Genomic signals of local adaptation and hybridization in Asian white birch

Gabriele Nocchi[1,2#], Jing Wang[3#], Long Yang[4], Junyi Ding[5,6], Ying Gao[4], Richard J. A. Buggs[1,2*], Nian Wang[5,6,7*]

[1]School of Biological and Behavioural Sciences, Queen Mary University of London, Mile End Road, London, E1 4NS, UK

[2]Royal Botanic Gardens Kew, Richmond, TW9 3AB, Surrey, UK

[3]Key Laboratory for Bio-resources and Eco-environment, College of Life Science, Sichuan University, Chengdu, 610064, China

[4]Agricultural Big-Data Research Center and College of Plant Protection, Shandong Agricultural University, Tai'an 271018, China

[5]State Forestry and Grassland Administration Key Laboratory of Silviculture in Downstream Areas of the Yellow River, College of Forestry, Shandong Agricultural University, Tai'an 271018, China

[6]Mountain Tai Forest Ecosystem Research Station of State Forestry and Grassland Administration, College of Forestry, Shandong Agricultural University, Tai'an, 271018, China.

[7]State Key Laboratory of Crop Biology, Shandong Agricultural University, Tai'an, 271018, China

[#]Contributed equally to this work

[*]Author for correspondence:

Richard Buggs: r.buggs@kew.org

Nian Wang: nian.wang@sdau.edu.cn

Abstract

Disentangling the numerous processes that affect patterns of genome-wide diversity in widespread tree species has important implications for taxonomy, conservation, and forestry. Here, we investigate the population genomic structure of Asian white birch (*Betula platyphylla*) in China and seek to explain it in terms of hybridization, demography and adaptation. 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 of *B. platyphylla* in north-western China and Central Russia is affected greatly by hybridization with *B. pendula*. Excluding these hybridized populations, *B. platyphylla* in China has a linear distribution from north-eastern to south-western China, along the edge of the inland mountainous region. Within this distribution, three genetic clusters are found, which we model as long diverged with subsequent episodes of gene flow. Patterns of co-variation between allele frequencies and environmental variables in *B. platyphylla* suggest the role of natural selection in the distribution of diversity at 7,609 SNPs of which 3,767 were significantly differentiated among the genetic clusters. 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 Eurasian *B. pendula*. We use the 7,609 environmentally associated SNPs to estimate the risk of non-adaptedness for each sequenced *B. platyphylla* individual under a scenario of future climate change, highlighting areas where populations may be under future threat from rising temperatures.

Keywords

Asian white birch, landscape genomics, climate change, local adaptation, hybridization

Introduction

Analysis of the population genomic diversity of species can help us to understand their past and predict their future. Past processes affecting present allele frequencies include hybridization, range contraction and expansion, population size changes, gene flow among populations, and natural selection. Understanding the effects of natural selection is particularly important for future predictions as the extent to which alleles are locally adaptive has important implications for population fitness under climate change (Sork et al., 2013; Savolainen et al., 2007; Aitken et al. 2008). Covariance between allele frequencies and environmental variables may be signatures of local adaptation but only if the effects of other past processes are accounted for (Sork et al., 2013; Wang & Bradburd, 2014). Hence, we need to gain the fullest possible understanding of all past processes affecting allele frequencies in populations before we can predict patterns of diversity that may be beneficial in future climates.

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 that seek to account for the effects of neutral processes (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).

The future of forest trees under climate change is critical carbon sequestration, ecosystem conservation and timber production. Forest trees with wide geographic distributions over spatially heterogeneous environments often show local adaptation with adaptive traits polygenically

74 determined (Savolainen et al., 2007; Alberto et al., 2013). Due to their long generation times, and
75 short dispersal distances relative to their range sizes, forest trees may struggle to adapt or migrate
76 under rapid climate change (Aitken et al., 2008; Keenan, 2015). The genus *Betula* (birch) is a case in
77 point, containing several tree species widespread in the Northern Hemisphere, comprising key
78 components of temperate forests (Ashburner and McAllister, 2013; Wang et al., 2016; Wang et al.,
79 2021). Understanding the genomic basis of adaptation in this genus can help to predict their
80 potential to survive under rapid climate change, helping to manage species conservation (Borrell et
81 al., 2019).

