable to fully assemble and circularize the chloroplast DNA of 77 samples, however we deemed the 309 complete sequences sufficient to assess the chloroplast haplotypes composition of both parklands and species. A median-joining plastid haplotype network identified five major chloroplast haplotypes, with lengths varying from 161,148 bp to 161,306 bp (Figure 2.5). We matched the five haplotypes to those previously identified using the PCR-RFLP method (Petit et al. 2002b). Four haplotypes (haplotypes 10, 11, 12a, 12b) displayed a particular point mutation in the trnD-trnT chloroplast fragment (Table S2.3E, Supporting Information) characteristic of haplotypes belonging to lineage "B", which hypothetically originated in a late Pleistocene glacial refugium located in the Atlantic side of the Iberian Peninsula (Cottrell et al., 2002; Petit et al., 2002a; Petit et al., 2002b). Similarly, haplotype 7 exhibited a particular point mutation in the trnT-trnF chloroplast fragment (Table S2.3E, Supporting Information) characteristic of haplotypes stemming from lineage "A", which originated in a Pleistocene glacial refugium located in the Balkans geographic area (Cottrell et al., 2002; Petit et al., 2002a; Petit et al., 2002b).



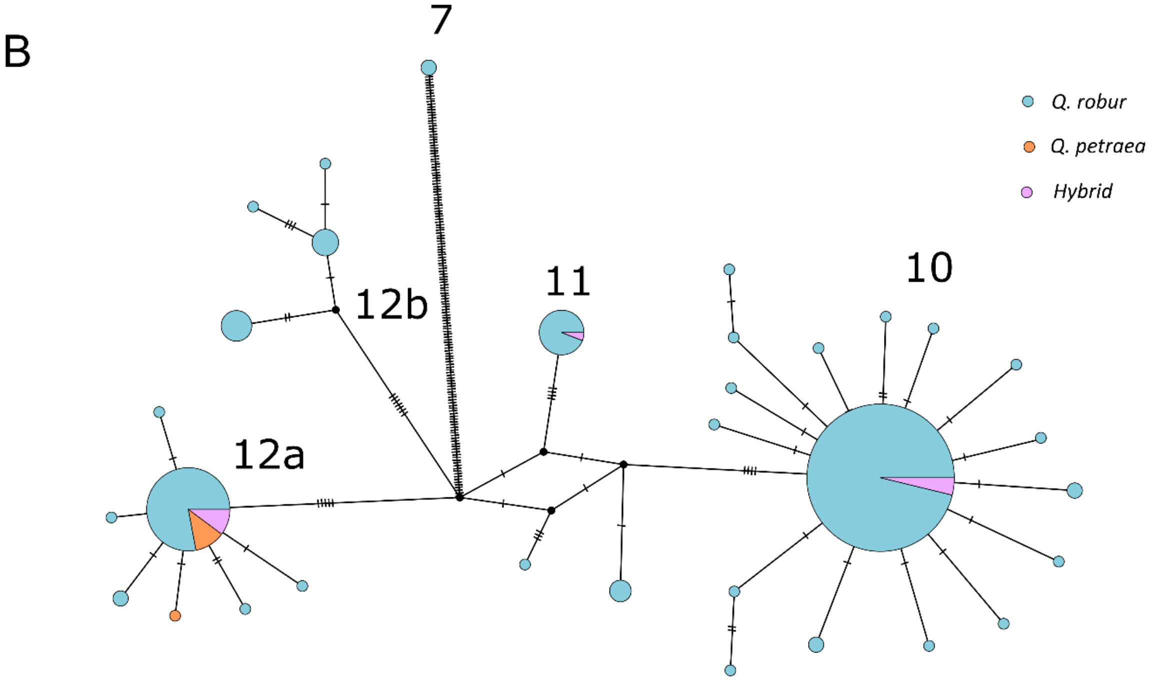


Figure 2.5. Median-joining chloroplast haplotype network of 287 *Q. robur*, 8 *Q. petraea* and 14 hybrid individuals. A) Colours represent parkland sites, and the size of nodes is proportional to number of individuals. Hatch marks along edges represent the number of single base mutations between nodes. Edges length does not carry any weight. The major nodes identified were matched to previously characterized haplotypes (Petit et al., 2002b). B) Colours represent species.

Haplotypes 10, 12a and 12b, described in Petit et al. (2002b), were identified according to the information provided by restriction digestion of the DT chloroplast fragment (Table S2.3A). Haplotype 11 displayed a 13 bp duplication in the TF chloroplast DNA fragments (Table S2.3D) characteristic of this haplotype (Cottrell et al., 2002). The remaining haplotype in our study was matched with either haplotype 7 or 26, two closely related Balkan haplotypes, according to the information provided by digestion of the DT and AS fragments (Table S2.3B, Supporting Information). We narrowed our choice to haplotype 7 due to previous reports of this haplotype in England (Cottrell et al., 2002); we excluded haplotype 26 as its presence was recorded only in the French alps and it is speculated to represent a post-colonization mutation of haplotype 7 (Petit et al., 2002a).

The two predominant haplotypes identified across the four sampled sites were haplotype 10 and 12 (Figure 2.6). Haplotype 10 is predominant at Sheen Wood and Hatchlands Park in the south-east. It is also well-represented at Langdale Wood; however, it is almost absent from Attingham Park, which is located further north-west in Shropshire. In contrast, haplotype 12 is nearly absent from the south-eastern sites while it is present at Langdale Wood, and it is the predominant haplotype at Attingham Park.

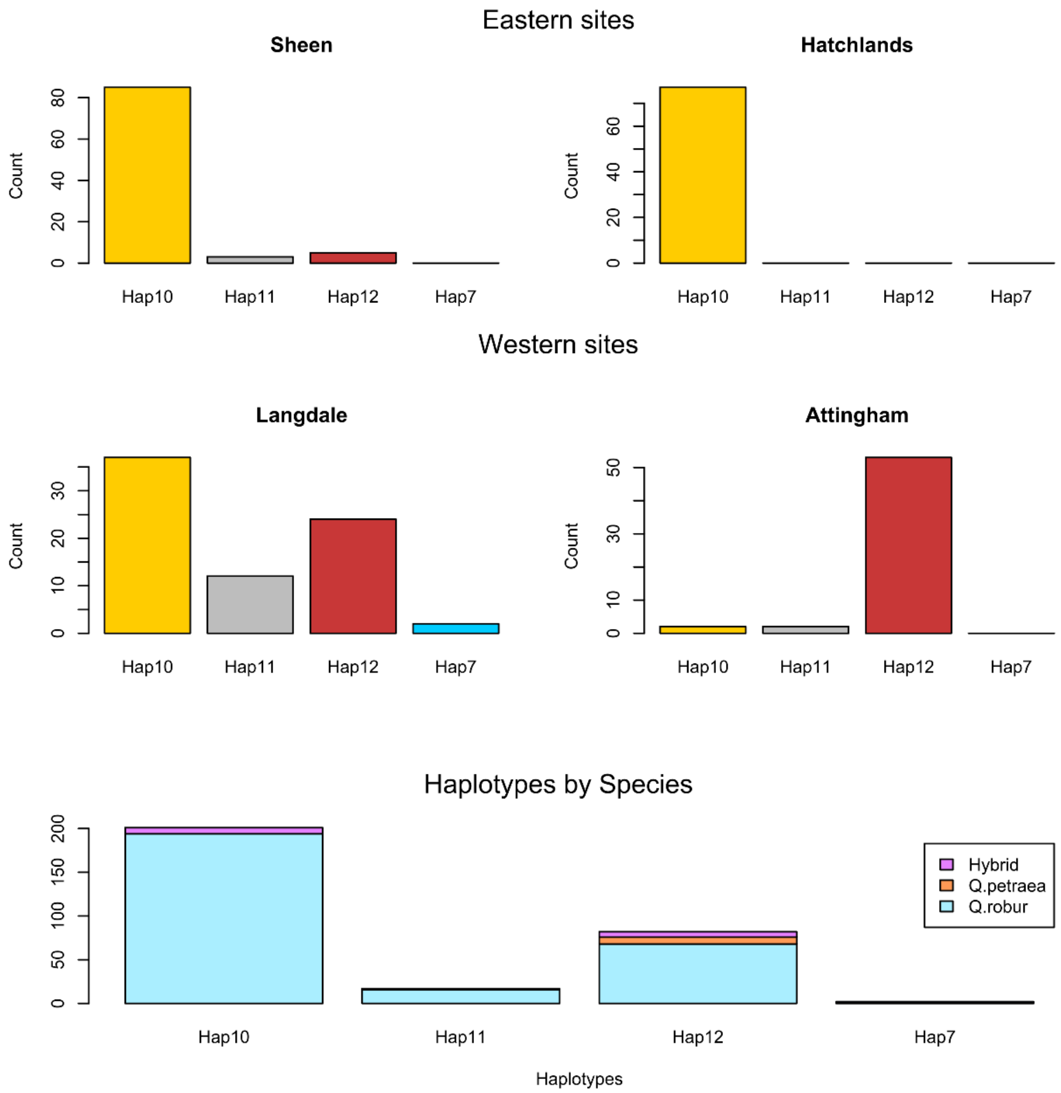


Figure 2.6. Distribution of the four chloroplast haplotypes identified between sampled parkland sites and species.

According to our ordinary kriging linear regression based on Cottrell et al. (2002) English and Welsh ancient woodlands data (Figure 2.7), Attingham Park (dominant haplotype 12) and Langdale Wood (dominant haplotype 10) did not occur in an area dominated (> 50%) by a single haplotype, but both sites are located in close proximity (< 25 km) to a matching dominant autochthonous haplotype patch. Hatchlands Park (dominant haplotype 10) and Sheen Wood (dominant haplotype 10) do not have the same dominant haplotypes as local ancient woodlands according to our interpolation; however, this area of England appears particularly fragmented and dominance regions for all three major British haplotypes (haplotypes 10, 11 and 12) are found in proximity (~50 km) to both sites (Figure 2.7). Haplotype 7 does not show any dominance region (Figure 2.7), unlike the interpolation in Lowe et al. (2004). That is because this haplotype is dominant (> 50% frequency) in only a single ancient woodland in the English and Welsh data retrieved from Cottrell et al. (2002), therefore no pixels in the map are assigned a value above 50% unless a very fine prediction grid is used in the interpolation. There was no distinction between species: most of the *Q. robur* individuals identified at Attingham Park shared the same haplotype of the *Q. petraea* individuals from the same site (Figure 2.6).

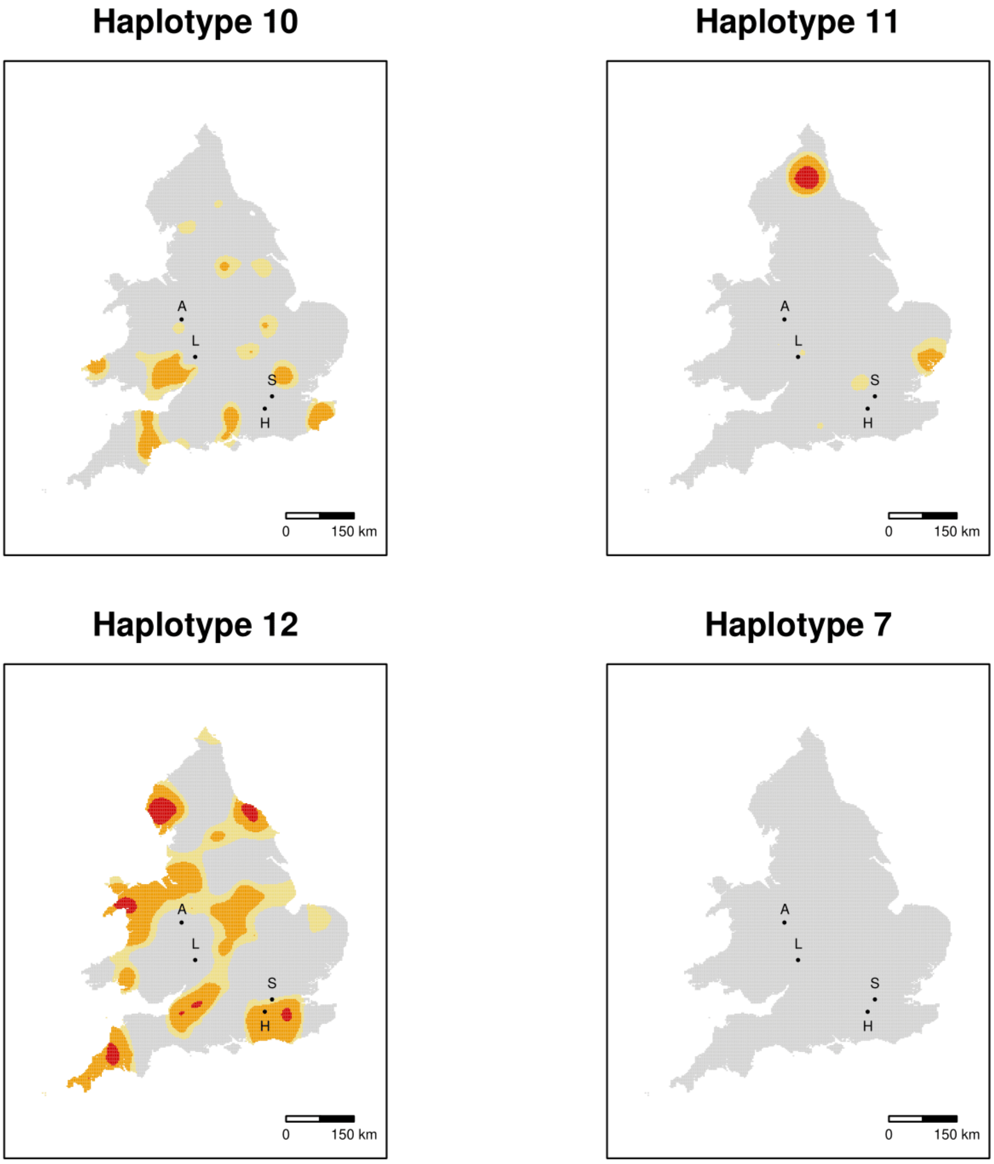


Figure 2.7. Haplotypes frequency contour regions in England and Wales estimated with ordinary kriging interpolation of the chloroplast data of 163 ancient woodlands from Cottrell et al. (2002). Yellow: > 50%. Orange: 60-70%. Red: > 70%. Black points represent the sites sampled in this study (A = Attingham Park, L = Langdale Wood, S = Sheen Wood, H = Hatchlands Park).

Discussion

Interspecific differentiation and population structure

We sequenced the whole genome of 386 oak trees across four British parkland sites and characterized over 50 million SNPs in British *Q. robur* and *Q. petraea*. Trees with evidence for hybridisation in their genomes generally exhibited predominantly the morphological traits of one of the parental species. Previous studies have found that oak hybrids often display uneven combinations of phenotypic characters of the parental species, especially where there has been recurrent backcrossing (Curtu, Gailing & Finkeldey, 2007; Beatty et al., 2016). Three individuals could represent first-generation (F1) hybrids in our samples (Figure 2.1, though additional analyses would be needed to confirm this hypothesis), while the other admixed trees showed evidence of backcrossing to *Q. robur*, likely due to the predominance of this species in the parklands surveyed (Lepais et al., 2009).

The genomes of *Q. robur* and *Q. petraea* were largely undifferentiated, corroborating previous results from sites across Europe (Barreneche et al., 1996; Bodenes et al., 1997; Coart et al., 2002; Leroy et al., 2019b; Lesur et al., 2018; Mariette et al., 2002; Saintagne et al., 2004; Scotti-Saintagne et al., 2004; Zanetto et al., 1994). We reported mean F_{st} of 0.155 between *Q. robur* and *Q. petraea* coding regions, close to recent estimates (mean F_{st} = ~0.13) by Lang et al. (2018), who studied populations in central and western Europe. We did not find any substantial difference between intergenic and genic F_{st} trends (Figure S2.2, Supporting Information), suggesting that there could be species-discriminant markers also in the neutral regions of the oak genome. We identified 81 genic regions, with strong species discriminatory power (Figure 2.3). Two of these genes were previously found in a study of genome-wide differentiation among four white oak species (*Q. robur*, *Q. petraea*, *Q. pyrenaica* and *Q. pubescens*) in France, which identified a total of 215 regions of differentiation, 133 of which contained genes (Leroy et al., 2019b). Our sample size of ten individuals of *Q. robur* and *Q. petraea* was very small, likely leading to false positives within our 81 regions. Future analyses with larger sample sizes including more locations and spanning a larger geographic area, including continental Europe, may reduce the number of F_{st} outliers detected between *Q. robur* and *Q. petraea*. Such a combined analysis would also greatly increase the applicability of the results across Europe.

The nuclear genomes showed no clear differentiation among the four parkland sites (Figure S2.4, Supporting Information), though we note that our sampling had small latitudinal range. If we had sampled more northerly and southerly parts of Britain, we may have found more among site variation. We identified 10 genes with, or in proximity to, signatures of recent positive selection in *Q. robur* based on both site frequency spectrum and linkage disequilibrium patterns (Figure S2.11, Supporting Information). The putative functions of these suggest that they may be involved in stress tolerance (Table S2.2, Supporting Information).

Population history and secondary contacts

Hybridization and extensive exchanges of chloroplast DNA between European oaks have been well documented (Dumolin-Lapegue, Kremer, & Petit, 1999; Lepais et al., 2009; Leroy et al., 2019b; Petit et al., 2002b; Petit et al., 1997) and most recent genetic studies based on ABC models have shown that massive secondary contacts between European white oak species have occurred recently after a long period of isolation, probably at the end of the last glacial period or at the start of the current interglacial period (Leroy et al., 2017; Leroy et al., 2019b). These contacts are thought to have homogenized the majority of the nuclear genome of European white oak species with the exception of some barrier regions accumulated during the long periods of isolation that preceded secondary contacts (Lepais et al., 2013; Leroy et al., 2019b; Petit et al., 2002a). Our estimates of past effective population sizes suggested that *Q. robur* and *Q. petraea* shared a similar history of population size decline during the last 2.5 million years followed by a recent postglacial re-expansion. Due to uncertainties about generation time and mutation rates in oaks, we do not know if the onset of this apparent decline was when Eurasian white oaks colonised Europe from America, or if it began due to Pleistocene glaciations (Cottrell et al., 2002; Leroy et al, 2017; Mazet et al., 2015; Petit et al., 2002a; Petit et al., 2002b). The steep increase in population sizes recorded in more recent history in both species using 8 haplotypes seems to reflect the post-glacial expansion of white oaks in Europe that is reported to have taken place after the last glacial maximum (Cottrell et al., 2002; Leroy et al, 2017;

Petit et al., 2002a; Petit et al., 2002b). Our estimation of cross-coalescence rates is concordant with the hypothesis that despite the extensive interspecific gene-flow in the late Pleistocene, the species were already fully diverged (Leroy et al., 2017; Leroy et al., 2019b), and their separation appear to have started roughly 8 Mya, in line with the latest oak phylogeny (Hipp et al., 2019).

Relatedness and parklands management

The differing levels of relatedness that we found among trees within parklands could reflect different management practices in different parklands in the past (Figure S2.6, Supporting Information). Trees in Attingham Park had particularly high relatedness and we identified two potential parent/offspring relationships, two pairs of full-sibs and numerous second- and third-degree relatives there. Trees in Langdale Wood and Sheen Wood had slightly lower levels of relatedness however potential full-sibs (five at Langdale Wood and three at Sheen Wood) and several second- and third-degree relatives can be found at each site. Hatchlands Park showed substantially less kinship. It is likely that sites with higher levels of relatedness had high natural regeneration in the past, and perhaps planting of locally sourced acorns.

British parkland chloroplast haplotypes

We detected four native British oak chloroplast haplotypes (Haplotypes 10, 11, 12 and 7) in the parklands we studied, and no haplotypes that could be attributed to planting of non-native seed stocks (Figure 2.5, Figure 2.6). The vast majority of our samples (> 99%) possess haplotypes thought to be derived from Iberian glacial refugia, but two individuals at Langdale Wood displayed haplotype 7 (Lineage “A”), from a Balkan refugium (Figure 2.6). This Balkan haplotype is rare in Britain but has had a long presence in the Forest of Dean and surrounding area, where it is the dominant haplotype of the oldest trees group (133-281 years old) and has been detected in a 320-year-old tree (Cottrell et al., 2004). Langdale Wood, where we identified this haplotype, is located approximately within 50 km from the Forest of Dean. It is not known if this haplotype reached the Welsh marches naturally from the Balkans through natural postglacial colonization or if human-mediated activity, such as the return of medieval knights from crusades, is responsible for the scattered presence of this haplotype in the north of France and around the Forest of Dean (Cottrell et al., 2002; Cottrell et al., 2004).

The chloroplast haplotypes present in the four parklands fit well with local ancient woodland haplotype distributions (Cottrell et al., 2002; Lowe et al., 2004). According to our kriging interpolation of dominant haplotypes in Britain, based on Cottrell et al. (2002) English and Welsh data, all four parkland sites surveyed are located within a short radius (25-50 km) of a matching autochthonous dominant haplotype region (Figure 2.7). This suggests that these parkland oaks derive from local seed sources, but we cannot fully exclude the possibility that seeds could have been imported from different regions of Britain with the same dominant haplotype, or even from Spain or France where these haplotypes are similarly abundant (Lowe et al., 2004; Petit et al., 2002a; Petit et al., 2002b). A strategy to further validate this finding would involve comparing global genomic relatedness levels between our samples and natural populations from Britain and populations from continental Europe. While data for parts of Europe is available (Leroy et al., 2019b), there is not, to our knowledge, any whole genome dataset for natural populations of Britain native oak species available to perform such comparison.

Concluding remarks

The large genomic dataset presented in this study provides new insights into the origins and past management of oaks in British parklands. This may assist future management decisions: if continuity is sought with the past, then local seed-sourcing will be appropriate, but for adaptation to future climates, seed sourcing from further afield may be needed. The genetic diversity present in these parkland populations may provide some resilience to climate change and may already have been drawn upon to some extent, via selective sweeps for which we found evidence in the *Q. robur* genome. A warmer and drier climate in future may favour *Q. petraea* ecology over *Q. robur*, so the proportion of *Q. petraea* and hybrid individuals may increase in future. An important health concern for British parklands is acute oak decline. All four of the parklands in this study are under long-term monitoring for this complex multifactorial syndrome (Denman et al., 2018). Further expansion of our dataset may give sample sizes sufficient to investigate whether or not susceptibility to this syndrome has a genetic basis.

Contributions

Nathan Brown sampled and phenotyped the trees, Tim Coker extracted DNA, William Plumb extracted DNA, Jonathan Stocks extracted DNA, Sandra Denman selected the sites and organised the sampling. Gabriele Nocchi (I, the author) performed all analyses and wrote the chapter. Richard Buggs obtained funding and oversaw the project.

Data Accessibility

DNA sequences: trimmed sequencing reads have been deposited in the European Nucleotide Archive under the Project Accession no. PRJEB30573 which can be found at https://www.ebi.ac.uk/ena/browser/view/PRJEB30573. Sample metadata are available in the supplementary spreadsheet provided (Supporting Information).

Chapter 3: Genomic signals of local adaptation in Asian White Birch

The research presented in this chapter is currently being submitted for publication and I will be the lead author.

Abstract

To survey the patterns of genome-wide diversity in a pan-Eurasian tree species complex and understand their underlying causes is a daunting task, but one with important implications for taxonomy, conservation, and forestry. Here, we investigate the population genetic structure of white birches in Eurasia and some of the processes that have driven their patterns of diversity in China. We generate whole genome sequence data from 83 individuals across the species range in China. Combining this with an existing dataset for 79 European and Russian white birches, we show a clear distinction between *B. pendula* and *B. platyphylla*, which have sometimes been lumped taxonomically. Genomic diversity in north-western China and Central Russia is affected greatly by hybridisation between these two species. Within *B. platyphylla* in China we give evidence for three lineages and hypothesise divergence times and past population sizes for them. We show patterns of co-variation between allele frequencies and environmental variables in *B. platyphylla*, suggesting the role of natural selection in the distribution of diversity at 7,609 SNPs. Interestingly, 3,767 of these SNPs were also statistically significant outliers in our population differentiation analysis. The putative adaptive SNPs are distributed throughout the genome and span 1,633 genic regions. Of these genic regions, 87 were previously identified as candidates for selective sweeps in *B. pendula*. We use the 7,609 environmentally associated SNPs to estimate the risk of non-adaptedness for each *B. platyphylla* individual under a scenario of future climate change, highlighting areas where populations may be under future threat from rising temperatures.

Introduction

Understanding the genomic basis of local adaptation may help to inform species conservation and predict population fitness under climate change (Sork et al., 2013). Local adaptation occurs when environmental conditions impose selective pressure on phenotypes related to local performance, such as survival and fecundity, resulting in the higher fitness of certain genotypes, which increase in local abundance and allele frequency controlling these traits will vary across the landscape (Fisher, 1930; Savolainen et al., 2007). Adaptation to environmental conditions can result from standing genetic variation or new mutations (Aitken et al. 2008). Local adaptation may be more obvious for species occupying a broad range as the environments are more likely to be spatially heterogeneous. However, spatial genetic variation can also result from neutral processes, such as genetic drift due to demographic history (Sork et al., 2013; Wang & Bradburd, 2014). Hence, controlling for neutral genetic structures helps to reduce false positives when aiming to identify outlier loci among populations involved in local adaptation.

Forest trees, usually with large population sizes and wide geographic distributions, are an ideal system to study local adaptation, as they occupy spatially heterogeneous environments, which often results in significant differentiation among populations and distinct genetic structures (Savolainen et al., 2007; Alberto et al., 2013). Rapid climate change threatens forest trees greatly, especially those with long generation times, as this translates to slower adaptation, and short dispersal, which would limit their ability to track and migrate to more suitable environments (Aitken et al., 2008; Keenan, 2015).

Recently, the genetic basis of local adaptation has been investigated for certain non-model species using high-throughput sequencing, high-resolution environmental data and novel analytical methods (Eckert et al., 2010; Ellegren, 2014). Environmental association analysis (EAA) correlates abiotic data with genomic data, such as SNP allele frequencies (Manel and Holderegger 2013; Sork et al. 2013). Such studies pinpoint SNPs showing signals of elevated differentiation among environments and showing significant correlations with one or more environmental variables (Rellstab et al., 2015), which are putative candidates involved in local adaptation (Hoban et al., 2016). Using EAA, we can also predict the possible degree of maladaptation of local populations to future climate change (Borrell et al., 2019; Rellstab et al., 2021; Rellstab et al., 2016).

Betula (birch), an important genus in the Northern Hemisphere, including both widespread and highly localized species, comprises key components of temperate forests (Ashburner and McAllister, 2013; Wang et al., 2016; Wang et al., 2021). Understanding the genomic basis of adaptation in this genus can help to predict their potential to survive under rapid climate change, helping to manage species conservation (Borrell et al., 2019).

In this study, we selected Asian white birch (*Betula platyphylla*), which belongs to the subgenus *Betula*, to understand its past demographical histories, current local adaptation and future potential to respond to climate change. Asian white birch is of great ecological and commercial value and commonly occurs in eastern Asia including the Far East Russia, China and Japan. In China, it has distributions in northwestern (NW), central (CE), northeastern (NE) and southwestern regions (SW) (Chen & Lou, 2019; Ashburner and McAllister, 2013). Asian white birch is a pioneer species, colonizing open habitats in grasslands, on roadsides and on the verge of forest (Ashburner and McAllister, 2013). We selected this species not only because of the above-mentioned values but also for the following reasons. First, it occupies spatial heterogeneous environments across its range with over 26 latitudinal degrees based on our collection records. Some populations distribute in SW China and its adjacent regions, with an altitude of over 3,700 meters whereas some populations distributing in NW China grow by riverside with a much higher annual evaporation than precipitation. Furthermore, the well assembled and annotated reference genomes of both silver birch (*B. pendula*) (Salojärvi et al., 2017) and Asian white birch (Chen et al., 2021) have been available, enhancing the potential to study genomic footprints of adaptation in Asian white birch. Asian white birch and silver birch are sister species (Wang et al., 2021) and some researchers have suggested that they could be a single species (Ashburner & McAllister, 2013). They diverged during the last glaciation before the LGM (approx. 36 000 years ago), according to previous research, but this inferred divergence time is probably underestimated as gene-flow after divergence was not included in the model and the inference is based on only 18 nuclear simple sequence repeats (SSRs) loci (Tsuda et al., 2017). More recently, a phylogenetic tree based on gene families was built and re-estimated the divergence time between *B. pendula* and *B. platyphylla* at 2.6 million years ago (Chen et al., 2021). A previous landscape genomic study of silver birch showed that variations around the genes involved in photoperiodic responses were highly associated with local environmental adaptation (Salojärvi et al., 2017). More recently, *B. platyphylla* was shown to be divided into five genetic clusters and many populations show substantial genetic admixtures between different genetic clusters (Chen & Lou, 2019). However, this investigation was also based on a small number (ten) of microsatellite markers, therefore very little is known about demography history and landscape genomics of Asian white birch until now despite its wide distribution in East Asia.

In this study, population stratification of *B. platyphylla* was first investigated using principal component analysis (PCA), F_{st} and fastSTRUCTURE analyses. We then used environmental niche modelling (ENM) to characterize *B. platyphylla* habitat under current and future climate. We coupled it with EAA to identify SNP markers associated with environmental clines after controlling for population structure. The annotated genome sequence of silver birch was used to predict the function of the identified adaptive SNPs and to locate the distribution of these adaptive SNPs along the genome. Finally, the risk of non-adaptedness (RONA) (Rellstab et al., 2016) was calculated for each individual under a scenario of climate of 2080-2100, which would inform species conservation strategies under conditions of future climate change. Together, this study generates new hypotheses regarding local adaptation in Asian white birch, which can be further tested in common garden trials or genetic modification analyses in future studies.

Materials and Methods

Sampling

Samples of Asian white birch (*B. platyphylla*) were collected from 74 naturally occurring populations throughout its distribution in China between 2016 and 2019. Sample identifiers and GPS locations are presented in Table S3.1 (Supporting Information). Dried cambial tissues were used for DNA extraction following a modified CTAB protocol (Wang et al., 2013). One or two samples per population were selected for resequencing, resulting in a total of 83 samples. Extracted DNA was assessed with 1% agarose gel and sent to BerryGenomics (Beijing, China) for library preparation and whole genome re-sequencing. Sequencing was performed on the Illumina NovaSeq6000 platform with 150 bp paired end read sequencing with approximate 11Gbp data obtained for each sample (average depth of ~25x).

Trimming and SNP filtering

The raw data was trimmed using Trimmomatic (Bolger et al., 2014) in paired-end mode with a required quality of 30. Reads shorter than 90 bp were discarded. Clean reads of the sequenced samples were aligned to a genome assembly of *B. pendula* (Salojärvi et al., 2017) using BWA-MEM v.0.7.17-r1188 algorithm in BWA (v0.7.17) with default parameters (Li & Durbin, 2009). Reads with non-specific matches were discarded. Alignments were converted from sequence alignment map (SAM) format to sorted, indexed binary alignment map (BAM) files (SAMtools v1.8) (Li et al., 2009). The MarkDuplicates tool from the Genome Analysis Tool Kit (GATK) (v 4.1.4.) was used to mark duplicates (DePristo et al., 2011; McKenna et al., 2010). Variant calling was performed the software Haplotype Caller in joint genotyping mode, available in GATK v4.0.8.1 (DePristo et al., 2011). We filtered for biallelic SNPs genotyped in all individuals with minor allele frequency greater than 0.05, quality by depth greater than 2, fisher strand test less than 60, root mean square mapping quality greater than 40, mapping quality rank sum test greater than -12, read position rank sum test greater than -8 and strand odds ratio less than 3. We pruned the filtered SNP set by linkage disequilibrium (r^2 > 0.4) using the "indep-pairphase" function of Plink (Chang et al., 2015) in windows of 50 and step of 5. We term this set of SNPs the China dataset.

Using the same method to the above we also assembled a second SNP dataset that included genomic read data from Salojarvi et al. (2017) for 78 *B. pendula* individuals from most of this species' European and Asian geographic range, and one *B. platyphylla* individual deliberately included from Russia. These individuals were included because *B. pendula* and *B. platyphylla* are sister species that are reported to hybridise (Tsuda et al., 2017). We term this set of SNPs the Eurasian dataset.

Population structure

We used *fastSTRUCTURE* (Raj, Stephens, & Pritchard, 2014) on the China dataset and also on the Eurasian dataset of SNPs to assign individuals to populations. We used a simple prior, 5-fold cross-validation and number of ancestral populations (K) from one to ten. We used the python function *chooseK*, which is the recommended tool for model selection in *fastSTRUCTURE* (Raj, Stephens & Pritchard, 2014), and the cross-validation profile, which identifies the model complexity for which the prediction error is minimized, to assess the models generated. The function *chooseK* suggests two values for K: one that maximises the log-marginal likelihood lower bound (LLBO) of the dataset (K*ε) and captures strong structure, and another one (Køc) which reports the model components that have a cumulative ancestry contribution of at least 99% and it is aimed at finding additional weak underlying structure (Raj, Stephens, & Pritchard, 2014). Individual trees were assigned to populations according to the value of the admixture coefficient (q) computed by *fastSTRUCTURE* at the chosen model complexity: $q \geq 0.9$ to belong to a population, while individuals with $q < 0.9$ were classified as admixed.

On the basis of the fastSTRUCTURE analysis of the Eurasian dataset, we excluded from the China dataset any individuals that had over 10% assignment to *B. pendula* and ran *fastSTRUCTURE* only on Chinese *B. platyphylla* based on 1,387,994 SNPs (after filtering for missing genotypes). Furthermore, we used the program sparse non-negative matrix factorization (*snmf*), part of LEA package (Caye et al., 2019), to perform a population differentiation test and identify F_{st} outlier SNPs between the clusters identified for *B. platyphylla* in China. *snmf* computes fixation also in the presence of admixed individuals in the sample (Martins et al., 2016) and performs a test similar to other F_{st}-outlier approaches (Lotterhos & Whitlock, 2014). Multiple testing was controlled by converting the p-values computed by *snmf* into q-values with the R package "qvalue" (Storey et al., 2015). Q-values represent the expected false discovery rate (FDR) associated with a given p-value (Storey and Tibshirani 2003). Finally, we selected candidate outlier SNPs with FDR < 1%.

Principal component analysis (PCA) was performed on the Eurasian dataset and the Chinese *B. platyphylla* dataset using *Plink v2.0* (Chang et al., 2015).

Population size history and species divergence

We inferred the historical changes in population size and the demographic history of the three *B. platyphylla* populations identified in China using a multiple sequentially Markovian coalescent approach implemented in the software MSMC (Schiffels & Durbin, 2014). We randomly chose two individuals for each population (excluding admixed individuals, q < 0.9 in fastSTRUCTURE) for this analysis, corresponding to four haplotypes. We used the biallelic SNPs set prior to minor allele frequency filtering, SNP call rate filtering and linkage disequilibrium pruning (28,058,885 SNPs in total). Statistical phasing was performed using Beagle 5.1 with default settings (Browning & Browning, 2007). MSMC outputs times and population sizes scaled by mutation rate per base pair per generation. To convert the output to real time and sizes it is necessary to divide estimates by the mutation rate and further multiply these by the generation time. We used the substitution rate 7.7×10^{-9} of peach (*Prunus persica*), and generation time of 40 years, as suggested for *B. pendula* in Salojarvi et al. (2017).

Population diversity and differentiation

We estimated linkage disequilibrium decay along the *B. platyphylla* genome with the tool *PopLDdecay* (Zhang, Dong, Xu, He, & Yang, 2018) separately for the different groups identified for *B. platyphylla* in China by *fastSTRUCTURE* at K = 3 (Figure 3.3). For this calculation we used the SNP set prior to linkage disequilibrium pruning but filtered to retain SNPs with minor allele frequency > 0.05. After excluding individuals admixed between the three populations (admixture coefficient < 0.9 in fastSTRUCTURE at K = 3), Weir and Cockerham F_{st} (Weir & Cockerham, 1984) was calculated between the three *B. platyphylla* populations on a per SNP site basis with the *vcftools* function *weir-fst-pop* (Danecek et al., 2011). We calculated nucleotide diversity π (Nei & Li, 1979) within each population using the *vcftools* (Danecek et al., 2011) function *window-pi* in non-overlapping windows of 5 kb. For the π calculation we used the SNPs set prior to minor allele frequency filtering and linkage disequilibrium pruning (28,058,885 SNPs in total).

Environmental niche modelling

We used environmental niche modelling (ENM) to characterize the present habitat and predict the future distribution of *B. platyphylla* in China and to identify the climatic variables influencing this species distribution. ENM analysis was performed by using the samples' geographic locations in this study and observation records of this species from 1970 to 2010 in China, sourced from the Global Biodiversity Information Facility (http://www.gbif.org). To mitigate the effects of spatial autocorrelation, we removed records within 0.2 degrees (approximately 22-25 km) from one another, which resulted in a total of 138 presence locations: 66 records were *B. platyphylla* from the 83 individuals sampled (excluding admixed *B. pendula* - *B. platyphylla* individuals) and another 72 records were from the GBIF. We downloaded 19 bioclimatic variables related to temperature and precipitation from the WorldClim database (www.worldclim.org) at 1km resolution (Hijmans, Cameron, Parra, Jones, & Jarvis, 2005) for the period 1970-2000 representing "current climate", as this is the estimated time period of establishment of the sampled trees and also corresponds to the time that most GBIF observations included in the model were recorded. We downloaded elevation data at the same resolution and used it to compute slope, aspect and four additional terrain related characteristics (topographic position index, terrain ruggedness index, terrain roughness and water flow direction) using the *raster* R package (Hijmans & Etten, 2012). In order to avoid overfitting, we selected environmental variables with Pearson’s correlation coefficient < 0.7, preferring annual rather than monthly or quarterly values, which resulted in 11 variables retained for the ENM (Table 3.1, Figure S3.1-3.2, Supporting Information). We assembled eight further datasets with the same 11 variables under four Shared Socioeconomic Pathways (SSPs) defined by the Intergovernmental Panel on Climate Change sixth Assessment (Masson-Delmotte et al., 2021) at each of two future time points (2041–2060 and 2081–2100). The ENMs were generated using Maxent (Phillips, Anderson, & Schapire, 2006) with 50 subsampled replicate 5000 iterations runs with 20% of observations left-out for cross-validation. Variables’ importance was assessed with jack-knife tests and multiple models were generated and evaluated using a test of omission rate and area under the receiver operating characteristic curve (AUC). An environment suitability threshold was defined by “maximum training sensitivity plus specificity,” which should optimize the trade-off between commission and omission errors (Borrell et al., 2019; Liu, Newell, & White, 2016).

Genome environment association analysis: identification of adaptive SNPs

We used *LFMM2* (Caye, Jumentier, Lepeule, & Francois, 2019) to test for associations between environmental variables and SNPs in 71 Chinese *B. platyphylla* individuals. *LFMM2* performs multivariate linear regressions to evaluate the association between a response matrix, corresponding to SNP frequencies of individuals, and a matrix of environmental variables. It combines the environmental fixed effects with latent effects, which are unobserved confounding effects due to population structure (Frichot et al., 2013; Caye, Jumentier, Lepeule, & Francois, 2019). *LFMM2* has a two steps approach: first it estimates the latent factors and then tests for the genotype-environment associations. Each SNP locus is tested separately using the latent scores as covariates to control for demographic history and population structure. The null hypothesis is that the environmental variables have no effects on the SNP frequency, which is tested using a student distribution with n−K−1 degrees of freedom (Caye, Jumentier, Lepeule, & Francois, 2019). *LFMM2* requires to specify the number (K) of latent factors to include in the model which was determined according to *fastSTRUCTURE* cross-validation schemes. We chose the more conservative K = 3 (see Results section) to run *LFMM2*, including all the 11 uncorrelated environmental variables used in the ENM (Table 3.1). Multiple testing was controlled by converting the p-values computed by *LFMM2* into q-values with the R package “qvalue” (Storey et al., 2015). Finally, we selected candidate adaptive SNPs with FDR < 1%.

For comparison, we also searched for signals of local adaptation with Samβada (Stucki et al., 2016). Samβada employs multivariate logistic regression to model probability of having a particular genotype given the environmental conditions at the sampled locations (Joost et al. 2007) and measures the environmental and molecular global and local spatial autocorrelation (Stucki et al., 2016). In Samβada a genotype is treated as a binary trait (presence/absence) therefore each biallelic SNP has three possible genotypes which are tested independently: homozygous reference, heterozygous, homozygous alternate (Stucki et al., 2017). Samβada compares each model with a constant model where the probabilities of the presence of a genotype is the same at each location in the landscape and is equal to its frequency in the dataset. The models are fitted with a maximum likelihood approach and significance is assessed with both log-likelihood ratio (G) and Wald tests (Stucki et al., 2016). Samβada allows the user to explicitly account for population structure, which is suggested to include in form of the significant axes of PCA on the ancestry coefficients, to avoid the collinearity of admixture coefficients (Stucki et al., 2016). The proposed approach in Samβada requires the user to build a model with one or several population structure variables and one environmental variable (the alternative model) and a corresponding null model including only the population structure variables. The significance of the environmental variable is calculated by comparison with the assumption that the G-score (twice the difference in log likelihood between the models) is distributed as X^2 (Stucki et al., 2016). We tested each environmental variable separately in Samβada, as suggested, because the aim of this program is to detect which SNP loci are potentially locally adapted rather than making predictions for the genotype of an individual based on its environment (Stucki et al., 2016). Q-values were computed from p-values with the same method used for LFMM2 (Storey et al., 2015). We allowed for co-ancestry by including the axes of the PCA on admixture coefficients computed with fastSTRUCTURE at K = 3. We considered a SNP locus a candidate if at least one of the three possible genotypes reported FDR < 1%.

Characterization of candidate adaptive SNPs

We performed two further principal component analyses of the Chinese *B. platyphylla* samples with *Plink* (Chang et al., 2015): one including only the putative adaptive SNPs (detected by *LFMM2* (K = 3) with FDR < 1%) and another only including a similar number of putatively neutral SNPs. The neutral set of 7,500 SNPs was generated by selecting SNP loci at random from those that reported a q-value above 0.4 across all environmental variables in the *LFMM2* analysis. Weir and Cockerham F_{st} between the three *B. platyphylla* populations was re-computed based on the adaptive and neutral SNP sets with the *vcftools* function *weir-fst-pop* (Danecek et al., 2011) after excluding admixed individuals (admixture coefficient < 0.9 in fastSTRUCTURE at K = 3).

In order to functionally annotate our putative adaptive SNPs, we performed a functional annotation of the whole *B. pendula* reference gene set provided in Salojärvi et al. (2017), retrieving gene ontology (GO) terms associated with the coding regions (Ashburner et al., 2000) using *Omics Box* (Conesa & Götz, 2008). We used *Omics Box* default functional annotation workflow which incorporates BLAST, run on the Viridiplantae non-redundant database, and Interproscan (Zdobnov & Apweiler, 2001). We then identified the genic regions overlapping our candidate adaptive SNPs with *bedtools* v2.28.0 (Quinlan & Hall, 2010) and performed a functional enrichment analysis of this adaptive gene set against the fully annotated reference gene set using Fisher's exact test in *Omics Box* (Conesa & Götz, 2008). We applied multiple testing corrections (FDR < 5%) and reduced the resulting significantly enriched GOs to the most specific term in the hierarchy for all three ontology levels: biological process, molecular function and cellular component (Ashburner et al., 2000).

Adaptation to future climate

We carried out risk of non-adaptedness (RONA) analysis using a slight modification of the method of Rellstab et al. (2016), implemented in the software *pyRona* (Pina-Martins, Baptista, Pappas & Paulo, 2018). RONA represents an estimate of the average change in allele frequencies at climate associated SNP loci required in a population to cope with future climatic conditions and is calculated separately for each environmental variable (Rellstab et al., 2016). RONA uses simple linear regressions of the selected current climate environmental variable and the alternative allele frequencies of the candidate loci identified in genome-environment association analyses (Rellstab et al., 2016). The regression coefficients of the environmentally associated loci are then used to predict the allele frequencies of the adaptive loci at a future climate value for the chosen environmental variable (Rellstab et al., 2016). The main improvement in the implementation by Pina-Martins, Baptista, Pappas & Paulo (2018) is that the RONA for an environmental variable is given by the weighted mean RONA of all relevant SNP loci for that environmental variable, weighted by the r^2 value of each locus regression, while the original method (Rellstab et al., 2016) uses unweighted means. We computed RONA on an individual sample basis for the seven uncorrelated environmental variables that are expected to change in the future, therefore excluding elevation and its derived measures. We used as future climate the profile ssp370 for 2080-2100 (Figure 3.5C). For each environmental variable, RONA was calculated by using the candidate adaptive SNP loci identified with *LFMM2* (FDR < 1%). For each individual tree we then calculated weighted mean RONA, giving each environmental variable RONA a weight equivalent to the contribution reported by that variable in the ENM (Table 3.2). Furthermore, we reported the maximum RONA for each individual, out of the seven variables for which it was calculated.

Results

SNP discovery

After filtering, the number of reads retained for each sample in the China dataset that we generated ranged between 66,016,635 and 134,129,672 and between 91.5% and 96.7% of these reads per individual were mapped to the reference genome (Table S3.1, Supporting Information). SNPs filtering of the 83 individuals resulted in 1,497,547 unlinked (r^2< 0.4) SNPs. After filtering, the reads retained for each sample in the Salojarvi et al. (2017) dataset were mapped to the reference genome. The combined Eurasian dataset resulted in 278,717 SNPs after filtering for missing genotypes. When the China dataset was restricted to pure *B. platyphylla* (71 individuals in SW, CE and NE China) we were left with 1,387,994 SNPs.

Population structure, diversity and differentiation

For the China dataset *fastSTRUCTURE* analysis with K = 2 separates the samples in north-western China (Xinjiang) from the other populations (Figure 3.1A). At K = 3 south-western and north-eastern populations are further separated and at K = 4 a central population is revealed (Figure 3.1C). The model that maximises the log-marginal likelihood lower bound (LLBO) of the data and that best explains additional weak underlying structure was K = 3, as suggested by the function "*chooseK*" (Figure S3.3, Supporting Information), therefore we produced a map of the individuals based on the K = 3 assignment (Figure 3.1D).

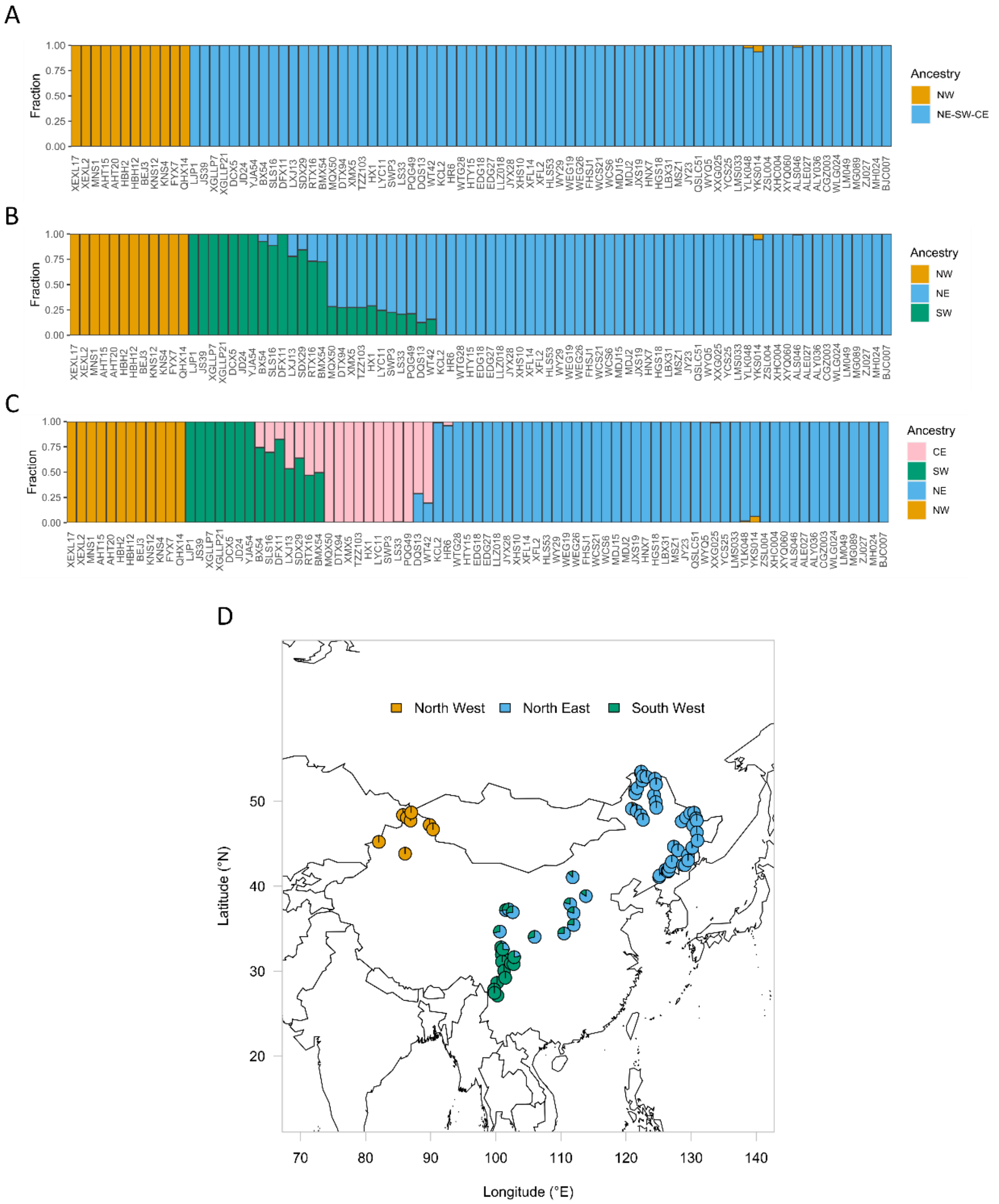


Figure 3.1. *fastSTRUCTURE* results for the China dataset, ordered by longitude. A) Bar plot of *fastSTRUCTURE* results of 83 sampled birch individuals, based on 1,497,547 unlinked SNPs ($r^2 < 0.4$) at K = 2, B) K=3 and C) K =4. D) Map of the individuals showing their genetic ancestry composition according to fastSTRUCTURE at the ideal model complexity, K = 3.

For the Eurasian dataset at K = 2, one cluster contains the *B. pendula* individuals from Salojarvi et al. (2017) and the other cluster contains 71 of the 83 individuals that we sequenced from China and the Russian *B. platyphylla* individual from Salojarvi et al. (2019). The 12 individuals we sampled in north-western China (Xinjiang) are hybrids between *B. pendula* and *B. platyphylla* (Figure 3.2A). At K = 3, the pattern was similar but a south-western China *B. platyphylla* cluster separates from a north-eastern China *B. platyphylla* cluster (Figure 3.2B). At K = 4, the only difference was that Irish and two Finnish *B. pendula* samples showed an unknown component (Figure 3.2C). The *fastSTRUCTURE* model that maximises the log-marginal likelihood lower bound (LLBO) of the data and that best explains additional weak underlying structure was K = 4 (Figure 3.4C-D), as suggested by the function choose K (Figure S3.4, Supporting Information). The LLBO curve and the cross-validation profile show that the marginal likelihood plateau at K = 2, and so does the prediction error, which stays within 1 standard error from K = 2 with added complexity (Figure S3.4A-B). These separations suggested by fastSTRUCTURE are reflected in the PCA (Figure S3.5, Supporting Information).

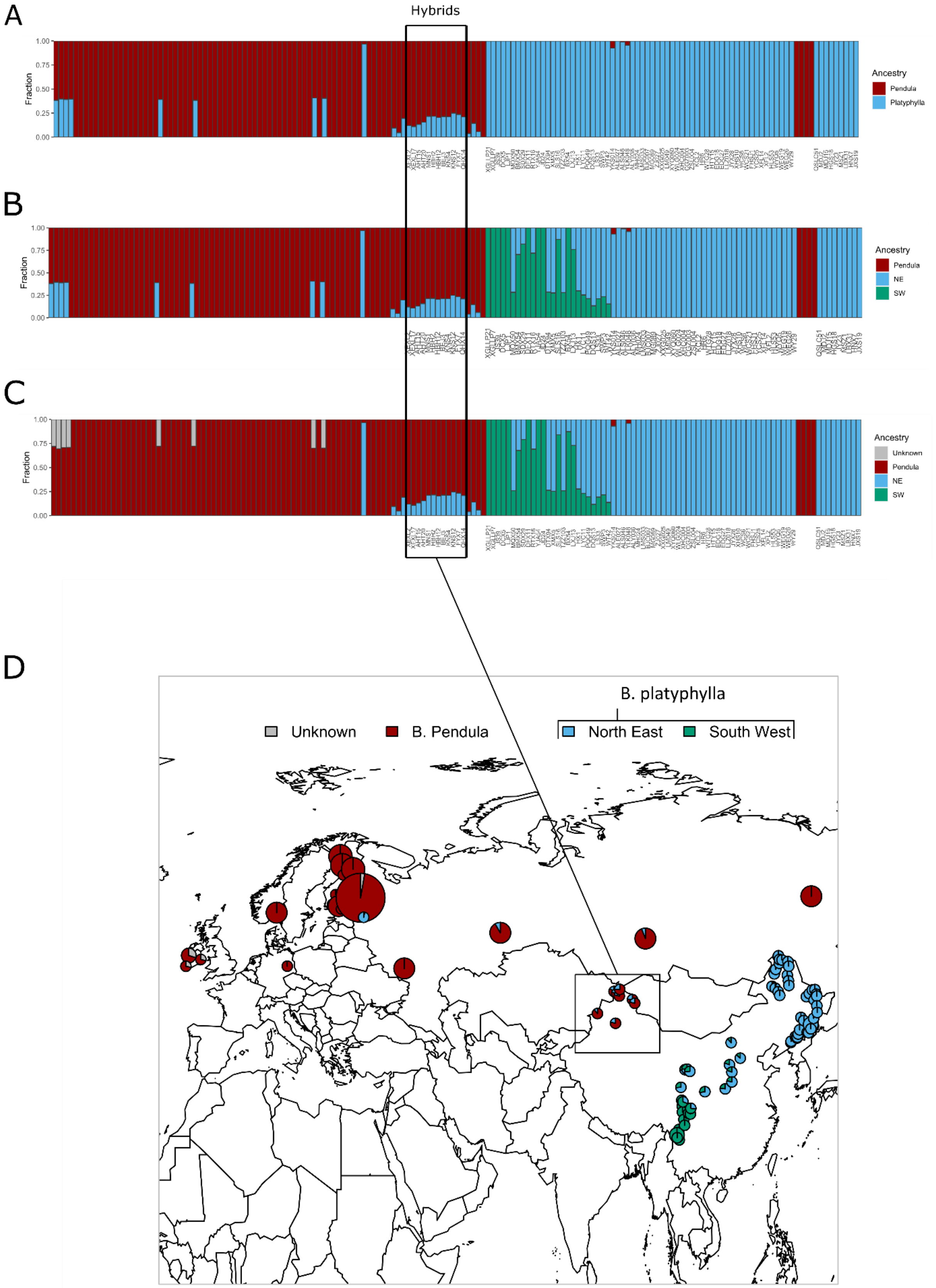


Figure 3.2. *fastSTRUCTURE* results for the Eurasian dataset, ordered by longitude. The size of the pies is proportional to the number of samples. The labelled bars correspond to individuals sampled in this study, while the unlabelled bars are the samples added from Salojarvi et al. (2017) **A)** Bar plot

of *fastSTRUCTURE* results of 162 *B. pendula* and *B. platyphylla* individuals, based on 278,717 unlinked SNPs ($r^2 < 0.4$) at K = 2, **B)** K=3 and **C)** K =4. **D)** Map of the individuals showing their genetic ancestry composition according to fastSTRUCTURE at K = 4.

In *fastSTRUCTURE* restricted to the 71 pure *B. platyphylla* individuals in south-western, central and north-eastern China, the model that maximises the log-marginal likelihood lower bound (LLBO) of the data was K = 2 and that that best explains additional weak underlying structure was K = 3, as suggested by the function "*chooseK*" (Figure S3.6A, Supporting Information). The *fastSTRUCTURE* cross-validation score profile shows that the prediction error is lowest at K = 3 (Figure S3.6B, Supporting Information). At K = 2 the north-eastern population is separated from the south-western population and the individuals in central China appear admixed between these two populations, while at K = 3 the central population is revealed (Figure 3.3).

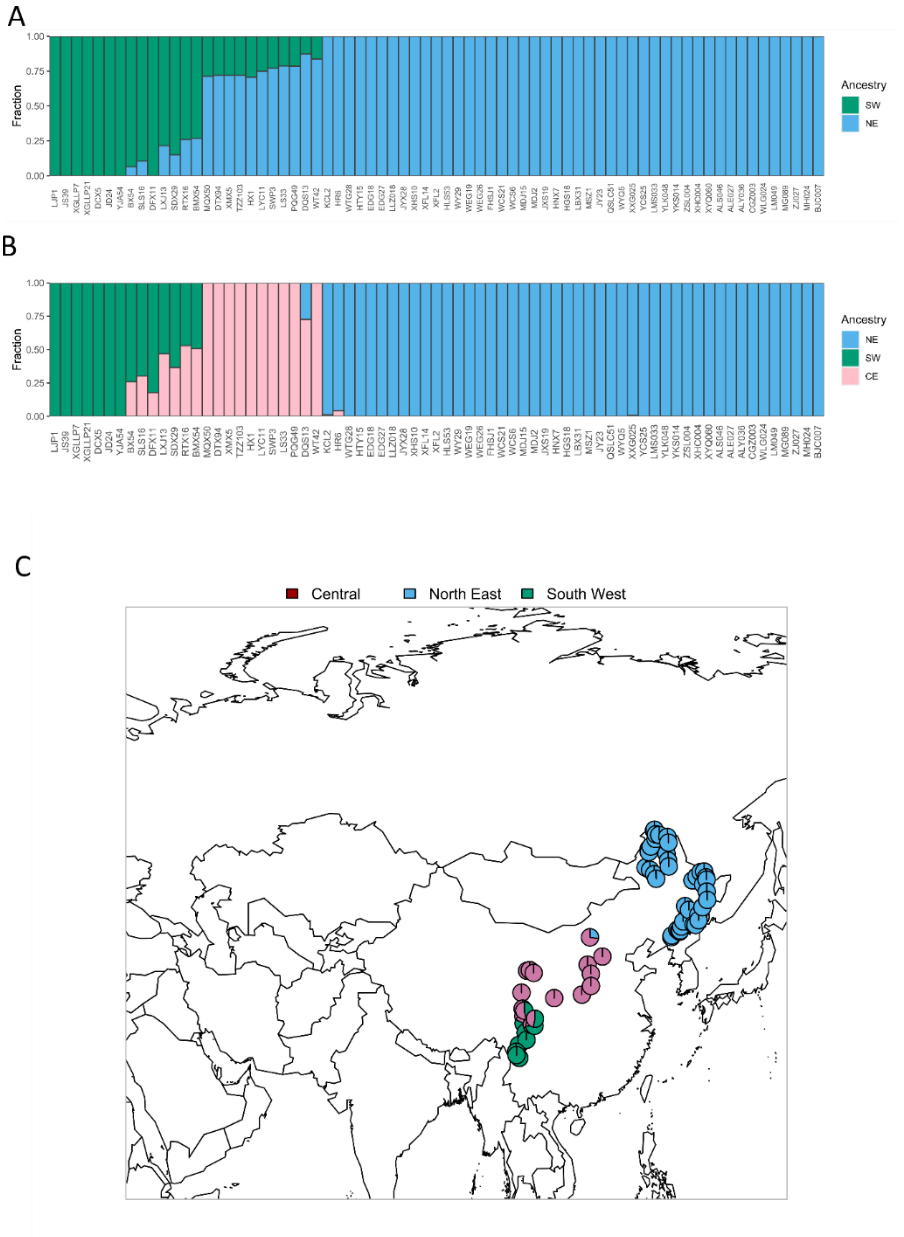


Figure 3.3. *fastSTRUCTURE* results for Chinese *B. platyphylla*, ordered by longitude. **A)** Bar plot of *fastSTRUCTURE* results of 71 *B. platyphylla* individuals, based on 1,387,994 unlinked SNPs ($r^2 < 0.4$) at K = 2, and **B)** K=3. **C)** Map of the sampled individuals showing their genetic ancestry composition according to fastSTRUCTURE at K = 3.

In PCA on the 71 pure *B. platyphylla* individuals, PC1 explained ~33% of the total variance and separated the north-eastern, central and south-western populations (Figure S3.7, Supporting Information). PC2 (15% of total variance) also separates the three populations while PC3 (8% of total variance) scatters the individuals of the north-eastern population. The explained variance levelled-off after PC3 (Figure S3.7D, Supporting Information).

Due to the differentiation between the individuals located in central China and those in the North-East shown by the PCA (Figure S3.7, Supporting Information) and the recommendation by *fastSTRUCTURE* cross-validation scheme, we assigned individuals to populations according to the admixture coefficients computed with *fastSTRUCTURE* at K = 3. To support this choice there are numerous records of birch pollen in central China dated at the LGM and throughout the Holocene (Cao et al., 2015), which coupled with the predicted suitable habitat for this species at the LGM and mid-Holocene and SSR based structure analyses of *B. platyphylla* populations (Chen & Lou, 2019), suggest that *B. platyphylla* have been stably present in central China for at least 22,000 years. In total, ten individuals were assigned to the central population (CE), seven to the south-western population (SW), 46 to the north-eastern population (NE) and the remaining eight individuals were classified as admixed (Figure 3.3B-C). Based on this assignment, *snmf* identified 17,218 outlier SNPs with FDR < 1% (Figure S3.8, Supporting Information).

Linkage disequilibrium decays rapidly in *B. platyphylla* populations and reaches background levels at approximately 50 kb (Figure S3.9, Supporting Information). Assessment of pairwise F_{st} between populations showed that the highest levels of differentiation are between the north-eastern and south-western population or the central population and south-western populations (Figure S3.10, Supporting Information). On the other side, the central and the north-eastern population reported significantly lower mean pairwise F_{st} (Figure S3.10, Supporting Information).

Nucleotide diversity π in windows of 5kb shows a slight increase in populations from south to north geographically, reporting mean of 0.0057 (~0.6%) in the south-western population, 0.0075 (~0.7%) in the central population and 0.0083 (~0.8%) in the north-eastern population (Figure S3.11-3.12, Supporting Information).

Species population size and separation history

Multiple sequentially Markovian coalescent (MSMC) plots suggest a long-term history of population size decline throughout the Pleistocene followed by expansion in more recent history (Figure 3.4A).

Estimation of cross-coalescence rates with MSMC suggests that these population have been fully separated for the last 500,000 years (Figure 3.4B).

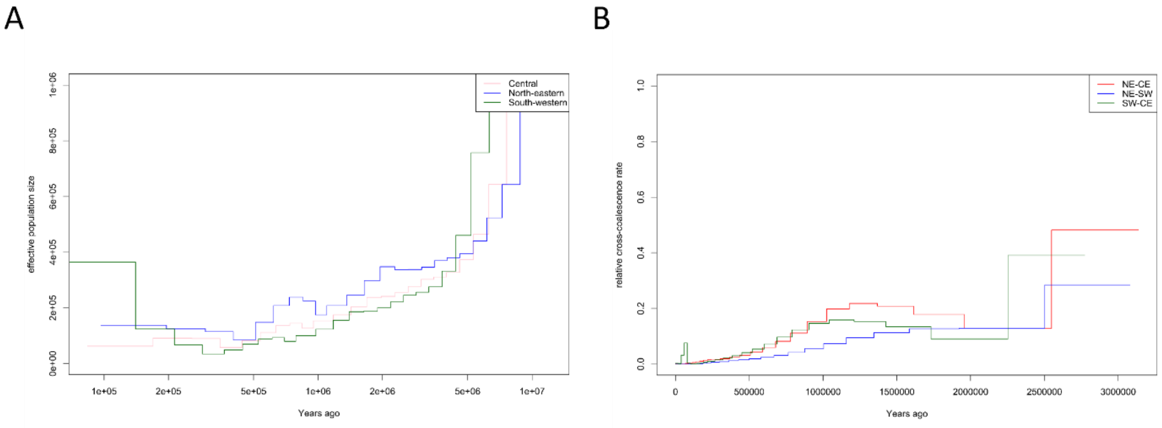


Figure 3.4. Effective population size and separation history estimated with the MSMC method. Generation time: 40 years. Mutation-rate: 7.7×10^{-9}. A) Estimates based on 4 haplotypes for each population. B) Relative cross-coalescence rate (CCR) ratio per year. The CCR ratio is a measure of divergence and represents the ratio of between-populations over within-populations coalescence rate. Values close to 0 indicate that populations have diverged, while values close to 1 indicate that populations have not yet diverged as between and within populations coalescence rates are equal.

Environmental niche modelling

Environmental niche models were built based on eleven uncorrelated environmental variables (Table 3.1, Figure S3.1, Supporting Information) and were based on B. platyphylla samples' locations and GBIF observations of this species (Figure 3.5).

Climatic Variable ID	Description	Retained
AMT	Annual Mean Temperature	Yes
MDR	Mean Diurnal Range (Mean of monthly (max temp - min temp))	Yes
ISO	Isothermality (BIO2/BIO7) (×100)	Yes
TS	Temperature Seasonality (standard deviation ×100)	No
MaxTWM	Max Temperature of Warmest Month	No
MINTCM	Min Temperature of Coldest Month	No
TAR	Temperature Annual Range (BIO5-BIO6)	No
MTWQ	Mean Temperature of Wettest Quarter	Yes
MTDQ	Mean Temperature of Driest Quarter	No
MTWARMQ	Mean Temperature of Warmest Quarter	No
MTCOLDQ	Mean Temperature of Coldest Quarter	No
AP	Annual Precipitation	Yes
PWM	Precipitation of Wettest Month	No
PDM	Precipitation of Driest Month	No
PS	Precipitation Seasonality (Coefficient of Variation)	Yes
PWQ	Precipitation of Wettest Quarter	No
PDQ	Precipitation of Driest Quarter	Yes
PWARMQ	Precipitation of Warmest Quarter	No
PCOLDQ	Precipitation of Coldest Quarter	No
ALT	Elevation	No
SLO	Slope	Yes
ASP	Aspect	Yes
TPI	Topographic Position Index (TPI)	Yes
TRI	Terrain Ruggedness Index (TRI)	No
TR	Terrain Roughness	No
WFD	Water flow direction	Yes

Table 3.1. The 26 climatic variables downloaded from www.worldclim.org for 1970-2000. 11 uncorrelated variables (correlation coefficient < 0.7) were retained for the ENM and GEA.

Jack-knife tests suggested that the variables “aspect” and “water-flow direction” did not play a relevant contribution either singly or in combination, so they were excluded (Figure S3.13-15, Supporting Information). The final Maxent model with the nine remaining variables reported high mean test AUC (0.913 ± 0.017) and low mean test omission rate (0.13, $p < .001$) at a maximum training sensitivity plus specificity logistic threshold of 0.2 (Figure 3.5A-B, Figure S3.16-3.17, Supporting Information). Two variables, “annual precipitation” and “mean temperature of wettest quarter", together contributed to > 50% of the predicting performance of the model, defined as the increase in regularized gain added (or subtracted) to the contribution of the corresponding variable, over each model iteration (Table 3.2).

Variable	Percent contribution	Permutation importance
AP	34.4	31.3
MTWQ	21.9	17.6
PDQ	12	12.1
SLO	10.4	2.6
TPI	7.4	0.7
ISO	4.5	10.9
MDR	3.3	1.5
AMT	3.2	21.2
PS	3	2.2

Table 3.2. The nine variables retained in the final maxent model. Percent contribution: in each iteration of training, the increase in regularized gain is added to the contribution of the corresponding variable or subtracted from it if the change to the absolute value of lambda is negative. Permutation importance: the values of each variable (in turn) on training presence and background data are randomly permuted. The model is then re-evaluated on the permuted data, and the resulting drop in training AUC is shown in the table, normalized to percentages.

Response curves for the final Maxent model are available in Supporting Information. Future projections show an overall contraction of the *B. platyphylla* habitat throughout China, particularly noticeable in central China and in the north-east (Figure 3.5C). Currently suitable environments located at higher elevations seem to contract less in the future (Figure 3.5C). The predicted habitat reduces between 19 and 30 % in 2040-2060, and between 17 and 34 % in 2080-2100 (Figure 3.5C), compared to the present. In the most adverse scenario (ssp370), *B. platyphylla* habitat in China may be reduced by approximately a third of its current extent by 2080-2100 (Figure 3.5C).

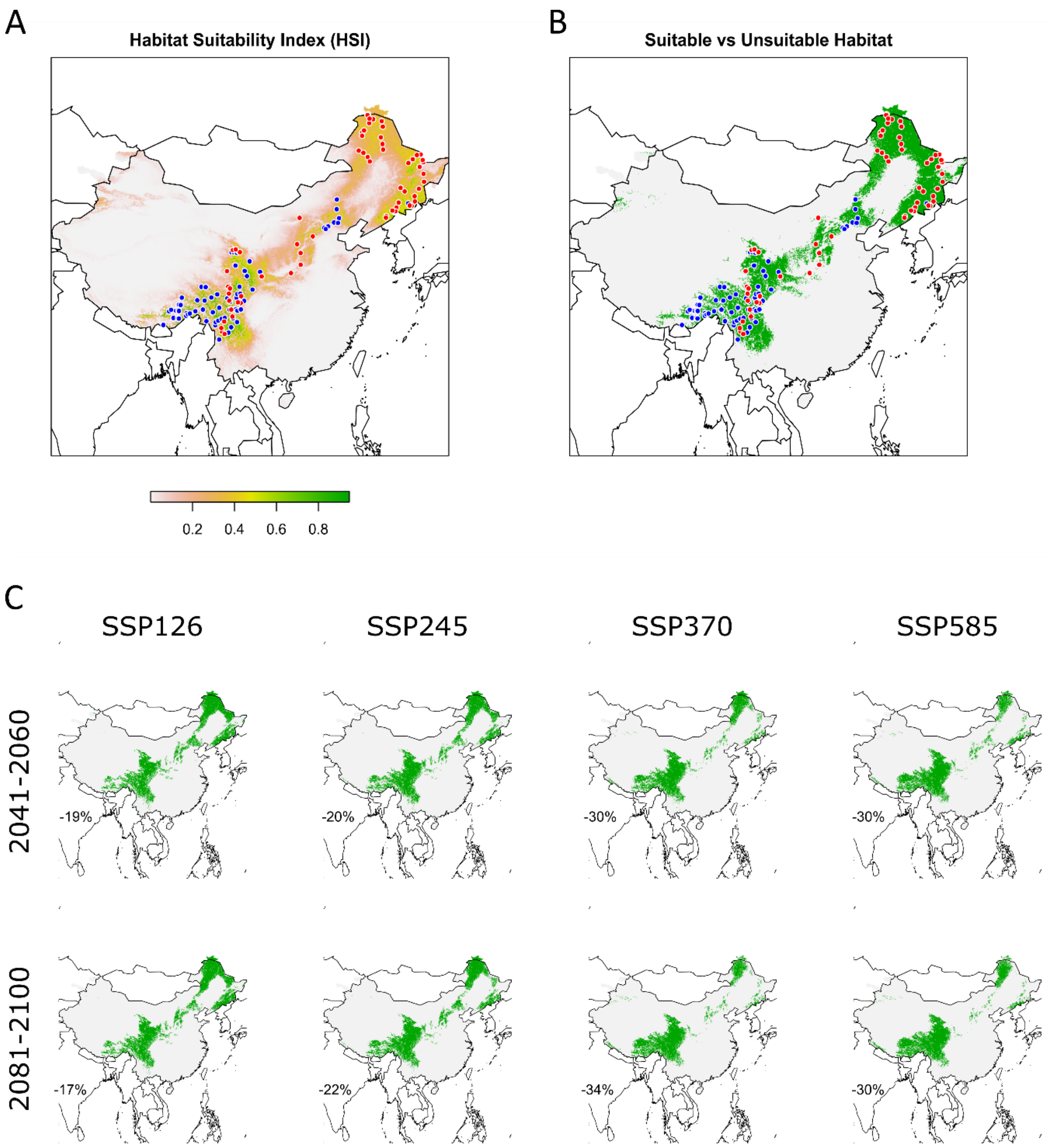


Figure 3.5. Environmental niche model (ENM) of *B. platyphylla* in China. Red points represent the individuals sampled in this study included in the model, blue points represent observations downloaded from the GBIF (http://www.gbif.org). **A)** Present time ENM showing the habitat suitability index (HSI) throughout China using a coloured scale from 0 (white) to 1 (green). **B)** Binomial representation of the HSI model, using the maximum training sensitivity plus specificity threshold of 0.2. **C)** Binomial HSI projections (> 0.2) of *B. platyphylla* under four different future climate scenarios (columns: ssp126, ssp245, ssp370 and ssp585) at two different time points (rows: 2041-2060 and 2080-2100). The percentage in each plot shows the decrease in suitable environment area compared with the current habitat.

Genome-environment association analysis: identification and characterization of putatively adaptive SNPs

We choose a q-value threshold of 0.01 (FDR < 1%) to select candidate associations in the LFMM2 analysis on 1,387,994 SNPs ($r^2 < 0.4$) restricted to the 71 B. platyphylla individuals identified with fastSTRUCTURE (excluding the admixed individuals in NW China).

In total, 7,609 SNPs showed significant associations with one or more environmental variables in *LFMM2* (K = 3) using our criteria (Figure S3.18, Supporting Information), for a total of 11,304 associations. In more details: 4,643 SNP were associated to one environmental variable, 2,424 SNPs to two environmental variables, 384 SNPs to three environmental variables, 133 SNPs to four environmental variables, 21 SNPs to five environmental variables and four SNPs to six environmental variables. In particular, 3,767 of the 7,609 putative adaptive SNPs identified *LFMM2* also resulted significant (FDR < 1%) in the *snmf* F_{st} outlier test. Samβada detected a low number of associations at FDR < 1%, with only 14 significant SNPs (Figure S3.19, Supporting Information), therefore only *LFMM2* candidates were considered in the subsequent analyses.

The distribution of the putative environmentally associated SNPs detected with LFMM2 shows a large variability across the variables tested (Figure S3.20, Supporting Information): particularly, “isothermality” and “mean diurnal range” reported a much larger number of candidates than the other variables, with 5,679 and 2,794 putative SNP-environment associations identified for these two variables, respectively. These were followed by “annual precipitation” (1,059 SNPs hits), “precipitation of driest quarter” (735 SNPs hits), “annual mean temperature” (391 SNPs hits), “flow direction” (263 SNPs hits), “mean temperature of wettest quarter” (227 SNPs hits), “TPI” (104 SNPs hits), “slope” (49 SNPs hits) and “precipitation seasonality” (3 SNPs hits). The variable “aspect” did not report any SNP association with an FDR < 1% in *LFMM2*. There was no significant correlation between the number of identified SNP association per variable and the percentage of contribution of each variable to the ENM (Pearson's test p > 0.05).

The distribution of the candidate adaptive SNPs in the *B. pendula* reference genome shows that they are spread across the entire genome and present on all linkage groups (Figure S3.20A, Supporting

Information). The chromosome that reported the largest number of candidates is chromosome 2 (738 SNPs), while the lowest number was recorded for chromosome 6 (344 SNPs). The number of SNPs identified per chromosome does not appear correlated with chromosome size (Pearson's test $p > 0.05$). A hotspot is visible in the second half of chromosome 7 (Figure S3.20A, Supporting Information).

PCA based only on the 7,609 adaptive SNPs showed a steep discrepancy in the variance explained by PC1 compared to that of the other PCs (Figure S3.21, Supporting Information). Furthermore, PCA based on only adaptive SNPs does not show a clear separation between the central and north-eastern population in the first three PCs, while it separates the south-western population (Figure S3.21, Supporting Information). When using only 7,500 putatively "neutral" SNPs in PCA, the separation between populations is slightly weaker compared to that observed with the whole data set, and individuals in each population are more scattered along the PCA axes (Figure S3.22, Supporting Information).

The pairwise F_{st} patterns of differentiation between populations based only the 7,609 adaptive SNPs and only on neutral loci reflected those observed with the whole SNPs set, however loci under selection showed substantially higher F_{st} compared to those recorded for neutral markers and for the whole SNP set (Figure S3.23-3.24, Supporting Information).

We further searched for the mRNA regions in the *B. pendula* reference genome (Salojärvi et al., 2017) containing at least one candidate adaptive SNP out of the 7,609 candidates, and we identified 1,633 genic regions.

A functional enrichment analysis of the identified genic regions reported significant hits (FDR < 5%) at all three gene ontology levels: biological process (3 hits), cellular component (1 hit) and molecular function (3 hits) (Figure S3.25, Supporting Information).

Risk of non-adaptedness to future conditions (RONA)

We calculated RONA for seven environmental variables projected under the future climate profile ssp370, identified as the worst-case scenario for *B. platyphylla* for the years 2080 – 2100 in terms of habitat reduction (Figure 3.6, Table S3.2 Supporting Information). The most represented environmental variables in terms of number of associations in the RONA computation were “isothermality” (5,679 SNPs, mean r^2 = 0.459), “mean diurnal range” (2,794 SNPs, mean r^2 = 0.0157) and “annual precipitation” (1,059 SNPs, mean r^2 = 0.0976) (Table S3.1, Supporting Information). The variables that reported higher average r^2 of adaptive SNPs were “isothermality” (5,679 SNPs, mean r^2 = 0.459), “annual mean temperature” (391 SNPs, mean r^2 = 0.195), and “mean temperature of wettest quarter” (227 SNPs, mean r^2 = 0.185) (Table S3.2, Supporting Information).

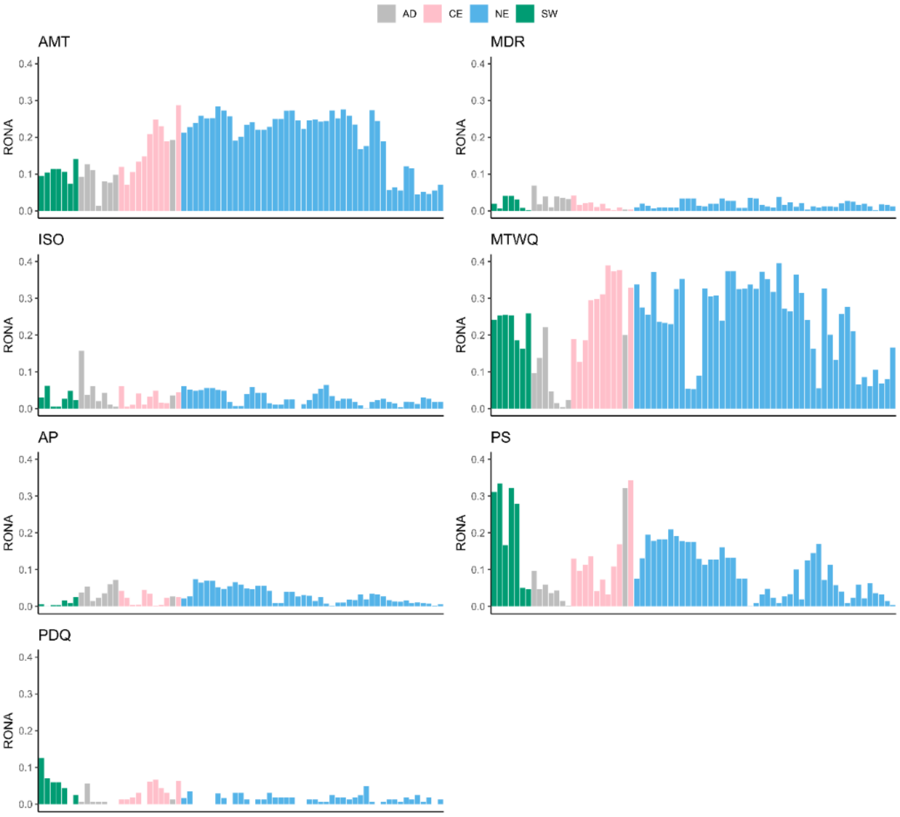


Figure 3.6. Risk of non-adaptedness (RONA) for the 71 *B. platyphylla* individuals**.** RONA calculated independently for each of the seven climatic variables projected to change in the future. The future climate profile ssp370 at 2080-2100 was used for RONA calculation. AD: admixed individuals. AMT: annual mean temperature. MDR: mean diurnal range. ISO: isothermality. MTWQ: mean temperature of wettest quarter. AP: annual precipitation. PS: precipitation seasonality. PDQ: precipitation of driest quarter.

There was large variability in the expected allele frequency changes required to match future conditions across environmental variables. “Annual mean temperature” and “mean temperature of wettest quarter” reported notably higher mean individual RONA compared to all the other variables, being 0.1773 and 0.2314 respectively (Table S3.2, Supporting Information).

Weighted mean RONA scores were overall relatively low across individuals, with mean of 0.08 (sd = 0.03) and maximum of 0.1455, that was reported for an individual in the north-eastern population (HTY15) (Figure 3.7, Table S3.3 Supporting Information). The maximum RONA scores show greater variability across individuals, with mean of 0.25 (sd = 0.09) and maximum of 0.3956, that was recorded for an individual in the north-eastern population (JY23, “mean temperature of wettest quarter”) (Figure 3.7, Table S3.3, Supporting Information).

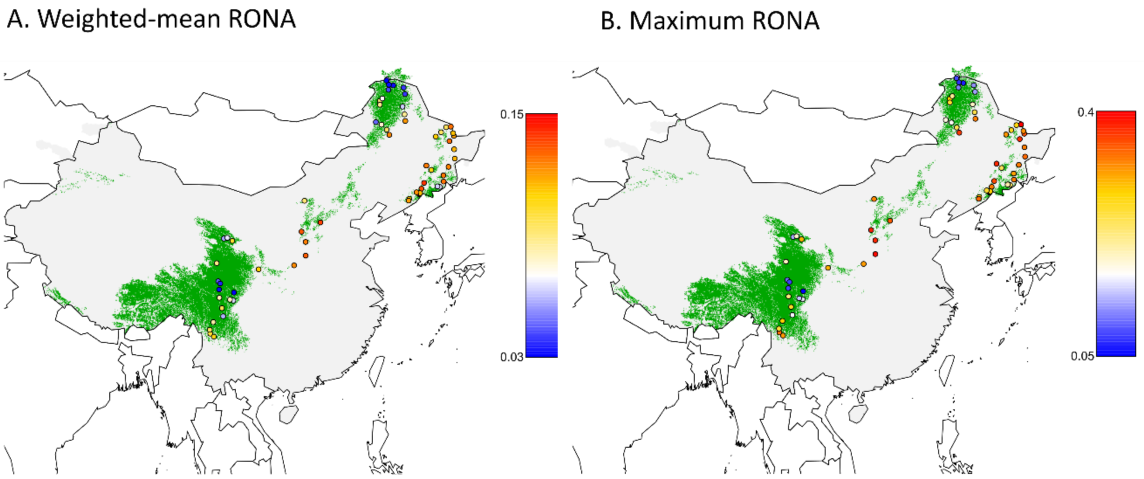


Figure 3.7. Mean-weighted RONA and MAX RONA. **A)** Map showing weighted mean RONA for the 71 *B. platyphylla* individuals, out of the seven environmental variables included in the analysis. Each variable RONA was given a weight equivalent to its percent contribution in the ENM. **B)** Map showing maximum RONA per individual, out of the seven environmental variables tested. Green shading represents suitable environment in 2080-2100 under profile ssp370.

There were some differences between the three Chinese *B. platyphylla* populations in regard to the environmental variables that will require steeper allele frequency changes (in ssp370 at 2080 – 2100) (Table S3.4, Supporting Information): although all populations showed similar RONA patterns across the environmental variables tested, with the largest mean RONA (averaged over the individuals of each population) reported for temperature derived variables ("annual mean temperature" and "mean temperature of wettest quarter") rather than precipitation-derived variables (Table S3.4, Supporting Information), the south-western population, however, reported much higher mean RONA for "precipitation seasonality" (0.21) compared to the other two populations that had relatively low mean RONA for this specific environmental variable (Table S3.4, Supporting Information).

Discussion

Pan Eurasian population genetic structure of white birch

Our study clearly differentiated between *B. pendula* and *B. platyphylla* across Eurasia, with strong separation between the two in PCA and fastSTRUCTURE analyses. We found evidence for extensive hybridisation between these two species in northwest China, and to a lesser extent in central Russia. Populations in northwest China that were initially identified as *B. platyphylla* turned out to be mainly *B. pendula* with introgression from *B. platyphylla*. A previous microsatellite study of *B. pendula* and *B. platyphylla* across Eurasia (Tsuda et al., 2017) showed admixture between the two species in central Asia, but east of Europe this study only included populations from Russia and Japan. In a microsatellite study of white birches in China, Chen and Lou (2019) found populations in northwest China to be highly differentiated from other populations, and to be of high genetic diversity. However, as no *B. pendula* were included in their study, they could not pick up a signature of hybridisation in these populations. Instead, they attributed their high differentiation and diversity to a northern glacial refugium for white birch in the Altay mountains.

Genetic structure of white birch within China and signals of local adaptation

We performed a genomic study on Asian white birch across its distribution in China. The collected samples are spanning 26.4 latitudinal and 48.9 longitudinal degrees. Based on fastSTRUCTURE and PCA analyses, we detected three genetic clusters in Asian white birch corresponding to samples collected from NE, SW and central China. As expected, we detected spatial population structures and genetic admixtures based on genome wide SNPs. Unexpectedly, all samples from north-western China showed strong evidence of admixture between *B. pendula* from Europe and *B. platyphylla* from north-eastern China. This indicates that Xinjiang region appear to be a hybrid zone between these two species*,* therefore we removed the NW population of *B. platyphylla* for downstream analyses.

Furthermore, we identified a set of SNPs associated with various environmental variables, which reflect signatures of local adaptation. Based on these, we inferred the risk of maladaptation for each individual in the context of future climate change. To our knowledge, this is one of the most comprehensive studies investigating the genomic basis of local adaptation of forest tree species that spans such a vast geographical area so far.

We next examined the divergence history of the three different lineages of *B. platyphylla* in China. Assuming the generation of white birch is 40 years, our MSMC results show that that the three population have been separated throughout the last 3 million years and reached full separation roughly 500,000 years ago. Gene flow appears to have accompanied their divergence throughout the last 3 million years (Figure 3.4). In accordance with this, some individuals show genetic admixture between the SW and central clusters and between the NE and central clusters. For instance, several hybrids were detected between the SW and CE lineages (Figure 3.3). The areas between the SW and the CE lineages or between the CE lineages and the NE lineages are suitable for Asian white birch as evidenced by ENMs, online records and our own field observations (Figure 3.5). Consistent with a previous phylogeographic study of Asian white birch showing that genetic diversity of Asian white birch increased along latitude based on SSR (Chen & Lou, 2019), our result confirmed that genomic diversity from NE is higher than that from CE which is higher than that from SW (Figure S3.11-3.12, Supporting Information), therefore seemingly contradicting the typical expansion from the southern refugia model (Harrison et al., 2001). However, even in case of northward expansion from the south there are other factors, such as the less severe geographical barriers to pollen movement in northern China compared to the southern QTB, that could still produce the observed nucleotide diversity pattern.

Ecological niche models indicated that mean temperature of wettest quarter and annual mean precipitation were the most important predictors of *B. platyphylla* distribution and both north-eastern China and the south-western Tibetan plateau currently provide the most suitable environments for this species in China. Future climate projections show an overall decline of the species range and environmental suitability throughout of its current distribution in China, with suitable habitats persisting mostly in areas of higher elevation (Figure 3.5).

To decipher the genetic basis of local environmental adaptation of *B. platyphylla*, we performed a genome-wide environmental association study. We found signatures of natural selection at 7,609 SNPs with *LFMM2*, which represents 0.5% of the SNPs tested. Interestingly, 3,767 of these SNPs were also identified as statistically significant F_{st} outliers between the NE, CE and SW lineages, reinforcing the confidence in their involvement in local adaptation. In total, we identified 17,218 outlier SNPs between the NE, CE and SW lineage, which may include additional adaptive loci associated to climatic or environmental variables that we did not measure, as well as neutral loci due to drift.

PCA based only on adaptive SNPs showed that the NE and central populations formed an overlapping cluster whereas based on neutral SNPs the two populations separated (Figure S3.21-3.22, Supporting Information). This is possibly due to the relatively lower diversifying selection acting on the two populations and the existence of gene flow which can counteract the impact of diversifying selection (Guichoux et al., 2013; Nosil et al., 2009). F_{st} analyses suggested that the level of differentiation is much higher at loci under selection rather than neutral loci, as expected, and the SW populations appears to be highly differentiated from the others at adaptive loci (Figure S3.23, Supporting Information).

As linkage disequilibrium decays rapidly in *B. platyphylla* populations (Figure S3.9, Supporting Information), we searched for the genic regions intersecting the identified adaptive SNPs and identified 1,633 mRNA regions, according to the available *B. pendula* annotation (Salojärvi et al., 2017). Functional enrichment analysis of the identified adaptive genes suggested that these genes were enriched for growth and environmental stress response (Figure S3.25, Supporting Information), although further studies and functional experiments are needed to confirm these findings. Among the significantly enriched categories, the most interesting in view of climate adaptation were: “regulation of response to stimulus”, “DNA helicase activity” which are enzymes involved in DNA repair and have a crucial role as caretakers of the plant genome against environmental damages (Sami et al., 2021) and “inorganic cation transmembrane transporter activity”, as potassium, the most abundant inorganic cation in plant cells, it is essential for plant growth and development and it has been shown to have a major role in resistance to drought, salinity and fungal infections (Sharma, Dreyer & Ridelsberger, 2013). The 1,633 identified adaptive genes also included several light-response and growth-related genes, according to our OmicsBox annotation, as well as members of all the three GO categories significantly enriched in the putative genes under selection detected by Solojarvi et al. (2017): transmembrane receptor protein tyrosine kinase, peptidyl-histidine phosphorylation and axis specification. Furthermore, 87 of the 1,633 genic regions detected in this study were identified as putative selective sweeps in *B. pendula* (Salojarvi et al., 2017). Even though it would be interesting to explore all the identified genes singularly, this goes beyond the extent of this work and further experimental validation would still be necessary to confirm their role in growth and environmental stress response.

In line with similar studies in other species (Jordan et al., 2017; Rellstab et al., 2016; Borrell et al., 2019; Pina-Martins et al., 2018), local adaptation appears to be highly polygenic in *B. platyphylla*, therefore it is likely that maintaining the standing variation and adaptive diversity may be a better solution to aid future climate adaptation, rather than focusing on raising the frequencies of a specific small set of adaptive SNPs, as it was proposed already (Jordan et al. 2017). Two variables, "mean diurnal range" and "isothermality" reported a substantially larger number of significant SNPs association compared to the other climatic variables (Figure S3.20), which may have arisen due to the spatial correlation of these environmental variables with the genetic structure across the sampling range (Figure S3.1, Supporting Information) therefore these results should be interpreted with caution and further validation is necessary; however even though the adaptive SNPs identified for these two variables certainly contain many neutral loci arisen due to drift, we cannot exclude an highly polygenic adaptation to these variables with many loci with low effects involved. We did not identify any significant correlation between the strength of selection, intended as the number of associated SNPs, and the percentage of contribution of each climatic variable in our ENM, differently from similar studies (Borrell et al., 2019). However, we note that it is not a logical necessity as the variables with higher discriminatory power in the ENM could be limiting species ranges either because they lack adaptation (Borrell et al., 2019), or contrarily they could be subjected to strong adaptation but at a global level throughout the species range, rather than locally. Therefore, the lack of variation in these variables across the sampling range prevents the detection of these adaptive signals with methods designed for local selection, such as *LFMM2.* Other methods such as genome scans for selective sweeps (Alachiotis, Stamatakis, & Pavlidis, 2012; Pavlidis, Živković, Stamatakis, & Alachiotis, 2013), could be used to identify regions of the genome subjected to global selective processes.

To investigate how the samples of *B. platyphylla* will respond to rapid future climate change, we calculated RONA for the 71 *B. platyphylla* individuals according to the implementation by Pina-Martins et al. (2018). Overall, RONA was relatively low across the environmental variables tested

(Figure 3.6), with the exception of “annual mean temperature” and “mean temperature of wettest quarter”, very likely reflecting the larger relative projected change of temperature derived variables compared with that of environmental variables related to precipitation throughout China. The expected allelic frequencies changes are similar to those reported in other studies (Borrell et al., 2019; Jordan et al., 2017; Pina-Martins et al., 2019; Rellstab et al., 2016). Weighted mean RONA incorporated the ENM with EAA and generally remains low and it is always below 0.2 (Figure 3.7). We also reported the maximum RONA per individual out of the seven variables tested (Figure 3.7) because even if a population, in this case an individual, has “low” average RONA for a given future projection, it might still be its highest RONA that will determine how much it will need to shift its allelic frequencies in order respond to future selective pressure (Pina-Martins, Baptista, Pappas & Paulo, 2018). Maximum RONA across individuals showed significantly higher values, as expected. Interestingly, the maximum RONA plot also shows particularly high values for four individuals in the south-western population located at the very southern edge of this population distribution, reported for precipitation seasonality which is expected to have a much larger relative change at these locations (Figure 3.7). Overall, both weighted mean and maximum RONA match well with our ENM for ssp370 2080-2100, with lower RONA values reported within the inferred projected suitable environment (Figure 3.7).

The RONA method is of course a huge simplification, and its limitations are accurately described in the original publication (Rellstab et al., 2016), therefore our results should be viewed considering such caveats. In addition, RONA has been originally designed to be used with population allelic frequencies rather than individuals, as this results in only three possible data points for the RONA regression which simplifies this method even further. However due to the nature of our sampling, this was to our knowledge the best approach. Grouping the individuals in populations for the RONA calculation would be problematic in this study as it would require averaging climatic conditions across different locations, which would result in an even greater simplification due to the broad geographic range of our study and the high heterogeneity of some of the environmental variables tested.