82 In this study, we selected Asian white birch (*Betula platyphylla*), which belongs to the subgenus
83 *Betula*, to understand its past demographical histories, current local adaptation, and future potential
84 to respond to climate change. Asian white birch is of great ecological and commercial value. It
85 commonly occurs in eastern Asia including the Far East Russia, China, and Japan. In China, it has
86 distributions in northwestern (NW), central (CE), northeastern (NE) and southwestern regions (SW)
87 (Chen & Lou, 2019; Ashburner and McAllister, 2013). Asian white birch is a pioneer species,
88 colonizing open habitats in grasslands, on roadsides and on the verge of forest (Ashburner and
89 McAllister, 2013). We selected this species not only because of the above-mentioned values but also
90 for the following reasons. First, it occupies spatial heterogeneous environments across its range with
91 over 26 latitudinal degrees based on our collection records. Some populations distribute in SW China
92 and its adjacent regions, with an altitude of over 3,700 meters whereas some populations
93 distributing in NW China grow by riverside with a much higher annual evaporation than precipitation.

94 Some taxonomists place *B. platyphylla* into the same species as silver birch, *B. pendula* (Ashburner &
95 McAllister, 2013), which is widespread in Europe and Russia. Others see them as sister species (Chen
96 et al., 2021; Wang et al., 2021; Salojärvi et al., 2017; Tsuda et al., 2017), with estimates of divergence
97 times differing between 36,000 years (Tsuda et al., 2017) and 2.6 million years (Chen et al., 2021).
98 Well assembled and annotated reference genomes now exist for both *B. pendula* (Salojärvi et al.,

2017) and *B. platyphylla* (Chen et al., 2021) but the latter was not available when we commenced the present study. A large population genomic dataset is publicly available for *B. pendula*, analysis of which identified two genetic clusters with some inter-mixed populations and showed that variations around the genes involved in photoperiodic responses were highly associated with local environmental adaptation (Salojärvi et al., 2017). The genetic structure of *B. platyphylla* has been investigated using ten microsatellite markers, showing five genetic clusters within its range, with substantial gene flow among them (Chen & Lou, 2019).

In this present study, we generate whole genome sequence data for *B. platyphylla* across its range in China to investigate its population genetic structure. We ask if *B. platyphylla* is differentiated from *B. pendula*, and if so, the extent to which hybridization among them affects *B. platyphylla* populations. We investigate the genetic structure within nonhybridized *B. platyphylla*, modelling how this has been affected by population divergence, population size change, and gene flow. We also model the environmental niche of *B. platyphylla* and how this may affect its range under future climates. We use EAA to identify SNP markers in *B. platyphylla* associated with environmental clines after controlling for population structure. We calculate the risk of non-adaptedness (RONA) (Rellstab et al., 2016) for each individual under a scenario of climate of 2080-2100. Together, this study generates new hypotheses regarding local adaptation in Asian white birch, which can be further tested in common garden trials or trans-genetic 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 S1 (Supporting Information). Dried cambial tissues were used for DNA

123 extraction following a modified CTAB protocol (Wang et al., 2013). One or two samples per
124 population were selected for resequencing, resulting in a total of 83 samples. Extracted DNA was
125 assessed with 1% agarose gel and sent to BerryGenomics (Beijing, China) for library preparation and
126 whole genome re-sequencing. Sequencing was performed on the Illumina NovaSeq6000 platform
127 with 150 bp paired end read sequencing with approximate 11Gbp data obtained for each sample
128 (average depth of ~25x). We also downloaded genomic read data from Salojarvi et al. (2017) for 78
129 *B. pendula* individuals from most of this species' European and Asian geographic range and one *B.*
130 *platyphylla* individual from Russia.