It is not straightforward to determine when RONA is too high to result in a lag between allele frequencies and adaptation. A previous study on *Fagus sylvatica* have observed allele frequencies changes of 0.1 – 0.2 per decade (Jump, Hunt, Martinez-Izquierdo, & Peñuelas, 2006), but this estimate was based on amplified fragment length polymorphism (AFLP) molecular markers rather than SNPs. A more recent study on Swiss stone pine, *Pinus cembra,* based on SNPs markers has investigated the observed allele frequencies shifts over time by looking at the differences in SNPs frequencies between adults and juvenile cohorts at seven locations in Switzerland, and validated the observed shifts with forward-in-time simulations (Dauphin et al., 2020). This analysis reported an average observed rate of allele frequency shifts of 1.26 x 10^{-2} per generation (*P. cembra* generation time = 40 years) at neutral SNPs, and a slightly lower estimate for adaptive SNP, however neutral and adaptive loci generally behaved similarly. Based on this estimate we can hypothesize that SNPs frequencies changes < 0.05 may be matched by populations naturally within a couple of generations, whereas larger changes, such as those predicted for temperature derived variables (Figure 3.6), are unlikely to be achieved naturally by long-living forest trees with long-generation times, even when they exhibit high levels of standing genetic variation. This reinforces the importance of considering some of the proposed conservation strategies, such as assisted gene flow or assisted migration (Borrell et al., 2019).

Concluding remarks

This study provides insights into both the evolution and ecology of Asian White Birch, *B. platyphtylla*, in China. We showed that *B. platyphylla* and *B. pendula* are better considered distinct species, as their differentiation is clear at the genomic level. We identified three distinct lineages of *B. platyphylla* in China distributed along a latitudinal gradient from the north-east to the south-west, and a hybrid zone between *B. platyphylla/B. pendula* in north-western China. The three *B.*

platyphylla lineages appear to have diverged prior LGM, suggesting that this species survived the last glaciation in multiple local glacial refugia. Our species distribution model shows that *B. platyphylla* predilects mountainous environments with cool summers and moderate levels of precipitations. Future prediction shows significant reduction on the suitable habitat of this species in China under every future scenario tested, up to a third reduction compared to the current extent by 2080-2100. We identified signals of local adaptation throughout the genome of this species, suggesting that climate adaptation is a polygenic mechanism. We estimated the degree of maladaptation to inform conservation strategies. Despite the limitation of our method, we showed that rapidly rising temperature, particularly in the summer months, poses a risk to this species and our environmental niche model suggests that current habitats located in central China and in the north-eastern region adjacent to the Changbai mountains will likely become unsuitable by 2080-2100; on the other hand, current habitats located at higher elevations, such as the south-western Tibetan plateau and the region surrounding the Greater Khingan mountains range in the north-east will likely remain suitable. It is possible that populations outside the predicted suitable environment may be able to adapt, however climate, particularly temperature, might change to an extent that adaptation will not be possible and assisted migration might be the only option possible to rescue some populations. The estimated required changes in allelic frequencies at adaptive loci do not appear as concerning within the predicted suitable environment, however with the current dataset, based on individuals' allelic frequencies, it is difficult to estimate with confidence an allelic frequency change threshold after which natural populations will fail to adapt on time naturally and will require assisted gene flow from donor pre-adapted populations.

Contributions

Nian Wang organised and performed the sampling, extracted DNA, arranged sequencing and performed mapping and SNPs calling on both the China and Eurasian datasets. Gabriele Nocchi (I, the author) performed all subsequent analyses and wrote the chapter. Richard Buggs oversaw the project.

Chapter 4: Conclusions

Overall thesis contribution

In this thesis I have presented population genetics analyses of two large temperate forest tree whole genome re-sequence datasets. The first included 386 oak individuals of species *Q. robur*, *Q. petraea* and *Q. robur x Q. petraea* distributed across four British managed parkland environments, and the second included 83 Asian white birch (*B. platyphylla*) individuals from 74 natural populations scattered throughout the known distribution of this species in China.

In the oak project I have assembled a huge SNPs dataset, including over two million high-quality SNPs after filtering, which makes it the biggest whole genome re-sequencing dataset for Britain native oak species to my knowledge. I analysed the population genetic structure of oak trees across four British parkland environments, estimated the levels of admixture between *Q. robur* and *Q. petraea* at the sampled locations and identified hybrids between the two species. Furthermore, I assessed the patterns of genetic diversity between English and sessile oak, and I was able to identify some genomic regions significantly differentiated between the two species. I further explored the demographic and divergence history of both species and searched the *Q. robur* genome for signatures of recent strong natural selection. I assessed the levels of genetic relatedness between individuals within each parkland and linked it to past planting practices. Finally, I took advantage of the huge data generated to explore chloroplast DNA variation by assembling *de novo* the complete chloroplast sequences of most of the oak individuals sampled and matching them with previously characterized oak haplotypes. To my knowledge, I provided the first complete chloroplast sequences for all four Britain native oak haplotypes. I used the chloroplast data to determine whether the oak individuals sampled derive from native and local seed stock, by comparing the distribution of chloroplast haplotypes in the four parklands with that of the major oak haplotypes across Europe and Britain ancient natural woodlands, mapped in previous studies.

In the Asian white birch project, our collaborators have assembled a large whole genome dataset including over a million high-quality SNPs after filtering, spanning a very large geographic area encompassing China entirely. By analysing this data, I have assessed the population genetic structure of this species in China and estimated its levels of admixture with the closely related sister species *B. pendula,* by adding additional data of this species from a large Eurasian study. I was able to identify a hybrid zone between these two species in north-western China. I further assessed Chinese *B. platyphylla* genetic diversity, and the demographic and divergence history of the different lineages identified in China. Furthermore, I mapped the present distribution of this species in China and identified the climatic variables influencing its geographic range the most, and in turn I used this information to predict the future distribution of this species. Finally, I performed a large environmental association analysis in the *B. platyphylla* Chinese populations, encompassing over 26 latitudinal gradients, that led to the identification of over 7,000 SNPs associated to climate. I used the identified adaptive loci to predict the degree of maladaptation to future conditions for this species across China, highlighting regions that may require the aid of conservation strategies, such as assisted migration or assisted gene flow, to reduce adaptational lag in view of the ongoing rapid climate change.

The Oak project

The oak project led to the development of important genomics resources for research on Britain's native oak species and provided a descriptive preliminary analysis of the genetic structure of oak trees across four British parklands affected by acute oak decline (AOD), at both nuclear and chloroplast DNA levels.

Comparison with other oak datasets

Around the world similarly large oak genomics datasets including hundreds of individuals across large geographic areas have been assembled in recent years, however none with a coverage as deep as the one presented in this thesis, which included over 2 million SNPs after quality filtering and linkage disequilibrium pruning.

Guichoux et al. (2012) used a SNP assay in 855 oak trees from six mixed *Q. robur* and *Q. petraea* stands in the North of France, to generate a set of 262 SNPs after validation. Rellstab et al. (2016) recently employed a targeted pooled sequencing approach to assemble a large dataset based on 1,400 oak individuals of species *Q. robur, Q. petraea* and *Q. pubescens* from 71 populations scattered across Switzerland, including 3,576 SNPs after filtering. Pina-Martins et al. (2018) used genotype by sequencing (GBS) to assemble a set of over 1,996 SNPs after filtering, based on 95 *Q. suber* individuals from 17 location scattered across the entire range of this species in Southern Europe and North Africa. Gugger et al. (2020) sequenced the DNA of 436 adult valley oak trees, *Q. lobata,* across this species entire range in California using GBS and assembled a set of over 11,000 SNPs spread throughout the genome of this species. Martins et al. (2018) assembled a panel of 103 individuals from 17 populations of *Q. rugosa* scattered throughout this species geographic range in Mexico, including 3,534 SNPS identified through GBS. Another large oak dataset has been assembled recently for the evergreen Asian oak species *Q. aquifolioides,* based on over 500 individuals from 60 populations scattered across the distribution of this species in south-western and central China, which were genotyped for 381 SNPs and eight microsatellites (Du et al., 2020). The only oak dataset close to ours in terms of SNPs number to my knowledge is the one assembled by Leroy et al. (2019b). This included over 30 million SNPs based on 71 individuals of four oak species sampled in the south-west of France, 20 *Q. robur*, 13 *Q. petraea, 18 Q. pubescens* and 20 *Q. pyrenaica,* however this dataset was generated via pool-sequencing.

Species differentiation

In the oak project, I made use of the large SNPs panel assembled to first assess the patterns of differentiation and hybridization between *Q. robur* and *Q. petraea*. The entire dataset assembled was not ideal for this type of study, as there was a contrasting number of *Q. robur* compared to *Q.*

petraea, which was limited to ten individuals in only one of the parklands (Attingham Park). To mitigate this limitation, I restricted this analysis to ten *Q. robur* and ten *Q. petraea* individuals from the same parkland site. Furthermore, in order to find F_{st} outlier loci between the two species, I employed an innovative method as the classical null infinite island model (Beaumont & Nichols, 1996) does not fit well in cases where the number of populations and migration rate are low, therefore I preferred not to use common tools based on this model such as FDIST (Beaumont & Balding, 2004; Beaumont & Nichols, 1996) or Bayescan (Foll & Gaggiotti, 2008). I tried to mitigate the number of false positives in this analysis by using a window-based approach: rather than considering each outliers SNPs detected, I searched for genomic windows enriched in F_{st} outliers. This analysis led to the discovery of over 600 SNPs fixed between the species, and 81 genic regions enriched in outliers showing high interspecific differentiation; however, the *Q. robur* and *Q. petraea* genomes generally exhibited high-similarity and low-levels of differentiation overall, in accordance with previous SNP based recent research (Lang et al., 2018; Leroy et al., 2019b), and several hybrids were identified.

Similarly, the Leroy et al. (2019b) study assessed the differentiation patterns and divergence history between *Q. robur*, *Q. petraea, Q. pubescens* and *Q. pyrenaica*. This research confirmed previous findings (Leroy et al., 2017), suggesting that extensive secondary contacts between these four oak species have occurred recently, possibly at the beginning of the last interglacial period, after a long period of isolation. This, together with pre and post zygotic selection (Lepais et al., 2013), appear to have eroded most species-specific genetic structures between these four oak species except those at barrier loci. A few highly differentiated regions were identified in the research by Leroy et al. (2019b), but the genomes of these species generally showed low levels of differentiation, with F_{st} patterns between *Q. robur* and *Q. petraea* similar to those reported in this thesis, and comparable to those usually reported between within species populations (Roux et al., 2016), such as in *B. platyphylla* (Figure S3.10, Supporting Information). Lang et al. (2018) also recently assessed the diversity between *Q. robur* and *Q. petraea* in central and western Europe based on about 12,500

SNPs, further confirming little differentiation overall, and reported a mean F_{st} estimate very close to ours. Furthermore, our mean F_{st} estimate is also very close to that reported in Guichoux et al. (2012), that identified between 13 and 74 outlier loci between *Q. robur* and *Q. petraea* based on 262 SNPs.

An interesting finding related to oak species differentiation also emerged from the study by Rellstab et al. (2016), that assessed the environmental differences between the habitat of *Q. robur, Q. petraea* and *Q. pubescens* in Switzerland. Although reporting significant differences across all the abiotic variables tested, further analysis showed that *Q. robur* and *Q. pubescens* habitats are clearly ecologically differentiated, while no environmental factor could clearly distinguish *Q. petraea* habitat from that of the other two species. This suggests that *Q. petraea* is less specialized, with which may explain why it often occurs in sympatry with *Q. robur*. As Rellstab et al. (2016) noted, this could be beneficial as it could mean that *Q. petraea* has higher flexibility to adapt to changing environmental conditions. In addition, the Rellstab et al. (2016) study aimed to detect genomic signals of local adaptation, both on a global interspecific level as well as within each species, leading to the identification of numerous candidate adaptive SNPs and genes. Interestingly, this research has identified seven genes associated to the same environmental factors across all three species, which strengthened the evidence of the involvement of these loci in local adaptation. In addition, the function of these candidate genes appeared related to the associated environmental factor. Rellstab et al. (2016) further assessed the risk of non-adaptedness of these Swiss oak populations to a future climate scenario based on the identified adaptive loci and warned that the required allelic frequency changes required to match future climate are unlikely to be achieved based on standing genetic variation alone, given the long generation time of oaks.

Population structure

In addition to the above, I provided an initial descriptive analysis of British parkland *Q. robur* genomics: I showed that there is not apparent strong geographically correlated structure at nuclear DNA level between the genomes of individuals among the four sites sampled, similarly to the study on *Q. suber* by Pina-Martins et al. (2018) and *Q. rugosa* by Martins et al. (2018). Pina-Martins et al. (2018) identified only a weak east-west differentiation across *Q. suber* European range and identified 249 SNPs associated to environmental variables. Similarly, Martins et al. (2018) identified only two weakly differentiated genetic groups following a longitudinal gradient for *Q. rugosa* in Mexico and detected 97 SNPs associated with climate. Differently, similar studies in other oak species have detected stronger geographically correlated within species structures (Du et al., 2020; Gugger et al., 2020), however I have to note that the oak sampling in this thesis encompassed a much smaller latitudinal and longitudinal range. Gugger et al. (2020) identified two major genetic groups correlated with geography in *Q. lobata*: a southern California cluster around Tejon and a larger cluster including all individuals from central and northern California (Gugger et al., 2020). Furthermore, this study assessed local adaptation at multiple spatial ranges, state-wide and regionally, and led to the identification of over 600 SNPs with signature of natural selection. Interestingly, many loci reported significant associations at multiple spatial scales, possibly reflecting either independent selection on the same loci or a global selective process. On the other hand, some loci were only identified associated in a single spatial scale. Even more interestingly several putative adaptive SNPs identified in this study were also detected in Mexican oak (Martins et al., 2018), *Q. rugosa,* suggesting parallel adaptation may be at play within the oak genus, an hypothesis that already emerged from the study by Rellstab et al. (2016). Another study that identified within species structure in oak is the one by Du et al. (2020), that identified two clear genetic clusters following an east-west gradient in the Asian *Quercus aquifolioides* and detected several genome-environment associations, including some lineage specific signals of local adaptation.

In addition to assessing *Q. robur* within species population structure, I identified ten coding regions significantly depleted in diversity in its genome, showing signals of recent strong global selection, which included known stress and growth regulators according to the available annotation (Plomion et al., 2018). Furthermore, I calculated genomic relatedness levels within each parkland and linked it to possible past planting practices, with higher relatedness levels perhaps due to greater promotion of natural regeneration.

Chloroplast haplotypes

Finally, I took advantage of the huge data generated to assemble the full chloroplast sequences for most of the individuals sampled *de novo*, providing the full sequences for all four British native oak chloroplast haplotypes. The analysis of the chloroplast data suggested that the oak trees across all four parklands derive from British seed stock, as no individuals could be matched to foreign haplotypes. I proposed a method, inspired by previous research (Lowe et al., 2004), to estimate whether the populations sampled derive from local trees or seed stock. This analysis suggests that trees in these four parklands are likely to derive from local stock, which is reinforced by the levels of genetic relatedness which show that natural regeneration, or planting of on-site acorns, may have been encouraged in the past.

AOD

The data set that I worked on has also been used for a metagenomics study. Louise Gathercole, another PhD student, examined the microbial composition of oak leaves of 421 oak trees across five parkland sites (including all four parklands sampled in this thesis) affected by AOD, including symptomatic trees, asymptomatic trees and trees in remission. I contributed towards this project by extracting all the sequencing reads that did not map to the *Q. robur* reference genome for each oak individual and removing any possible human contamination by mapping to the human reference genome, to assemble a dataset suitable for metagenomics analysis of oak leaves. This study concluded that AOD-associated bacteria may be part of the healthy oak leaf microbiome, as there was not significant difference in the abundance of the putative AOD-causative bacteria among AOD status categories, however there were significant differences in their abundances among sites (Gathercole et al., 2021).

Future directions for the Oak project

The oak dataset was originally assembled to assess whether AOD (Denman & Webber, 2009; Denman et al., 2014), which is widespread across the four sampled parklands, has a genetic component. The ultimate aim was the identification of loci linked to resistance to this very complex abiotic and polymicrobial disease. However, my dataset was inadequate to identify candidate loci for two reasons.

First, it is difficult to categorically classify oak trees based on AOD symptoms, as in addition to asymptomatic trees showing no signs of infection and clearly affected oak trees displaying stem bleeds, some individuals enter into a remission stage after contracting the disease. It is unclear whether this remission stage is due to some sort of host resistance or not (Brown et al, 2016; Denman et al., 2014). Among oak trees that go into remission, some remain asymptomatic, while others are re-infected by AOD. The causative factors of AOD were not fully understood (Denman et al., 2018), which complicates the identification of possibly resistant trees through inoculation, as it is not known whether some trees can prevent AOD infection altogether, or if resistance might simply translate to better ability to fight-off the infection and enter into permanent remission. One of the approaches that I proposed and started developing, was to transform the categorical classification of AOD (healthy, symptomatic, remission) into a continuous phenotype. In order to do this, a variety of phenotypic descriptors related to tree health and AOD should be recorded and could be used in a machine learning algorithm, to create an AOD damage score for each tree. A similar approach has been developed recently (Finch et al., 2021). A continuous phenotype would increase the statistical power to identify whether AOD resistance has a heritable genetic component altogether, and if so, it would indeed facilitate genome-wide association studies (GWAS).

The second issue that prevented me from going forward with the study of the genetic basis of AOD was the sample size of 386 individuals, which is insufficient for such an ambitious GWAS, particularly with poorly characterized categorical phenotypes. My dataset is the initial tranche of data in a much larger AOD genomics study, some more of which has already taken place. Over the past four years, members of Prof. Richard Buggs lab have extended the oak dataset presented in this thesis by sequencing the whole genome of an additional 1,151 oak trees of both species, *Q. robur* and *Q. petraea*, from over 60 sites encompassing Britain entirely, thanks to generous funding from DEFRA, as part of Action Oak. These trees have been phenotyped with respect to AOD symptoms, diameter at breast height (DBH), canopy transparency, number of *Agrilus biguttatus* exit holes, and number of stem bleeds.

We hope that the expansion of this dataset, which now includes 1,561 oak trees, will allow to assess whether AOD resistance has a genetic component altogether and if that is the case, we hope that this sample size will give enough statistical power to perform a GWAS aimed at identifying variants associated with increased resistance to this complex disease. This would help conservation by allowing to identify, through genomics prediction, trees to be used for future plantings given their higher resistance to AOD. On the other hand, if AOD heritability is found to be very low and environmental factors appear to play a deciding role in AOD susceptibility and tolerance, it would still be a major finding as it would point towards a conservation strategy that concentrates on modifying the trees micro-environment, rather than focused on breeding programs. This project is inspired by similar research carried on *Fraxinus excelsior* (European ash tree) by Prof. Richard Buggs group during the last decade.

In a 2016 study (Harper et al., 2016), the group used transcriptomic data to identify molecular markers for tolerance of *F. excelsior* to ash dieback, a chronic tree disease caused by the fungal pathogen *Hymenoscyphus fraxineus* (Harper et al, 2016). Over 100 Danish ash trees with wide range of susceptibilities to ash dieback were scored for dieback damage and RNA sequenced. The identified SNPs and gene expression markers (GEMs) were used (independently) with the damaged scores in linear models aimed at identifying significant associations to ash die back tolerance. The most significant markers were functionally annotated, and many appeared to be related to MADS box transcription factors, suggesting a role of this family of genes in susceptibility to ash dieback. The top SNP and the top two GEMs markers were used as PCR target and tested to predict the phenotype in a test panel of over 50 trees, first singularly and then combined. The combined predicted score was highly indicative of the trees' phenotypes.

This project was followed by the study by Sollars et al. (2017) that sequenced and assembled the genome of a low-heterozygosity ash tree. Genetic diversity was assessed by mapping the sequencing data of additional 37 individuals to the newly assembled reference. A set of genome-wide SNPs was identified after mapping: over 50% of these were located within or close to genes. The population structure was inferred based on SNPs and revealed four distinct plastid network haplotypes.

This study also aimed to identify improved markers for susceptibility to ash dieback. RNA sequencing data from the previous publication (Harper et al., 2016), was re-analyzed by mapping it to the newly assembled ash genome and the GWAS was re-performed. The top GWAS markers were functionally annotated confirming to be associated with MADS box transcription factors genes. Following the hypothesis that MADS box transcription factors may be involved in disease tolerance through the modulation of secondary metabolites (Sollars et al., 2017), the metabolite composition of the leaves of 10 trees was analyzed with liquid chromatography. The results of this analysis suggested a correlation between iridoid glycoside and susceptibility to ash dieback, as higher levels of this metabolite were found in the genotypes with high susceptibility to *H. fraxineus*. More recently, Stocks et al. (2021) sequenced the genome of over 1,000 ash trees scored for ash die back damage using a pool-sequencing approach and identified 3,149 SNPs associated with ash dieback damage. Furthermore, the study successfully used this data to train a genomic prediction model which was able to predict tree health with high accuracy, therefore showing the potential of genomic prediction to aid breeding aimed at conserving highly polygenic traits.

In plants, GWAS were first successfully applied to crop species, such as maize and rice, due to their obvious economic importance and the early availability of reference genome assemblies (Huang and Han, 2014). In one of the first next generation GWAS on rice (Huang et al., 2010) over 500 landraces were genotyped for over 3 million genome-wide SNPs. Diversity between the landraces was assessed and their phylogenetic relationship determined based on SNPs. Finally, GWAS were conducted and led to the identification of 80 associations for 14 agronomic traits.

Rice and other important food crops have seen a large increase in the number of resources developed in recent years to aid the development of GWAS for important traits. A more recent example is the investigation by Shrestha et al. (2018), which relied on available SNP data to identify genes for Manganese (Mn) toxicity tolerance in 271 genotypes of *Oryza sativa L* and successfully identified six SNPs associated with Mn leaf damage and shoot Mn concentration. Similar GWAS have been conducted in many other economically plant species and many examples are available, such as in Chinese cotton and in the model plant organism *Arabidopsis thaliana* (Horton et al., 2014; Su et al., 2018). In forest trees research the need to apply high throughput sequencing technologies and GWAS in the identification of genes for important traits is augmented by generally long generation times, which make traditional methods such as bi-parental QTL mapping less feasible, in addition to the obvious gain in resolution given by analyzing genome-wide diversity rather than limiting the analysis to a set of known markers. Several GWAS have been performed in forest trees in recent years, such as in the investigation by Uchiyama et al. (2013) which focused on wood property and quantity of male strobili of *Cryptomeria japonica*, one of the most important forest trees in Japan. This study identified 6 SNPs significantly associated with these traits. Fruit trees have also been the subject of GWAS usually aimed at identifying variants associated with fruit quality, such as in *Pyrus pyrifolia* (Japanese pear) and in *Prunus persica* (peach) (Minamikawa et al., 2018). Interestingly, *P. persica* is closely related to *Q. robur*: previous research reported high levels of orthology between the oak and peach transcriptomes (Lesur et al., 2015) as well as macrosynteny, so that the peach genome has been used to aid the chromosome anchoring of the oak genome scaffolds that did not map to the dense SNP-based oak linkage map in the work of Plomion et al. (2018). A recently conducted GWAS in *P. persica* (Cao et al., 2016) characterized over 4 million SNPs in a population of over 100 individuals, in order to find association with 12 agronomic traits. Three different linear models were tested to identify association with the traits of interest: a general linear model, a general linear model with correction for population structure and a mixed linear model with population structure correction. This GWAS led to identification of several candidate genes for traits such as fruit shape, acidity level, fruit hairiness, flesh colour, soluble solid content and weight (Cao et al., 2016). The success of these and similar research efforts reinforced the potential of genomics and derived methods in the identification of markers for economically and ecologically important plant traits, such as disease resistance.

Therefore, if additional phenotypic data is collected, numerous GWAS for other traits of interests for foresters and ecologists could stem from the extended Oak dataset that stemmed from this thesis, now including 1,561 individuals. In addition, this large oak panel could be used to design an EEA aimed at identifying climate-adaptive alleles in view of the ongoing climate change and inform conservation strategies. A warmer and drier future climate may favour *Q. petraea* ecology over *Q. robur*, therefore adaptive introgression among these species could play an important role in the future. Such study could also include existing whole genome sequencing datasets of oak from Europe (Leroy et al., 2019b), to cover a larger latitudinal range. These analyses could inform breeding programs to preserve oaks, particularly in southern Britain, which is the area mostly affected by AOD and more severely threatened by rising temperature.

Finally, the expanded dataset spans a larger latitudinal range, and includes a more balanced number of individuals of both oak species. This may allow for new studies of species differentiation with greatly improved confidence and transferability of results. In particular this could allow to detect *Q. robur* and *Q. petraea* interspecific outlier loci with more confidence decreasing false positives and could lead to the identification of a core set of genomic regions always differentiated between the two species independently of location.

The Asian White Birch project

The birch project led to several conclusions in the contexts of evolution and ecology of Asian White Birch. First, I found strong genetic differentiation between *B. platyphylla* and the closely related sister species silver birch, *B. pendula*, from Europe. These two species have been often lumped taxonomically as sub-species (Ashburner & McAllister, 2013), however my analysis strongly suggests that they should be treated as different species, as the levels of differentiation between *B. pendula* and *B. platyphylla* populations are significantly higher than between *B. platyphylla* Chinese populations. Furthermore, I identified a hybrid zone between these two species in north-western China in the area including and adjacent to the Altay mountains in the Xinjiang region: all the individuals sampled in this area displayed approximately 80% *B. pendula* and 20% *B. platyphylla* genetic composition, suggesting that this region is a contact point between *B. pendula* from Eurasia and *B. platyphylla* from north-eastern China. Interestingly, hybrid zones for other tree species have been found in north-western and northern China. *Populus × jrtyschensis*, a hybrid poplar species distributed in this region of China, appears to have resulted from the admixture between *P. nigra* from Europe and *P. laurifolia* from Asia converging in this region (Jiang et al., 2016). The Altay region in northern Xinjian is also a natural hybrid zone between two other *Populus* species, the ecologically distinct *P. alba* and *P. tremula* (Zeng et al., 2016). North and north-eastern China are a contact point between two Asian oak species, *Quercus liaotungensis* which has more of a western distribution in China, and *Q. mongolica,* which is more diffused in north-eastern China, North-Korea and Japan (Zeng et al., 2011). Similar to the findings of the study by Tsuda et al. (2017), that identified a hybrid zone for eastern and western *Betula* species pairs, *B. pendula/B. platyphylla* and *B. pubescens/B. ermanii*, in southern Siberia around Lake Baikal, this region was also identified as a contact point between three willows species: *Salix dasyclados* and *S. viminalis* from Europe and western Asia and

S. schwerinii from Eastern Asia (Fogelqvist et al., 2015). Similarly, a hybrid zone between Siberian spruce, *Picea obovate*, from Asia, and Norway spruce, (*Picea abies),* from Europe, was found centred around the Urals (Tsuda et al., 2016).

Within China, I found evidence of three lineages of *B. platyphylla* differentiated along a latitudinal gradient, in north-eastern, central, and south-western China. The central population may have resulted from the admixture between the north-eastern and south-western population. The levels of differentiation between the central and north-eastern population are significantly lower than those between the central and south-western population, probably reflecting less severe geographical barriers to geneflow in northern China, compared to the south-western Himalayan plateau. The population divergence analysis suggested that the three identified lineages have been fully separated for at least 500,000 years, therefore their divergence appear to have occurred prior to the LGM, suggesting that *B. platyphylla* survived the last glaciation in multiple local glacial refugia in China, which is in line with fossil pollen records and ENM dated at the LGM (Cao et al., 2015; Chen & Lou, 2019).

The environmental niche model presented in this thesis agrees with the observed distribution of this species in China (Chen & Lou, 2019) and is similar to the distribution model reported for *B. platyphylla* in Duan et al. (2014). Annual levels of precipitation and mean temperature of wettest quarter, which correspond to the summer season in China, are important predictors of this species suitable environment. Overall, the ENM suggests that *B. platyphylla* prefers mountainous environments characterized by mild summers and moderate levels of precipitation throughout the year, which is consistent with the known ecology of this species (Ashburner & McAllister, 2013). Future predictions of *B. platyphylla* distribution show a significant reduction of its habitat in China compared to the present, mainly due to rising temperature, from an estimated minimum reduction of 17% up to a maximum of 34% by 2080-2100, in the worst-case scenario tested.

I detected numerous associations between *B. platyphylla* SNPs and environmental variables, with two variables in particular reporting a much larger number of significant SNPs. This suggests that local adaptation is highly polygenic and involves loci distributed throughout the genome of this species. This is in line with similar studies in other tree species, which reported a similar percentage of identified putatively adaptive SNPs (~ 0.5% in this analysis) out of all the SNPs tested and a highly polygenic nature of climate adaptation (Borrell et al., 2019; Dauphin et al., 2020; Jordan et al., 2017; Pina-Martins et al., 2018; Rellstab et al., 2016; Yeaman et al., 2016). This may suggest that maintaining standing genetic variation and adaptive diversity might be of greater importance to support local adaptation than attempting to raise the frequencies of a specific set of adaptive alleles.

Furthermore, I used a recently developed approach, RONA (Pina-Martins, Baptista, Pappas & Paulo, 2018; Rellstab et al., 2016) to infer the degree to which each individual allele frequencies at adaptive loci deviate for the optimum required at a future climate scenario for 2080-2100. This analysis predicted that populations in central China and in the area adjacent to the Changbai mountains in the north-east might be more severely threatened by climate change, particularly by rising temperature, with required frequencies changes exceeding expectations that can be theoretically reached naturally (Jordan et al., 2017). This is reflected by the predicted future suitable environment (Figure 3.5), which shows drastic reduction in these two regions of China, under every socioeconomic pathway tested. It is indeed plausible that some environmental variables are limiting *B. platyphylla* range precisely because they lack adaptation, or it is also possible that some climatic variable will change to an extent that adaptation will not be possible anymore, and some regions of China where *B. platyphylla* is currently found, may irreversibly become unsuitable in the future therefore assisted migration might be the only option available to rescue some populations. The Himalayan plateau and the greater Khingan range and surrounding area, and in general environments at higher elevations, appear less threatened by rising temperature and appear to remain suitable for *B. platyphylla* under every future climate scenario tested. Consequently, the estimated adaptive loci allelic frequency changes required to match future climate for the individuals at these locations appear less concerning. Interestingly, similar to the findings of a landscape genomics study on *Quercus aquifolioides* (Du et al., 2020), which is distributed in south-western and central China, the south-western population reported on average higher RONA scores for "precipitation of driest quarter" compared to the central and north-eastern lineages (Figure 3.6). Precipitation of driest quarter are expected to increase in the Himalayan plateau in the future, whereas are expected to slightly decrease in central and north-eastern China.

The RONA method is itself a simplification (Rellstab et al., 2016), and was originally designed to be used with population allelic frequencies. As I discussed in Chapter 3, due to our sampling it was only possible to calculate RONA on an individual basis. Therefore, this analysis relies on the assumption that the individuals sampled are good representatives of their population of provenance allelic frequencies. The differences in RONA between populations is determined by two main factors: the magnitude of the environmental change for each population and the deviation of the current allele frequencies of each population from the association model (Rellstab et al., 2016). There is another method, termed genetic offset, which is used to estimate the risk of maladaptation to future conditions (Fitzpatrick and Keller, 2015). However, the most recent implementation of this method (Gain & Francois, 2021) aims to give a global measure of maladaptedness by including all the environmental variables tested together as well as all genotyped SNPs in the model. Therefore, I preferred the RONA implementation as it is based solely on previously identified adaptive SNPs and allows to assess each environmental variable separately, which is advantageous as these can have significantly different importance in limiting a species geographic range (Table 3.2).

Future directions for research on Asian White Birch

The sampling design of this study allowed to detect genomics signals of local adaptation, however having sampled a single individual for the vast majority of the populations limited the choice of tools that I could use for EAA. A superior design would involve sampling a smaller number of populations scattered throughout the same geographic area but including more individuals in each population. This would allow to use population allelic frequencies in the RONA calculation, rather than individuals' frequencies, which are limiting as can take only three values (0, 0.5 or 1) in linear regressions. This would give a more reliable RONA estimate which in turn would better inform conservation, by more precisely detecting natural forests that might require conservation strategies such as assisted gene flow to aid future climate adaptation, particularly due to rising temperature.

Having more than a single individual per population would also enable to use landscape genomics methods based on outlier F_{st} loci between populations, such as Bayescan (Foll & Gaggiotti, 2008) followed by Bayenv (Gunter & Coop, 2013), to detect association between environmental variables and SNPs (Borrell et al., 2019). This would allow to select only adaptive SNPs consistently identified by more than one tool to increase confidence in the results and decrease the number of false positives.

For comparison, I performed the environmental association analysis also using another software, Samβada (Joost et al., 2007; Stucki et al., 2017), discussed in Chapter 3. However, the negative side of this approach is that is overly conservative when the population structure is included in the model, as it aims to identify strong associations between SNPs and environment by testing each climatic variable separately, rather than aiming to predict the genotype of an individual based on its environment, like LFMM2. Samβada identified only 14 candidates at the same significance level (FDR < 1%) of LFMM2, also possibly due to a highly polygenic mechanism of adaptation to the environment, with many SNPs with tiny effects which might not result as significant with some conservative approaches such as Samβada.

Interesting research that could follow up would be to investigate the convergent and divergent local adaptation between *B. platyphylla* and the closely related *B. pendula*, as previous work has shown that adaptation to certain environmental variables can be constrained and repeatable genetically, with key genes playing non-redundant roles even in distantly related tree species (Yeaman et al., 2016). This would involve first the design of an EAA on *B. pendula* aimed at detecting signals of local adaptation like it was done for *B. platyphylla* in the work presented in this thesis, and subsequently it would be possible to compare adaptation signals between the two species and find overlapping associations, between both SNPs loci and genes. Given the results from previous research that compared convergent local adaption between lodgepole pine and interior spruce and identified between 10 and 18% of putative locally adaptive genes evolving convergently (Yeaman et al., 2016), I would expect a greater proportion of convergent adaptive loci for the *B. platyphylla/B. pendula* complex given their closer phylogenetic relationship. However, the highly polygenic mechanism of adaptation to some environmental variable may translate to greater genetic redundancy, and therefore exhibit less repeatability of genetic evolution even across closely related species (Yeaman et al., 2015). Another interesting finding that emerged from the comparative study by Yeaman et al. (2016), regards the role of macro and micro genetic rearrangements, such as chromosomal fusions, translocations and gene duplications, and that of clustered genetic structures, in local adaptation (Yeaman et al., 2016; Yeaman, 2022). From previous preliminary analysis on plant and animal species (Yeaman et al., 2016), it appears that duplicated genes are more likely to be involved in local adaptation. This could mean that gene duplication could be a way to increase genetic flexibility, and could facilitate convergent genotypic evolution (Yeaman et al., 2016). The role of pleiotropy in the repeatability of local adaptation is also another topic that could be investigated further, through comparative adaptation studies (Yeaman et al., 2021).

Ultimately, in order to verify the adaptative capabilities of *B. platyphylla*, genomics approaches should be coupled with reciprocal transplant garden experiments. These experiments involve growing different ecotypes among reciprocal habitats and have been the gold standard to study local adaptation over the last century (Cheplick, 2015; Johnson et al., 2021). With reciprocal transplant experiments, it is possible to distinguish between genetic and environmental control of phenotypic variation and to verify whether local adaptation is actually occurring. Local populations should perform better in their habitat than foreign populations, if they are locally adapted (Johnson et al., 2021). Traditionally these experiments measured traits related to morphology, growth and physiology (Johnson et al., 2021), and by coupling these with landscape genomics methods lead to a deeper understanding of the ecology of adaptation, by connecting the function of genes associated with environmental variables in EAAs to actual phenotypic variation, and to subsequently identify alleles that give rise to phenotypes better suited for certain environments. Reciprocal transplant experiments however come with several limitations and challenges, particularly when the species of interest is long-lived and has a long generation time, as it is often the case in forest trees (Manzanedo et al., 2019). First, reciprocal transplants are not always possible due to practical, legal, ethical or ecological reasons: a primary concern that arise from moving plant material is the spread of pests and other infectious diseases, as well as the spread of invasive genotypes (Manzanedo et al., 2019). These issues can be avoided with the use of common gardens, which can also be set up within forests, however these may limit the realism of the experiment by removing the fine-scale effects of the environment and by excluding any intra and interspecific competition (Manzanedo et al., 2019). In addition, reciprocal transplant experiments in long-lived species often require significant labour and need to last for at least a few decades, as it was shown that short-term experiments can produce biased results in long-lived species and may fail to detect local adaptation (Bennington et al., 2012; Manzanedo et al., 2019).

To conclude, the analysis on Asian White Birch presented in this thesis provided important information on both ecology and evolution of this species from a genomic perspective. Further research may also expand this dataset to include more individuals from the natural populations sampled to increase confidence in the analysis, and subsequently investigate the functional potential of the identified adaptive genic regions in more depth and compare it with existing results from similar studies on other tree species.

Supporting Information

Chapter 2

Table S2.1. Summary table of the 81 gene models within or flanking the top 1% F_{st} outlier-enriched 10kb windows between *Q. robur* and *Q. petraea*. Protein definition retrieved from Plomion et al. (2018).

ID	Type	Start	End	Gene ID	Protein Definition
Qrob_Chr01	mRNA	24923692	24927355	ID=Qrob_T0001470.2	Hypothetical protein
Qrob_Chr01	mRNA	37595498	37602288	ID=Qrob_T0532570.2	Uncharacterized conserved protein [Function unknown]
Qrob_Chr01	mRNA	38691337	38694866	ID=Qrob_T0184060.2	NADH dehydrogenase (ubiquinone) Fe-S protein 1
Qrob_Chr01	mRNA	38697663	38699156	ID=Qrob_T0184050.2	NADH dehydrogenase (ubiquinone) Fe-S protein 1
Qrob_Chr01	mRNA	55066440	55068599	ID=Qrob_T0611170.2	Genomic DNA, chromosome 3, tac clone: k13n2-related
Qrob_Chr02	mRNA	17419146	17420228	ID=Qrob_T0454110.2	Protein LTV1
Qrob_Chr02	mRNA	28382238	28397120	ID=Qrob_T0299620.2	Predicted ATP-dependent RNA helicase FAL1 involved in rRNA maturation DEAD-box superfamily [Translation ribosomal structure and biogenesis].
Qrob_Chr02	mRNA	28400301	28401018	ID=Qrob_T0299610.2	Ataxin-2 C-terminal region
Qrob_Chr02	mRNA	28404084	28407207	ID=Qrob_T0299600.2	PPR repeat
Qrob_Chr02	mRNA	28987060	28998888	ID=Qrob_T0089890.2	DNA-directed RNA polymerase IV and V subunit 2
Qrob_Chr02	mRNA	29002128	29005412	ID=Qrob_T0089880.2	Chorismate mutase
Qrob_Chr02	mRNA	34712751	34715369	ID=Qrob_T0437960.2	Phospholipase d
Qrob_Chr02	mRNA	42655612	42657345	ID=Qrob_T0589320.2	GRAS domain family
Qrob_Chr02	mRNA	44205331	44205970	ID=Qrob_T0377200.2	NA
Qrob_Chr02	mRNA	44213324	44216710	ID=Qrob_T0377190.2	NA
Qrob_Chr02	mRNA	46990675	46991783	ID=Qrob_T0609890.2	Prohibitin 1
Qrob_Chr02	mRNA	46996815	47001368	ID=Qrob_T0609880.2	Glyceraldehyde-3-phosphate dehydrogenase (phosphorylating)

Qrob_Chr02	mRNA	47003403	47004571	ID=Qrob_T0609870.2	Anion-transporting ATPase
Qrob_Chr02	mRNA	49365511	49368425	ID=Qrob_T0709860.2	Reticulon-related (plant)
Qrob_Chr02	mRNA	49383529	49390427	ID=Qrob_T0709880.2	NA
Qrob_Chr02	mRNA	50880650	50885782	ID=Qrob_T0528660.2	Protein of unknown function
Qrob_Chr02	mRNA	50889161	50893177	ID=Qrob_T0528650.2	GTP-binding protein SEC4 small G protein superfamily and related Ras family GTP-binding proteins [Signal transduction mechanisms Intracellular trafficking secretion and vesicular transport]
Qrob_Chr02	mRNA	50904986	50905468	ID=Qrob_T0528630.2	NA
Qrob_Chr02	mRNA	53983164	53985503	ID=Qrob_T0700420.2	FAR1 DNA-binding domain
Qrob_Chr02	mRNA	53992359	54004961	ID=Qrob_T0700440.2	Starch phosphorylase
Qrob_Chr02	mRNA	92431869	92436245	ID=Qrob_T0282290.2	Bile acid: Na+ symporter, BASS family
Qrob_Chr03	mRNA	40167291	40167596	ID=Qrob_T0745850.2	NA
Qrob_Chr03	mRNA	40169165	40169425	ID=Qrob_T0745840.2	NA
Qrob_Chr03	mRNA	40171988	40176838	ID=Qrob_T0745830.2	Solute carrier family 35 (UDP-sugar transporter), member A1/2/3
Qrob_Chr04	mRNA	27997801	27998563	ID=Qrob_T0399160.2	RING-H2 zinc finger protein RHA1
Qrob_Chr04	mRNA	41651662	41655844	ID=Qrob_T0308170.2	Leucine-rich repeat protein [Function unknown]. [Cytoskeleton]
Qrob_Chr04	mRNA	41665218	41666719	ID=Qrob_T0308190.2	NA
Qrob_Chr05	mRNA	16208833	16211432	ID=Qrob_T0178490.2	Calmodulin-binding family protein
Qrob_Chr05	mRNA	17212822	17227633	ID=Qrob_T0178670.2	DNA repair/transcription protein
Qrob_Chr05	mRNA	17232249	17239413	ID=Qrob_T0178680.2	RNA polymerase II-associated factor 1
Qrob_Chr05	mRNA	2032128	2035537	ID=Qrob_T0649050.2	(S)-2-hydroxy-acid oxidase.
Qrob_Chr05	mRNA	2041278	2044530	ID=Qrob_T0649030.2	(S)-2-hydroxy-acid oxidase.
Qrob_Chr05	mRNA	22675622	22677555	ID=Qrob_T0583160.2	Ankyrin repeat-containing protein
Qrob_Chr05	mRNA	22684391	22686918	ID=Qrob_T0583150.2	Heat stress transcription factor a-6a-related

Qrob_Chr05	mRNA	22704507	22708866	ID=Qrob_T0013730.2	Indeterminate-domain 12 protein
Qrob_Chr05	mRNA	22714427	22722619	ID=Qrob_T0013740.2	Hypoxia-inducible factor-asparagine dioxygenase.
Qrob_Chr05	mRNA	23925445	23929495	ID=Qrob_T0523430.2	Bromo-adjacent homology (bah) domain-containing protein-related
Qrob_Chr05	mRNA	23941803	23946877	ID=Qrob_T0523410.2	Serine-threonine kinase receptor-associated protein
Qrob_Chr05	mRNA	28012467	28013501	ID=Qrob_T0449720.2	Leucine-rich repeat receptor-like protein kinase
Qrob_Chr05	mRNA	7906041	7912235	ID=Qrob_T0683210.2	26S proteasome regulatory subunit N7
Qrob_Chr06	mRNA	21260908	21264170	ID=Qrob_T0005740.2	Carbamoyl-phosphate synthase (glutamine-hydrolyzing)
Qrob_Chr06	mRNA	27193253	27195122	ID=Qrob_T0256040.2	NA
Qrob_Chr06	mRNA	28622710	28630237	ID=Qrob_T0747960.2	Arm repeat superfamily protein
Qrob_Chr06	mRNA	40963979	40968076	ID=Qrob_T0413630.2	BTB/POZ domain/ NPH3 family
Qrob_Chr06	mRNA	40974308	40977654	ID=Qrob_T0413610.2	Mitochondrial outer membrane protein porin 2-related
Qrob_Chr07	mRNA	37924676	37927169	ID=Qrob_T0131310.2	BTB/POZ domain
Qrob_Chr07	mRNA	37930043	37933882	ID=Qrob_T0131320.2	Homeobox protein transcription factors
Qrob_Chr07	mRNA	4037164	4038276	ID=Qrob_T0379860.2	CCR4-associated factor 1 homolog 11-related
Qrob_Chr07	mRNA	4038746	4042140	ID=Qrob_T0379870.2	Alpha, alpha-trehalose-phosphate synthase [udp-forming] 10-related
Qrob_Chr07	mRNA	41623546	41636343	ID=Qrob_T0090900.2	Letm1-like protein
Qrob_Chr07	mRNA	46049058	46056801	ID=Qrob_T0407530.2	1-phosphatidylinositol-4-phosphate 5-kinase
Qrob_Chr07	mRNA	7062267	7067123	ID=Qrob_T0265770.2	L-ascorbate peroxidase
Qrob_Chr08	mRNA	10656567	10658485	ID=Qrob_T0132290.2	WWE domain
Qrob_Chr08	mRNA	25970299	25973975	ID=Qrob_T0441200.2	3-deoxy-manno-octulosonate cytidylyltransferase
Qrob_Chr08	mRNA	25978580	25979250	ID=Qrob_T0441190.2	Dynein light chain type 1-like protein
Qrob_Chr08	mRNA	50962263	50985338	ID=Qrob_T0437680.2	Rabconnectin-related

Qrob_Chr08	mRNA	52158757	52169697	ID=Qrob_T0631150.2	Region in Clathrin and VPS / Golgi CORVET complex core vacuolar protein 8
Qrob_Chr08	mRNA	52173071	52176597	ID=Qrob_T0631140.2	Pyruvate dehydrogenase e1 component, alpha subunit
Qrob_Chr08	mRNA	52181642	52185377	ID=Qrob_T0631130.2	Golgi transport complex COD1 protein [Intracellular trafficking secretion and vesicular transport]
Qrob_Chr08	mRNA	62286008	62293369	ID=Qrob_T0413010.2	NA
Qrob_Chr08	mRNA	62295320	62296804	ID=Qrob_T0413000.2	Histone H4
Qrob_Chr08	mRNA	62298760	62300062	ID=Qrob_T0412990.2	Histone H4
Qrob_Chr08	mRNA	62746427	62753205	ID=Qrob_T0605490.2	Leukocyte receptor cluster (LRC) member 8
Qrob_Chr08	mRNA	62756677	62757416	ID=Qrob_T0605480.2	Chlorophyll A-B binding protein
Qrob_Chr08	mRNA	62783709	62786511	ID=Qrob_T0605470.2	ABC transporter g family member 1-related
Qrob_Chr08	mRNA	62794403	62797526	ID=Qrob_T0211590.2	U4/U6 small nuclear ribonucleoprotein PRP3
Qrob_Chr09	mRNA	26419438	26420323	ID=Qrob_T0285940.2	PPR repeat family
Qrob_Chr09	mRNA	26422819	26423321	ID=Qrob_T0285930.2	NA
Qrob_Chr09	mRNA	41156949	41176337	ID=Qrob_T0489020.2	Beige/beach-related
Qrob_Chr10	mRNA	48860710	48861429	ID=Qrob_T0644820.2	Predicted E3 ubiquitin ligase [Posttranslational modification protein turnover chaperones]
Qrob_Chr10	mRNA	48864519	48865241	ID=Qrob_T0644810.2	E3 ubiquitin-protein ligase RNF144
Qrob_Chr11	mRNA	15021331	15022690	ID=Qrob_T0104460.2	Cytochrome c6
Qrob_Chr11	mRNA	33429757	33451483	ID=Qrob_T0010570.2	PI-3-kinase-related kinase SMG-1
Qrob_Chr11	mRNA	51079119	51086901	ID=Qrob_T0251500.2	Inorganic phosphate transporter [Inorganic ion transport and metabolism]
Qrob_Chr11	mRNA	51094139	51094775	ID=Qrob_T0251490.2	U3 small nucleolar RNA-associated protein 18
Qrob_Chr12	mRNA	32347782	32354615	ID=Qrob_T0149640.2	BRCA1-associated RING domain protein 1

Table S2.2. Summary table of *Q. robur* candidate genes under recent selection detected with both SweeD and OmegaPlus.

ID	Start	End	Gene ID	Description
Qrob_Chr01	21470613	21471928	ID=Qrob_T0000290.2	Hydrophobic seed protein. Protein expressed on the seed surface that seem to play a role in seed survival by regulating water-uptake and affecting pathogen attachment and penetration (Gijzen et al., 1999).
Qrob_Chr01	47849967	47852265	ID=Qrob_T0041720.2	Beta/Galactosidase/2 related
Qrob_Chr02	9550828	9552258	ID=Qrob_T0433250.2	Vacuolar H+-ATPase, a proton pump found in organelle membrane involved in ion, metabolites and pH homeostasis in plants that appears to cover a crucial role in plant tolerance to salt induced stress (Golldack & Dietz, 2001; Padmanaban et al., 2004; Ratajczak, 2000; Zhang et al., 2012).
Qrob_Chr03	36770831	36774503	ID=Qrob_T0102900.2	Zinc-finger proteins. Transcription factors that regulate plant growth and are thought to be involved in the response to environmental stresses such as salinity, cold and draught (Han et al., 2020).
Qrob_Chr03	49666659	49673966	ID=Qrob_T0170450.2	Copines, a class of calcium-dependent phospholipid-binding proteins linked to disease resistance and acclimation in *A. thaliana* and *Triticum aestivum* (wheat) (Jambunathan & McNellis, 2003; Zou et al., 2016; Zou, Ding, Liu, & Hua, 2017).
Qrob_Chr09	13945795	13946619	ID=Qrob_T0542880.2	Cytoskeletal regulator Flightless-I
Qrob_Chr11	5403212	5411610	ID=Qrob_T0080510.2	NA
Qrob_Chr11	44537895	44539098	Name=Qrob_T0158900.2	Tyrosine kinases, key transmembrane receptors involved in signal transduction and linked to plant growth and response to both biotic and abiotic

				stresses (Miyamoto et al., 2019)
Qrob_Chr11	38960405	38963858	ID=Qrob_T0699490.2	Patellin proteins, phosphatidylinositol membrane transfer proteins seemingly involved in development, response against salt stress and immunity against certain viruses in plants (Peiro et al., 2014; Zhou et al., 2019).
Qrob_Chr12	33369924	33377321	ID=Qrob_T0662860.2	Cytoskeletal regulator Flightless-I

Table S2.3. Length variants in base pairs and point mutations detected in cpDNA fragments of five identified haplotypes. Key features used to match haplotypes are underlined. A) DT fragment digested with TaqI. B) AS fragment digested with HinfI. C) CD fragment digested with TaqI. D) TF fragment digested with AluI. E) Point mutations in the DT and TF fragments, identified with AluI and CfoI, respectively.

A.

Haplotypes (Petit et al., 2002b)	**Lineage**	**DT1**	**DT2**	**DT3**	**DT4**
10	B	571	389	288	215
11	B	571	389	288	215
12	B	571	389	287	215
7,26	A	571	389	215	211

B.

Haplotypes (Petit et al., 2002b)	**Lineage**	**AS1**	**AS2**	**AS3**	**AS4**	**AS5**	**AS6**
10	B	677	569	516	370	310	211
11	B	677	569	516	370	310	210
12	B	677	569	516	370	310	211
7,26	A	649	569	516	370	310	210

C.

Haplotypes (Petit et al., 2002b)	**Lineage**	**CD1**	**CD2**	**CD3**	**CD4**	**CD5**	**CD6**
10	B	972	658	533	322	291	259

11	B	972	658	533	322	291	259
12	B	972	658	533	322	291	259
7,26	A	972	658	533	322	291	259

D.

Haplotypes (Petit et al., 2002b)	**Lineage**	**TF1**	**TF2**	**TF3**	**TF4**	**TF5**	**TF6**
10	B	981	665	89	65	45	30
11	B	981	678	89	65	45	30
12	B	981	665	89	65	45	30
7,26	A	980	664	89	65	45	30

E.

Haplotypes (Petit et al., 2002b)	**Lineage**	**DT-AluI**	**TF-CfoI**
10	B	Yes	No
11	B	Yes	No
12	B	Yes	No
7, 26	A	No	Yes

Figure S2.1. Results of the fastSTRUCTURE analysis. A) Log-marginal likelihood lower bound (LLBO) of the data for K from 1 to 10 for the 386 individuals' dataset including both species. B) Cross-validation error calculated using 5-fold cross-validation for K from 1 to 10 for the 386 individuals' dataset including both species, with standard error bars. C) Log-marginal likelihood lower bound (LLBO) of the data for K from 1 to 10 for the 360 *Q. robur* individuals. D) Cross-validation error calculated using 5-fold cross-validation for K from 1 to 10 for the 360 *Q. robur* individuals, with standard error bars.

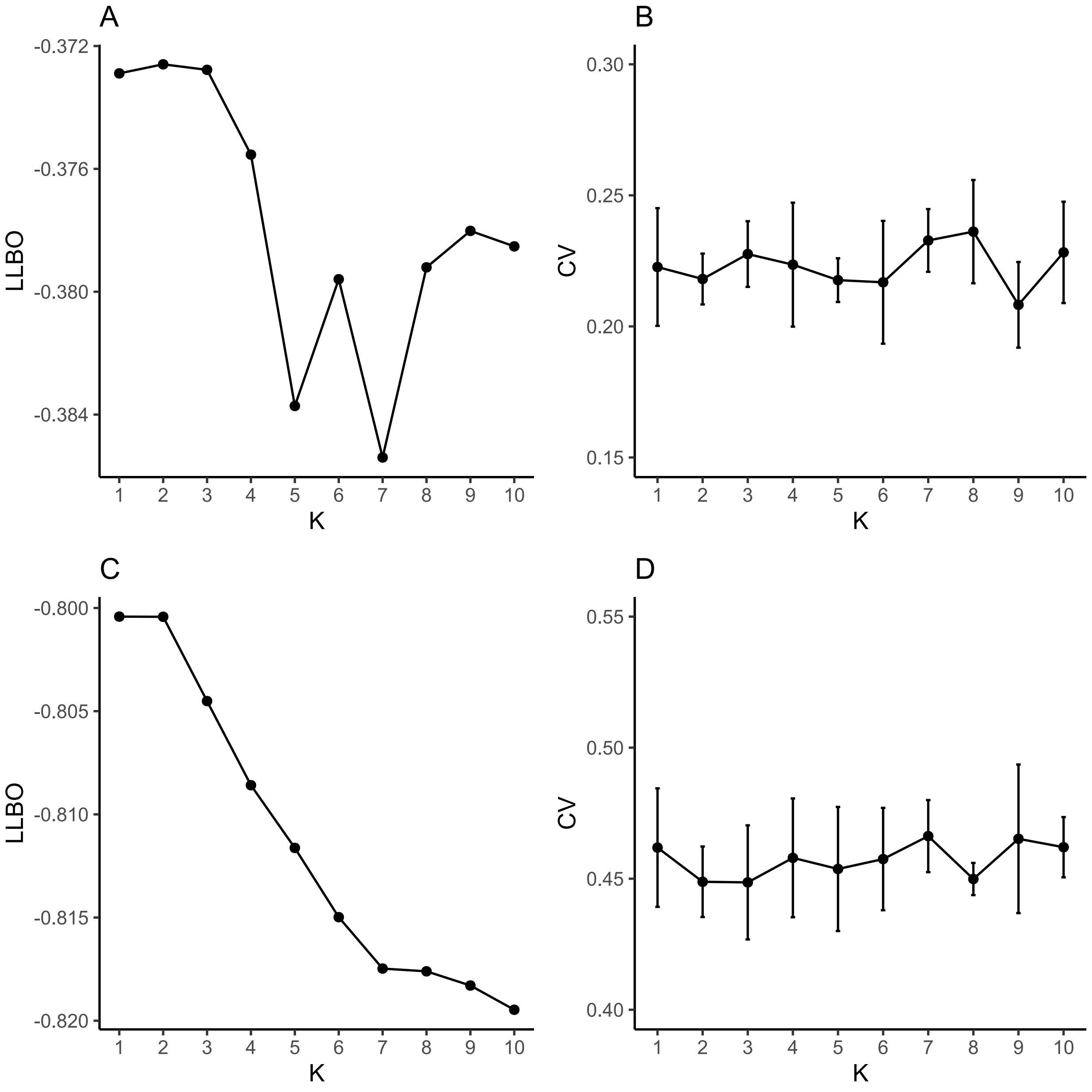


Figure S2.2. Genome-wide, genic and intergenic interspecific F_{st} distribution between *Q. robur* and *Q. petraea* based on 914,242, 223,319 and 690,923 SNPs, respectively. Genome-wide mean, median and standard deviation: 0.158, 0.111 and 0.156. Genic regions mean, median and standard deviation: 0.155, 0.111 and 0.153. Intergenic regions mean, median and standard deviation: 0.159, 0.111 and 0.157.

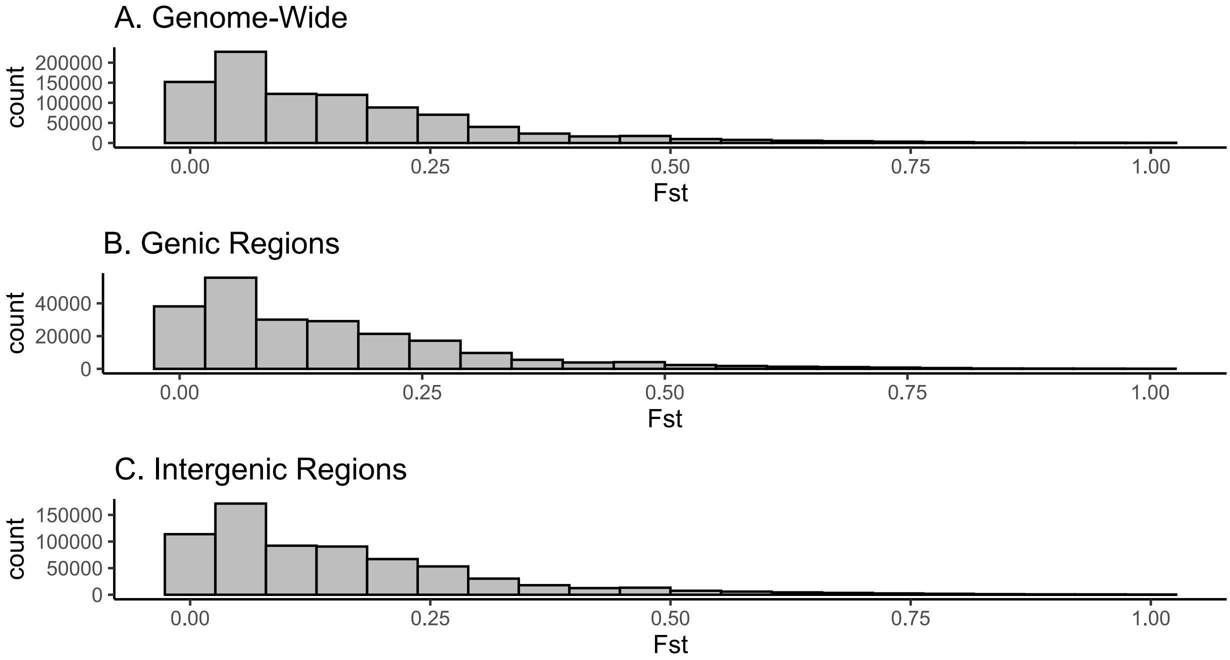


Figure S2.3. F_{st}/Heterozygosity distribution computed with the R package “fsthet”, based on 914,242 genome-wide SNPs between 10 *Q. robur* and 10 *Q. petraea* individuals. Red lines delimit the 0.99 confidence envelope.

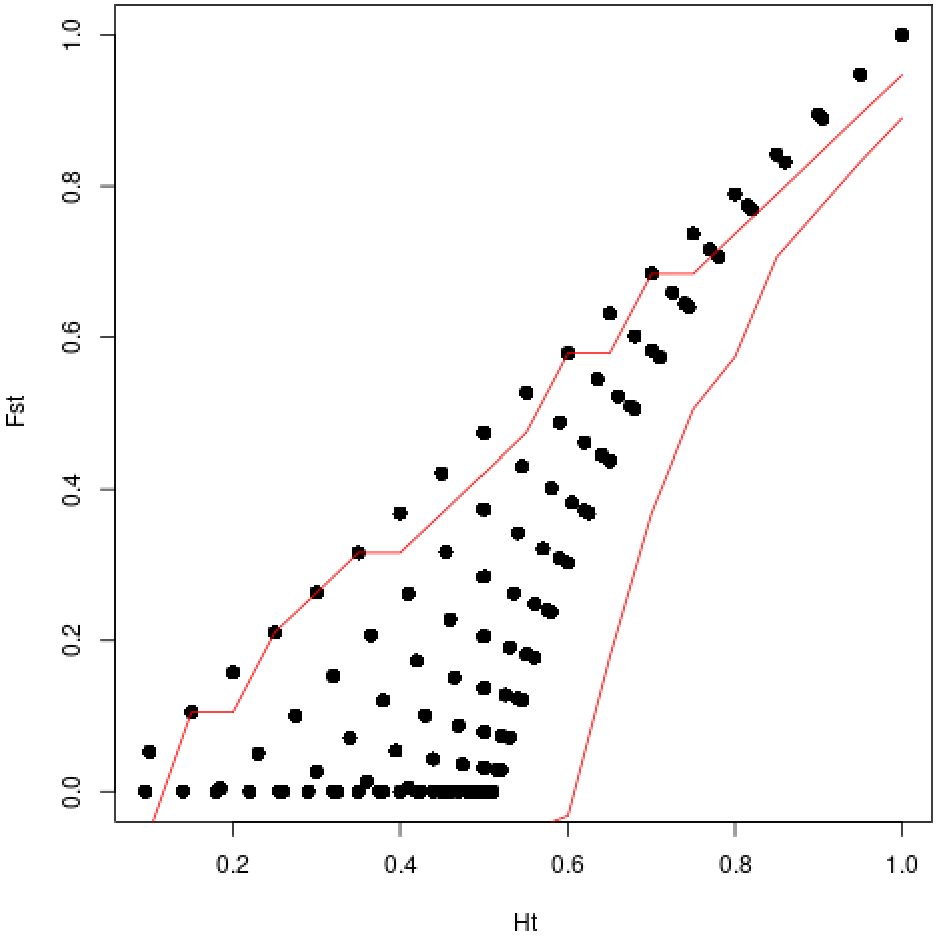


Figure S2.4. PCA of 360 *Q. robur* individual based on 839,911 unlinked SNPs ($r^2 < 0.4$). Colours represent sites. A) PC1 against PC2. B) PC1 against PC3. C) PC2 against PC3. D) Eigenvalues of the computed principal components.

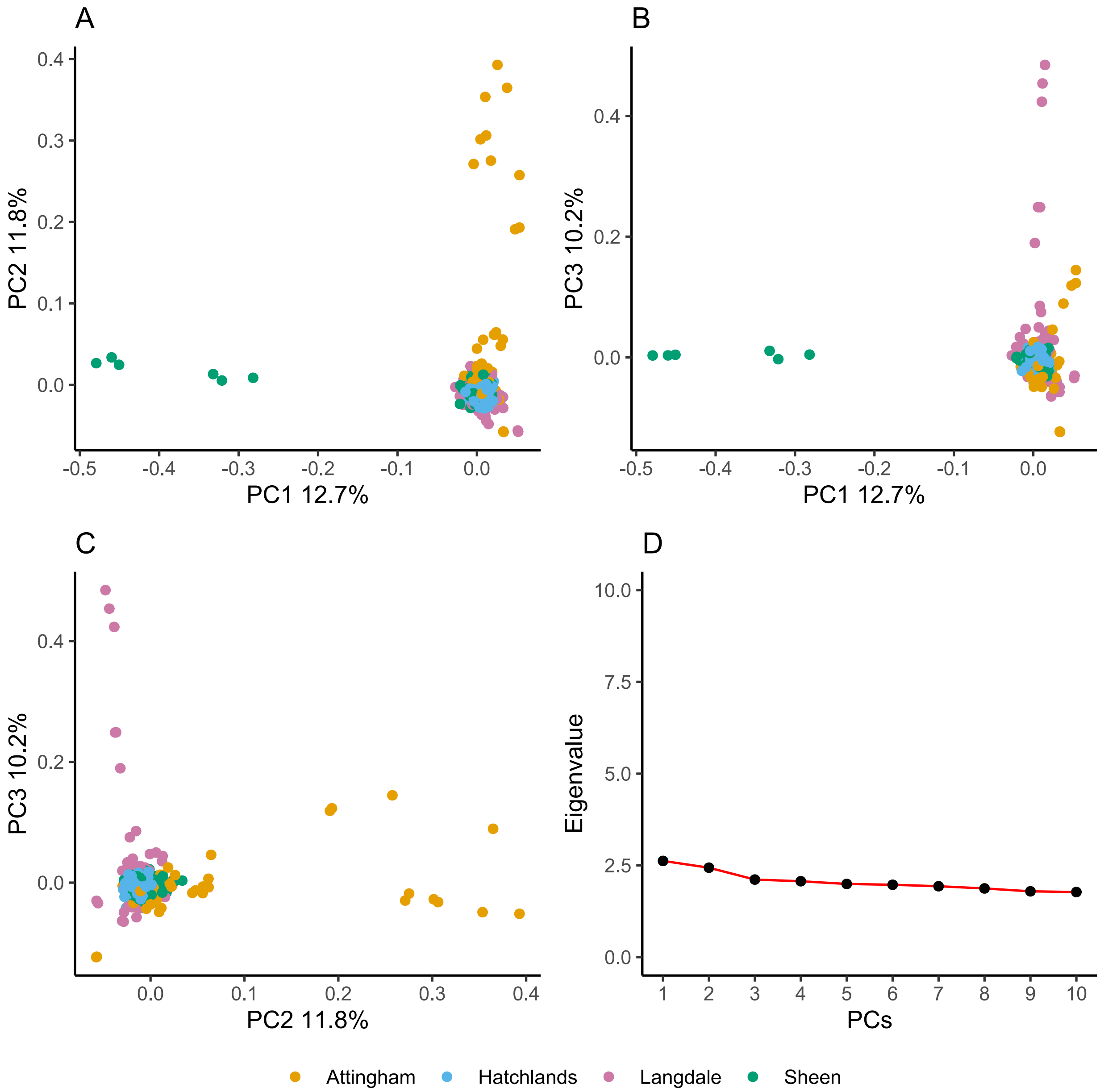


Figure S2.5. Distribution of genomic relatedness (kin) between 360 *Q. robur* trees based on 839,911 unlinked ($r^2 < 0.4$) SNPs with minor allele frequencies above 0.05.

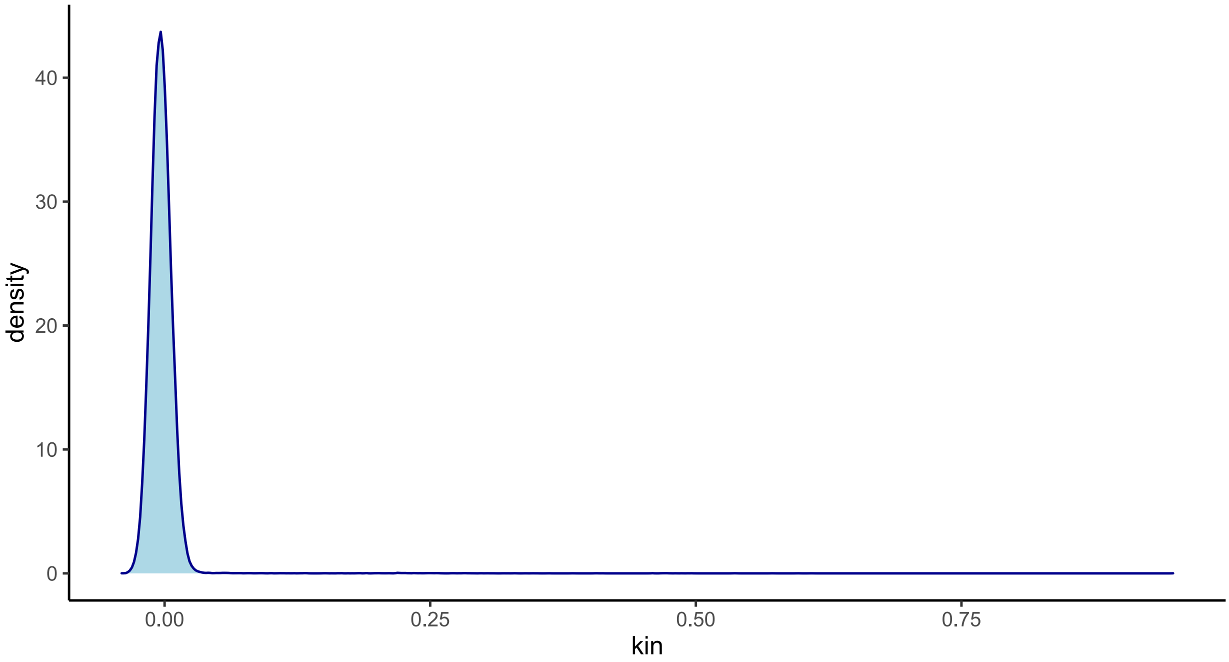


Figure S2.6. Marker based genomic relatedness within parkland sites, calculated according to the formulas in vanRaden (2008). Diagonals represent self-to-self relatedness.

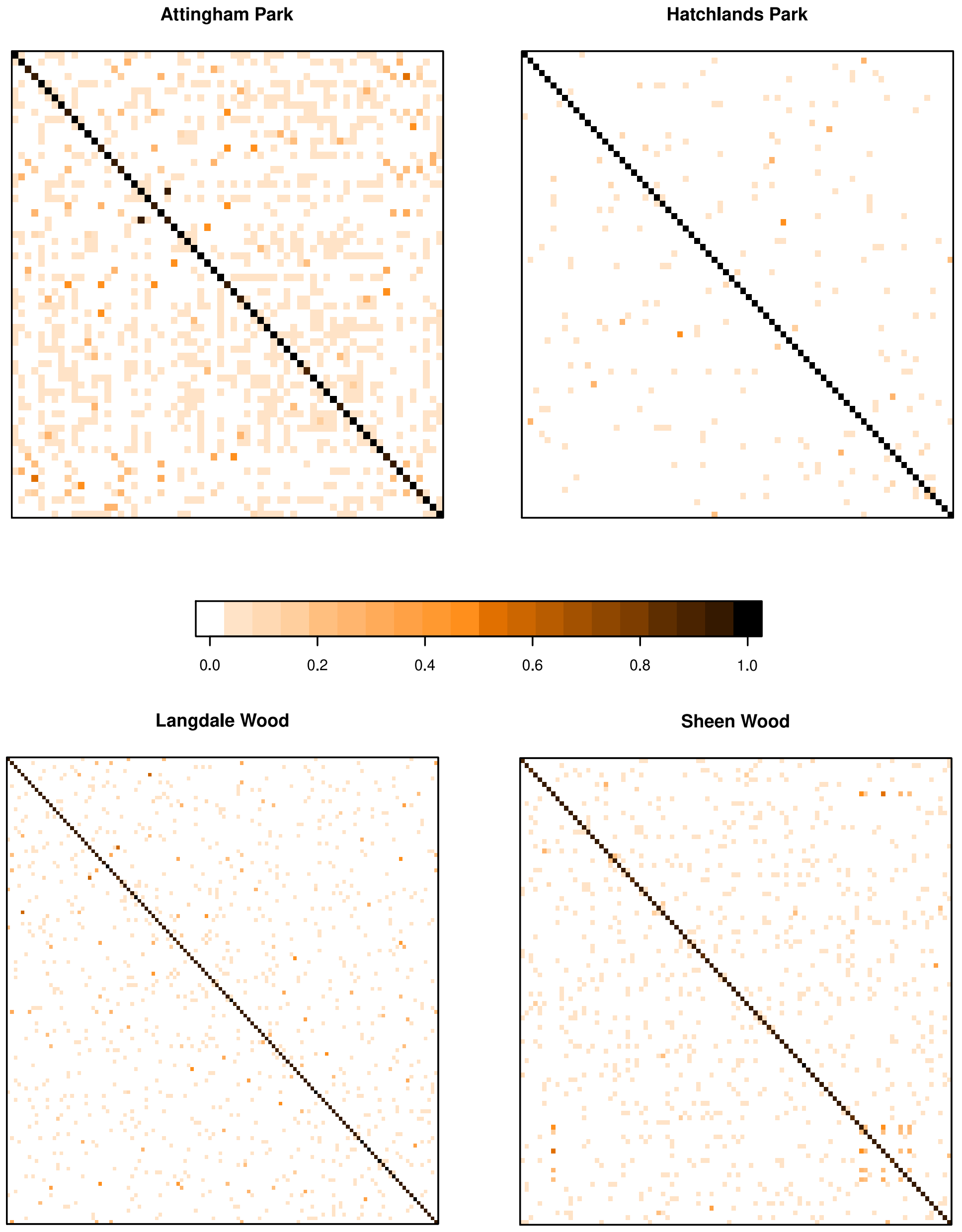


Figure S2.7. PCA of 261 unrelated (kin < 0.05) *Q. robur* individuals based on 839,911 SNPs ($r^2 < 0.4$). Colours represent sites. A) PC1 against PC2. B) PC1 against PC3. C) PC2 against PC3. D) Plot of eigenvalues for the computed principal components.

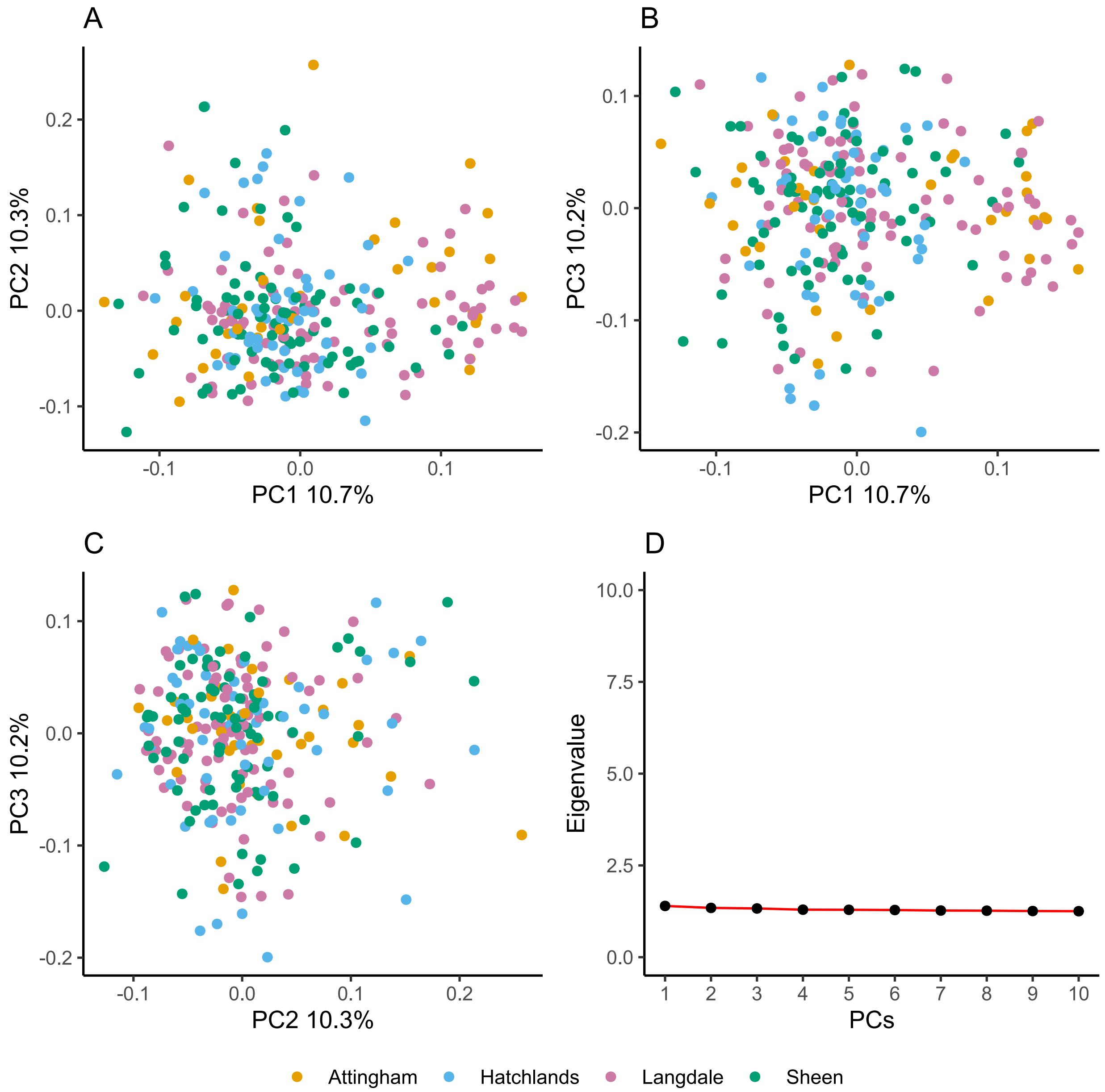


Figure S2.8. Linkage disequilibrium (LD) decay in the Q. robur genome. A) LD decay, estimated with Plink r^2 function. Points are 500 bases apart. B) Linkage disequilibrium block size distribution.

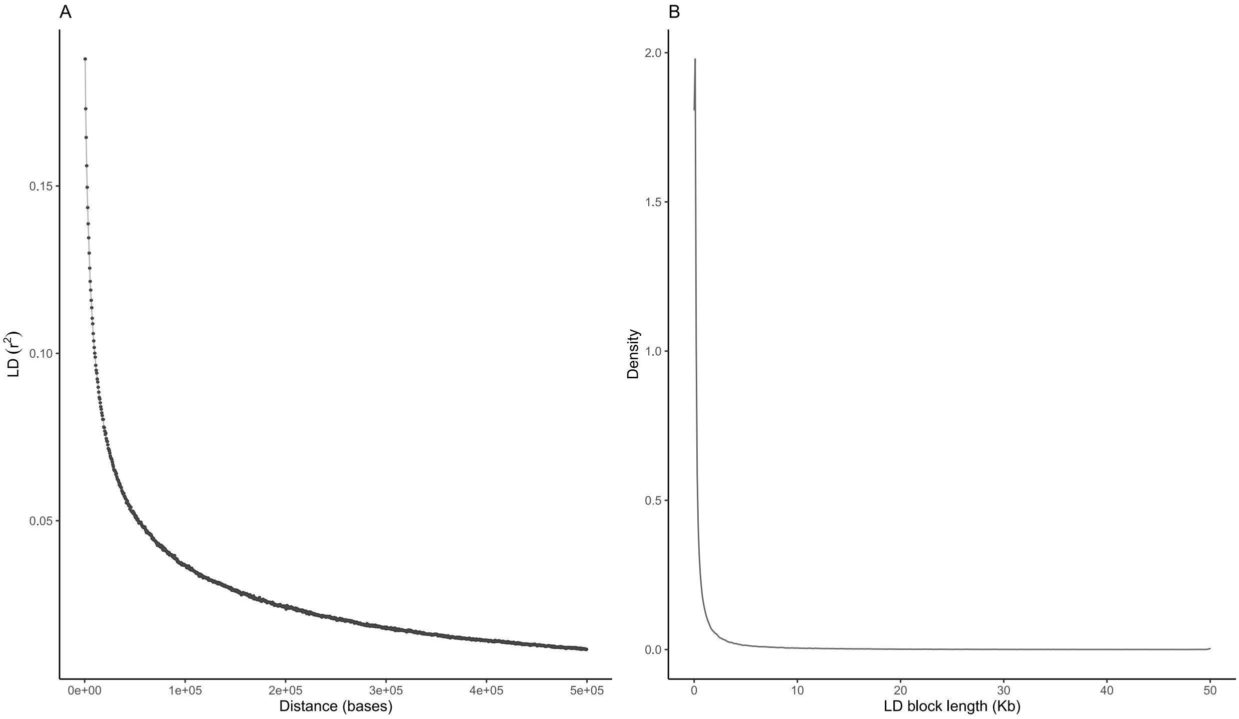


Figure S2.9. Pairwise nucleotide diversity π bar plots computed in windows of 5,000 bp across the *Q. robur* genome, based on 360 individuals.

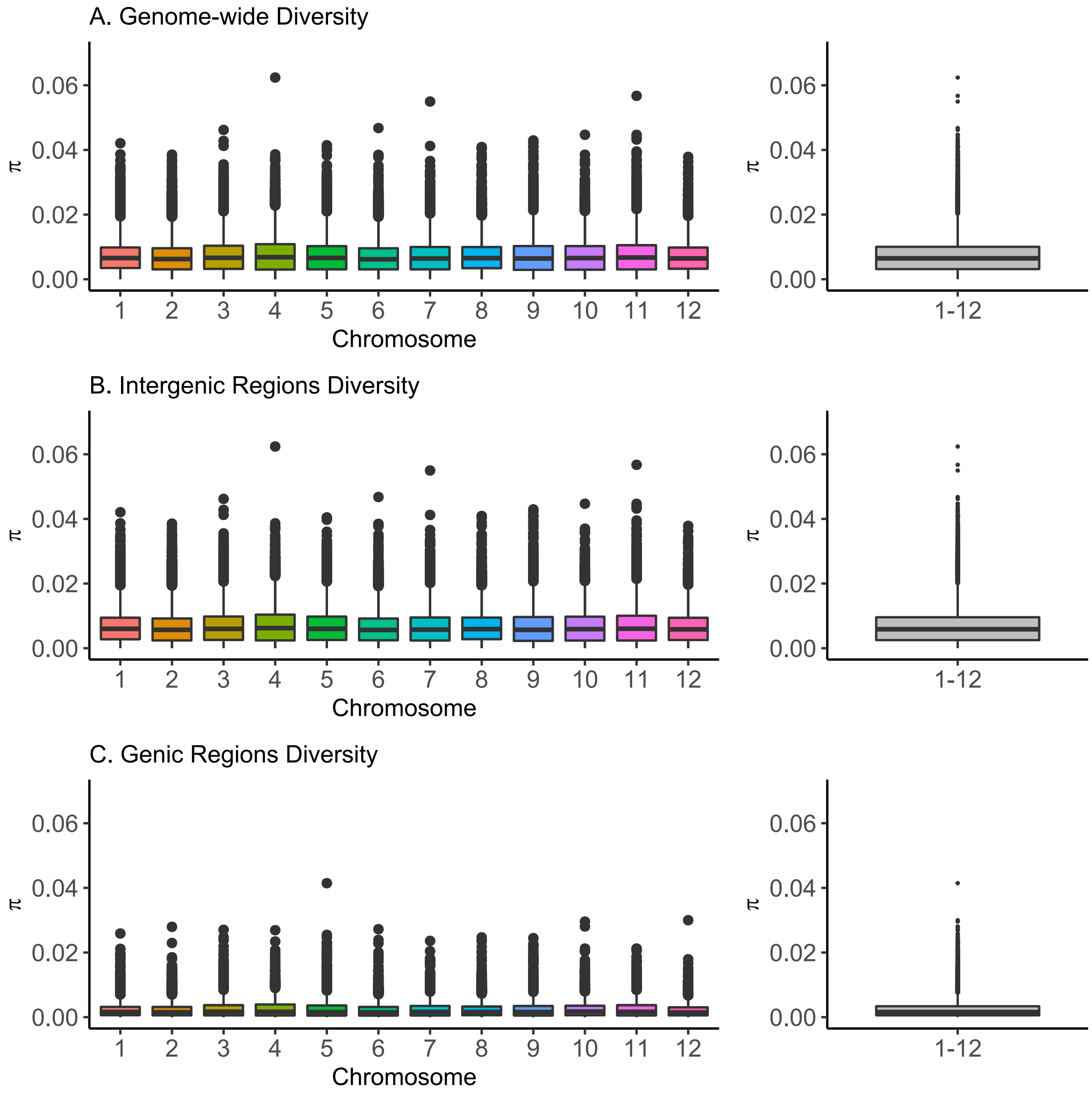


Figure S2.10. Pairwise nucleotide diversity π bar plots within sites, computed in windows of 5,000 bp across the *Q. robur* genome, based on 360 individuals.

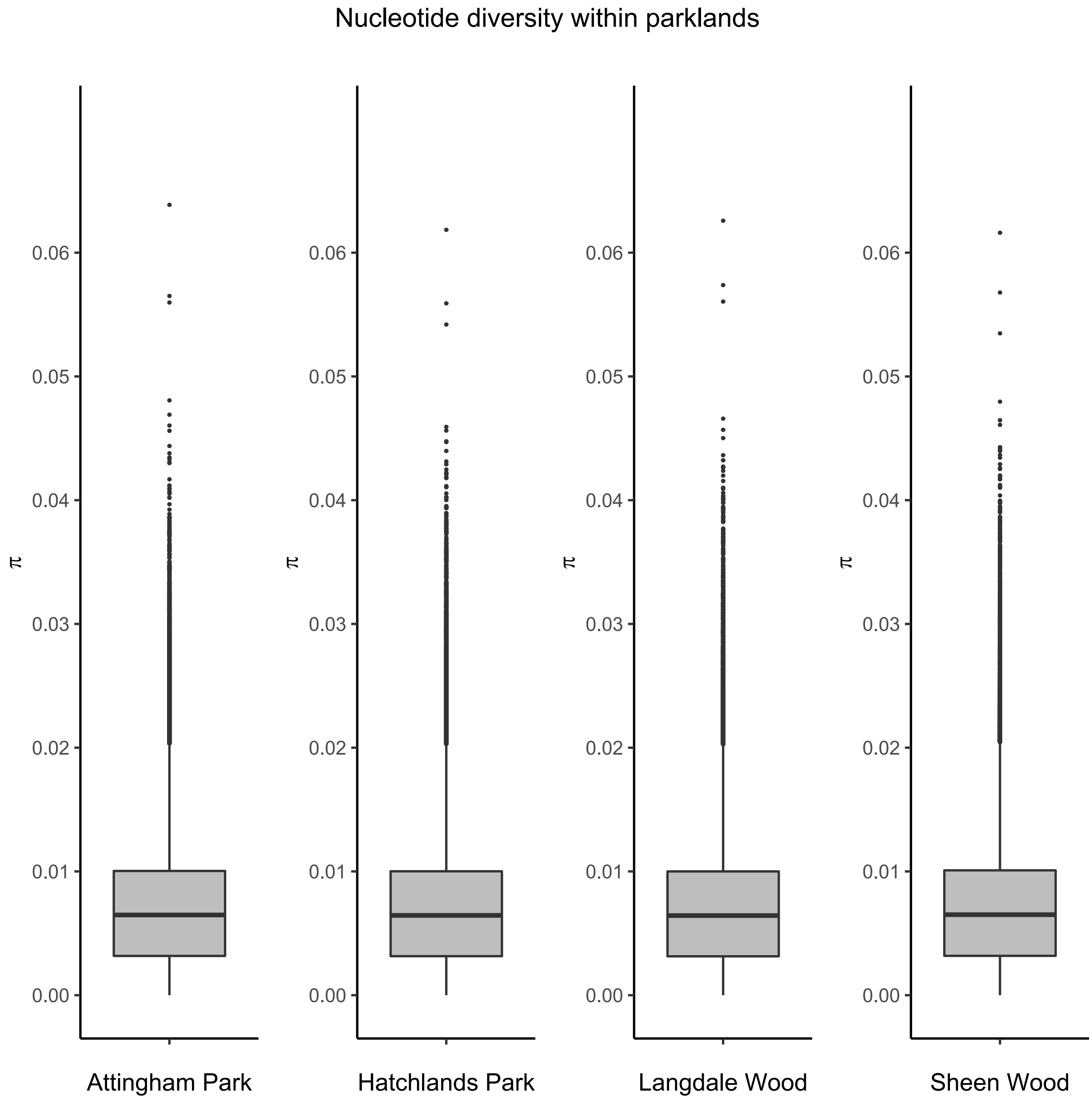
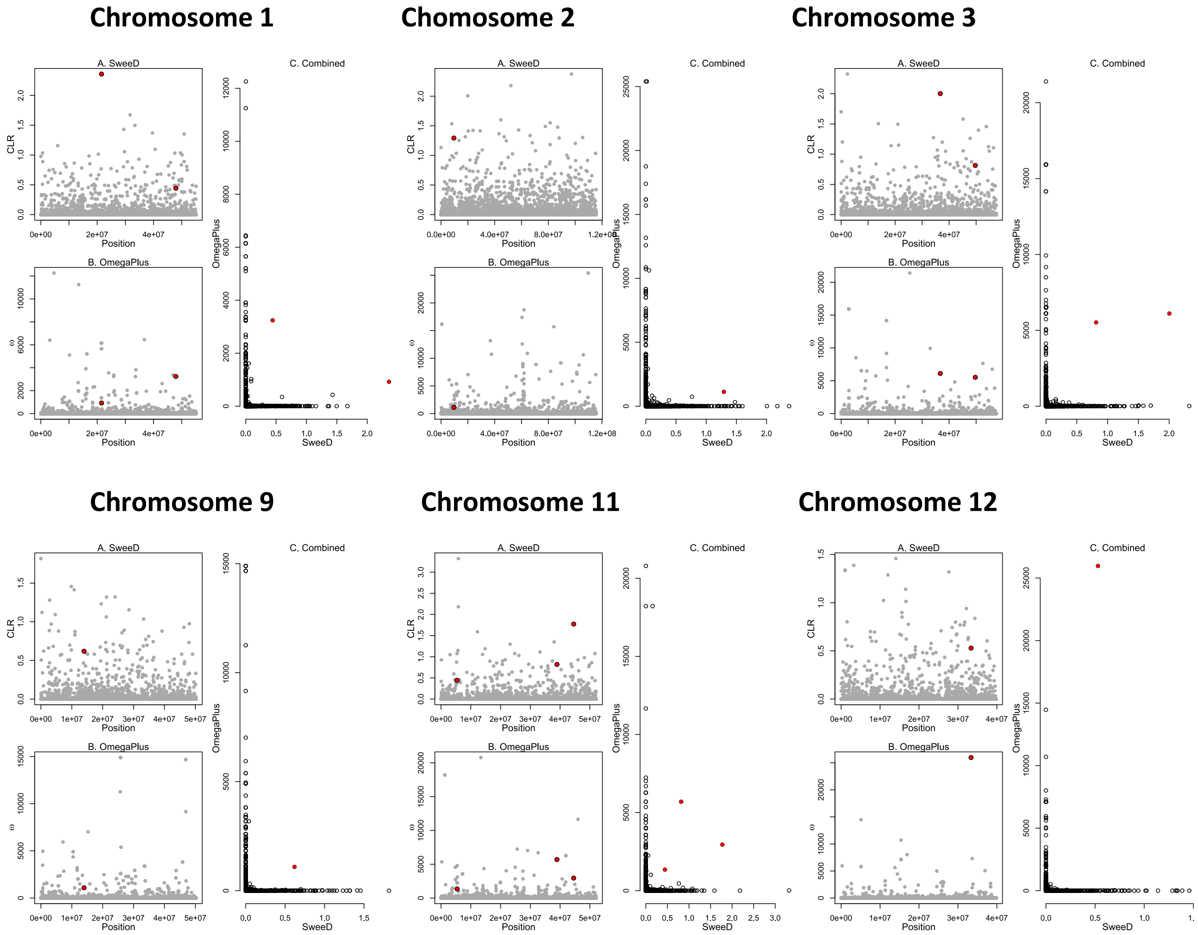


Figure S2.11. Selective sweep scan output for chromosomes with common outliers between SweeD and OmegaPlus. (A-B) The x axis denotes the base pair position on the chromosome, and the y axis shows the CLR and ω statistic computed with SweeD and OmegaPlus, respectively. (C) Combined plot for SweeD and OmegaPlus. Red dots represent outliers in common between SweeD and OmegaPlus (p < .01).



Chromosome 1

Chomosome 2

Chromosome 3

Chromosome 9

Chromosome 11

Chromosome 12

Chapter 2 Supporting Spreadsheet. Oak samples metadata.