131

132 **Trimming and SNP filtering**

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

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

151

152 **Population structure**

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

166 On the basis of the fastSTRUCTURE analysis of the Eurasian dataset, we excluded from the China
167 dataset any individuals that had over 10% assignment to *B. pendula* and ran *fastSTRUCTURE* only on
168 Chinese *B. platyphylla* based on 1,387,994 SNPs (after filtering for missing genotypes). Furthermore,
169 we used the program sparse non-negative matrix factorization (*snmf*), part of LEA package (Caye et
170 al., 2019), to perform a population differentiation test and identify F_{st} outlier SNPs between the
171 clusters identified for *B. platyphylla* in China. *snmf* computes fixation also in the presence of

172 admixed individuals in the sample (Martins et al., 2016) and performs a test similar to other F_{st}-
173 outlier approaches (Lotterhos & Whitlock, 2014). Multiple testing was controlled by converting the
174 p-values computed by *snmf* into q-values with the R package "qvalue" (Storey et al., 2015). Q-values
175 represent the expected false discovery rate (FDR) associated with a given p-value (Storey and
176 Tibshirani 2003). Finally, we selected candidate outlier SNPs with FDR < 1%.

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

179

180 **Demographic histories**

181 We inferred the demographic histories of the three identified populations of Asian white birch using
182 a coalescent simulation-based method implemented in fastsimcoal2 v 2.6.0.3 (Excoffier et al. 2013).
183 The pairwise two-dimensional joint SFSs (2D-SFSs) among the three populations were constructed
184 from the posterior probabilities of sample allele frequencies using ANGSD v 0.934 (Korneliussen et
185 al., 2014). A total of 14 models were evaluated assuming divergence with or without gene flow and a
186 constant or changed population size after divergence (Figure S1, Table S2, Supporting Information).
187 All models show the NE diverging first, followed by the split of the CE and SW populations (Figure S1,
188 Supporting Information). The global maximum likelihood estimates for all demographic parameters
189 under each model were obtained from 60 independent runs, with 100,000 coalescent simulations
190 per likelihood estimates and 40 cycles of the likelihood maximization algorithm. The models were
191 compared based on the maximum value of likelihood over the 60 independent runs using the
192 Akaike's weight. The model with the maximum Akaike's weight value was chosen as the optimal one.
193 Confidence intervals were generated by performing parametric bootstrapping with 100 bootstrap
194 replicates, and with 30 independent runs in each bootstrap. We used a mutation rate of 9.5×10^{-9}
195 per site per year (Salojärvi et al. 2017) and a generation time of 15 years to convert the coalescent
196 scaled time into a speculative time scale in years.

197 In addition, we inferred the historical change in effective population sizes using Multiple Sequentially
198 Markovian Coalescent approach (MSMC v2.0.0) (Schiffels & Durbin, 2014) for each of the three
199 populations, respectively. For each population, we randomly chosen seven individuals and applied
200 MSMC to phased whole-genome sequences from four (eight haplotypes) individuals. In total, we had
201 35 different individual configurations for haplotype analyses in each lineage. We ran MSMC on all
202 individual configurations and estimated medians and standard deviations of effective population
203 sizes changes across time.

204

205 **Population diversity and differentiation**

206 We estimated linkage disequilibrium decay along the *B. platyphylla* genome with the tool
207 *PopLDdecay* (Zhang, Dong, Xu, He, & Yang, 2018) separately for the different populations identified
208 for *B. platyphylla* in China by *fastSTRUCTURE* at K = 3 (Figure 3). For this calculation we used the SNP
209 set prior to linkage disequilibrium pruning but filtered to retain SNPs with minor allele frequency >
210 0.05. After excluding individuals admixed between the three populations (admixture coefficient < 0.9
211 in fastSTRUCTURE at K = 3), Weir and Cockerham F_{st} (Weir & Cockerham, 1984) was calculated
212 between the three *B. platyphylla* populations on a per SNP site basis with the *vcftools* function *weir-*
213 *fst-pop* (Danecek et al., 2011). We calculated nucleotide diversity π (Nei & Li, 1979) within each
214 population using the *vcftools* (Danecek et al., 2011) function *window-pi* in non-overlapping windows
215 of 5 kb. For the π calculation we used the SNPs set prior to minor allele frequency filtering and
216 linkage disequilibrium pruning (28,058,885 SNPs in total).