Sample	Site	Species (based on morphology)	Age	fastStructure
EC10	Attingham	Q. robur	M	Hybrid
T22	Attingham	Q. robur	M	Hybrid
EC1	Attingham	Q. robur	M	Hybrid
EC5	Attingham	Q. robur	M	Hybrid
ED1	Attingham	Q. robur	M	Petraea
EB1	Attingham	Q. robur	M	Robur
B71	Attingham	Q. robur	M	Robur
EH6	Attingham	Q. robur	M	Robur
Y80	Attingham	Q. robur	M	Robur
EE1	Attingham	Q. robur	M	Robur
ED3	Attingham	Q. robur	M	Robur
ED5	Attingham	Q. robur	M	Robur
EG4	Attingham	Q. robur	M	Robur
FD2	Attingham	Q. robur	M	Robur
EH10	Attingham	Q. robur	M	Robur
Y28	Attingham	Q. robur	M	Robur
FA2	Attingham	Q. robur	M	Robur
Y7	Attingham	Q. robur	M	Robur
FH1	Attingham	Q. robur	M	Robur
Y6	Attingham	Q. robur	M	Robur
B23	Hatchlands	Q. robur	M	Hybrid
C9	Hatchlands	Q. robur	M	Hybrid
z_17	Hatchlands	Q. robur	M	Robur
B95	Hatchlands	Q. robur	M	Robur
T4	Hatchlands	Q. robur	M	Robur
aaa16	Hatchlands	Q. robur	M	Robur
z_34	Hatchlands	Q. robur	M	Robur
aaa98	Hatchlands	Q. robur	M	Robur
B8	Hatchlands	Q. robur	M	Robur
B27	Hatchlands	Q. robur	M	Robur
z_22	Hatchlands	Q. robur	M	Robur
C8	Hatchlands	Q. robur	M	Robur
C81	Hatchlands	Q. robur	M	Robur
aaa18	Hatchlands	Q. robur	M	Robur
BG4	Hatchlands	Q. robur	M	Robur
z_31	Hatchlands	Q. robur	M	Robur
C76	Hatchlands	Q. robur	M	Robur
C79	Hatchlands	Q. robur	M	Robur
z_7	Hatchlands	Q. robur	M	Robur
z_49	Hatchlands	Q. robur	M	Robur
BG1	Hatchlands	Q. robur	M	Robur
z_33	Hatchlands	Q. robur	M	Robur
T29	Hatchlands	Q. robur	M	Robur
z_51	Hatchlands	Q. robur	M	Robur
C92	Hatchlands	Q. robur	M	Robur
C74	Hatchlands	Q. robur	M	Robur
Y36	Hatchlands	Q. robur	M	Robur
z_37	Hatchlands	Q. robur	M	Robur
z_97	Hatchlands	Q. robur	M	Robur
C45	Hatchlands	Q. robur	M	Robur
C12	Hatchlands	Q. robur	M	Robur
BF3	Hatchlands	Q. robur	M	Robur
C15	Hatchlands	Q. robur	M	Robur
z_35	Hatchlands	Q. robur	M	Robur
Y31	Hatchlands	Q. robur	M	Robur
aaa25	Hatchlands	Q. robur	M	Robur
BF1	Hatchlands	Q. robur	M	Robur
BA1	Hatchlands	Q. robur	M	Robur
AA11	Langdale	Q. robur	M	Hybrid
AE7	Langdale	Q. robur	M	Robur
X3	Langdale	Q. robur	M	Robur
AA8	Langdale	Q. robur	M	Robur
AF7	Langdale	Q. robur	M	Robur
AF10	Langdale	Q. robur	M	Robur

AH7	Langdale	Q. robur	M	Robur
FB8	Langdale	Q. robur	M	Robur
AH11	Langdale	Q. robur	M	Robur
Y55	Langdale	Q. robur	M	Robur
z_81	Langdale	Q. robur	M	Robur
AG4	Langdale	Q. robur	M	Robur
FF11	Langdale	Q. robur	M	Robur
AB5	Langdale	Q. robur	M	Robur
FG9	Langdale	Q. robur	M	Robur
aaa99	Langdale	Q. robur	M	Robur
AG5	Langdale	Q. robur	M	Robur
aaa88	Langdale	Q. robur	M	Robur
FE11	Langdale	Q. robur	M	Robur
z_93	Langdale	Q. robur	M	Robur
z_121	Langdale	Q. robur	M	Robur
AB12	Langdale	Q. robur	M	Robur
AA5	Langdale	Q. robur	M	Robur
AF6	Langdale	Q. robur	M	Robur
FE7	Langdale	Q. robur	M	Robur
FG6	Langdale	Q. robur	M	Robur
AH1	Langdale	Q. robur	M	Robur
GB1	Langdale	Q. robur	M	Robur
AD2	Langdale	Q. robur	M	Robur
AF4	Langdale	Q. robur	M	Robur
AH2	Langdale	Q. robur	M	Robur
z_79	Langdale	Q. robur	M	Robur
FB7	Langdale	Q. robur	M	Robur
aaa111	Langdale	Q. robur	M	Robur
AH9	Langdale	Q. robur	M	Robur
GF1	Langdale	Q. robur	M	Robur
AB8	Langdale	Q. robur	M	Robur
z_115	Langdale	Q. robur	M	Robur
aaa112	Langdale	Q. robur	M	Robur
AG7	Langdale	Q. robur	M	Robur
AD9	Langdale	Q. robur	M	Robur
AG9	Langdale	Q. robur	M	Robur
AC10	Langdale	Q. robur	M	Robur
AC4	Langdale	Q. robur	M	Robur
AA6	Langdale	Q. robur	M	Robur
AB10	Langdale	Q. robur	M	Robur
AE10	Langdale	Q. robur	M	Robur
AE8	Langdale	Q. robur	M	Robur
aaa67	Langdale	Q. robur	M	Robur
z_102	Langdale	Q. robur	M	Robur
z_62	Langdale	Q. robur	M	Robur
z_65	Langdale	Q. robur	M	Robur
AD12	Langdale	Q. robur	M	Robur
z_78	Langdale	Q. robur	M	Robur
z_77	Langdale	Q. robur	M	Robur
z_86	Langdale	Q. robur	M	Robur
AD6	Langdale	Q. robur	M	Robur
z_117	Langdale	Q. robur	M	Robur
AC7	Langdale	Q. robur	M	Robur
AG6	Langdale	Q. robur	M	Robur
z_89	Langdale	Q. robur	M	Robur
AD4	Langdale	Q. robur	M	Robur
FG7	Langdale	Q. robur	M	Robur
FC8	Langdale	Q. robur	M	Robur
aaa94	Langdale	Q. robur	M	Robur
z_114	Langdale	Q. robur	M	Robur
AG11	Langdale	Q. robur	M	Robur
AH6	Langdale	Q. robur	M	Robur
FA11	Langdale	Q. robur	M	Robur
AA3	Langdale	Q. robur	M	Robur
AH4	Langdale	Q. robur	M	Robur
AC11	Langdale	Q. robur	M	Robur
AB2	Langdale	Q. robur	M	Robur
AC2	Langdale	Q. robur	M	Robur
z_82	Langdale	Q. robur	M	Robur
aaa96	Langdale	Q. robur	M	Robur
z_110	Langdale	Q. robur	M	Robur

AA9	Langdale	Q. robur	M	Robur
z_104	Langdale	Q. robur	M	Robur
z_66	Langdale	Q. robur	M	Robur
z_118	Langdale	Q. robur	M	Robur
z_108	Langdale	Q. robur	M	Robur
FF8	Langdale	Q. robur	M	Robur
z_73	Langdale	Q. robur	M	Robur
GD2	Langdale	Q. robur	M	Robur
AA10	Langdale	Q. robur	M	Robur
z_109	Langdale	Q. robur	M	Robur
z_105	Langdale	Q. robur	M	Robur
AG1	Langdale	Q. robur	M	Robur
T15	Langdale	Q. robur	M	Robur
AA2	Langdale	Q. robur	M	Robur
FC12	Langdale	Q. robur	M	Robur
aaa113	Langdale	Q. robur	M	Robur
z_75	Langdale	Q. robur	M	Robur
FE8	Langdale	Q. robur	M	Robur
z_92	Langdale	Q. robur	M	Robur
AC6	Langdale	Q. robur	M	Robur
AF1	Langdale	Q. robur	M	Robur
z_107	Langdale	Q. robur	M	Robur
AD10	Langdale	Q. robur	M	Robur
AD1	Langdale	Q. robur	M	Robur
FE9	Langdale	Q. robur	M	Robur
AC12	Langdale	Q. robur	M	Robur
z_123	Langdale	Q. robur	M	Robur
AF11	Langdale	Q. robur	M	Robur
AE1	Langdale	Q. robur	M	Robur
GF2	Langdale	Q. robur	M	Robur
FF6	Langdale	Q. robur	M	Robur
AD8	Langdale	Q. robur	M	Robur
AB1	Langdale	Q. robur	M	Robur
z_106	Langdale	Q. robur	M	Robur
FH7	Langdale	Q. robur	M	Robur
FD7	Langdale	Q. robur	M	Robur
z_74	Langdale	Q. robur	M	Robur
AF5	Langdale	Q. robur	M	Robur
AA1	Langdale	Q. robur	M	Robur
BD5	Sheen	Q. robur	M	Robur
B49	Sheen	Q. robur	M	Robur
X76	Sheen	Q. robur	M	Robur
B66	Sheen	Q. robur	M	Robur
C94	Sheen	Q. robur	M	Robur
T14	Sheen	Q. robur	M	Robur
B77	Sheen	Q. robur	M	Robur
B53	Sheen	Q. robur	M	Robur
B41	Sheen	Q. robur	M	Robur
B96	Sheen	Q. robur	M	Robur
B48	Sheen	Q. robur	M	Robur
B72	Sheen	Q. robur	M	Robur
B99	Sheen	Q. robur	M	Robur
Y14	Sheen	Q. robur	M	Robur
X51	Sheen	Q. robur	M	Robur
B43	Sheen	Q. robur	M	Robur
C42	Sheen	Q. robur	M	Robur
C35	Sheen	Q. robur	M	Robur
B57	Sheen	Q. robur	M	Robur
B84	Sheen	Q. robur	M	Robur
B52	Sheen	Q. robur	M	Robur
EA11	Attingham	Q. robur	OM	Hybrid
B81	Attingham	Q. petraea	OM	Petraea
EF4	Attingham	Q. robur	OM	Robur
EF3	Attingham	Q. robur	OM	Robur
C21	Attingham	Q. robur	OM	Robur
C26	Attingham	Q. robur	OM	Robur
EG1	Attingham	Q. robur	OM	Robur
Y27	Attingham	Q. robur	OM	Robur
X53	Attingham	Q. robur	OM	Robur
EG6	Attingham	Q. robur	OM	Robur
EG5	Attingham	Q. robur	OM	Robur

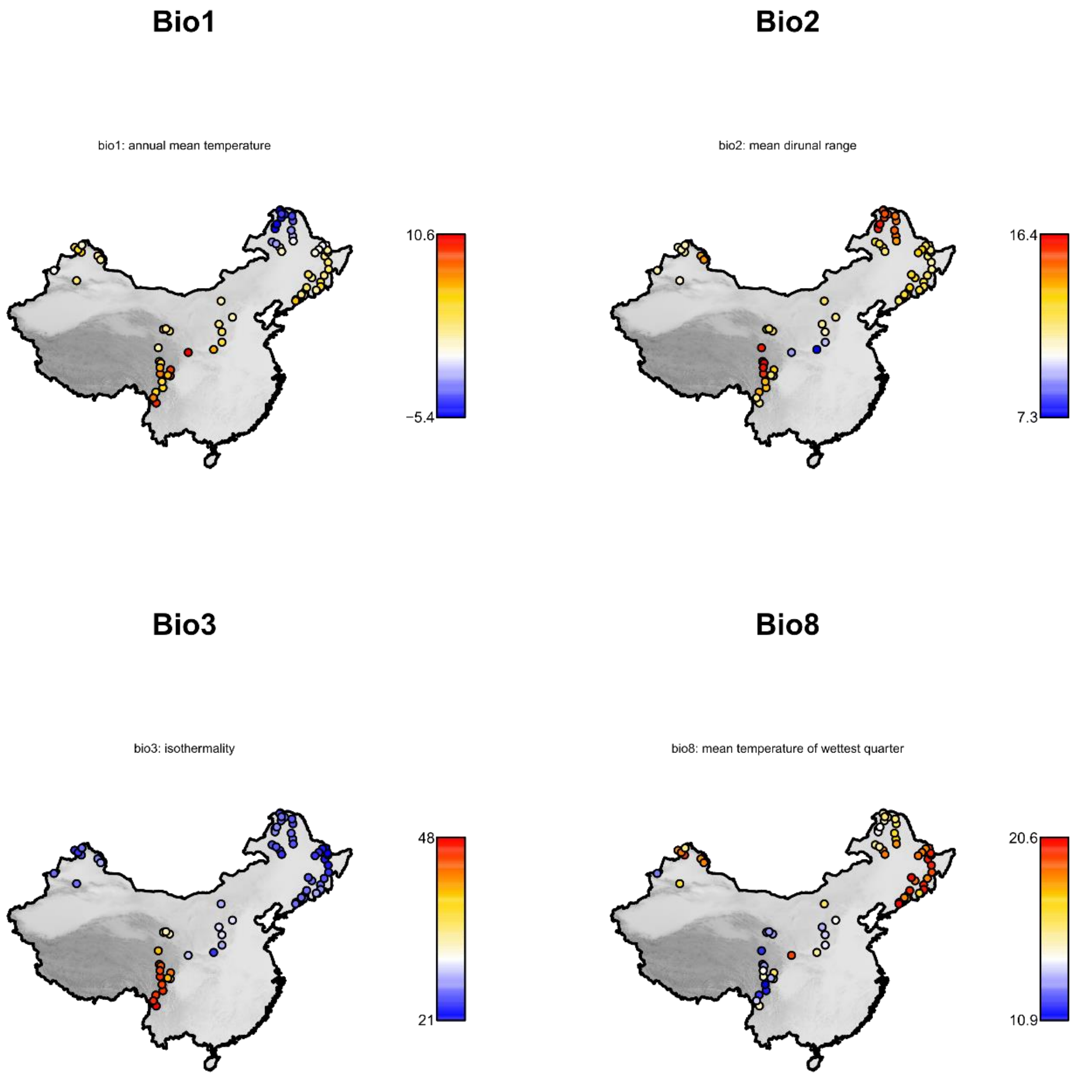
X45	Attingham	Q. robur	OM	Robur
X49	Attingham	Q. robur	OM	Robur
EE10	Attingham	Q. robur	OM	Robur
C27	Attingham	Q. robur	OM	Robur
B75	Attingham	Q. robur	OM	Robur
EC9	Attingham	Q. robur	OM	Robur
B64	Attingham	Q. robur	OM	Robur
C69	Hatchlands	Q. petraea	OM	Hybrid
B1	Hatchlands	Q. petraea	OM	Hybrid
T1	Hatchlands	Q. robur	OM	Hybrid
BA2	Hatchlands	Q. robur	OM	Robur
B36	Hatchlands	Q. robur	OM	Robur
C7	Hatchlands	Q. robur	OM	Robur
z_9	Hatchlands	Q. robur	OM	Robur
BA4	Hatchlands	Q. robur	OM	Robur
B33	Hatchlands	Q. robur	OM	Robur
B34	Hatchlands	Q. robur	OM	Robur
BB2	Hatchlands	Q. robur	OM	Robur
T21	Hatchlands	Q. robur	OM	Robur
z_39	Hatchlands	Q. robur	OM	Robur
C25	Hatchlands	Q. robur	OM	Robur
C19	Hatchlands	Q. robur	OM	Robur
C33	Hatchlands	Q. robur	OM	Robur
C22	Hatchlands	Q. robur	OM	Robur
AF12	Hatchlands	Q. robur	OM	Robur
BD4	Hatchlands	Q. robur	OM	Robur
C13	Hatchlands	Q. robur	OM	Robur
X59	Hatchlands	Q. robur	OM	Robur
AH12	Hatchlands	Q. robur	OM	Robur
Y52	Hatchlands	Q. robur	OM	Robur
BD2	Hatchlands	Q. robur	OM	Robur
C14	Hatchlands	Q. robur	OM	Robur
C5	Hatchlands	Q. robur	OM	Robur
C3	Hatchlands	Q. robur	OM	Robur
z_26	Hatchlands	Q. robur	OM	Robur
z_27	Hatchlands	Q. robur	OM	Robur
C11	Hatchlands	Q. robur	OM	Robur
aaa8	Hatchlands	Q. robur	OM	Robur
BE4	Hatchlands	Q. robur	OM	Robur
B97	Hatchlands	Q. robur	OM	Robur
BH4	Hatchlands	Q. robur	OM	Robur
T2	Hatchlands	Q. robur	OM	Robur
BE2	Hatchlands	Q. robur	OM	Robur
C4	Hatchlands	Q. robur	OM	Robur
BC2	Hatchlands	Q. robur	OM	Robur
BC1	Hatchlands	Q. robur	OM	Robur
AD3	Langdale	Q. robur	OM	Robur
BD9	Sheen	Q. robur	OM	Hybrid
BD12	Sheen	Q. robur	OM	Hybrid
BA7	Sheen	Q. robur	OM	Robur
T9	Sheen	Q. robur	OM	Robur
BE9	Sheen	Q. robur	OM	Robur
BC10	Sheen	Q. robur	OM	Robur
BG9	Sheen	Q. robur	OM	Robur
BF6	Sheen	Q. robur	OM	Robur
C66	Sheen	Q. robur	OM	Robur
B93	Sheen	Q. robur	OM	Robur
BD11	Sheen	Q. robur	OM	Robur
BF5	Sheen	Q. robur	OM	Robur
T6	Sheen	Q. robur	OM	Robur
BC9	Sheen	Q. robur	OM	Robur
BC12	Sheen	Q. robur	OM	Robur
BA5	Sheen	Q. robur	OM	Robur
BH11	Sheen	Q. robur	OM	Robur
Y5	Sheen	Q. robur	OM	Robur
K1	Sheen	Q. robur	OM	Robur
BD6	Sheen	Q. robur	OM	Robur
C59	Sheen	Q. robur	OM	Robur
BC5	Sheen	Q. robur	OM	Robur
BE10	Sheen	Q. robur	OM	Robur
BD10	Sheen	Q. robur	OM	Robur

BF8	Sheen	Q. robur	OM	Robur
BA12	Sheen	Q. robur	OM	Robur
C91	Sheen	Q. robur	OM	Robur
C65	Sheen	Q. robur	OM	Robur
X71	Sheen	Q. robur	OM	Robur
BC11	Sheen	Q. robur	OM	Robur
BG5	Sheen	Q. robur	OM	Robur
Y29	Sheen	Q. robur	OM	Robur
BC8	Sheen	Q. robur	OM	Robur
B61	Sheen	Q. robur	OM	Robur
C83	Sheen	Q. robur	OM	Robur
T31	Sheen	Q. robur	OM	Robur
B91	Sheen	Q. robur	OM	Robur
BF10	Sheen	Q. robur	OM	Robur
BE12	Sheen	Q. robur	OM	Robur
C93	Sheen	Q. robur	OM	Robur
C98	Sheen	Q. robur	OM	Robur
BF7	Sheen	Q. robur	OM	Robur
BE11	Sheen	Q. robur	OM	Robur
C95	Sheen	Q. robur	OM	Robur
C55	Sheen	Q. robur	OM	Robur
BH10	Sheen	Q. robur	OM	Robur
BA8	Sheen	Q. robur	OM	Robur
BA11	Sheen	Q. robur	OM	Robur
X15	Sheen	Q. robur	OM	Robur
C86	Sheen	Q. robur	OM	Robur
Y75	Sheen	Q. robur	OM	Robur
T13	Sheen	Q. robur	OM	Robur
B67	Sheen	Q. robur	OM	Robur
B55	Sheen	Q. robur	OM	Robur
BF9	Sheen	Q. robur	OM	Robur
BH5	Sheen	Q. robur	OM	Robur
BB5	Sheen	Q. robur	OM	Robur
BG11	Sheen	Q. robur	OM	Robur
X17	Sheen	Q. robur	OM	Robur
B89	Sheen	Q. robur	OM	Robur
T34	Sheen	Q. robur	OM	Robur
B87	Sheen	Q. robur	OM	Robur
X8	Sheen	Q. robur	OM	Robur
C82	Sheen	Q. robur	OM	Robur
B44	Sheen	Q. robur	OM	Robur
X25	Sheen	Q. robur	OM	Robur
T28	Sheen	Q. robur	OM	Robur
T7	Sheen	Q. robur	OM	Robur
T8	Sheen	Q. robur	OM	Robur
Y48	Sheen	Q. robur	OM	Robur
X7	Sheen	Q. robur	OM	Robur
C68	Sheen	Q. robur	OM	Robur
C87	Sheen	Q. robur	OM	Robur
Y30	Sheen	Q. robur	OM	Robur
B46	Sheen	Q. robur	OM	Robur
EH3	Attingham	Q. petraea	SM	Hybrid
X56	Attingham	Q. robur	SM	Hybrid
X38	Attingham	Q. petraea	SM	Petraea
EB7	Attingham	Q. petraea	SM	Petraea
Y33	Attingham	Q. petraea	SM	Petraea
EC7	Attingham	Q. petraea	SM	Petraea
EB8	Attingham	Q. petraea	SM	Petraea
ED7	Attingham	Q. petraea	SM	Petraea
B98	Attingham	Q. robur	SM	Petraea
Y24	Attingham	Q. robur	SM	Petraea
Y8	Attingham	Q. robur	SM	Robur
B62	Attingham	Q. robur	SM	Robur
B83	Attingham	Q. robur	SM	Robur
Y13	Attingham	Q. robur	SM	Robur
EC3	Attingham	Q. robur	SM	Robur
FA1	Attingham	Q. robur	SM	Robur
C51	Attingham	Q. robur	SM	Robur
EG2	Attingham	Q. robur	SM	Robur
X18	Attingham	Q. robur	SM	Robur
B92	Attingham	Q. robur	SM	Robur

EB12	Attingham	Q. robur	SM	Robur
FC2	Attingham	Q. robur	SM	Robur
C16	Attingham	Q. robur	SM	Robur
ED12	Attingham	Q. robur	SM	Robur
Y66	Attingham	Q. robur	SM	Robur
EE2	Attingham	Q. robur	SM	Robur
Y23	Attingham	Q. robur	SM	Robur
B94	Attingham	Q. robur	SM	Robur
EE11	Attingham	Q. robur	SM	Robur
X39	Attingham	Q. robur	SM	Robur
X64	Attingham	Q. robur	SM	Robur
EC12	Attingham	Q. robur	SM	Robur
Y1	Attingham	Q. robur	SM	Robur
EG10	Attingham	Q. robur	SM	Robur
EA2	Attingham	Q. robur	SM	Robur
B68	Attingham	Q. robur	SM	Robur
Y67	Attingham	Q. robur	SM	Robur
EC2	Attingham	Q. robur	SM	Robur
B88	Attingham	Q. robur	SM	Robur
EF2	Attingham	Q. robur	SM	Robur
X12	Attingham	Q. robur	SM	Robur
C28	Attingham	Q. robur	SM	Robur
Y2	Attingham	Q. robur	SM	Robur
EA5	Attingham	Q. robur	SM	Robur
z_48	Hatchlands	Q. robur	SM	Robur
C24	Hatchlands	Q. robur	SM	Robur
z_46	Hatchlands	Q. robur	SM	Robur
z_72	Langdale	Q. robur	SM	Hybrid
z_84	Langdale	Q. robur	SM	Robur
AF3	Langdale	Q. robur	SM	Robur
z_119	Langdale	Q. robur	SM	Robur
AH5	Langdale	Q. robur	SM	Robur
aaa103	Langdale	Q. robur	SM	Robur
AD11	Langdale	Q. robur	SM	Robur
BH6	Sheen	Q. robur	SM	Robur
BB9	Sheen	Q. robur	SM	Robur
BB7	Sheen	Q. robur	SM	Robur
BD8	Sheen	Q. robur	SM	Robur

Chapter 3

Figure S3.1. The 11 uncorrelated environmental variables selected for ENM and EAA.



Bio1

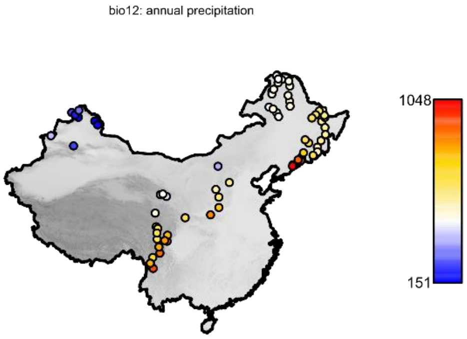
Bio2

Bio3

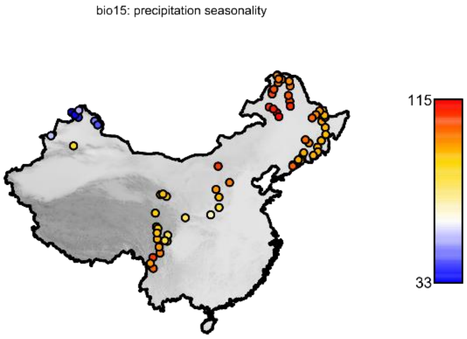
Bio8

bio1: annual mean temperature
10.6
−5.4
bio2: mean dirunal range
16.4
7.3
bio3: isothermality
48
21
bio8: mean temperature of wettest quarter
20.6
10.9

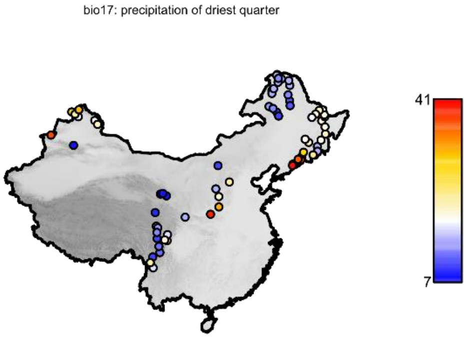
Bio12



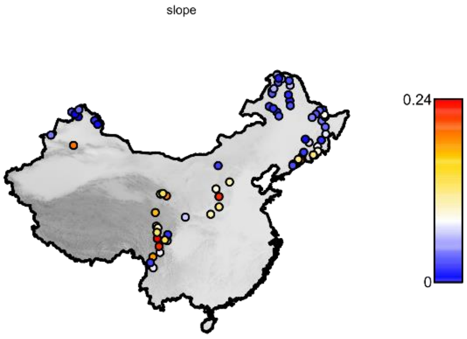
Bio15



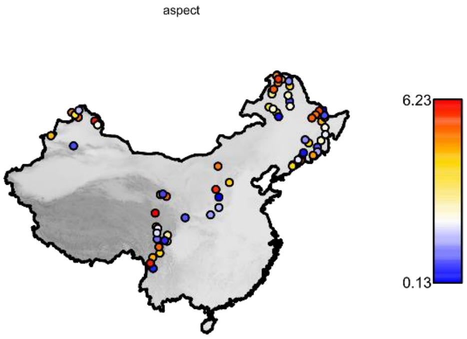
Bio17



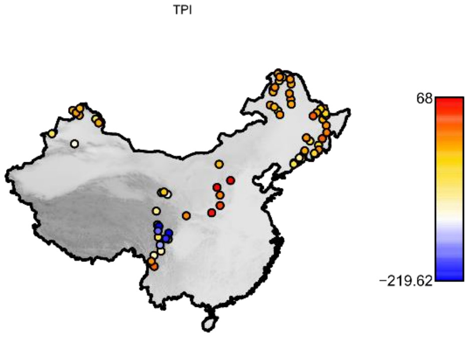
Slope



Aspect



TPI



Flowdir

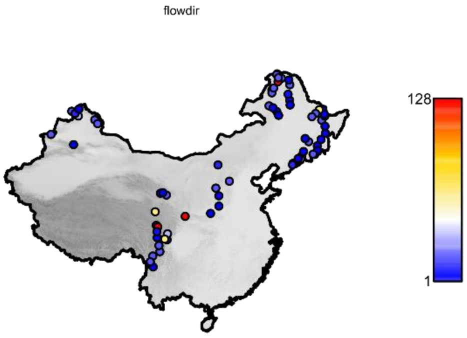


Figure S3.2. Correlation between the original 26 climatic variables.

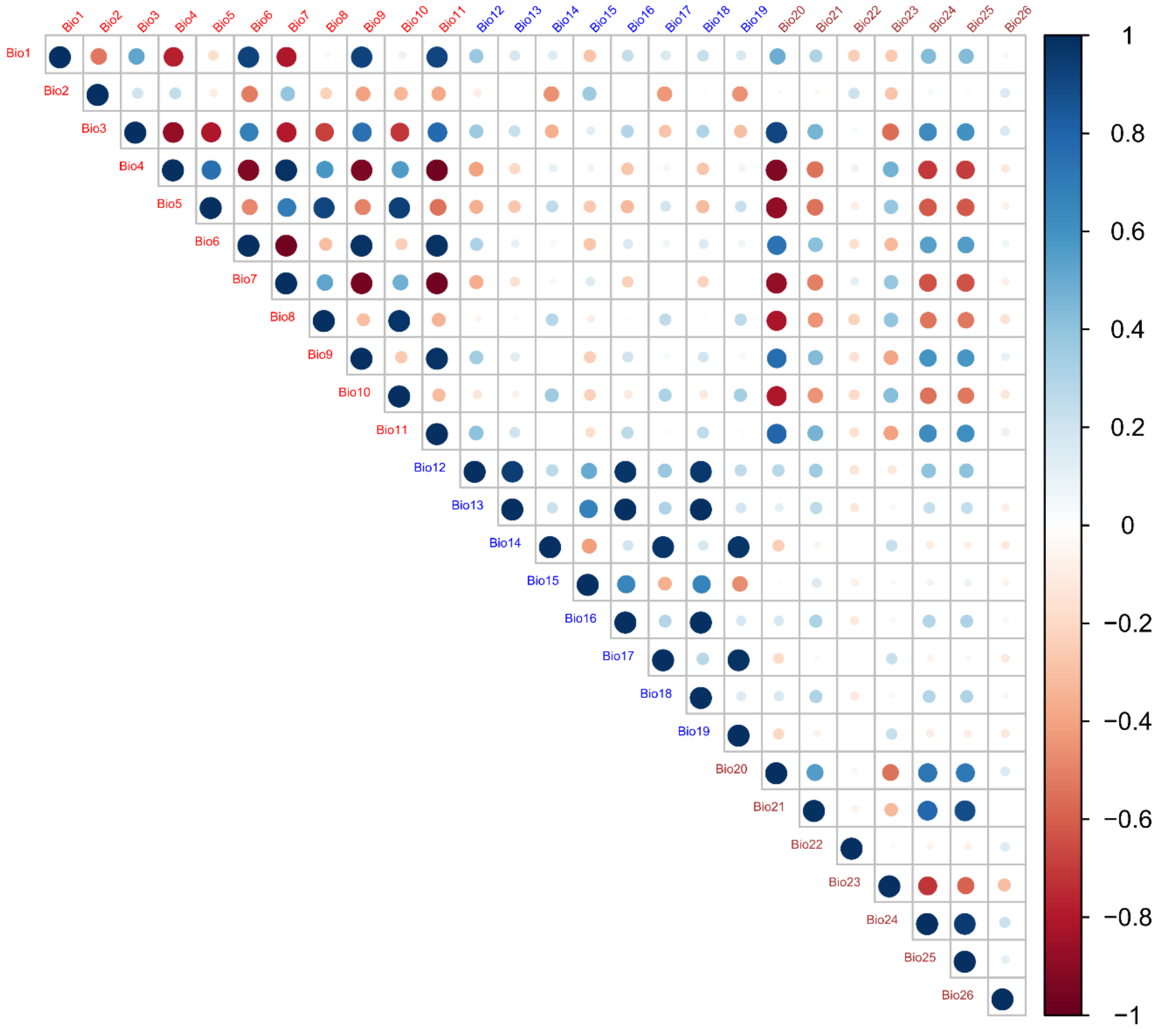


Figure S3.3. Results of the fastSTRUCTURE analysis including 83 birch individuals. A) Log-marginal likelihood lower bound (LLBO) of the data for Ks from 1 to 10. B) Cross-validation error calculated using 5-fold cross-validation for Ks from 1 to 10, with standard error bars.

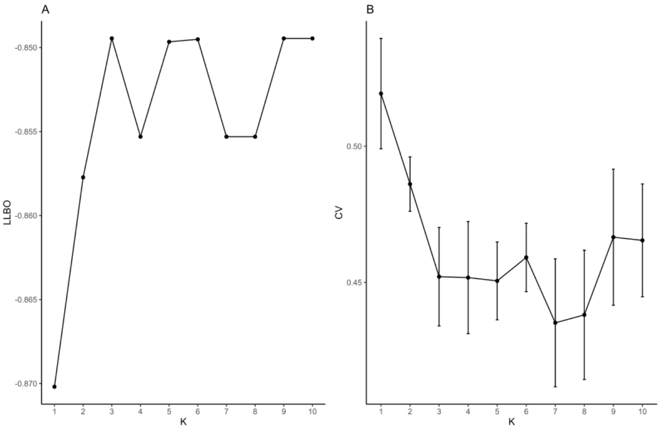


Figure S3.4. Results of the fastSTRUCTURE analysis including 162 *B. pendula* and *B. platyphylla* individuals. A) Log-marginal likelihood lower bound (LLBO) of the data for Ks from 1 to 10. B) Cross-validation error calculated using 5-fold cross-validation for Ks from 1 to 10, with standard error bars.

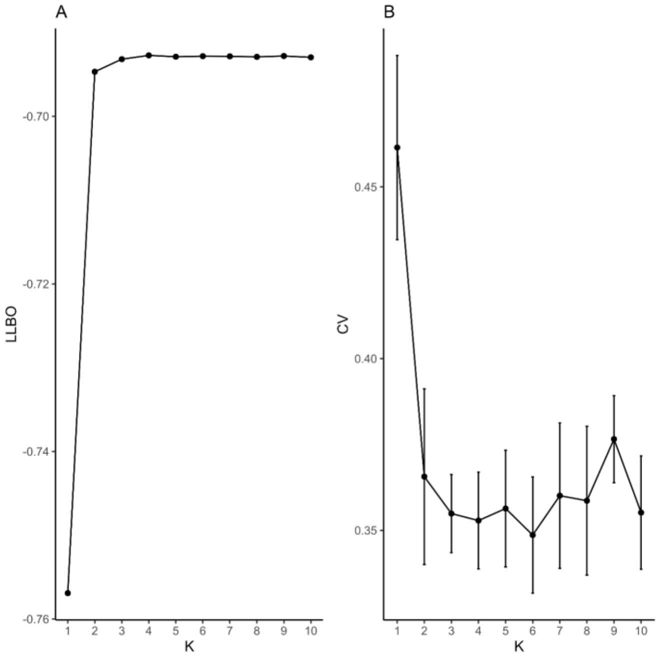


Figure S3.5. PCA of 162 *B. pendula* and *B. platyphylla* individuals based on 278,717 unlinked SNPs (r^2 < 0.4). Colours represent populations assignments. A) PC1 against PC2. B) PC1 against PC3. C) PC2 against PC3. D) Eigenvalues of the computed principal components.

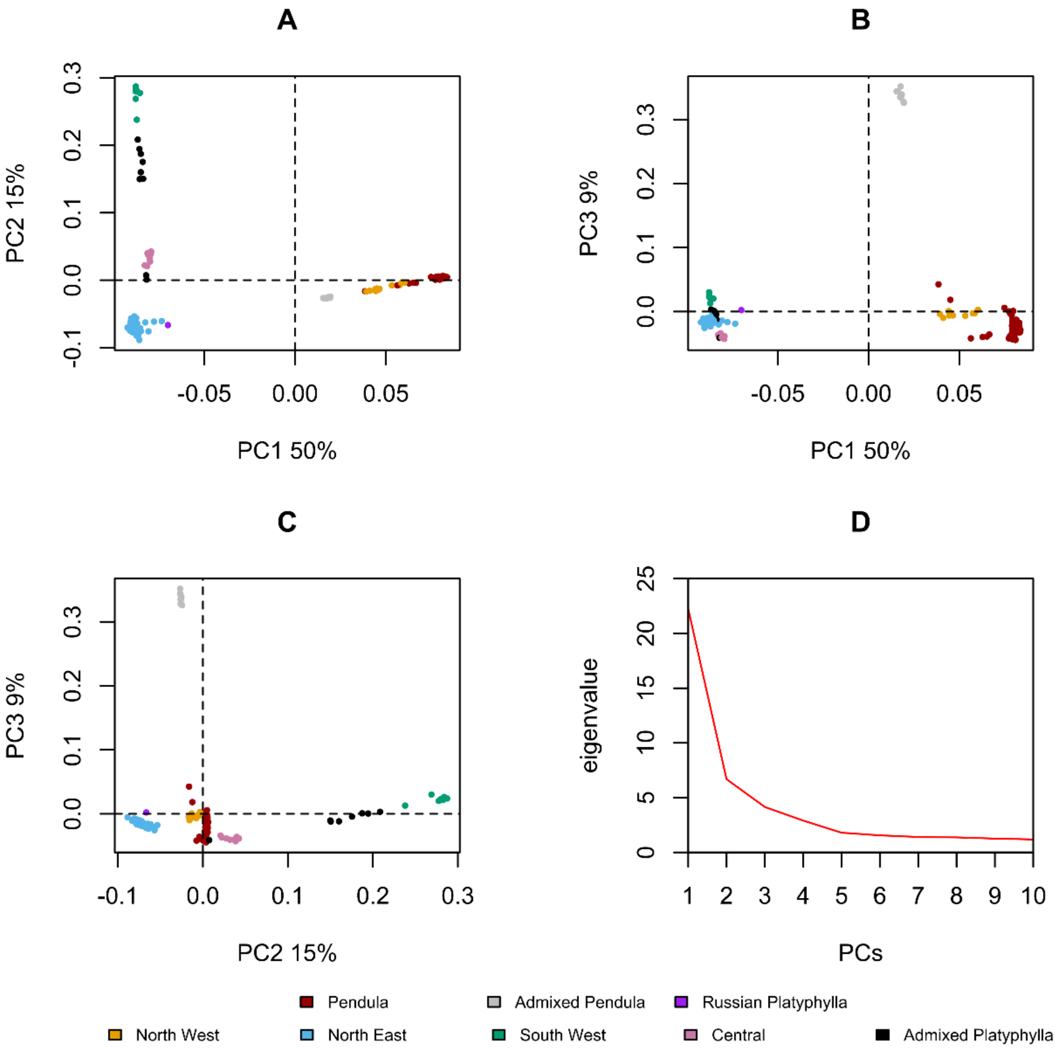


Figure S3.6. Results of the fastSTRUCTURE analysis including 71 *B. platyphylla* individuals. A) Log-marginal likelihood lower bound (LLBO) of the data for Ks from 1 to 10. B) Cross-validation error calculated using 5-fold cross-validation for Ks from 1 to 10, with standard error bars.

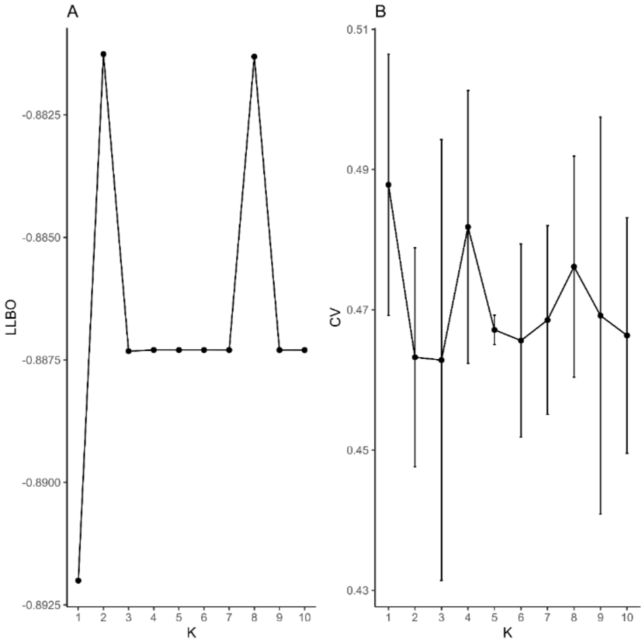


Figure S3.7. PCA of 71 *B. platyphylla* individuals based on 1,387,994 unlinked SNPs ($r^2 < 0.4$). Colours represent fastSTRUCTURE populations assignments at K = 3. A) PC1 against PC2. B) PC1 against PC3. C) PC2 against PC3. D) Eigenvalues of the computed principal components.

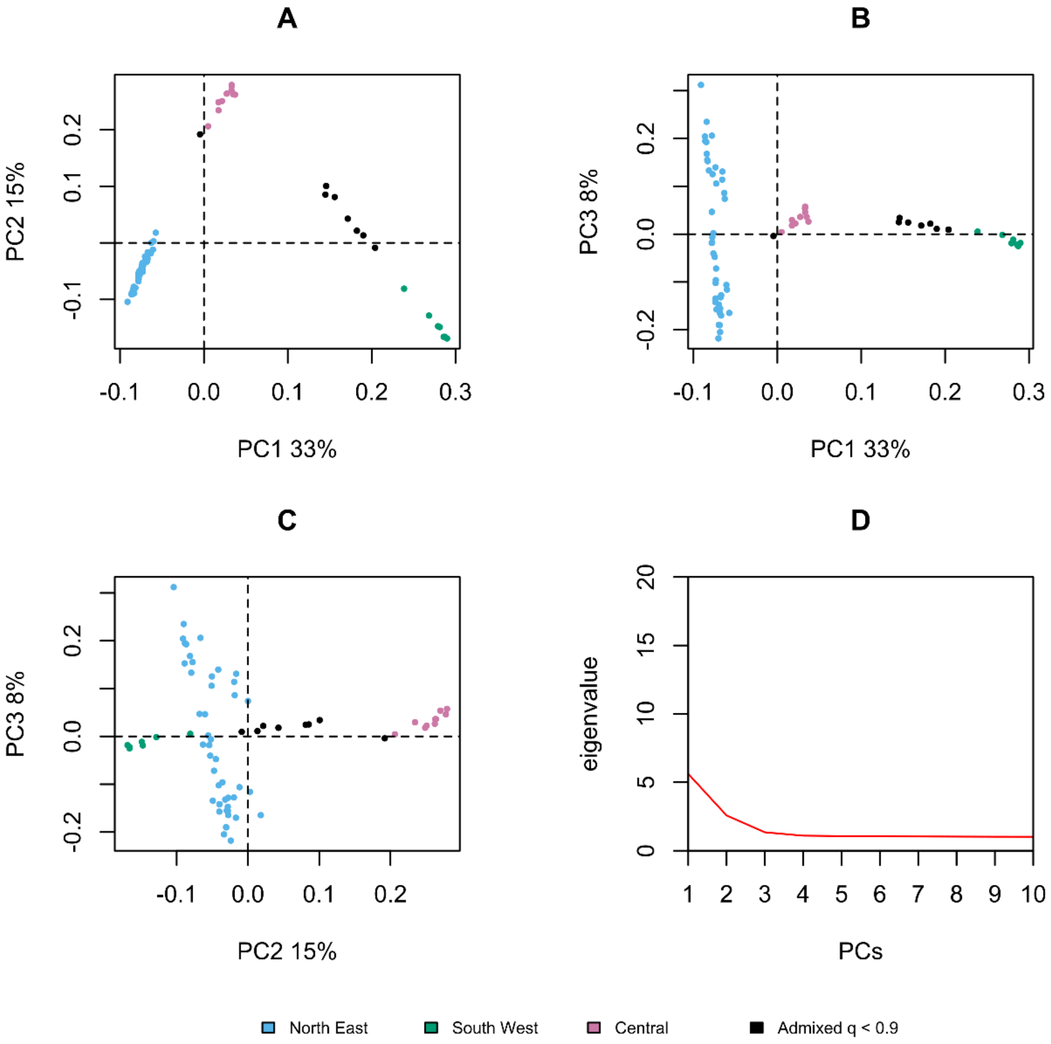
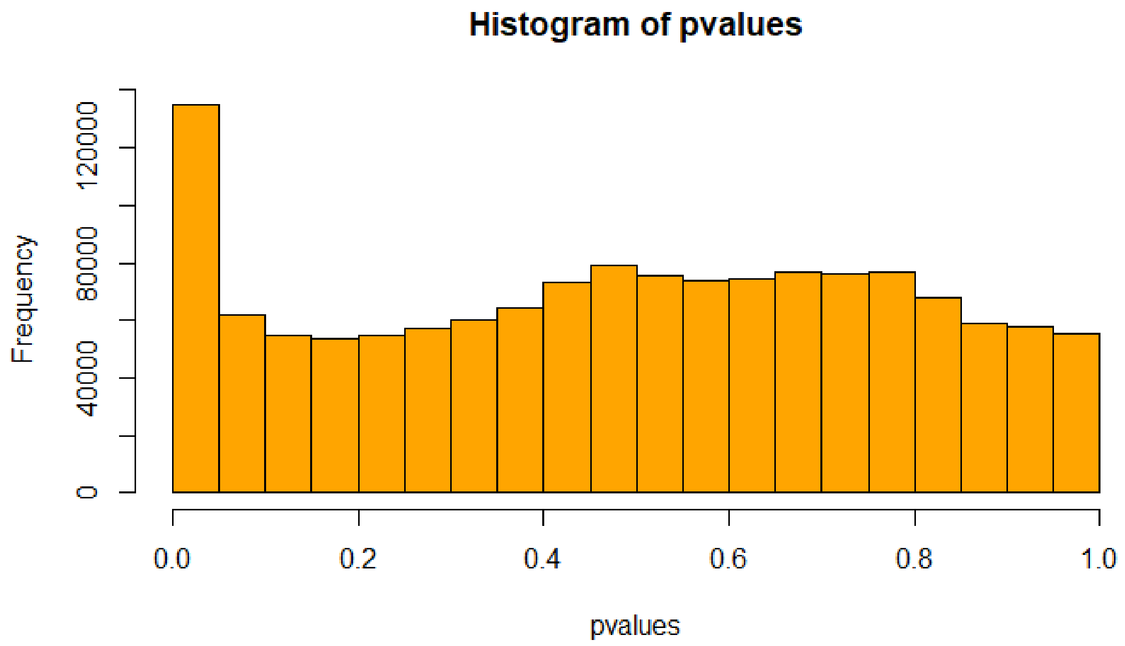


Figure S3.8. *snmf* Fst outlier test. A) Histogram of outlier test p-values. B) -log10 of p-value for each SNP.

A)



B)

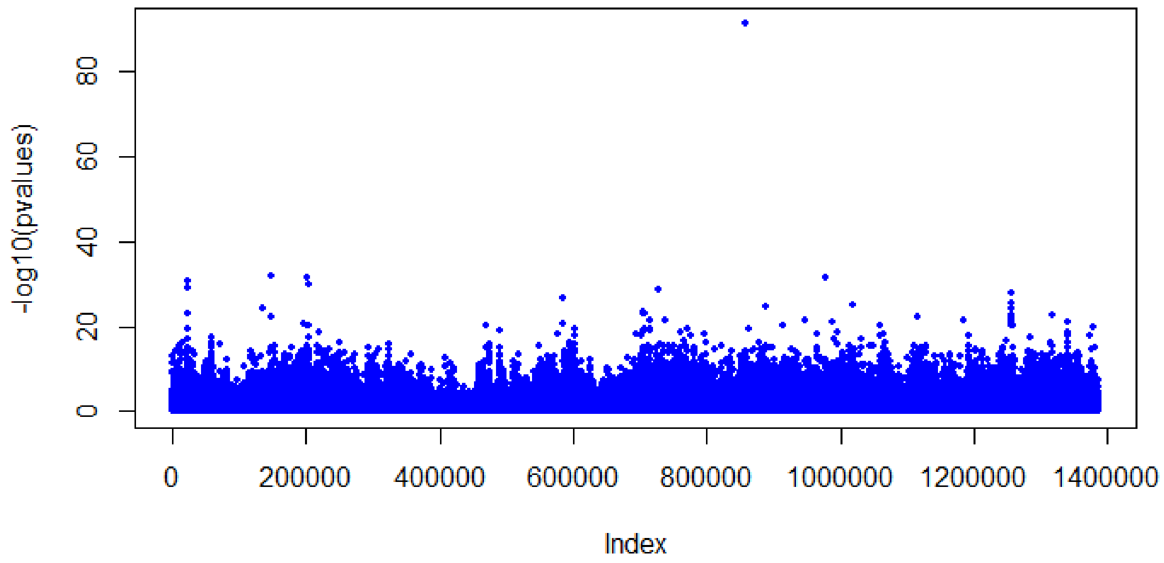


Figure S3.9. Linkage disequilibrium (LD) decay in *B. platyphylla* populations, calculated with the tool PopLDDecay (Zhang, Dong, Xu, He, & Yang, 2018).

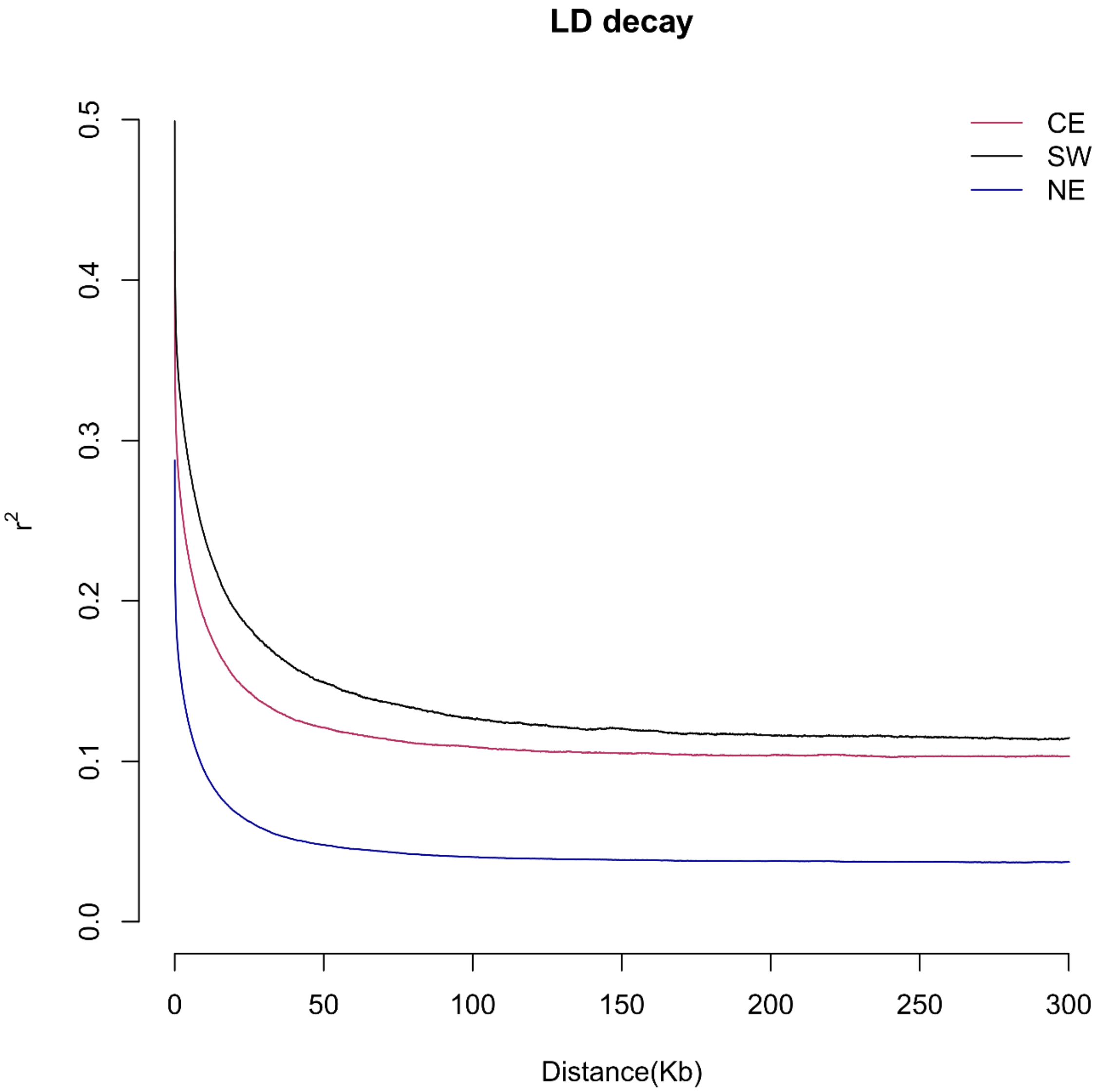


Figure S3.10. Genome-wide F_{st} (by SNP site) distribution between *B. platyphylla* populations, based on 1,387,994 SNPs and excluding admixed individuals. Populations' assignment based on *fastSTRUCTURE* at K = 3, excluding admixed individuals (q < 0.9).

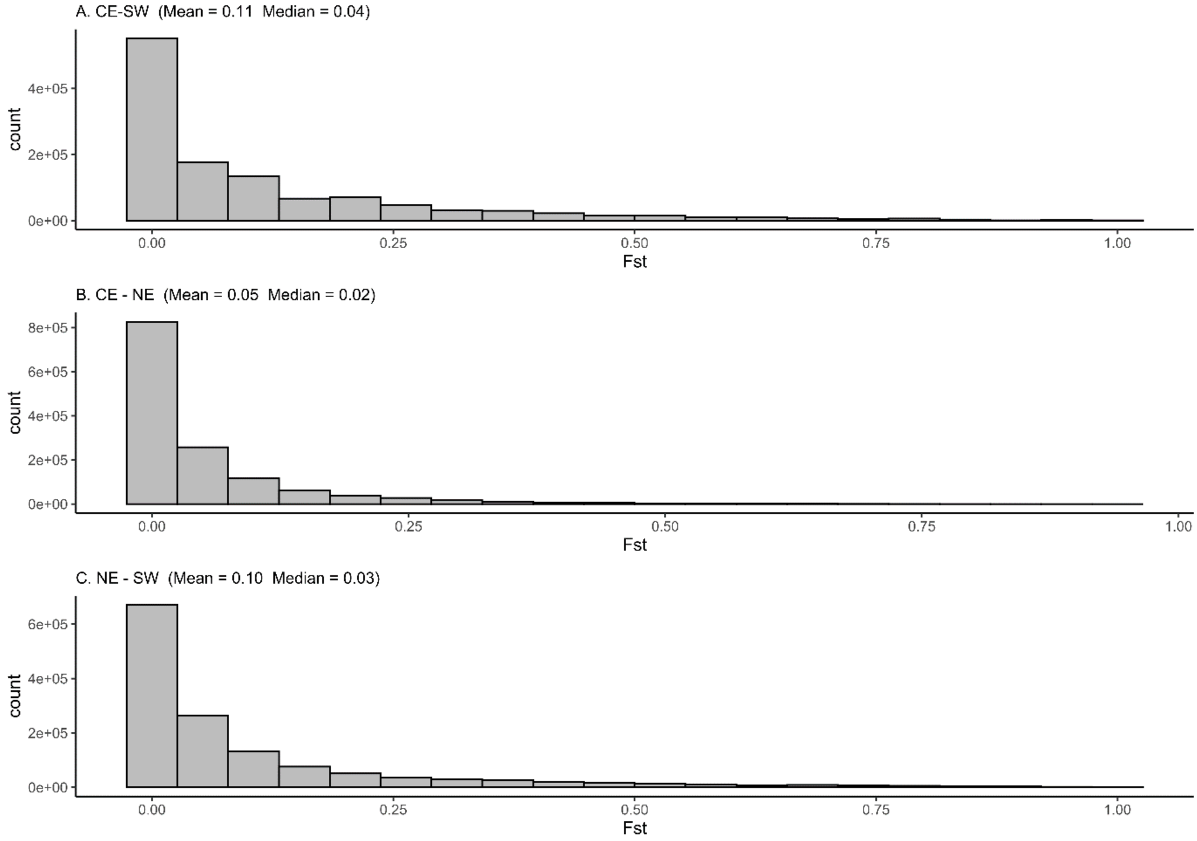


Figure S3.11. Pairwise nucleotide diversity π bar plots per population, computed in windows of 5,000 bp across the 14 *B. platyphylla* chromosomes. Populations according to *fastSTRUCTURE* at K = 3, excluding admixed individuals (q < 0.9).

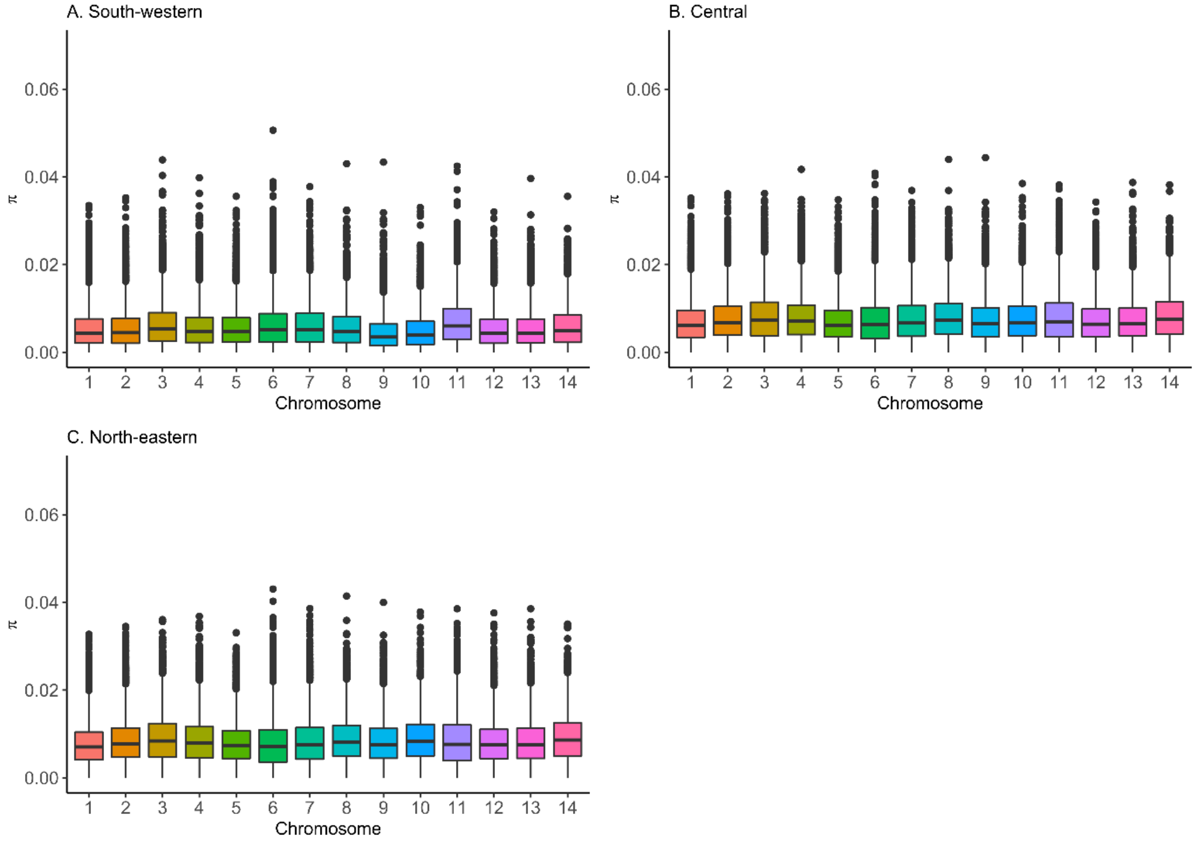


Figure S3.12. Pairwise nucleotide diversity π bar plots per population, computed in windows of 5,000 bp across the entire *B. platyphylla* genome. Populations according to *fastSTRUCTURE* at K = 3, excluding admixed individuals (q < 0.9).

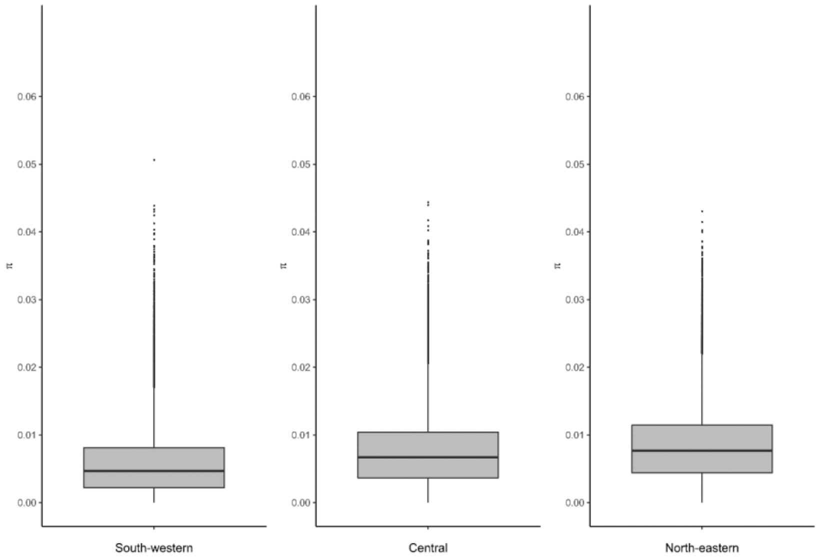


Figure S3.13. Jack-knife test of variable importance, using training gain for the Maxent model including 11 variables.

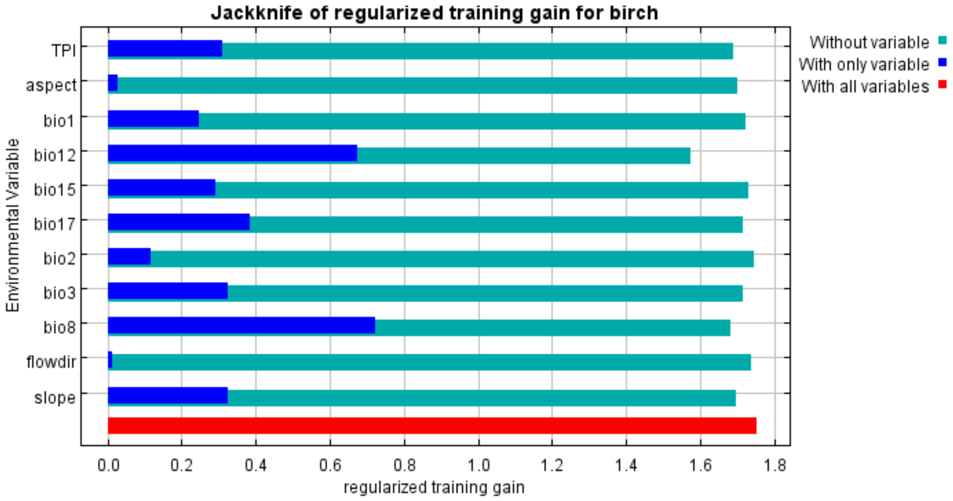


Figure S3.14. Jack-knife test of variable importance, using test gain instead of training gain for the Maxent model including 11 variables.

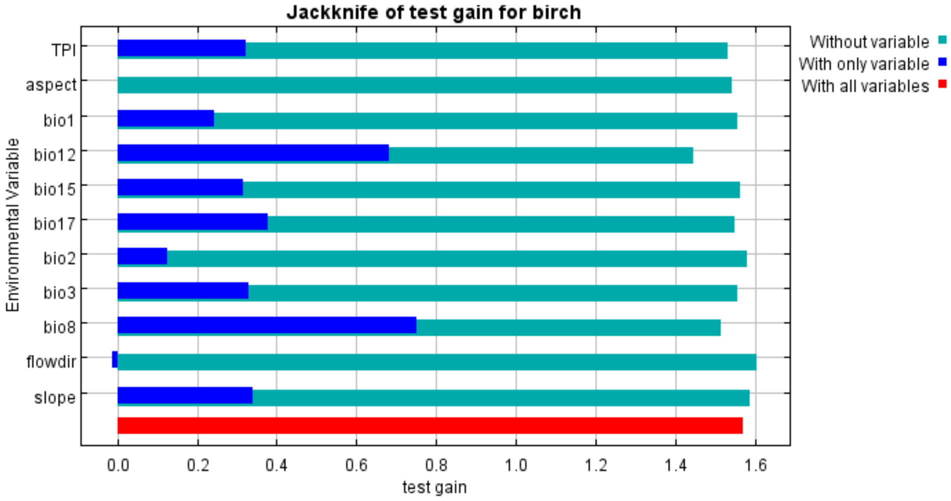


Figure S3.15. Jack-knife test of variable importance, using AUC on test data for the Maxent model including 11 variables.

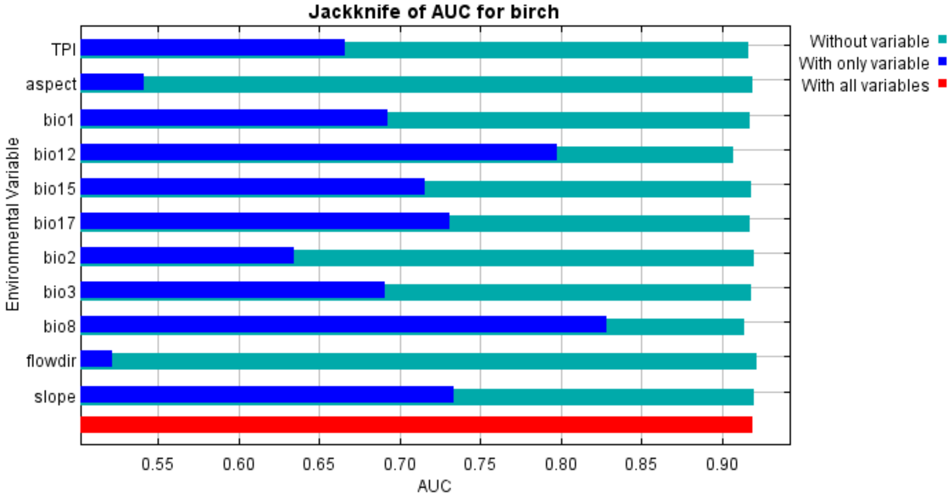


Figure S3.16. Omission rate and predicted area as a function of the cumulative threshold for the final Maxent model including nine variables (Table 3.2), averaged over 50 replicate runs.

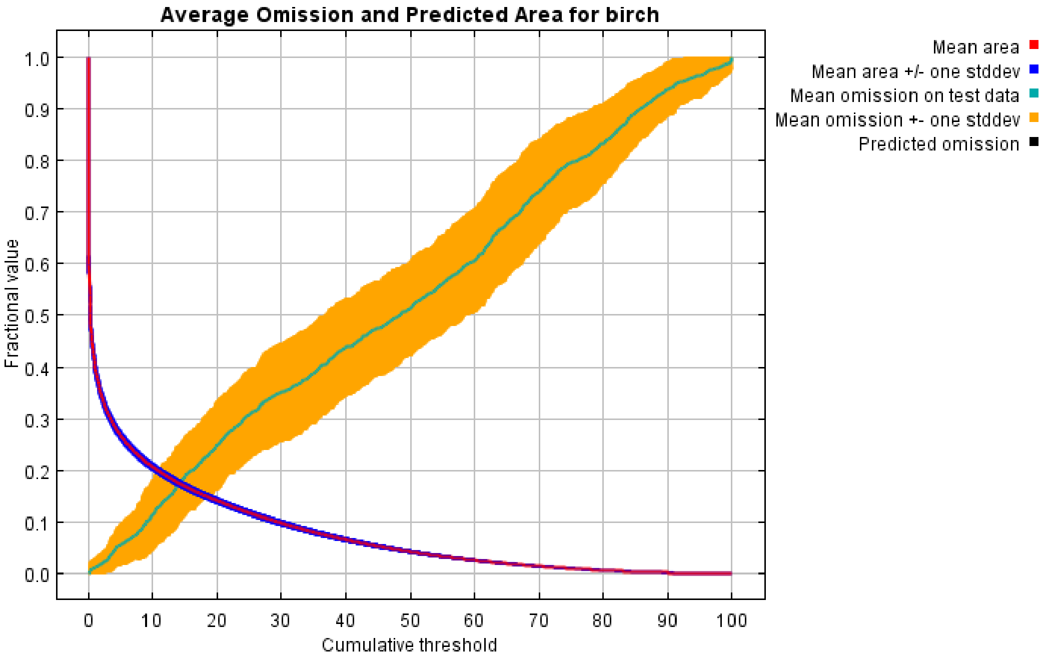


Figure S3.17. Receiver Operator Characteristic (ROC) curve for the final Maxent model including nine variables (Table 3.2), averaged over 50 replicate runs. The average test AUC for the replicate runs is 0.913, and the standard deviation is 0.017.

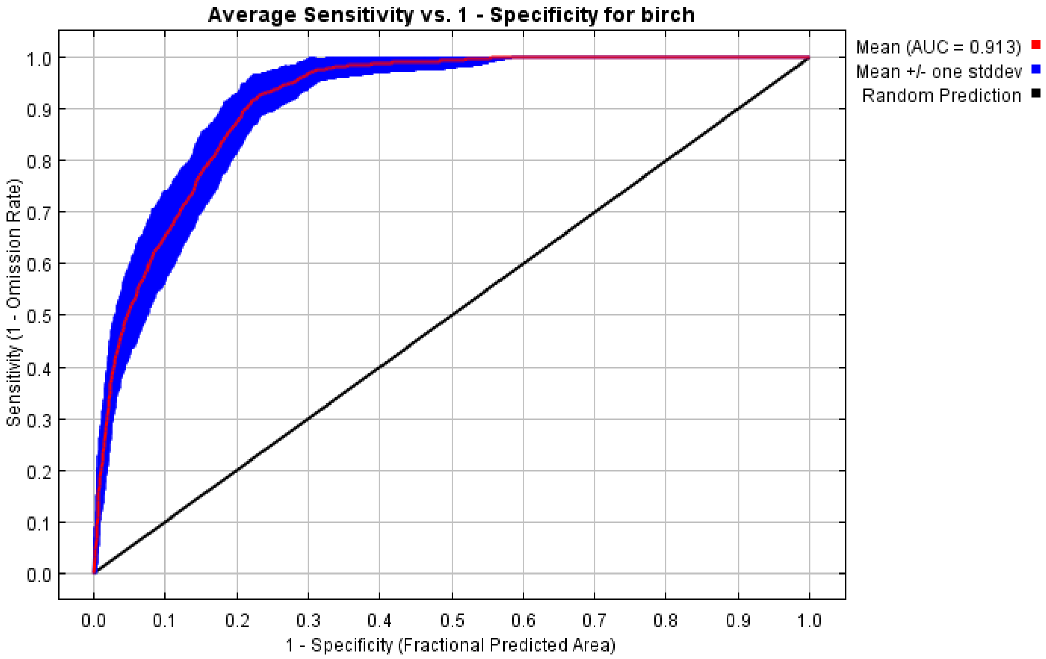
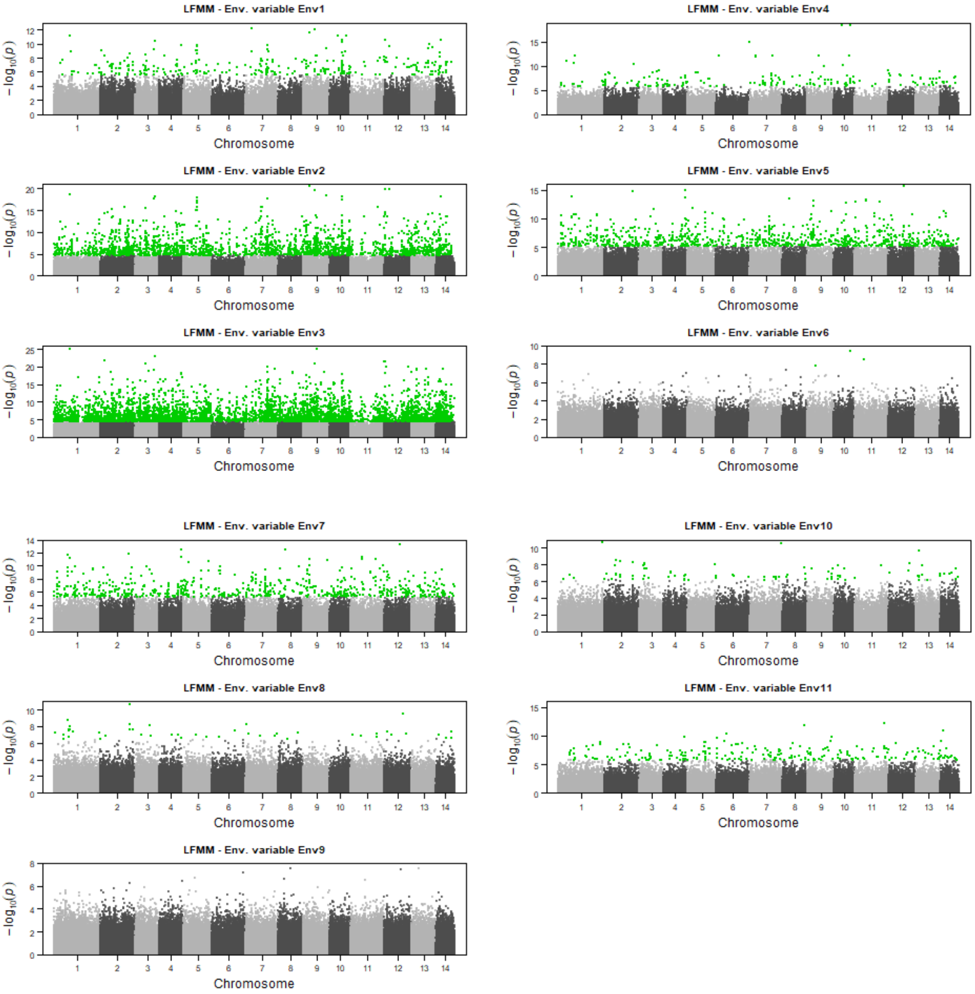


Figure S3.18. LFMM2 results. A) P-values Manhattan plot of the *LFMM2* analysis with K = 3. Green points are SNP with corresponding q-value (FDR) < 1%. B) Histogram of p-values across environmental variables, suggesting that the rate of false positive is well controlled. The histogram of significance values is expected to be flat with a peak near 0. Total number of SNPs tested = 1,387,994.

A)



B)

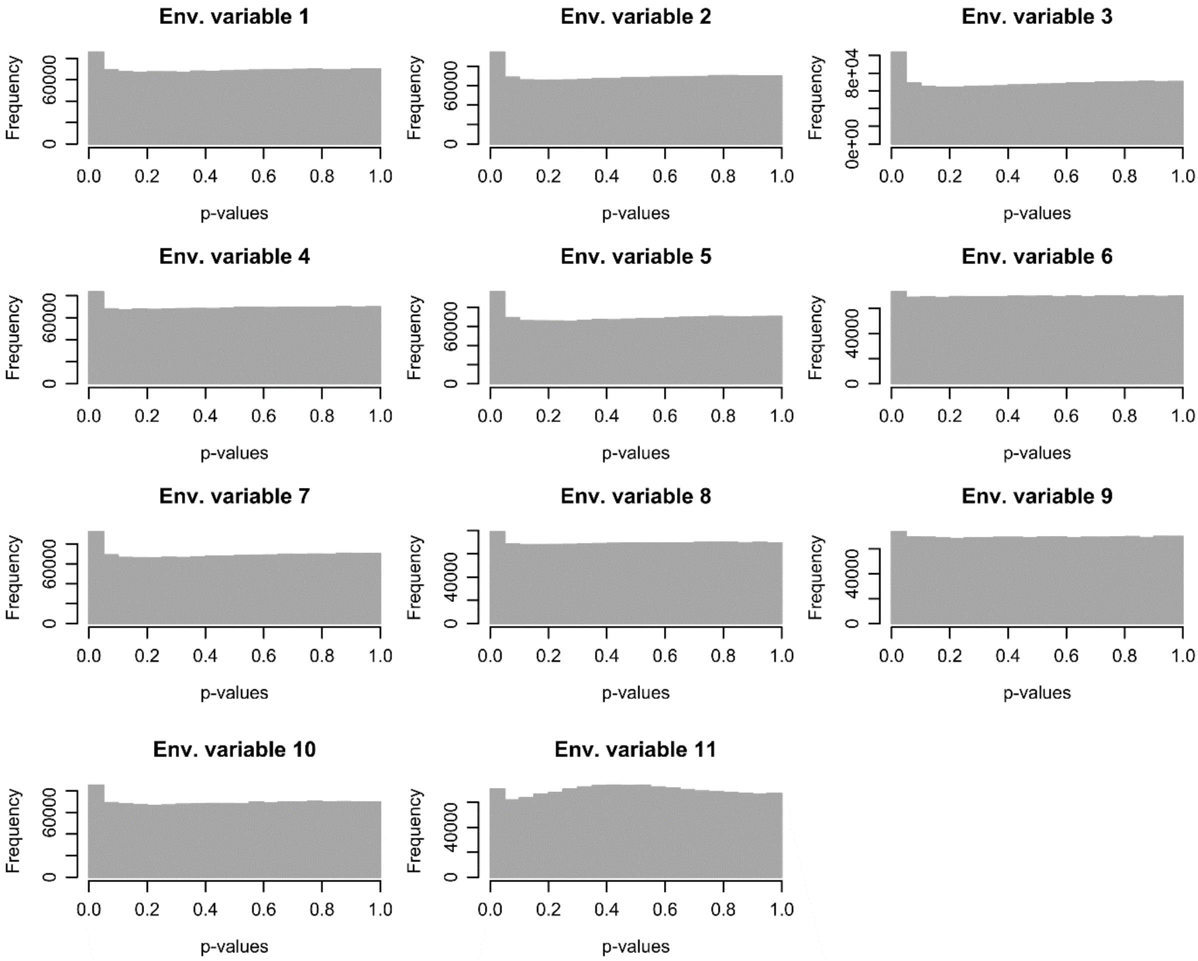

Env. variable 1
Frequency
0
60000
0.0
0.2
0.4
0.6
0.8
1.0
p-values
Env. variable 2
Frequency
0
60000
0.0
0.2
0.4
0.6
0.8
1.0
p-values
Env. variable 3
Frequency
0e+00
8e+04
0.0
0.2
0.4
0.6
0.8
1.0
p-values
Env. variable 4
Frequency
0
60000
0.0
0.2
0.4
0.6
0.8
1.0
p-values
Env. variable 5
Frequency
0
60000
0.0
0.2
0.4
0.6
0.8
1.0
p-values
Env. variable 6
Frequency
0
40000
0.0
0.2
0.4
0.6
0.8
1.0
p-values
Env. variable 7
Frequency
0
60000
0.0
0.2
0.4
0.6
0.8
1.0
p-values
Env. variable 8
Frequency
0
40000
0.0
0.2
0.4
0.6
0.8
1.0
p-values
Env. variable 9
Frequency
0
40000
0.0
0.2
0.4
0.6
0.8
1.0
p-values
Env. variable 10
Frequency
0
60000
0.0
0.2
0.4
0.6
0.8
1.0
p-values
Env. variable 11
Frequency
0
40000
0.0
0.2
0.4
0.6
0.8
1.0
p-values

Figure S3.19. P-values Manhattan plot of the Samβada analysis, only for the variables that reported significant hits. Green points are SNP with corresponding q-value (FDR) < 0.01. Total number of SNPs tested = 1,387,994.

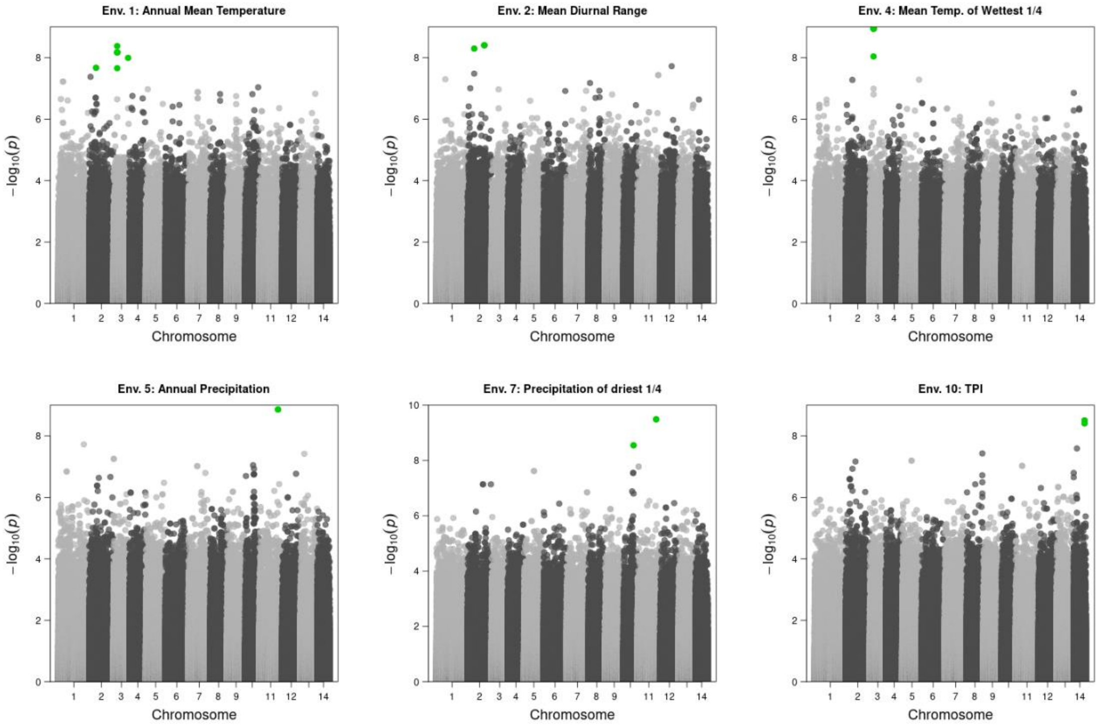


Figure 3.20. EAA results. **A)** Distribution of the 7,609 putatively adaptive SNPs identified with LFMM2 across the reference genome. **B)** Distribution of the 11,304 SNP-environment associations detected (q < 0.01 in LFMM2) across environmental variables.

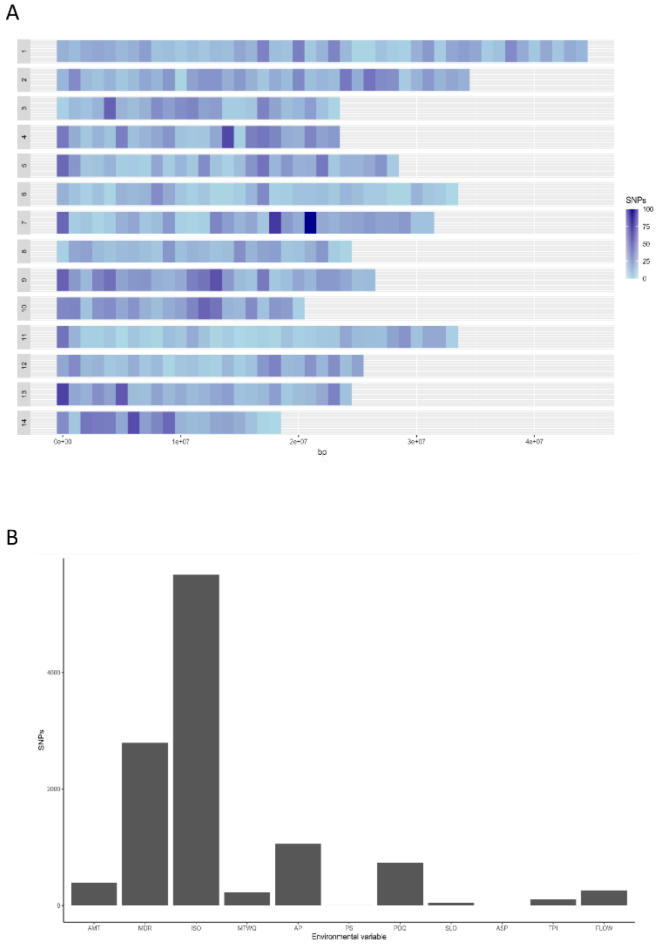


Figure S3.21. PCA of 71 *B. platyphylla* individuals based on the 7,609 putatively adaptive SNPs identified in this study. Colours represent fastSTRUCTURE populations assignments at K = 3. A) PC1 against PC2. B) PC1 against PC3. C) PC2 against PC3. D) Eigenvalues of the computed principal components.

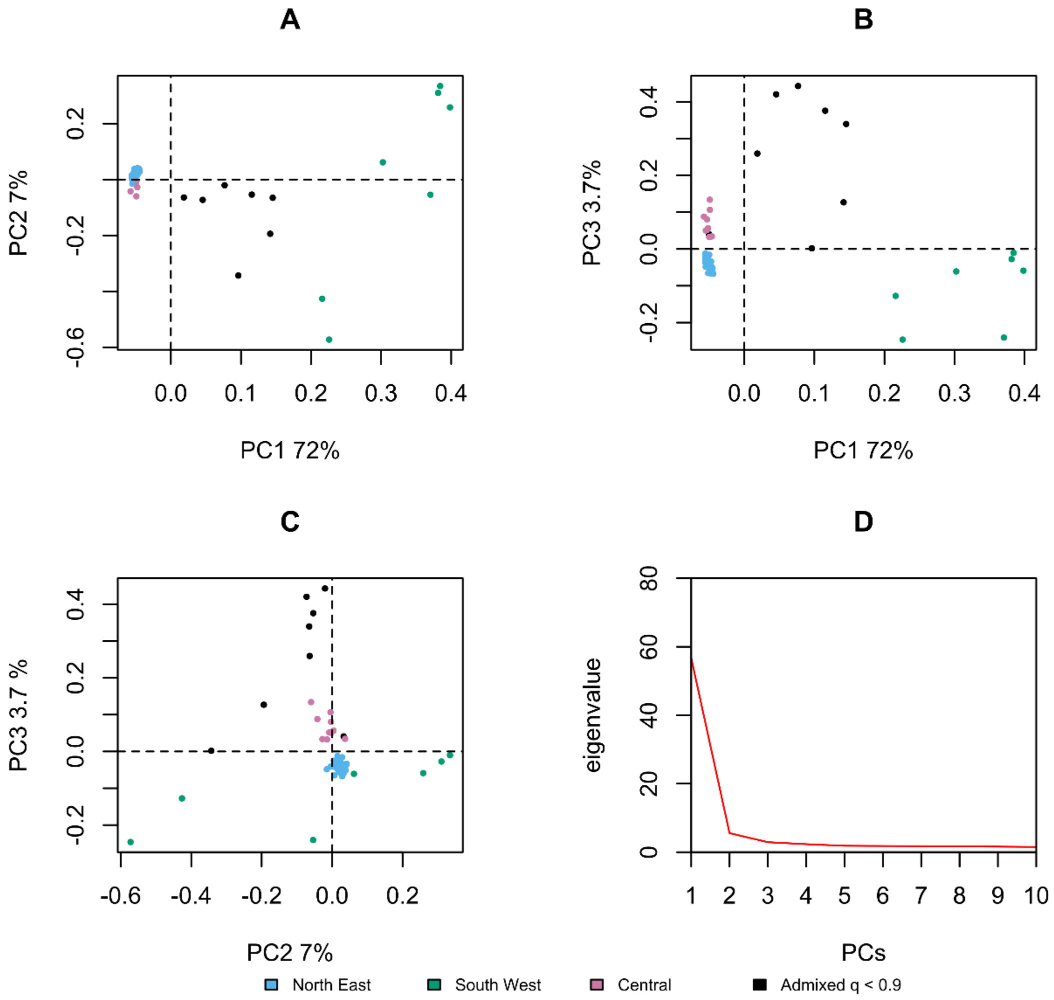


Figure S3.22. PCA of 71 *B. platyphylla* individuals based on 7,500 putatively “neutral” SNPs. Colours represent fastSTRUCTURE populations assignments at K = 3. A) PC1 against PC2. B) PC1 against PC3. C) PC2 against PC3. D) Eigenvalues of the computed principal components.

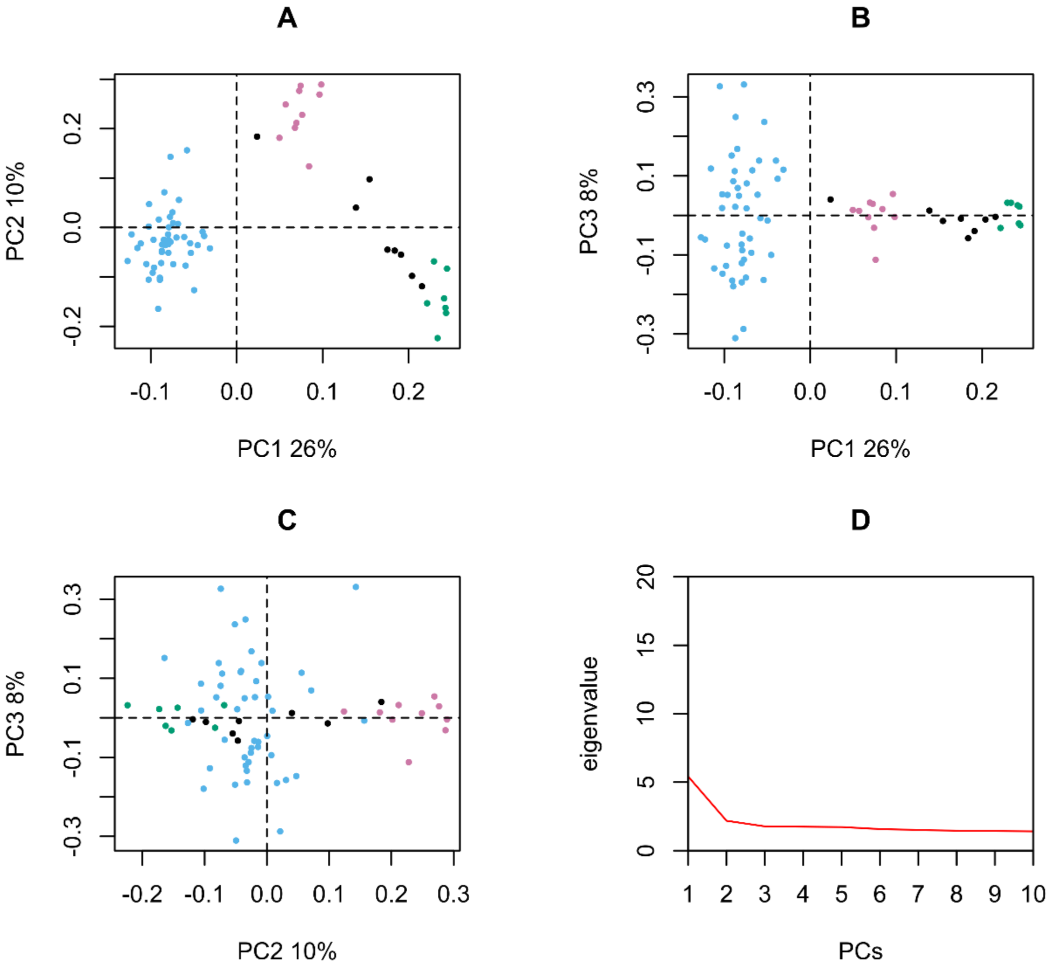


Figure S3.23. F_{st} (by SNP site) distribution between *B. platyphylla* populations, based on 7,609 adaptive SNPs and excluding admixed individuals. Populations' assignment based on *fastSTRUCTURE* at K = 3, excluding admixed individuals (q < 0.9).

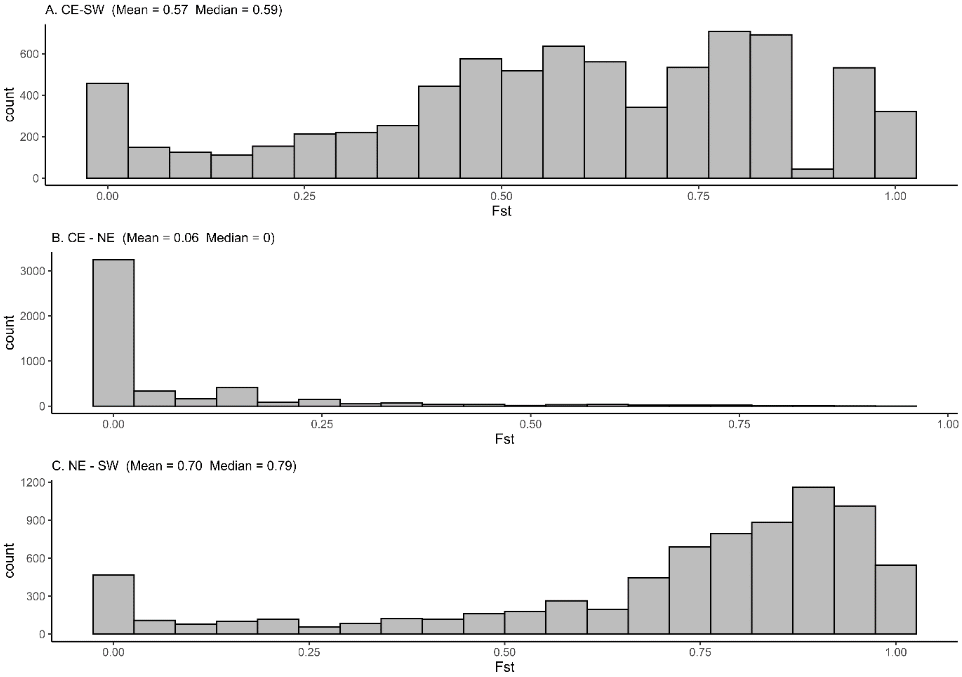


Figure S3.24. F_{st} (by SNP site) distribution between *B. platyphylla* populations, based on 7,500 neutral SNPs and excluding admixed individuals. Populations' assignment based on *fastSTRUCTURE* at K = 3, excluding admixed individuals (q < 0.9).

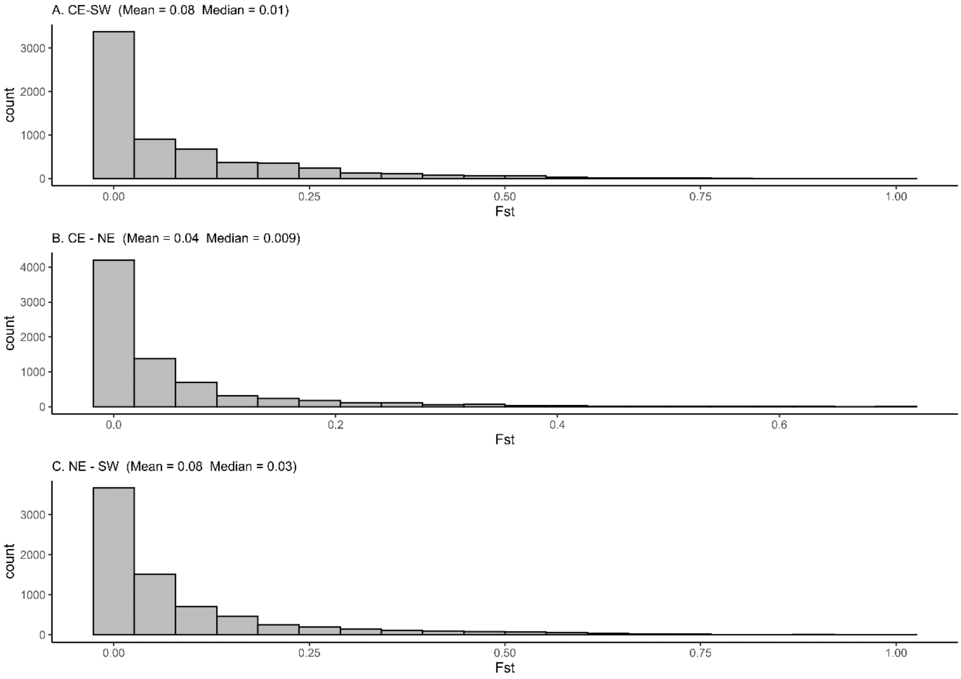


Figure S3.25. Omics Box functional enrichment analysis results (FDR < 5%) of the 1,633 genes spanned by the 7,609 adaptive SNPs identified with *LFMM2*. The plot includes all three ontology levels. The GOs retrieved for each gene region have been reduced to the most specific term in the gene ontology hierarchy.

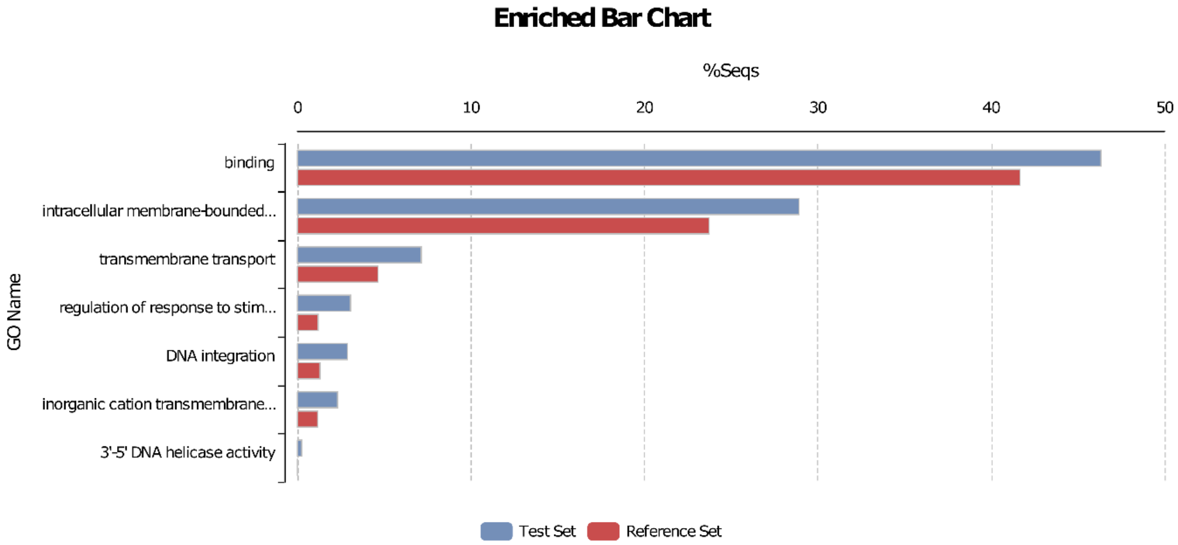


Table S3.1. Detailed information on sampling and resequencing. This table was made by Nian Wang.

Population code	*Latitude*	*Longitude*	*Number of clean reads*	*Number of mapped reads*	*Percentage of mapped reads (%)*
AHT15	48.3777	85.7447	95108278	91773480	96.5
AHT20	48.3795	85.7479	79444471	76620148	96.4
ALE027	51.2818	121.4191	96472026	93032815	96.4
ALS046	50.9	121.4212	78941821	76179630	96.5
ALY036	51.5469	121.7336	75425352	72546335	96.2
BEJ3	47.7249	86.9159	92422883	89176947	96.5
BJC007	53.4901	122.3499	70706776	68192740	96.4
BMX54	32.8011	100.8116	87728865	84442989	96.3
BX54	30.8295	102.7403	82641706	79579298	96.3
CGZ003	51.9907	124.617	86423518	82807447	95.8
DCX5	28.6167	100.1922	74270014	71367883	96.1
DFX11	31.1147	100.9644	80531220	74391328	92.4
DQS13	41.0516	111.8089	83821719	80586874	96.1
DTX94	37.186	101.5427	88645240	85119253	96.0
EDG18	41.7171	126.4611	68634376	65886090	96.0
EDG27	41.717	126.4618	84619385	77580250	91.7
FHSJ1	44.2193	127.9786	93948831	90247913	96.1
FYX7	47.2135	89.843	73588524	70957171	96.4
HBH12	48.0726	86.342	86639208	83393478	96.3
HBH2	48.0727	86.3415	93353556	89974372	96.4
HGS18	47.1663	130.2832	86411628	83065783	96.1
HLS53	42.4833	129.0528	83804019	81096539	96.8
HNX7	46.3334	130.886	89061012	86065739	96.6
HR6	41.2565	125.1582	134,129,672	129136963	96.3
HTY15	41.8068	126.3612	74296931	71749626	96.6
HX1	34.0296	105.9652	84099348	80988071	96.3
JD24	29.2119	101.4489	79669091	75939795	95.3
JS39	27.4391	99.8034	75823319	72834926	96.1
JXS19	45.3755	130.9656	74515373	71741428	96.3
JY23	48.6551	130.4574	92096901	88942754	96.6
JYX28	42.292	126.7209	74378176	71889202	96.7
KCL2	41.0696	125.0931	120,208,514	115,947,440	96.5
KNS12	48.6767	87.014	81694447	77800819	95.2
KNS4	48.6787	87.0135	78621478	75626395	96.2
LBX31	47.7191	130.8627	73488854	70960134	96.6
LJP1	27.1222	100.2572	74489677	69887415	93.8
LLZ018	41.8415	126.5507	68977737	66455786	96.3
LM049	52.8803	123.1261	95296931	91875251	96.4
LMS033	48.3454	122.2915	69164718	66810351	96.6
LS33	36.8358	111.9603	79043181	76301165	96.5
LXJ13	31.6495	102.821	101980534	97912512	96.0

LYC11	34.4317	110.4852	72037777	69435165	96.4
MDJ15	44.5341	130.1711	80833056	77707495	96.1
MDJ2	44.5343	130.1708	82740819	79510087	96.1
MG089	52.4248	122.5113	80846449	77864358	96.3
MH024	53.3786	122.2587	68798644	66348179	96.4
MNS1	43.8274	86.0682	80978331	78011246	96.3
MQX50	34.6597	100.6272	83453311	80557376	96.5
MSZ1	47.9875	130.759	85739346	82621706	96.4
PQG49	37.89	111.4306	78183522	75533247	96.6
QHX14	46.6841	90.3548	95738066	92386580	96.5
QSLC51	48.5909	129.8668	81216632	78177595	96.3
RTX16	32.5793	101.0815	73673220	71079464	96.5
SDX29	31.9529	100.9266	70954809	68184929	96.1
SLS16	30.9248	102.3128	67,482,194	64922238	96.2
SWP3	35.4288	111.9715	74598260	71796850	96.2
TZZ103	36.9395	102.6141	80145267	77185210	96.3
WCS21	44.6561	127.3484	85343984	82003059	96.1
WCS6	44.6563	127.3468	97872192	94574167	96.6
WEG19	43.6511	129.516	82097483	79232118	96.5
WEG26	43.6511	129.516	71811997	69041972	96.1
WLG024	52.6535	124.4441	67993333	65198125	95.9
WT42	38.831	113.8389	83683234	80511518	96.2
WTG28	41.9269	126.0903	92797096	89223554	96.1
WY29	43.0257	129.5483	72726663	70308073	96.7
WYQ5	48.0796	129.2231	78521085	75804535	96.5
XEXL17	45.2187	82.0478	73114971	70656562	96.6
XEXL2	45.2199	82.0454	85742705	82188619	95.9
XFL14	42.5305	128.7583	98912997	94705944	95.7
XFL2	42.5303	128.7587	89848329	86157227	95.9
XGLLP21	27.8293	99.7441	70443682	67688473	96.1
XGLLP7	27.8293	99.7441	78361504	74936652	95.6
XHC004	49.9104	124.5655	102022698	98582755	96.6
XHS10	42.8874	127.0478	80830112	77985804	96.5
XMX5	37.2623	101.9801	92847104	89270911	96.1
XXG025	47.8172	122.6181	88881295	85900807	96.6
XYQ060	50.7067	124.3103	77914626	75171347	96.5
YCS25	47.6333	128.5499	71287688	65225061	91.5
YJA54	30.0472	101.2854	80072173	76742373	95.8
YKS014	49.1213	120.9051	104383304	99384140	95.2
YLK048	48.8535	121.6293	86034800	83023217	96.5
ZJ027	52.9555	122.5723	66,016,635	63643574	96.4
ZSL004	49.2359	124.6643	75063914	72555543	96.7

Table S3.2. RONA results for each of the 71 *B. platyphylla* individual for seven environmental variables, under the future climate profile ssp370 at 2080-2100. The second row shows the number of SNPs identified associated with *LFMM2* and therefore included in the RONA calculation for each environmental variable.

Variable	**AMT**	**MDR**	**ISO**	**MTWQ**	**AP**	**PS**	**PDQ**
SNPs	391	2794	5679	227	1059	3	735
ALE027	0.0644	0.0273	0.0144	0.277	0.0138	0.009	0.0126
ALS046	0.0564	0.0205	0.0176	0.2569	0.0174	0.0406	0.0126
ALY036	0.0558	0.0251	0.0039	0.2109	0.0123	0.0232	0.0063
BJC007	0.0713	0.0116	0.0186	0.1656	0.0054	0.0034	0.0126
BMX54	0.098	0.0319	0.0054	0.0239	0.0718	0.0026	0
BX54	0.0923	0.0686	0.1573	0.096	0.0381	0.0968	0.0063
CGZ003	0.1212	0.0162	0.0186	0.0666	0.0165	0.0595	0.0189
DCX5	0.1061	0.0309	0.0271	0.1854	0.0163	0.2788	0.044
DFX11	0.111	0.0395	0.0613	0.2209	0.0141	0.0597	0.0063
DQS13	0.1931	0.0039	0.0363	0.2004	0.0267	0.3215	0.0126
DTX94	0.0714	0.0159	0.0054	0.1264	0.024	0.0974	0.0126
EDG18	0.2518	0.0093	0.0563	0.2356	0.0695	0.1824	0
EDG27	0.2523	0.0093	0.0563	0.2332	0.0694	0.1824	0
FHSJ1	0.2281	0.033	0.0152	0.239	0.0424	0.1606	0.0312
HGS18	0.2464	0.016	0.0238	0.3724	0.0288	0.0325	0.0126
HLS53	0.2335	0.014	0.0391	0.089	0.0484	0.1292	0.0126
HNX7	0.2231	0.0336	0.0118	0.3261	0.0315	0.0094	0
HR6	0.2276	0.0201	0.0517	0.2754	0.0271	0.1306	0.0345
HTY15	0.2581	0.0076	0.051	0.3718	0.0645	0.1782	0
HX1	0.1485	0.0098	0.0107	0.2984	0.0451	0.0414	0
JD24	0.0738	0.0079	0.0488	0.1624	0.0098	0.0504	0
JS39	0.1041	0.0063	0.0617	0.2531	0.0011	0.3347	0.0702
JXS19	0.2457	0.0358	6e-04	0.3371	0.0267	0.0013	0
JY23	0.2448	0.038	0.0651	0.3956	0.0076	0.0094	0.0063
JYX28	0.2725	0.0083	0.0491	0.3257	0.0477	0.1913	0.0174
KCL2	0.2121	0.0095	0.0611	0.3379	0.0212	0.0759	0.0172
LBX31	0.2485	0.0116	0.0404	0.3515	0.0152	0.0478	0.0126
LJP1	0.0951	0.0191	0.0308	0.2418	0.0054	0.3116	0.1259
LLZ018	0.2834	0.0083	0.0511	0.2297	0.0528	0.2101	0.0294
LM049	0.0439	0.0121	0.0126	0.0612	0.0106	0.0626	0.0251
LMS033	0.2346	0.0037	0.0182	0.2406	0.0174	0.1257	0.0189
LS33	0.2297	0.0015	0.016	0.3736	0.0038	0.1086	0.044
LXJ13	0.0145	0.0096	0.0206	0.0469	0.0234	0.0358	0.0062
LYC11	0.2093	0.0187	0.0322	0.3113	0.0348	0.0729	0.0616
MDJ15	0.2721	0.0078	0.0244	0.325	0.0392	0.0761	0.0186
MDJ2	0.2726	0.0078	0.0244	0.3263	0.0392	0.0761	0.0186
MG089	0.0518	0.0017	0.0311	0.1053	0.0098	0.0356	0.0063
MH024	0.0558	0.0157	0.0184	0.081	0.0021	0.0151	0

MQX50	0.1198	0.0418	0.0604	0.189	0.0424	0.1299	0.0126
MSZ1	0.2427	0.0085	0.054	0.3173	0.0256	0.0234	0.0063
PQG49	0.1895	0.0093	0.0152	0.3765	0.0234	0.1692	0.0312
QSLC51	0.2733	0.0157	0.0344	0.2716	0.0016	0.027	0.0251
RTX16	0.0767	0.0351	0.0113	0.0044	0.0604	0.0153	0
SDX29	0.0796	0.0396	0.0427	0.0153	0.0353	0.0436	0.0063
SLS16	0.1267	0.0172	0.0373	0.1384	0.0539	0.0471	0.0565
SWP3	0.2477	0.0069	0.0491	0.3891	0.0022	0.0322	0.0669
TZZ103	0.1335	0.0222	0.0407	0.2946	0.0048	0.1358	0.0314
WCS21	0.2496	0.0275	0.0114	0.374	0.0098	0.1328	0.0186
WCS6	0.2498	0.0274	0.0114	0.3741	0.0098	0.1328	0.0186
WEG19	0.2203	0.0193	0.0429	0.3058	0.0555	0.128	0.0126
WEG26	0.2203	0.0193	0.0429	0.3067	0.0555	0.128	0.0126
WLG024	0.1163	0.0199	0.0185	0.0856	0.0098	0.0218	0.0126
WT42	0.2875	0.0039	0.0444	0.3282	0.0256	0.3431	0.063
WTG28	0.2398	0.0142	0.049	0.2558	0.0745	0.1958	0
WY29	0.2403	0.0124	0.0584	0.3272	0.0473	0.1136	0
WYQ5	0.2511	0.0231	0.0214	0.2645	0.0103	0.0334	0.0189
XFL14	0.1901	0.0339	0.0073	0.0544	0.0658	0.1758	0.0315
XFL2	0.201	0.0337	0.0073	0.0529	0.0598	0.1758	0.0315
XGLLP21	0.1147	0.041	0.0057	0.2543	0.0033	0.3226	0.0599
XGLLP7	0.1147	0.0406	0.0057	0.2547	0.0033	0.1667	0.0598
XHC004	0.245	0.0117	0.0233	0.201	0.0326	0.1132	0
XHS10	0.2566	0.033	0.0182	0.3528	0.0549	0.1781	0
XMX5	0.1059	0.0212	0.0102	0.1855	0.0041	0.1126	0.0188
XXG025	0.2757	0.0113	0.0274	0.364	0.0104	0.1005	0.0189
XYQ060	0.1896	0.0104	0.027	0.1329	0.0268	0.0582	0.0063
YCS25	0.2582	0.0218	0.0272	0.3147	0.0185	0.0191	0.0126
YJA54	0.141	0.0012	0.0232	0.259	0.025	0.0471	0.0251
YKS014	0.1758	0.0083	0.001	0.0549	0.0191	0.1696	0.0496
YLK048	0.1673	0.0123	0.0088	0.1623	0.0327	0.1448	0.0249
ZJ027	0.0455	0.0181	0.0273	0.068	0.0078	0.0331	0.0189
ZSL004	0.2739	0.0125	0.0178	0.3271	0.0349	0.0717	0.0063
Mean	0.1773	0.0189	0.0305	0.2314	0.0281	0.1073	0.0207
Min r^2	1e-04	0	0	1e-04	0	0.136	0
Max r^2	0.3956	0.2522	0.8405	0.5437	0.2643	0.1809	0.1398
Average r^2	0.1959	0.0157	0.459	0.1854	0.0977	0.1627	0.0214

Table S3.3. Weighted-mean RONA and maximum RONA of the 71 *B. platyphylla* individuals, relative to the future climate profile ssp370 in 2080-2100.

ID	mean RONA	max RONA
ALE027	0.086	0.277
ALS046	0.0829	0.2569
ALY036	0.0664	0.2109
BJC007	0.0525	0.1656
BMX54	0.0418	0.098
BX54	0.0609	0.1573
CGZ003	0.0359	0.1212
DCX5	0.0796	0.2788
DFX11	0.0771	0.2209
DQS13	0.0877	0.3215
DTX94	0.0528	0.1264
EDG18	0.1117	0.2518
EDG27	0.111	0.2523
FHSJ1	0.1028	0.239
HGS18	0.1257	0.3724
HLS53	0.0622	0.2335
HNX7	0.111	0.3261
HR6	0.1069	0.2754
HTY15	0.1455	0.3718
HX1	0.1065	0.2984
JD24	0.055	0.1624
JS39	0.0979	0.3347
JXS19	0.1119	0.3371
JY23	0.1243	0.3956
JYX28	0.1297	0.3257
KCL2	0.116	0.3379
LBX31	0.1158	0.3515
LJP1	0.1025	0.3116
LLZ018	0.1093	0.2834
LM049	0.0295	0.0626
LMS033	0.0889	0.2406
LS33	0.1212	0.3736
LXJ13	0.0265	0.0469
LYC11	0.1197	0.3113
MDJ15	0.1206	0.325
MDJ2	0.1209	0.3263
MG089	0.0381	0.1053
MH024	0.0268	0.081
MQX50	0.0842	0.189
MSZ1	0.1096	0.3173
PQG49	0.1293	0.3765
QSLC51	0.0907	0.2733

RTX16	0.032	0.0767
SDX29	0.0284	0.0796
SLS16	0.077	0.1384
SWP3	0.128	0.3891
TZZ103	0.0983	0.2946
WCS21	0.1226	0.374
WCS6	0.1226	0.3741
WEG19	0.1227	0.3058
WEG26	0.123	0.3067
WLG024	0.0358	0.1163
WT42	0.1335	0.3431
WTG28	0.1189	0.2558
WY29	0.124	0.3272
WYQ5	0.0905	0.2645
XFL14	0.0621	0.1901
XFL2	0.0596	0.201
XGLLP21	0.096	0.3226
XGLLP7	0.0903	0.2547
XHC004	0.0825	0.245
XHS10	0.1356	0.3528
XMX5	0.0635	0.1855
XXG025	0.1203	0.364
XYQ060	0.0589	0.1896
YCS25	0.1064	0.3147
YJA54	0.0915	0.259
YKS014	0.0432	0.1758
YLK048	0.0732	0.1673
ZJ027	0.0293	0.068
ZSL004	0.1173	0.3271
Mean	0.0896	0.2533
Max	0.1455	0.3956

Table S3.4. RONA of each population for the seven environmental variables tested, under ssp370 at 2080-2100. Each population RONA was calculated by averaging the RONA of the individuals belonging to that population, according to *fastSTRUCTURE* at K = 3 and excluding admixed individuals.

	Central	**South-western**	**North-eastern**	**All**
AMT	0.17	0.10	0.20	0. 17
MDR	0.01	0.02	0.01	0. 01
ISO	0.02	0.02	0.02	0. 03
MTWQ	0.28	0.23	0.24	0. 23
AP	0.02	0.009	0.03	0. 02
PS	0.12	0.21	0.09	0. 10
PDQ	0.03	0.05	0.01	0. 02

Final Maxent model response curves

There are two plots for each variable included in the final Maxent model. The first plots show how the predicted probability of presence changes as each environmental variable is varied, keeping all other environmental variables at their average sample value. The second plots show the predicted probability of presence in a Maxent model created using only the corresponding variable.

Figure S3.26. AMT response curves.

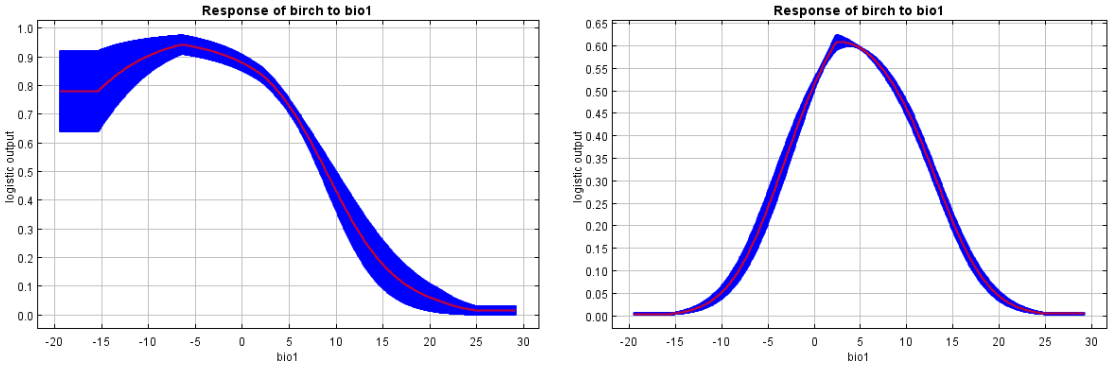


Figure S3.27. MDR response curves.

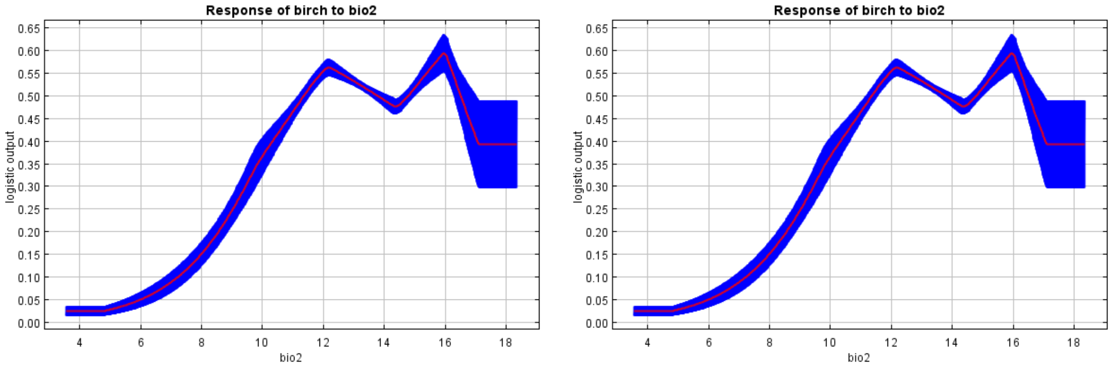


Figure S3.28. ISO response curves.

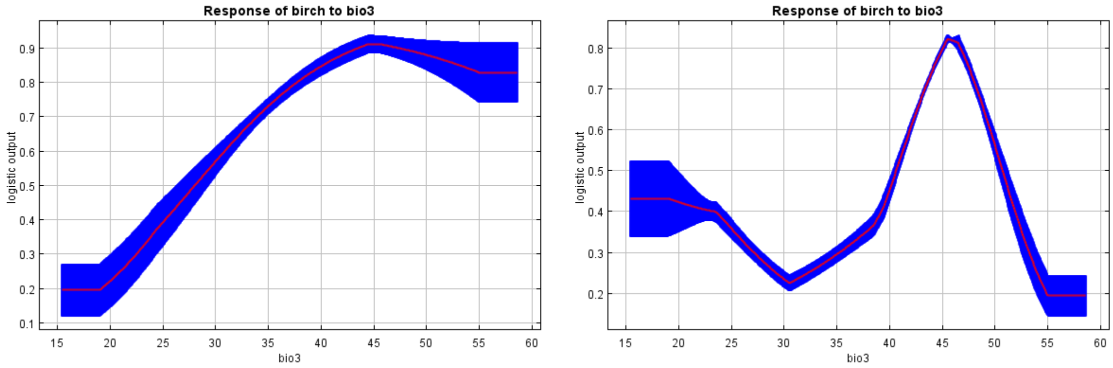


Figure S3.29. MTWQ response curves.

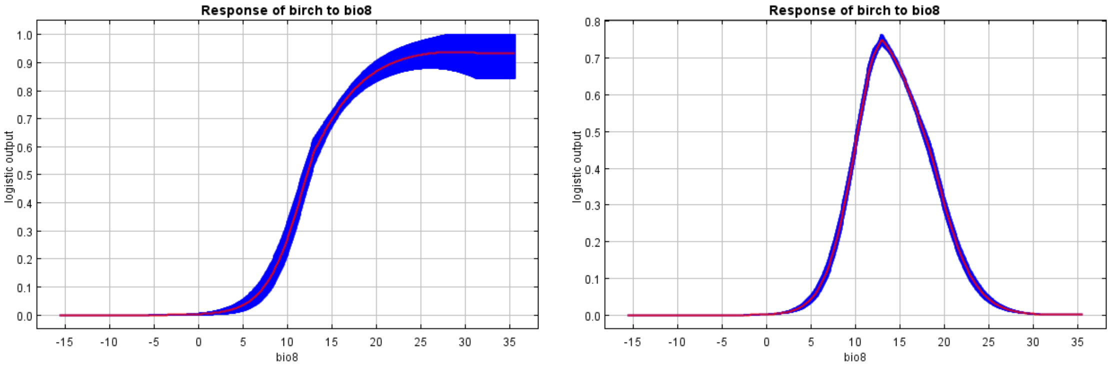


Figure S3.30. AP response curves.

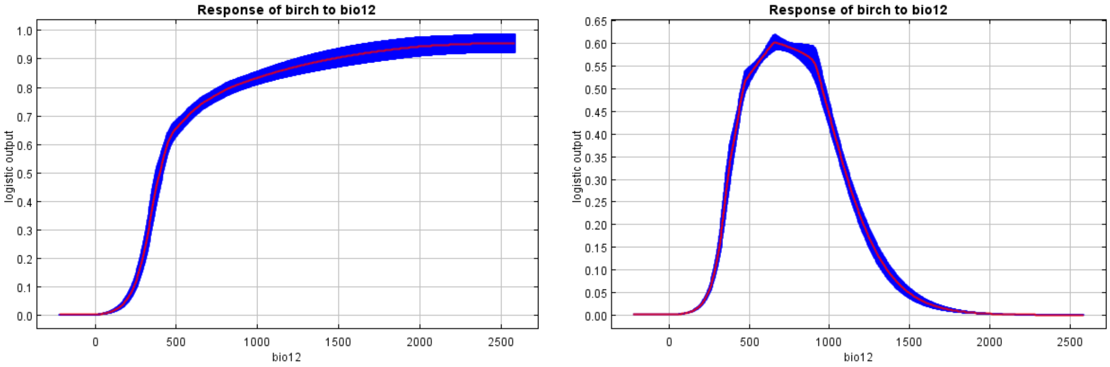


Figure S3.31. PS response curves.

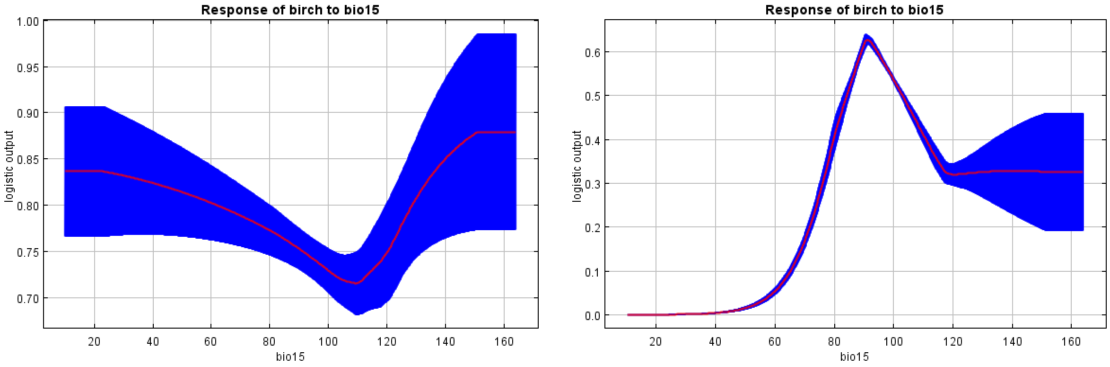


Figure S3.32. PDQ response curves.

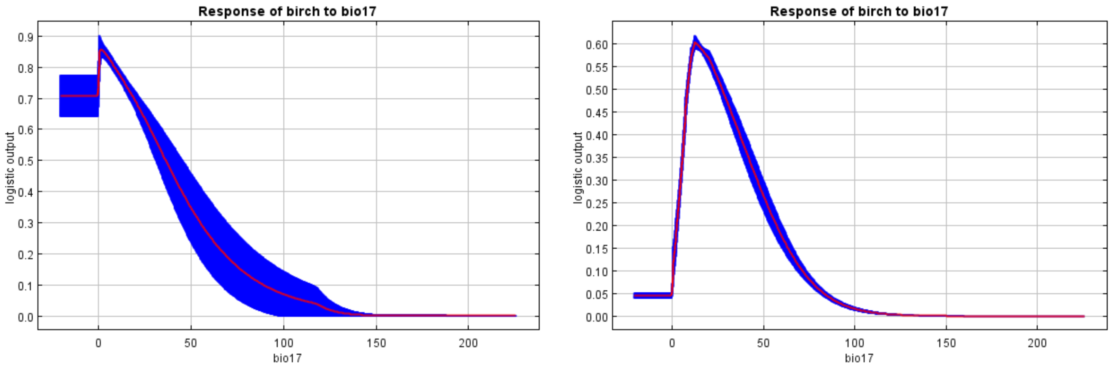


Figure S3.33. SLOPE response curves.

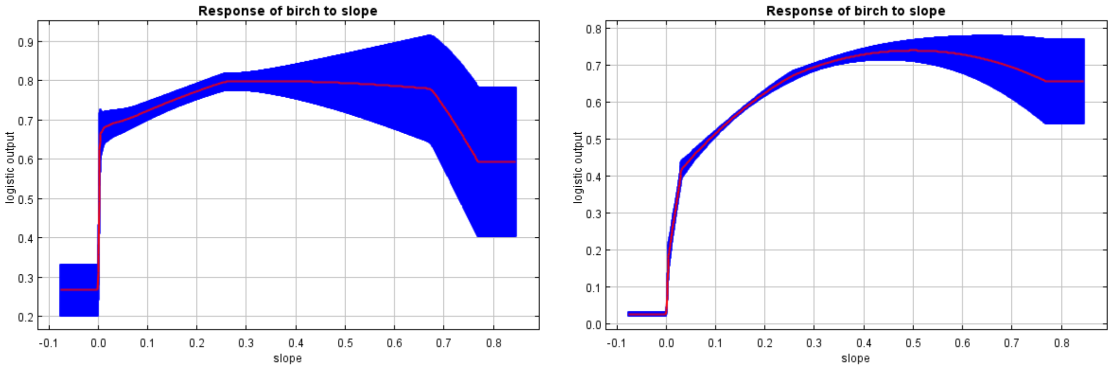
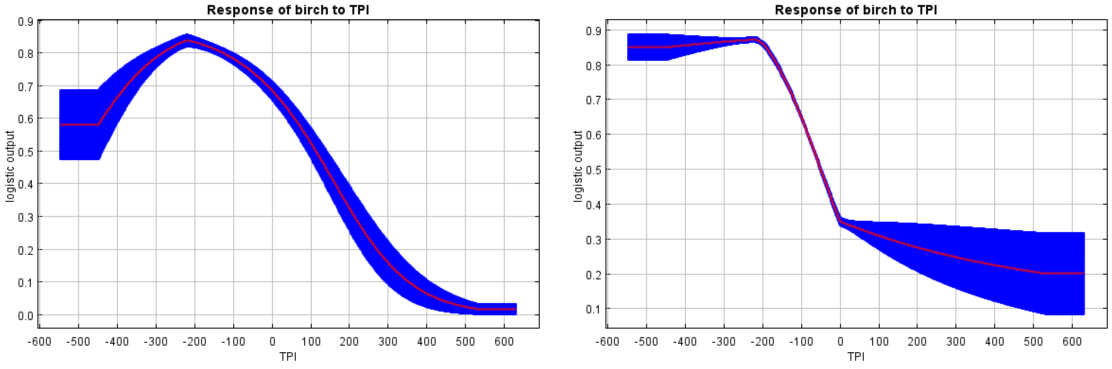


Figure S3.34. TPI response curves.



References

1. Ågren, J., Oakley, C., Lundemo, S., & Schemske, D. (2016). Adaptive divergence in flowering time among natural populations of *Arabidopsis thaliana*: Estimates of selection and QTL mapping. *Evolution*, *71*(3), 550-564.

2. Aitken, S., & Whitlock, M. (2013). Assisted Gene Flow to Facilitate Local Adaptation to Climate Change. *Annual Review Of Ecology, Evolution, And Systematics*, *44*(1), 367-388.

3. Aitken, S., Yeaman, S., Holliday, J., Wang, T., & Curtis-McLane, S. (2008). Adaptation, migration or extirpation: climate change outcomes for tree populations. *Evolutionary Applications*, *1*(1), 95-111.

4. Alachiotis, N., Stamatakis, A., & Pavlidis, P. (2012). OmegaPlus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets. *Bioinformatics*, *28* (17), 2274-2275.

5. Alberto, F., Aitken, S., Alía, R., González-Martínez, S., Hänninen, H., & Kremer, A., … Savolainen, O. (2013). Potential for evolutionary responses to climate change – evidence from tree populations. *Global Change Biology*, *19*(6), 1645-1661.

6. Anderson E. (1949). Introgressive hybridization. New York: Wiley & Sons.

7. Ashburner, K., & McAllister, H. (2013). Botanical Magazine Monograph: The Genus Betula: A Taxonomic Revision of Birches. Erscheinungsort nicht ermittelbar: Royal Botanic Gardens.

8. Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., & Cherry, J., … Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, *25*(1), 25-29.

9. Bandelt, H., Forster, P., & Röhl, A. (1999). Median-joining networks for inferring intraspecific phylogenies*. Molecular Biology and Evolution*, *16* (1), 37–48.

10. Bao, L., Kudureti, A., Bai, W., Chen, R., Wang, T., Wang, H., & Ge, J. (2015). Contributions of multiple refugia during the last glacial period to current mainland populations of Korean pine (Pinus koraiensis). *Scientific Reports*, *5*(1).

11. Barreneche, T., Bodénès, C., Lexer, C., Trontin, J., Fluch, S., Streiff, R., Plomion, C., Roussel, G., Steinkellner, H., Burg, K., Favre, J., Glössl, J. and Kremer, A. (1998). A genetic linkage map of Quercus robur L. (pedunculate oak) based on RAPD, SCAR, microsatellite, minisatellite, isozyme and 5S rDNA markers. *TAG Theoretical and Applied Genetics*, *97*(7), 1090-1103.

12. Barreneche, T., Casasoli, M., Russell, K., Akkak, A., Meddour, H., Plomion, C., Villani, F. and Kremer, A. (2004). Comparative mapping between Quercus and Castanea using simple-sequence repeats (SSRs). *TAG Theoretical and Applied Genetics*, *108*(3), 558-566.

13. Barreneche, T., N. Bahrman, & A. Kremer. (1996) Two-dimensional gel electrophoresis confirms the low level of genetic differentiation between *Quercus robur L.* and *Quercus petraea* (Matt.) Liebl. *Forest Genetics*, 3, 89–92.

14. Battlay, P., Yeaman, S., & Hodgins, K. (2021). Pleiotropy drives repeatability in the genetic basis of adaptation. *bioRxiv*

15. Beatty, G., Montgomery, W., Spaans, F., Tosh, D., & Provan, J. (2016). Pure species in a continuum of genetic and morphological variation: sympatric oaks at the edge of their range. *Annals Of Botany*, *117* (4), 541-549.

16. Beatty, G., Philipp, M., & Provan, J. (2010). Unidirectional hybridization at a species' range boundary: implications for habitat tracking. *Diversity And Distributions*, *16*(1), 1-9.

17. Beaumont, M. A., & Nichols R. A. (1996). Evaluating loci for use in the genetic analysis of population structure. Proceedings of the Royal Society of London. *Series B: Biological Sciences*, *263*, 1619–1626.

18. Beaumont, M., & Balding, D. (2004). Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, *13*(4), 969-980.

19. Becker, M., Gruenheit, N., Steel, M., Voelckel, C., Deusch, O., & Heenan, P. et al. (2013). Hybridization may facilitate in situ survival of endemic species through periods of climate change. *Nature Climate Change*, *3*(12), 1039-1043.

20. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *57* (1), 289–300.

21. Bennington, C. C., Fetcher, N., Vavrek, M. C., Shaver, G. R., Cummings, K. J., & McGraw, J. B. (2012). Home site advantage in two long-lived arctic plant species: Results from two 30-year reciprocal transplant studies. *Journal of Ecology*, *100*, 841–851.

22. Berger, A., & Loutre, M. (2010). Modelling the 100-kyr glacial–interglacial cycles. *Global And Planetary Change*, *72*(4), 275-281.

23. Bhagwat, S., & Willis, K. (2008). Species persistence in northerly glacial refugia of Europe: a matter of chance or biogeographical traits? *Journal Of Biogeography*, *35*(3), 464-482.

24. Birks, H., & Willis, K. (2008). Alpines, trees, and refugia in Europe. *Plant Ecology & Diversity*, *1* (2), 147-160.

25. Bodénès, C., Chancerel, E., Ehrenmann, F., Kremer, A. and Plomion, C. (2016). High-density linkage mapping and distribution of segregation distortion regions in the oak genome. *DNA Research*, *23*(2), 115-124.

26. Bodénès, C., Chancerel, E., Gailing, O., Vendramin, G., Bagnoli, F., Durand, J., Goicoechea, P., Soliani, C., Villani, F., Mattioni, C., Koelewijn, H., Murat, F., Salse, J., Roussel, G., Boury, C., Alberto, F., Kremer, A. and Plomion, C. (2012). Comparative mapping in the Fagaceae and beyond with EST-SSRs. *BMC Plant Biology*, *12*(1):153.

27. Bodénès, C., Joandet, S., Laigret, F. and Kremer, A. (1997). Detection of genomic regions differentiating two closely related oak species Quercus petraea (Matt.) Liebl. and Quercus robur L. *Heredity*, *78*(4), 433-444.

28. Bolger, A., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114-2120.

29. Borrell, J., Zohren, J., Nichols, R., & Buggs, R. (2019). Genomic assessment of local adaptation in dwarf birch to inform assisted gene flow. *Evolutionary Applications*, *13*(1), 161-175.

30. Brady, C., Arnold, D., McDonald, J. and Denman, S. (2017). Taxonomy and identification of bacteria associated with acute oak decline. *World Journal of Microbiology and Biotechnology*, *33*(7):143.

31. Brendel, O., Le Thiec, D., Scotti-Saintagne, C., Bodénès, C., Kremer, A. and Guehl, J. (2007). Quantitative trait loci controlling water use efficiency and related traits in Quercus robur L. *Tree Genetics & Genomes*, *4*(2), 263-278.

32. Brennan, A., Woodward, G., Seehausen, Ol., Muñoz-Fuentes, V., Moritz, C., Guelmami, A., Abbott, R. & Edelaar, P. (2014). Hybridization due to changing species distributions: Adding problems or solutions to conservation of biodiversity during global change? *Evolutionary ecology research*, *16*, 475-491.

33. Brewer, S., Cheddadi, R., de Beaulieu, J., & Reille, M. (2002). The spread of deciduous Quercus throughout Europe since the last glacial period. *Forest Ecology And Management*, *156*(1-3), 27-48.

34. Brown, N., Jeger, M., Kirk, S., Xu, X. and Denman, S. (2016). Spatial and temporal patterns in symptom expression within eight woodlands affected by Acute Oak Decline. *Forest Ecology and Management*, *360*, 97-109.

35. Browning, S. R., & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, *81*, 1084–1097

36. Burger, R., & Lynch, M. (1995). Evolution and Extinction in a Changing Environment: A Quantitative-Genetic Analysis. *Evolution*, *49*(1), 151.

37. Buschiazzo, E., Ritland, C., Bohlmann, J. & Ritland, K. (2012). Slow but not low: genomic comparisons reveal slower evolutionary rate and higher dN/dS in conifers compared to angiosperms. *BMC Evolutionary Biology*, 12, 8.

38. Butlin, R. K., Bridle, J. R., Kawata, M. (2003). Genetics and the boundaries of species' distributions. In: Blackburn T, Gaston KJ, editors. Macroecology: Concepts and Consequences. Oxford: Blackwell Science, 274–295.

39. Butlin, R., Galindo, J., & Grahame, J. (2008). Sympatric, parapatric or allopatric: the most important way to classify speciation? *Philosophical Transactions Of The Royal Society B: Biological Sciences*, *363*(1506), 2997-3007.

40. Campbell, R. K. (1979). Genecology of Douglas-fir in a watershed in the Oregon cascades. *Ecology*, *60*, 1036–1050.

41. Cao, K., Zhou, Z., Wang, Q., Guo, J., Zhao, P., Zhu, G., Fang, W., Chen, C., Wang, X., Wang, X., Tian, Z. and Wang, L. (2016). Genome-wide association study of 12 agronomic traits in peach. *Nature Communications*, *7*, 13246.

42. Cao, X., Herzschuh, U., Ni, J., Zhao, Y., & Böhmer, T. (2015). Spatial and temporal distributions of major tree taxa in eastern continental Asia during the last 22,000 years. *The Holocene*, *25*(1), 79-91.

43. Casasoli, M., Mattioni, C., Cherubini, M. and Villani, F. (2001). A genetic linkage map of European chestnut (Castanea sativa Mill.) based on RAPD, ISSR and isozyme markers. *TAG Theoretical and Applied Genetics*, *102*(8), 1190-1199.

44. Case, T. J., & Taper, M. L. (2000). Interspecific competition, environmental gradients, gene flow, and the coevolution of species' borders. *American Naturalist*, *155*, 583–605.

45. Caye, K., Deist, T., Martins, H., Michel, O., & François, O. (2016). TESS3: fast inference of spatial population structure and genome scans for selection. *Molecular Ecology Resources*, *16*(2), 540-548.

46. Caye, K., Jumentier, B., Lepeule, J., & François, O. (2019). LFMM2: Fast and Accurate Inference of Gene-Environment Associations in Genome-Wide Studies. *Molecular Biology and Evolution*, *36*(4), 852-860.

47. Chang C. C., Chow, CC, Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015) Second-generation PLINK: rising to the challenge of larger and ricer datasets. *GigaScience*, *4*(1)

48. Chen, S., Wang, Y., Yu, L., Zheng, T., Wang, S., & Yue, Z. et al. (2021). Genome sequence and evolution of *Betula platyphylla*. *Horticulture Research*, *8*(1).

49. Chen, T., & Lou, A. (2019). Phylogeography and paleodistribution models of a widespread birch (*Betula platyphylla Suk.*) across East Asia: Multiple refugia, multidirectional expansion, and heterogeneous genetic pattern. *Ecology And Evolution*, *9*(13), 7792-7807.

50. Cheplick, G. P. (2015). Approaches to plant evolutionary ecology. Oxford University Press. 312.

51. Clark, P., Dyke, A., Shakun, J., Carlson, A., Clark, J., & Wohlfarth, B. et al. (2009). The Last Glacial Maximum. *Science*, *325*(5941), 710-714.

52. Coart, E., Lamote, V., De Loose, M., Van Bockstaele, E., Lootens, P. and Roldán-Ruiz, I. (2002). AFLP markers demonstrate local genetic differentiation between two indigenous oak species [*Quercus robur L.* and *Quercus petraea (Matt.) Liebl.*] in Flemish populations. *Theoretical and Applied Genetics*, *105* (2), 431-439.

53. Comes, H., & Kadereit, J. (1998). The effect of Quaternary climatic changes on plant distribution and evolution. *Trends In Plant Science*, *3* (11), 432-438.

54. Conesa, A., & Götz, S. (2008). Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics. *International Journal of Plant Genomics*, 1-12.

55. Cottrell, J., Munro, R., Tabbener, H., Gillies, A., Forrest, G., Deans, J., & Lowe, A. (2002). Distribution of chloroplast DNA variation in British oaks (Quercus robur and Q. petraea): the influence of postglacial colonisation and human management. *Forest Ecology and Management*, *156* (1-3), 181-195.

56. Cottrell, J., Samule, C., & Sykes, R. (2004). The species and chloroplast DNA haplotype composition of oakwoods in the Forest of Dean planted between 1720 and 1993. *Forestry*, *77* (2), 99-106.

57. Coyne, J., & Orr, H. (2004). Speciation. Sunderland, Mass.: Sinauer Associates.

58. Curtu, A., Gailing, O., & Finkeldey, R. (2007). Evidence for hybridization and introgression within a species-rich oak (*Quercus* spp.) community. *BMC Evolutionary Biology*, *7* (1), 218.

59. Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M., … Durbin, R. and 1000 Genomes Project Analysis Group. (2011). The Variant Call Format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158.

60. Dauphin, B., Rellstab, C., Schmid, M., Zoller, S., Karger, D., & Brodbeck, S., Guillaume, F., & Gugerli, F. (2020). Genomic vulnerability to rapid climate warming in a tree species with a long generation time. *Global Change Biology*, *27*(6), 1181-1195.

61. Davey, J., Hohenlohe, P., Etter, P., Boone, J., Catchen, J., & Blaxter, M. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, *12*(7), 499-510.

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

63. De Jong, PC. (1993=. An introduction to Betula: its morphology, evolution, classification and distribution, with a survey of recent work. In: D Hunt, ed. Proceedings of the IDS Betula symposium, 2–4 October 1992. International Dendrology Society. Richmond, UK.

64. DeHond, P., & Campbell, C. (1989). Multivariate analyses of hybridization between Betula cordifolia and B. populifolia (Betulaceae). *Canadian Journal Of Botany*, *67* (8), 2252-2260.

65. Denk T., Grimm G.W., Manos P.S., Deng M., Hipp A.L. (2017). An Updated Infrageneric Classification of the Oaks: Review of Previous Taxonomic Schemes and Synthesis of Evolutionary Patterns. In: Gil-Pelegrín E., Peguero-Pina J., Sancho-Knapik D. (eds) Oaks Physiological Ecology. Exploring the Functional Diversity of Genus Quercus L. *Tree Physiology*, vol 7. Springer, Cham.

66. Denman, S. and Webber, J. (2009). Oak declines: new definitions and new episodes in Britain. *Quarterly Journal of Forestry*, *103*(4), 285-290

67. Denman, S., Brady, C., Kirk, S., Cleenwerck, I., Venter, S., Coutinho, T. and De Vos, P. (2012). Brenneria goodwinii sp. nov., associated with acute oak decline in the UK. *International Journal of Systematic and Evolutionary Microbiology*, *62*(10), 2451-2456.

68. Denman, S., Brown, N., Kirk, S., Jeger, M. and Webber, J. (2014). A description of the symptoms of Acute Oak Decline in Britain and a comparative review on causes of similar disorders on oak in Europe. *Forestry*, *87*(4), 535-551.

69. Denman, S., Doonan, J., Ransom-Jones, E., Broberg, M., Plummer, S., Kirk, S., Scarlett, K., Griffiths, A., Kaczmarek, M., Forster, J., Peace, A., Golyshin, P., Hassard, F., Brown, N., Kenny, J. and McDonald, J. (2018). Microbiome and infectivity studies reveal complex polyspecies tree disease in Acute Oak Decline. *The ISME Journal*, *12*(2), 386-399.

70. Denman, S., Plummer, S., Kirk, S., Peace, A. and McDonald, J. (2016). Isolation studies reveal a shift in the cultivable microbiome of oak affected with Acute Oak Decline. *Systematic and Applied Microbiology*, *39*(7), 484-490.

71. DePristo, M. A., Banks, E., Poplin, R., Garimella, K., Maguire, J., Hartl, C., … Daly, M. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43* (5), 491-498.

72. Dierckxsens, N., Mardulyn, P., & Smits, G. (2017). NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Research*, *45* (4), 28.

73. Ding, J., Hua, D., Borrell, J., Buggs, R., Wang, L., & Wang, F. et al. (2021). Introgression between Betula tianshanica and Betula microphylla and its implications for conservation. *Plants, People, Planet*, 3 (4), 363-374.

74. Du, F., Wang, T., Wang, Y., Ueno, S., & Lafontaine, G. (2020). Contrasted patterns of local adaptation to climate change across the range of an evergreen oak, Quercus aquifolioides. *Evolutionary Applications*, *13*(9), 2377-2391.

75. Duan, R., Kong, X., Huang, M., Fan, W., & Wang, Z. (2014). The Predictive Performance and Stability of Six Species Distribution Models. *Plos ONE*, *9*(11), e112764.

76. Dumolin-Lapègue, S., Demesure, B., Fineschi, S., Le Corre, V., & Petit, R. (1997). Phylogeographic Structure of White Oaks Throughout the European Continent. *Genetics*, *146*, 1475-1487.

77. Durand, J., Bodénès, C., Chancerel, E., Frigerio, J., Vendramin, G., Sebastiani, F., Buonamici, A., Gailing, O., Koelewijn, H., Villani, F., Mattioni, C., Cherubini, M., Goicoechea, P., Herrán, A., Ikaran, Z., Cabané, C., Ueno, S., Alberto, F., Dumoulin, P., Guichoux, E., de Daruvar, A., Kremer, A. and Plomion, C. (2010). A fast and cost-effective approach to develop and map EST-SSR markers: oak as a case study. *BMC Genomics*, *11*(1):570.

78. Eaton, E., Caudullo, G., Oliveira, S., & de Rigo, D. (2016). Quercus robur and Quercus petraea in Europe: distribution, habitat, usage and threats. In: San-Miguel-Ayanz, J., de Rigo, D., Caudullo, G., Houston Durrant, T., Mauri, A. (Eds.), *European Atlas of Forest Tree Species* (pp. 160-163). Luxembourg.

79. Eckert, A., van Heerwaarden, J., Wegrzyn, J., Nelson, C., Ross-Ibarra, J., González-Martínez, S., & Neale, D. (2010). Patterns of Population Structure and Environmental Associations to Aridity Across the Range of Loblolly Pine (*Pinus taeda L*., *Pinaceae*). Genetics, *185* (*3*), 969-982.

80. Ekblom, R., & Galindo, J. (2010). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, *107*(1), 1-15.

81. Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends In Ecology & Evolution*, *29*(1), 51-63.

82. Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V., & Foll, M. (2013). Robust Demographic Inference from Genomic and SNP Data. *Plos Genetics*, *9* (10), e1003905.

83. Exposito-Alonso, M., Vasseur, F., Ding, W., Wang, G., Burbano, H., & Weigel, D. (2017). Genomic basis and evolutionary potential for extreme drought adaptation in Arabidopsis thaliana. *Nature Ecology & Evolution*, *2*(2), 352-358.

84. Finch, J., Brown, N., Beckmann, M., Denman, S., & Draper, J. (2021). Index measures for oak decline severity using phenotypic descriptors. *Forest Ecology And Management*, *485*, 118948.

85. Fisher, R. (1930). The genetical theory of natural selection. Oxford: Oxford University Press.

86. Fitzpatrick, M. C., & Keller, S. R. (2015). Ecological genomics meets community-level modelling of biodiversity: Mapping the genomic landscape of current and future environmental adaptation. *Ecology Letters*, *18*(1), 1–16.

87. Flanagan, S., & Jones, A. (2017). Constraints on the F_{ST}–Heterozygosity Outlier Approach. *Journal of Heredity*, *108* (5), 561-573.

88. Fogelqvist, J., Verkhozina, A. V., Katyshev, A. I., Pucholt, P., Dixelius, C., Rönnberg-Wästljung, A. C., Lascoux, M., & Berlin, S. (2015). Genetic and morphological evidence for introgression between three species of willows. *BMC Evolutionary Biology*, *15*, 193.

89. Foll, M., & Gaggiotti, O. (2008). A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. *Genetics*, *180*(2), 977-993.

90. Frichot, E., Schoville, S., Bouchard, G., & François, O. (2013). Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models. *Molecular Biology and Evolution*, *30*(7), 1687-1699.

91. Gain, C., & François, O. (2021). LEA 3: Factor models in population genetics and ecological genomics with R. *Molecular Ecology Resources*.

92. Garcia-Ramos, G., & Kirkpatrick, M. (1997). Genetic models of adaptation and gene flow in peripheral populations. *Evolution*, *51*, 21–28.

93. Gathercole, L., Nocchi, G., Brown, N., Coker, T., Plumb, W., Stocks, J., Nichols, R., Denman, S., Buggs, R. (2021). Evidence for the widespread occurrence of bacteria implicated in acute oak decline from incidental genetic sampling. *Forests*, 12, 1683.

94. Gijzen, M., Miller, S., Kuflu, K., Buzzell, R., & Miki, B. (1999). Hydrophobic Protein Synthesized in the Pod Endocarp Adheres to the Seed Surface. *Plant Physiology*, *120*(4), 951-960.

95. Goczał, J., Oleksa, A., Rossa, R., Chybicki, I., Meyza, K., & Plewa, R. et al. (2020). Climatic oscillations in Quaternary have shaped the co-evolutionary patterns between the Norway spruce and its host-associated herbivore. *Scientific Reports*, *10*(1).

96. Golldack, D., & Dietz, K.-J. (2001). Salt-Induced Expression of the Vacuolar H+-ATPase in the Common Ice Plant Is Developmentally Controlled and Tissue Specific. *Plant Physiology*, *125*(4), 1643–1654.

97. Gómez, J., González-Megías, A., Lorite, J., Abdelaziz, M., & Perfectti, F. (2015). The silent extinction: climate change and the potential hybridization-mediated extinction of endemic high-mountain plants. *Biodiversity And Conservation*, *24*(8), 1843-1857.

98. Gugger, P., Fitz-Gibbon, S., Albarrán-Lara, A., Wright, J., & Sork, V. (2020). Landscape genomics of Quercus lobata reveals genes involved in local climate adaptation at multiple spatial scales. *Molecular Ecology*, *30*(2), 406-423.

99. Guichoux, E., Garnier-Géré, P., Lagache, L., Lang, T., Boury, C., & Petit, R. (2012). Outlier loci highlight the direction of introgression in oaks. *Molecular Ecology*, *22*(2), 450-462.

100. Guichoux, E., Lagache, L., Wagner, S., Léger, P., & Petit, R. (2011). Two highly validated multiplexes (12-plex and 8-plex) for species delimitation and parentage analysis in oaks (*Quercus* spp.). *Molecular Ecology Resources*, 11(3), 578-585.

101. Günther, T., & Coop, G. (2013). Robust Identification of Local Adaptation from Allele Frequencies. *Genetics*, *195*(1), 205-220.

102. Hallatschek, O., Hersen, P., Ramanathan, S., & Nelson, D. (2007). Genetic drift at expanding frontiers promotes gene segregation. *Proceedings Of The National Academy Of Sciences*, *104* (50), 19926-19930.

103. Hällfors, M., Vaara, E., Hyvärinen, M., Oksanen, M., Schulman, L., Siipi, H., & Lehvävirta, S. (2014). Coming to Terms with the Concept of Moving Species Threatened by Climate Change – A Systematic Review of the Terminology and Definitions. *Plos ONE*, *9*(7), e102979.

104. Hamilton, J., & Miller, J. (2015). Adaptive introgression as a resource for management and genetic conservation in a changing climate. *Conservation Biology*, *30*(1), 33-41.

105. Han, G., Lu, C., Guo, J., Qiao, Z., Sui, N., Qiu, N., & Wang, B. (2020). C2H2 Zinc Finger Proteins: Master Regulators of Abiotic Stress Responses in Plants. *Frontiers in Plant Science*, *11*.

106. Harper, A., McKinney, L., Nielsen, L., Havlickova, L., Li, Y., Trick, M., Fraser, F., Wang, L., Fellgett, A., Sollars, E., Janacek, S., Downie, J., Buggs, R., Kjær, E. and Bancroft, I. (2016). Molecular markers for tolerance of European ash (Fraxinus excelsior) to dieback disease identified using Associative Transcriptomics. *Scientific Reports*, *6*(1), 19335.

107. Harrison, R., & Larson, E. (2014). Hybridization, Introgression, and the Nature of Species Boundaries. Journal Of Heredity, 105(S1), 795-809.

108. Hewitt, G. (1996). Some genetic consequences of ice ages, and their role in divergence and speciation. *Biological Journal Of The Linnean Society*, *58*(3), 247-276.

109. Hewitt, G. (1999). Post-glacial re-colonization of European biota. *Biological Journal Of The Linnean Society*, *68*(1-2), 87-112.

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

111. Hijmans, R. J., & Etten, J. (2012). raster: Geographic analysis and modelling with raster data.

112. Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high-resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, *25*, 1965–1978.

113. Hipp, A., Manos, P., Hahn, M., Avishai, M., Bodénès, C., Cavender-Bares, J., … Valencia-Avalos, S. (2019). Genomic landscape of the global oak phylogeny. *New Phytologist*, *226* (4), 1198-1212.

114. Hoban, S., Kelley, J., Lotterhos, K., Antolin, M., Bradburd, G., & Lowry, D. et al. (2016). Finding the Genomic Basis of Local Adaptation: Pitfalls, Practical Solutions, and Future Directions. *The American Naturalist*, *188*(4), 379-397.

115. Holderegger, R., & Thiel-Egenter, C. (2009). A discussion of different types of glacial refugia used in mountain biogeography and phylogeography. *Journal Of Biogeography*, *36* (3), 476-480.

116. Horton, M., Bodenhausen, N., Beilsmith, K., Meng, D., Muegge, B., Subramanian, S., Vetter, M., Vilhjálmsson, B., Nordborg, M., Gordon, J. and Bergelson, J. (2014). Genome-wide association study of Arabidopsis thaliana leaf microbial community. *Nature Communications*, *5*(1), 5320.

117. Hu, Y., Zhao, L., Buggs, R., Zhang, X., Li, J., & Wang, N. (2019). Population structure of Betula albosinensis and Betula platyphylla: evidence for hybridization and a cryptic lineage. *Annals Of Botany*, *123* (7), 1179-1189.

118. Huang X., Wei X., Sang T., Zhao Q., Feng Q., … Han, B. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature Genetics*, *42*, 961–67

119. Huang, X. and Han, B. (2014). Natural Variations and Genome-Wide Association Studies in Crop Plants. *Annual Review of Plant Biology*, *65*(1), pp.531-551.

120. Ibrahim, K., Nichols, R., & Hewitt, G. (1996). Spatial patterns of genetic variation generated by different forms of dispersal during range expansion. *Heredity*, *77* (3), 282-291.

121. Jambunathan, N., & McNellis, T. W. (2003). Regulation of Arabidopsis COPINE 1 Gene Expression in Response to Pathogens and Abiotic Stimuli. *Plant Physiology*, *132*(3), 1370–1381.

122. Jensen, J., Larsen, A., Nielsen, L., & Cottrell, J. (2009). Hybridization between Quercus robur and Q. petraea in a mixed oak stand in Denmark. *Annals Of Forest Science*, *66*(7), 706-706.

123. Jiang, D., Feng, J., Dong, M., Wu, G., Mao, K., & Liu J. (2016). Genetic origin and composition of a natural hybrid poplar Populus x jrtyschensis from two distantly related species. *BMC Plant Biology*, *16*, 89.

124. Johnson, L., Galliart, M., Alsdurf, J., Maricle, B., Baer, S., & Bello, N. et al. (2021). Reciprocal transplant gardens as gold standard to detect local adaptation in grassland species: New opportunities moving into the 21st century. *Journal Of Ecology*.

125. Jones, E. W. (1959). Biological Flora of the British Isles. *The Journal of Ecology*. *47*(1), 169-222.

126. Joost, S., Bonin, A., Bruford, M. W., Després, L., Conord, C., Erhardt, G., & Taberlet, P. (2007). A spatial analysis method (SAM) to detect candidate loci for selection: Towards a landscape genomics approach to adaptation. Molecular Ecology, 16, 3955–3969.

127. Jordan, R., Hoffmann, A., Dillon, S., & Prober, S. (2017). Evidence of genomic adaptation to climate in Eucalyptus microcarpa: Implications for adaptive potential to projected climate change. *Molecular Ecology*, *26*(21), 6002-6020.

128. Jump, A. S., Hunt, J. M., Martinez-Izquierdo, J. A., & Peñuelas, J. (2006). Natural selection and climate change: Temperature-linked spatial and temporal trends in gene frequency in Fagus sylvatica. *Molecular Ecology*, *15*, 3469–3480.

129. Jurkšienė, G., Baranov, O., Kagan, D., Kovalevič-Razumova, O., & Baliuckas, V. (2019). Genetic diversity and differentiation of pedunculate (*Quercus robur*) and sessile (*Q. petraea*) oaks. *Journal Of Forestry Research*, *31*(6), 2445-2452.

130. Karlsdóttir, L., Hallsdóttir, M., Thórsson, Æ., & Anamthawat-Jónsson, K. (2009). Evidence of hybridisation between Betula pubescens and B. nana in Iceland during the early Holocene. *Review Of Palaeobotany And Palynology*, *156* (3-4), 350-357.

131. Katoh, K, & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30* (4), 772-780.

132. Keenan, R. (2015). Climate change impacts and adaptation in forest management: a review. *Annals Of Forest Science*, *72*(2), 145-167.

133. Kim, Y., & Nielsen, R. (2004). Linkage Disequilibrium as a Signature of Selective Sweeps. *Genetics*, *167* (3), 1513-1524.

134. Kleinschmit, J. (1993). Intraspecific variation of growth and adaptive traits in European oak species. Annales des Sciences *Forestières*, 50(Supplement), 166-185.

135. Kremer, A., Abbott, A., Carlson, J., Manos, P., Plomion, C., Sisco, P., Staton, M., Ueno, S. and Vendramin, G. (2012). Genomics of Fagaceae. *Tree Genetics & Genomes*, *8*(3), 583-610.

136. Kremer, A., Dupouey, J., Deans, J., Cottrell, J., Csaikl, U., Finkeldey, R., … & Badeau, V. (2002). Leaf morphological differentiation between Quercus robur and Quercus petraea is stable across western European mixed oak stands. *Annals Of Forest Science*, *59*(7), 777-787.

137. Lang, T., Abadie, P., Léger, V., Decourcelle, T., Frigerio, J., Burban, C., … Garnier-Géré, P. (2018). High-quality SNPs from genic regions highlight introgression patterns among European white oaks (*Quercus petraea* and *Q. robur*). *bioRxiv*

138. Leigh, J. W., & Bryant, D. (2015). PopART: Full-feature software for haplotype network construction. *Methods in Ecology and Evolution*, *6* (9), 1110–1116.

139. Lepais, O., Petit, R., Guichoux, E., Lavabre, J., Alberto, F., Kremer, A. and Gerber, S. (2009). Species relative abundance and direction of introgression in oaks. *Molecular Ecology*, *18*(10), 2228-2242.

140. Lepais, O., Roussel, G., Hubert, F., Kremer, A., & Gerber, S. (2013). Strength and variability of postmating reproductive isolating barriers between four European white oak species. *Tree Genetics & Genomes*, *9*(3), 841-853.

141. Lepoittevin, C., Bodénès, C., Chancerel, E., Villate, L., Lang, T., Lesur, I., Boury, C., Ehrenmann, F., Zelenica, D., Boland, A., Besse, C., Garnier-Géré, P., Plomion, C. and Kremer, A. (2015). Single-nucleotide polymorphism discovery and validation in high-density SNP

array for genetic analysis in European white oaks. *Molecular Ecology Resources*, *15*(6), 1446-1459.

142. Leroy, T., Louvet, J., Lalanne, C., Le Provost, G., Labadie, K., & Aury, J., … Kremer, A. (2019a). Adaptive introgression as a driver of local adaptation to climate in European white oaks. *New Phytologist*, *226* (4), 1171-1182.

143. Leroy, T., Rougemont, Q., Dupouey, J., Bodénès, C., Lalanne, C., Belser, C., … Plomion, C. (2019b). Massive postglacial gene flow between European white oaks uncovered genes underlying species barriers. *New Phytologist*, *226 (*4), 1183-1197.

144. Leroy, T., Roux, C., Villate, L., Bodénès, C., Romiguier, J., Paiva, J., Dossat, C., Aury, J., Plomion, C., & Kremer, A. (2017). Extensive recent secondary contacts between four European white oak species. *New Phytologist*, *214* (2), 865-878.

145. Lesur, I., Alexandre, H., Boury, C., Chancerel, E., Plomion, C., & Kremer, A. (2018). Development of Target Sequence Capture and Estimation of Genomic Relatedness in a Mixed Oak Stand. *Frontiers In Plant Science*, *9*.

146. Lesur, I., Le Provost, G., Bento, P., Da Silva, C., Leplé, J., Murat, F., Ueno, S., Bartholomé, J., Lalanne, C., Ehrenmann, F., Noirot, C., Burban, C., Léger, V., Amselem, J., Belser, C., Quesneville, H., Stierschneider, M., Fluch, S., Feldhahn, L., Tarkka, M., Herrmann, S., Buscot, F., Klopp, C., Kremer, A., Salse, J., Aury, J. and Plomion, C. (2015). The oak gene expression atlas: insights into Fagaceae genome evolution and the discovery of genes regulated during bud dormancy release. *BMC Genomics*, *16*(1), 112.

147. Levin, D. (2002). Hybridization and Extinction. *American Scientist*, *90*(3), 254.

148. Levin, D., Francisco-Ortega, J., & Jansen, R. (1996). Hybridization and the Extinction of Rare Plant Species. *Conservation Biology*, *10* (1), 10-16.

149. Levy G, Becker M, Duhamel D. (1992) A comparison of the ecology of pedunculate and sessile oaks: radial growth in the centre and Northwest of France. Forest Ecology and Management, 55, 51–63.

150. Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25* (14), 1754-1760.

151. Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, *475*, 493–496.

152. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25* (16), 2078-2079.

153. Liu, C., Newell, G., & White, M. (2016). On the selection of thresholds for predicting species occurrence with presence-only data. *Ecology and Evolution*, *6*, 337–348.

154. Liu, K. (1988). Quaternary history of the temperate forests of China. *Quaternary Science Reviews*, *7* (1), 1-20.

155. Lotterhos, K. E., & Whitlock, M. C. (2014). Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Molecular Ecology*, *23*(9), 2178–2192.

156. Loudon JC (1838). Arboretum et Fruticetum Brittanicum. Printed by A. Spottiswoode, London, 1717-1949

157. Lovell, J., MacQueen, A., Mamidi, S., Bonnette, J., Jenkins, J., & Napier, J., …, & Schmutz, J. (2021). Genomic mechanisms of climate adaptation in polyploid bioenergy switchgrass. *Nature*, *590*(7846), 438-444.

158. Lowe, A., Munro, R., Samule, S., & Cottrell, J. (2004). The utility and limitations of chloroplast DNA analysis for identifying native British oak stands and for guiding replanting strategy. *Forestry*, *77* (4), 335-347.

159. Lynch, M. & Lande, R. (1993). Evolution and extinction in response to environmental change. In: Kareiva PM, Kingsolver JG, Huey RB, editors. Biotic Interactions and Global Change. Sunderland, MA: Sinauer, 234–250.

160. Mallet, J. (2005). Hybridization as an invasion of the genome. *Trends In Ecology & Evolution*, *20*(5), 229-237.

161. Manel, S., & Holderegger, R. (2013). Ten years of landscape genetics. *Trends In Ecology & Evolution*, *28*(10), 614-621.

162. Manzanedo, R., Fischer, M., María Navarro-Cerrillo, R., & Allan, E. (2019). A new approach to study local adaptation in long-lived woody species: Virtual transplant experiments. *Methods In Ecology And Evolution*, 10(10), 1761-1772.

163. Mariette, S., Cottrell, J., Csaikl, U. M., Goikoechea, P., Konig, A., Lowe, A. J., ... Kremer, A. (2002). Comparison of levels of genetic diversity detected with AFLP and microsatellite markers within and among mixed Q. petraea (MATT.) LIEBL. and Q. robur L. stands. *Silvae genetica*, *51* (2/3), 72-79.

164. Martins, H., Caye, K., Luu, K., Blum, M. G. B., & François, O. (2016). Identifying outlier loci in admixed and in continuous populations using ancestral population differentiation statistics. *Molecular Ecology*, *25*(20), 5029–5042.

165. Martins, K., Gugger, P., Llanderal-Mendoza, J., González-Rodríguez, A., Fitz-Gibbon, S., & Zhao, J. et al. (2018). Landscape genomics provides evidence of climate-associated genetic variation in Mexican populations of *Quercus rugosa*. *Evolutionary Applications*, *11(*10), 1842-1858.

166. Masson-Delmotte, V., P. Zhai, A. Pirani, S.L., Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R., Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and Zhou, B. (2021). IPCC, 2021: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press. In Press.

167. Mazet, O., Rodríguez, W., Grusea, S., Boitard, S., & Chikhi, L. (2015). On the importance of being structured: instantaneous coalescence rates and human evolution--lessons for ancestral population size inference? *Heredity*, *116* (4), 362-371.

168. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., & Kernytsky, A., …, & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analysing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297-1303.

169. Meyer, H. V. (2020) plinkQC: Genotype quality control in genetic association studies.

170. Mimura, M., Aitken, S.N. (2007). Adaptive gradients and isolation-by-distance with postglacial migration in *Picea sitchensis*. *Heredity*, *99*, 22–24.

171. Minamikawa, M., Takada, N., Terakami, S., Saito, T., Onogi, A., Kajiya-Kanegae, H., Hayashi, T., Yamamoto, T. and Iwata, H. (2018). Genome-wide association study and genomic prediction using parental and breeding populations of Japanese pear (Pyrus pyrifolia Nakai). *Scientific Reports*, *8*(1), 11994.

172. Mishra, B., Gupta, D., Pfenninger, M., Hickler, T., Langer, E., & Nam, B., … Thines, M. (2018). A reference genome of the European beech (Fagus sylvatica L.). *Gigascience*, *7* (6).

173. Miyamoto, T., Uemura, T., Nemoto, K., Daito, M., Nozawa, A., Sawasaki, T., & Arimura, G. (2019). Tyrosine Kinase-Dependent Defense Responses Against Herbivory in Arabidopsis. *Frontiers in Plant Science*, *10*.

174. Montesinos-Navarro, A., Wig, J., Pico, F. X., & Tonsor, S. J. (2011). Arabidopsis thaliana populations show clinal variation in a climatic gradient associated with altitude. *New Phytologist*, *189*(1), 282–294.

175. Nagamitsu, T., Kawahara, T., & Kanazashi, A. (2006). Endemic dwarf birch Betula apoiensis (Betulaceae) is a hybrid that originated from Betula ermanii and Betula ovalifolia. *Plant Species Biology*, *21* (1), 19-29.

176. Nagaraj, S., Gasser, R. and Ranganathan, S. (2006). A hitchhiker's guide to expressed sequence tag (EST) analysis. *Briefings in Bioinformatics*, *8*(1), 6-21.

177. Nei, M., & Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, *76* (10), 5269–5273.

178. Nicotra, A., Atkin, O., Bonser, S., Davidson, A., Finnegan, E., & Mathesius, U. et al. (2010). Plant phenotypic plasticity in a changing climate. *Trends In Plant Science*, *15*(12), 684-692.

179. Nosil, P., Funk, D., & Ortiz-Barrientos, D. (2009). Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, *18*(3), 375-402.

180. Padmanaban, S., Lin, X., Perera, I., Kawamura, Y., & Sze, H. (2004). Differential Expression of Vacuolar H+-ATPase Subunit c Genes in Tissues Active in Membrane Trafficking and Their Roles in Plant Growth as Revealed by RNAi. *Plant Physiology*, *134*(4), 1514–1526.

181. Palme, A., Su, Q., Palsson, S., & Lascoux, M. (2004). Extensive sharing of chloroplast haplotypes among European birches indicates hybridization among Betula pendula, B. pubescens and B. nana. *Molecular Ecology*, *13* (1), 167-178.

182. Parelle, J., Zapater, M., Scotti-Saintagne, C., Kremer, A., Jolivet, Y., Dreyer, E., Brendel, O. (2007). Quantitative trait loci of tolerance to waterlogging in a European oak

(Quercus robur L.): physiological relevance and temporal effect patterns. *Plant Cell Environ*, *30*, 422–434

183. Parmesan, C. (2006). Ecological and Evolutionary Responses to Recent Climate Change. *Annual Review Of Ecology, Evolution, And Systematics*, *37*(1), 637-669.

184. Pavlidis, P., Živković, D., Stamatakis, A., & Alachiotis, N. (2013). SweeD: Likelihood-Based Detection of Selective Sweeps in Thousands of Genomes. *Molecular Biology and Evolution*, *30* (9), 2224-2234.

185. Pebesma, E. (2006). The Role of External Variables and GIS Databases in Geostatistical Analysis. *Transactions In GIS*, *10*(4), 615-632.

186. Peiro, A., Izquierdo-Garcia, A. C., Sanchez-Navarro, J. A., Pallas, V., Mulet, J. M., & Aparicio, F. (2014). Patellins 3 and 6, two members of the P lant P atellin family, interact with the movement protein of A lfalfa mosaic virus and interfere with viral movement. *Molecular Plant Pathology*, *15*(9), 881–891.

187. Petit, R. J., Bialozyt, R., Garnier-Gere, P., Hampe, A. (2004a). Ecology and genetics of tree invasions: from recent introductions to Quaternary migrations. *Forest Ecology and Management*, *197*, 117–137.

188. Petit, R. J., Bodenes, C., Ducousso, A., Roussel, G., & Kremer, A. (2004b). Hybridization as a mechanism of invasion in oaks. *The New Phytologist*, *161* (1), 151-164.

189. Petit, R. J., Brewer, S., Bordács, S., Burg, K., Cheddadi, R., Coart, E., … Kremer, A. (2002a). 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* (1-3), 49-74.

190. Petit, R. J., Csaikl, U. M., Bordács, S., Burg, K., Coart, E., Cottrell, J., … Kremer, A. (2002b). Chloroplast DNA variation in European white oaks. *Forest Ecology and Management*, *156* (1–3), 5–26.

191. Petit, R. J., Demesure, B., Dumolin, S. (1998) cpDNA and mtDNA Primers in Plants. In: Karp A., Isaac P.G., Ingram D.S. (Eds.), Molecular Tools for Screening Biodiversity Molecular Tools for Screening Biodiversity (pp. 256-261). Dordrecht: Springer.

192. Petit, R. J., Pineau, E., Demesure, B., Bacilieri, R., Ducousso, A., & Kremer, A. (1997). Chloroplast DNA footprints of postglacial recolonization by oaks. *Proceedings of the National Academy of Sciences of the United States of America*, *94*, 9996-10001.

193. Petit, R., Aguinagalde, I., de Beaulieu, J., Bittkau, C., Brewer, S., & Cheddadi, R. … Vendramin, G. G. (2003). Glacial Refugia: Hotspots But Not Melting Pots of Genetic Diversity. *Science*, *300* (5625), 1563-1565.

194. Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modelling of species geographic distributions*. Ecol. Modell.*, *190*, 231–259.

195. Pina-Martins, F., Baptista, J., Pappas, G., & Paulo, O. (2018). New insights into adaptation and population structure of cork oak using genotyping by sequencing. *Global Change Biology*, *25*(1), 337-350.

196. Plomion, C. and Fievet, V. (2013). Oak genomics takes off … and enters the ecological genomics era. *New Phytologist*, *199*(2), 308-310.

197. Plomion, C., Aury, J., Amselem, J., Alaeitabar, T., Barbe, V., Belser, C., Bergès, H., Bodénès, C., Boudet, N., Boury, C., Canaguier, A., Couloux, A., Da Silva, C., Duplessis, S., Ehrenmann, F., Estrada-Mairey, B., Fouteau, S., Francillonne, N., Gaspin, C., Guichard, C., Klopp, C., Labadie, K., Lalanne, C., Le Clainche, I., Leplé, J., Le Provost, G., Leroy, T., Lesur, I., Martin, F., Mercier, J., Michotey, C., Murat, F., Salin, F., Steinbach, D., Faivre-Rampant, P., Wincker, P., Salse, J., Quesneville, H. and Kremer, A. (2016). Decoding the oak genome: public release of sequence data, assembly, annotation and publication strategies. *Molecular Ecology Resources*, *16*(1), 254-265.

198. Plomion, C., Aury, J., Amselem, J., Leroy, T., Murat, S., Duplessis, S., Faye, S., Francillonne, N., Labadie, K., Le Provost, G., Lesur, I., Bartholome, J., Faivre-Rampant, P., Kohler, A., Leple, J., Chantret, N., Chen, J., Dievart, A., Alaeitabar, T., Barbe, V., Belser, C., Berges, H., Bodénès, C., Bogeat-Triboulout, m., Bouffaud, M., Brachi, B., Chancerel, E., Cohen, D., Couloux, A., Da Silva, C., Dossat, C., Ehrenmann, F., Gaspin, C., Grima-Pettenati, J., Guichoux, E., Hecker, A., Herrmann, S., Hugueney, P., Hummel, I., Klopp, C., Lalanne, C., Lascoux, M., Lasserre, E., Lemainque, A., Desprez-Loustau, M., Luyten, I., Madoui, M., Mangenot, S., Marchal, C., Maumus, F., Mercier, J., Michotey, C., Parnaud, O., Picault, N., Rouhier, N., Rue, O., Rustenholz, C., Salin, F., Soler, M., Tarkka, M., Velt, A., Zanne, A., Martin, F., Wincker, P., Quersneville, H., Kremer, A. and Salse, J. (2018). Oak genome reveals facets of long lifespan. *Nature Plants*, *4*, 440-452.

199. Prieto-Benítez, S., Morente-López, J., Rubio Teso, M., Lara-Romero, C., García-Fernández, A., Torres, E., & Iriondo, J. (2021). Evaluating Assisted Gene Flow in Marginal Populations of a High Mountain Species. *Frontiers In Ecology And Evolution*, *9*.

200. Provan, J., & Bennett, K. (2008). Phylogeographic insights into cryptic glacial refugia. *Trends In Ecology & Evolution*, *23* (10), 564-571.

201. Qiu, Y., Fu, C., & Comes, H. (2011). Plant molecular phylogeography in China and adjacent regions: Tracing the genetic imprints of Quaternary climate and environmental change in the world's most diverse temperate flora. *Molecular Phylogenetics And Evolution*, *59* (1), 225-244.

202. Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26* (6), 841-842.

203. Rackham, O. (1998) Trees and Woodland in a Cultural Landscape: The History of Woods in England. In: Sassa K. (eds) Environmental Forest Science. *Forestry Sciences*, 54. Springer, Dordrecht.

204. Raimbault, P. (1995) Physiological Diagnosis. *The proceedings of the second European congress in arboriculture*, *Societe Francaise d'Arboriculture*.

205. Raj, A., Stephens, M., & Pritchard, J. K. (2014). fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. *Genetics*, *197* (2), 573-589.

206. Ramos, A. M, Usie, A., Barbosa, P., Barros, P. M., Capote, T., Chaves, I., Simoes, F., Abreu, I., Carrasquinho, I., Faro, C., Guimaraes, J. B., Mendonca, D., Nobrega, F., Rodrigues,

L., Saibo, N. J. M., Varela, M. C., Egas, C., Matos, J., Miguel, C., M., Oliveira, M. M., Ricardo, C. P., Goncalves, S. (2018). The draft genome sequence of cork oak. *Scientific Data 5*, 180069

207. Ratajczak, R. (2000). Structure, function and regulation of the plant vacuolar H+-translocating ATPase. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, *1465*(1–2), 17–36.

208. Raymo, M. E. (1994). The Initiation of Northern Hemisphere Glaciation. *Annual Review of Earth Planet Sciences*, *22*, 353-383.

209. Regel, E. (1865). Bemerkungen über die Gattungen Betula und Alnus nebst Beschreibung einiger neuer Arten. Bulletin of Society of Naturalist (Moscou)38: 388–434.

210. Rehfeldt, G. E. (1995). Genetic-variation, climate models and the ecological genetics of Larix Occidentalis. *Forest Ecology and Management*, *78*, 21–37.

211. Rellstab, C., Bühler, A., Graf, R., Folly, C., & Gugerli, F. (2016b). Using joint multivariate analyses of leaf morphology and molecular-genetic markers for taxon identification in three hybridizing European white oak species (*Quercus* spp.). *Annals Of Forest Science*, *73*(3), 669-679.

212. Rellstab, C., Dauphin, B., & Exposito-Alonso, M. (2021). Prospects and limitations of genomic offset in conservation management. *Evolutionary Applications*, *14*(5), 1202-1212.

213. Rellstab, C., Gugerli, F., Eckert, A., Hancock, A., & Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology*, *24*(17), 4348-4370.

214. Rellstab, C., Zoller, S., Walthert, L., Lesur, I., Pluess, A., & Graf, R. et al. (2016). Signatures of local adaptation in candidate genes of oaks (*Quercus* spp.) with respect to present and future climatic conditions. *Molecular Ecology*, *25*(23), 5907-5924.

215. Reutimann, O., Gugerli, F., & Rellstab, C. (2020). A species-discriminatory single-nucleotide polymorphism set reveals maintenance of species integrity in hybridizing European white oaks (*Quercus* spp.) despite high levels of admixture. *Annals Of Botany*, *125*(4), 663-676.

216. Ricciardi, A., and Simberloff, D. (2009). Assisted colonization is not a viable conservation strategy. *Trends Ecol. Evol*, *24*, 248–253.

217. Richard, F., Millot, S., Gardes, M. and Selosse, M. (2005). Diversity and specificity of ectomycorrhizal fungi retrieved from an old-growth Mediterranean forest dominated by Quercus ilex. *New Phytologist*, *166*(3), 1011-1023

218. Rieseberg, L., Raymond, O., Rosenthal, D., Lai, Z., Livingstone, K., & Nakazato, T. et al. (2003). Major Ecological Transitions in Wild Sunflowers Facilitated by Hybridization. *Science*, *301*(5637), 1211-1216.

219. Roux, C., Fraïsse, C., Romiguier, J., Anciaux, Y., Galtier, N., & Bierne, N. (2016). Shedding Light on the Grey Zone of Speciation along a Continuum of Genomic Divergence. *PLOS Biology*, *14*(12), e2000234.

220. Saintagne, C., Bodénès, C., Barreneche, T., Pot, D., Plomion, C. and Kremer, A. (2004). Distribution of genomic regions differentiating oak species assessed by QTL detection. *Heredity*, *92* (1), 20-30.

221. Salojärvi, J., Smolander, OP., Nieminen, K., Rajaraman, S., Safronov, O., Safdari, P., … Kangasjarvi, J. (2017). Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch. *Nature Genetics*, *49*, 904–912.

222. Sami, A., Arabia, S., Sarker, R., & Islam, T. (2021). Deciphering the role of helicases and translocases: A multifunctional gene family safeguarding plants from diverse environmental adversities. *Current Plant Biology*, *26*, 100204.

223. Savolainen, O., Pyhäjärvi, T., & Knürr, T. (2007). Gene Flow and Local Adaptation in Trees. *Annual Review Of Ecology, Evolution, And Systematics*, *38*(1), 595-619.

224. Schenk, M., Thienpont, C., Koopman, W., Gilissen, L., & Smulders, M. (2008). Phylogenetic relationships in Betula (Betulaceae) based on AFLP markers. *Tree Genetics & Genomes*, *4* (4).

225. Schiffels, S., & Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 46, 919–925

226. Scotti-Saintagne, C., Mariette, S., Porth, I., Goicoechea, P., Barreneche, T., Bodénès, C., Burg, K., & Kremer, A. (2004). Genome Scanning for Interspecific Differentiation Between Two Closely Related Oak Species [Quercus robur L. and Q. petraea (Matt.) Liebl.]. *Genetics*, *168* (3), 1615-1626. https://doi.org/10.1534/genetics.104.026849

227. Sharma, T., Dreyer, I., & Riedelsberger, J. (2013). The role of K+ channels in uptake and redistribution of potassium in the model plant Arabidopsis thaliana. Frontiers In Plant Science, 4.

228. Shaw, K., Stritch, L., Rivers, M., Roy, S., Wilson, B., Govaerts R. (2014). The red list of Betulaceae. BGCI. Richmond. UK.

229. Shi, Y. (1986). Quaternary glaciation in China. *Quaternary Science Reviews*, *5*, 503-507.

230. Shrestha, A., Dziwornu, A., Ueda, Y., Wu, L., Mathew, B. and Frei, M. (2018). Genome-wide association study to identify candidate loci and genes for Mn toxicity tolerance in rice. *Plos One*, *13*(2).

231. Skvortsov, A. K. (2002). A new system of the genus Betula L. – the birch. *Bulletin of Moscow Society of Naturalist*, *107*, 73–76.

232. Sofiev, M., Siljamo, P., Ranta, H., & Rantio-Lehtimäki, A. (2006). Towards numerical forecasting of long-range air transport of birch pollen: theoretical considerations and a feasibility study. *International Journal Of Biometeorology*, *50*(6), 392-402.

233. Sollars, E. S. A., Harper, A. L., Kelly, L. J., Sambles, C. M., Ramirez-Gonzalez, R. H., Swarbreck, D., Kaithakottil, G., Cooper, E. D., Uauy, C., Havlickova, L., Worswick, G., Studholme, D. J., Zohren, J., Salmon, D., L., Clavijo, B. J., Li, Y., He, Z., Fellgett, A., McKinney, L. V., Nielsen, L. R., Douglas, G. C., Kjær, E. D., Downie, J. A., Boshier, D., Lee, S., Clark, J.,

Grant, M., Bancroft, I., Caccamo, M. and Buggs, R. J. A. (2017) Genome sequence and genetic diversity of European ash trees. *Nature, 541*, 212–216

234. Sork, V., Aitken, S., Dyer, R., Eckert, A., Legendre, P., & Neale, D. (2013). Putting the landscape into the genomics of trees: approaches for understanding local adaptation and population responses to changing climate. *Tree Genetics & Genomes*, *9*(4), 901-911.

235. Sork, V., Fitz-Gibbon, S., Puiu, D., Crepeau, M., Gugger, P., Sherman, R., Stevens, K., Langley, C., Pellegrini, M., & Salzberg, S. (2016). First Draft Assembly and Annotation of the Genome of a California Endemic Oak Quercus lobata Née (Fagaceae). *G3 (Bethesda)*, *6* (11), 3485-3495.

236. St Clair, J. B., Mandel, N. L., Vance-Boland, K. W. (2005). Genecology of Douglas-fir in western Oregon and Washington. *Annals Of Botany*, *96*, 1199–1214.

237. Stocks, J., Metheringham, C., Plumb, W., Lee, S., Kelly, L., Nichols, R., & Buggs, R. (2019). Genomic basis of European ash tree resistance to ash dieback fungus. *Nature Ecology & Evolution*, *3*(12), 1686-1696.

238. Storey, J. D., Bass, A. J., Dabney, A., & Robinson, D. (2015). qvalue: Q-value estimation for false discovery rate control. R package version 2.4.2.

239. Storey, J., & Tibshirani, R. (2003). Statistical significance for genome wide studies. *Proceedings Of the National Academy Of Sciences*, *100*(16), 9440-9445.

240. Stucki, S., Orozco-terWengel, P., Forester, B. R., Duruz, S., Colli, L., Masembe, C.…Consortium, N. (2017). High performance computation of landscape genomic models including local indicators of spatial association. *Molecular Ecology Resources*, *17*(5), 1072–1089.

241. Stucki, S., Orozco-terWengel, P., Forester, B. R., Duruz, S., Colli, L., Masembe, C.…Consortium, N. (2017). High performance computation of landscape genomic models including local indicators of spatial association. *Molecular Ecology Resources*, *17*(5), 1072–1089.

242. Su, J., Li, L., Zhang, C., Wang, C., Gu, L., Wang, H., Wei, H., Liu, Q., Huang, L. and Yu, S. (2018). Genome-wide association study identified genetic variations and candidate genes for plant architecture component traits in Chinese upland cotton. *Theoretical and Applied Genetics*, *131*(6), pp.1299-1314.

243. Tarkka, M., Herrmann, S., Wubet, T., Feldhahn, L., Recht, S., Kurth, F., Mailänder, S., Bönn, M., Neef, M., Angay, O., Bacht, M., Graf, M., Maboreke, H., Fleischmann, F., Grams, T., Ruess, L., Schädler, M., Brandl, R., Scheu, S., Schrey, S., Grosse, I. and Buscot, F. (2013). OakContigDF159.1, a reference library for studying differential gene expression in Quercus robur during controlled biotic interactions: use for quantitative transcriptomic profiling of oak roots in ectomycorrhizal symbiosis. *New Phytologist*, *199*(2), 529-540.

244. Thomas, C. (2015). Rapid acceleration of plant speciation during the Anthropocene. *Trends In Ecology & Evolution*, *30*(8), 448-455.

245. Thorsson, A. (2001). Morphological, Cytogenetic, and Molecular Evidence for Introgressive Hybridization in Birch. *Journal Of Heredity*, *92* (5), 404-408.

246. Todesco, M., Pascual, M., Owens, G., Ostevik, K., Moyers, B., & Hübner, S. et al. (2016). Hybridization and extinction. *Evolutionary Applications*, *9*(7), 892-908.

247. Truffaut, L., Chancerel, E., Ducousso, A., Dupouey, J., Badeau, V., Ehrenmann, F., & Kremer, A. (2017). Fine-scale species distribution changes in a mixed oak stand over two successive generations. *New Phytologist*, *215* (1), 126-139.

248. Tsuda, Y., Chen, J., Stocks, M., Källman, T., Sønstebø, J. H., Parducci, L., … Lascoux, M. (2016). The extent and meaning of hybridization and introgression between siberian spruce (*Picea obovata*) and norway spruce (*Picea abies*): cryptic refugia as steppingstones to the west? *Molecular Ecology*, *25* (12),2773–2789

249. Tsuda, Y., Semerikov, V., Sebastiani, F., Vendramin, G., & Lascoux, M. (2017). Multispecies genetic structure and hybridization in the *Betula* genus across Eurasia. *Molecular Ecology*, *26*(2), 589-605.

250. Uchiyama, K., Iwata, H., Moriguchi, Y., Ujino-Ihara, T., Ueno, S., Taguchi, Y., Tsubomura, M., Mishima, K., Iki, T., Watanabe, A., Futamura, N., Shinohara, K. and Tsumura, Y. (2013). Demonstration of Genome-Wide Association Studies for Identifying Markers for Wood Property and Male Strobili Traits in Cryptomeria japonica. *Plos One*, *8*(11), e79866.

251. Ueno, S., Le Provost, G., Léger, V., Klopp, C., Noirot, C., Frigerio, J., Salin, F., Salse, J., Abrouk, M., Murat, F., Brendel, O., Derory, J., Abadie, P., Léger, P., Cabane, C., Barré, A., de Daruvar, A., Couloux, A., Wincker, P., Reviron, M., Kremer, A. and Plomion, C. (2010). Bioinformatic analysis of ESTs collected by Sanger and pyrosequencing methods for a keystone forest tree species: oak. *BMC Genomics*, 11,650.

252. VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*, 4414–4423.

253. Wang, H., Yin, X., Yin, D., Li, L., & Xiao, H. (2019). Population genetic structures of two ecologically distinct species Betula platyphylla and B. ermanii inferred based on nuclear and chloroplast DNA markers. *Ecology And Evolution*, *9*(19), 11406-11419.

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

255. Wang, N., Kelly, L., McAllister, H., Zohren, J., & Buggs, R. (2021). Resolving phylogeny and polyploid parentage using genus-wide genome-wide sequence data from birch trees. *Molecular Phylogenetics And Evolution*, *160*, 107126.

256. Wang, N., McAllister, H., Bartlett, P., & Buggs, R. (2016). Molecular phylogeny and genome size evolution of the genus Betula (Betulaceae). *Annals Of Botany*, *117*(6), 1023-1035.

257. Wang, N., Thomson, M., Bodles, W., Crawford, R., Hunt, H., & Featherstone, A., Pellicer, J., Buggs, R. J. A. (2013). Genome sequence of dwarf birch (Betula nana) and cross-species RAD markers. *Molecular Ecology*, *22*(11), 3098-3111.

258. Wang, S., Yang, C., Zhao, X., Chen, S., & Qu, G. (2018). Complete chloroplast genome sequence of Betula platyphylla: gene organization, RNA editing, and comparative and phylogenetic analyses. *BMC Genomics*, *19*(1).

259. Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution, 38*, 1358–1370.

260. Whiteley, A., Fitzpatrick, S., Funk, W., & Tallmon, D. (2015). Genetic rescue to the rescue. *Trends In Ecology & Evolution*, *30*(1), 42-49.

261. Williams, Jr., J., & Arnold, M. (2001). Sources of Genetic Structure in the Woody PerennialBetula occidentalis. *International Journal Of Plant Sciences*, *162* (5), 1097-1109.

262. Wimmer, V., Albrecht, T., Auinger, H.-J., & Schön, C.C. (2012). Synbreed: a framework for the analysis of genomic prediction data using R. *Bioinformatics, 28* (15), 2086–2087.

263. Winkler, H. (1904). Betulaceae. Das Pflanzenreich19: 1–149.

264. Xing, Y., Liu, Y., Zhang, Q., Nie, X., Sun, Y., & Zhang, Z., … & Qin, L. (2019). Hybrid de novo genome assembly of Chinese chestnut (Castanea mollissima). *Gigascience*, *8* (9).

265. Yeaman, S. (2015). Local Adaptation by Alleles of Small Effect. *The American Naturalist*, *186*(S1), S74-S89.

266. Yeaman, S. (2022). Evolution of polygenic traits under global vs local adaptation. *Genetics*, *220*(1).

267. Yeaman, S., Hodgins, K., Lotterhos, K., Suren, H., Nadeau, S., & Degner, J. et al. (2016). Convergent local adaptation to climate in distantly related conifers. *Science*, *353*(6306), 1431-1433.

268. Zanetto A., Roussel G., & Kremer A. (1994). Geographic variation of inter-specific differentiation between Quercus robur L. and Quercus petraea (Matt.) Liebl. *Forest Genetics*, *1*, 111–123.

269. Zdobnov, E., & Apweiler, R. (2001). InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, *17*(9), 847-848.

270. Zeng, Y. F., Liao, W. J., Petit, R. J., & Zhang, D. Y. (2011). Geographic variation in the structure of oak hybrid zones provides insights into the dynamics of speciation. *Molecular Ecology*, *20*, 4995–5011.

271. Zeng, Y. F., Zhang, J. G., Duan, A. G., & Abuduhamit, A. (2016). Genetic structure of Populus hybrid zone along the Irtysh River provides insight into plastid-nuclear incompatibility. *Scientific Reports*, *6*, 2804.

272. Zhang, C., Dong, S., Xu, J., He, W., & Yang, T. (2018). PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics*, *35* (10), 1786-1788.

273. Zhou, H., Duan, H., Liu, Y., Sun, X., Zhao, J., & Lin, H. (2019). Patellin protein family functions in plant development and stress response. *Journal of Plant Physiology*, 234-235, 94-97.

274. Zou, B., Ding, Y., Liu, H., & Hua, J. (2017). Silencing of copine genes confers common wheat enhanced resistance to powdery mildew. *Molecular Plant Pathology*, *19*(6), 1343–1352.

275. Zou, B., Hong, X., Ding, Y., Wang, X., Liu, H., & Hua, J. (2016). Identification and analysis of copine/BONZAI proteins among evolutionarily diverse plant species. *Genome*, *59*(8), 565–573.