217

218 **Environmental niche modelling**

219 We used environmental niche modelling (ENM) to characterize the present habitat and predict the
220 future distribution of non-hybrid *B. platyphylla* in China and to identify the climatic variables

221 influencing this species distribution. ENM analysis was performed by using the samples' geographic
222 locations in this study and observation records of this species from 1970 to 2010 in China, sourced
223 from the Global Biodiversity Information Facility (http://www.gbif.org). To mitigate the effects of
224 spatial autocorrelation, we removed records within 0.2 degrees (approximately 22-25 km) from one
225 another, which resulted in a total of 138 presence locations: 66 records were *B. platyphylla* from the
226 83 individuals sampled (excluding admixed *B. pendula* - *B. platyphylla* individuals) and another 72
227 records were from the GBIF (excluding NW China as all the trees that we had sampled there were
228 hybrids – see Results). We downloaded 19 bioclimatic variables related to temperature and
229 precipitation from the WorldClim database (www.worldclim.org) at 1km resolution (Hijmans,
230 Cameron, Parra, Jones, & Jarvis, 2005) for the period 1970-2000 representing "current climate", as
231 this is the estimated time period of establishment of the sampled trees and also corresponds to the
232 time that most GBIF observations included in the model were recorded. We downloaded elevation
233 data at the same resolution and used it to compute slope, aspect and four additional terrain related
234 characteristics (topographic position index, terrain ruggedness index, terrain roughness and water
235 flow direction) using the *raster* R package (Hijmans & Etten, 2012). In order to avoid overfitting, we
236 selected environmental variables with Pearson's correlation coefficient < 0.7, preferring annual
237 rather than monthly or quarterly values, which resulted in 11 variables retained for the ENM (Table
238 S3, Figure S2-3, Supporting Information). We assembled eight further datasets with the same 11
239 variables under four Shared Socioeconomic Pathways (SSPs) defined by the Intergovernmental Panel
240 on Climate Change sixth Assessment (Masson-Delmotte et al., 2021) at each of two future time
241 points (2041–2060 and 2081–2100). The ENMs were generated using Maxent (Phillips, Anderson, &
242 Schapire, 2006) with 50 subsampled replicate 5000 iterations runs with 20% of observations left-out
243 for cross-validation. Variables' importance was assessed with jack-knife tests and multiple models
244 were generated and evaluated using a test of omission rate and area under the receiver operating
245 characteristic curve (AUC). An environment suitability threshold was defined by "maximum training

246 sensitivity plus specificity," which optimizes the trade-off between commission and omission errors
247 (Liu, Newell, & White, 2016).

248

249 **Genome environment association analysis: identification of adaptive SNPs**

250 We used *LFMM2* (Caye, Jumentier, Lepeule, & Francois, 2019) to test for associations between
251 environmental variables and SNPs in 71 Chinese non-hybrid *B. platyphylla* individuals. *LFMM2*
252 performs multivariate linear regressions to evaluate the association between a response matrix,
253 corresponding to SNP frequencies of individuals, and a matrix of environmental variables. It
254 combines the environmental fixed effects with latent effects, which are unobserved confounding
255 effects due to population structure (Frichot et al., 2013; Caye, Jumentier, Lepeule, & Francois, 2019).
256 *LFMM2* has a two steps approach: first it estimates the latent factors and then tests for the
257 genotype-environment associations. Each SNP locus is tested separately using the latent scores as
258 covariates to control for demographic history and population structure. The null hypothesis is that
259 the environmental variables have no effects on the SNP frequency, which is tested using a student
260 distribution with $n-K-1$ degrees of freedom (Caye, Jumentier, Lepeule, & Francois, 2019). *LFMM2*
261 requires to specify the number (K) of latent factors to include in the model which was determined
262 according to *fastSTRUCTURE* cross-validation schemes. We chose the more conservative K = 3 (see
263 Results section) to run *LFMM2*, including all the 11 uncorrelated environmental variables used in the
264 ENM (Table S3, Supporting Information). Multiple testing was controlled by converting the p-values
265 computed by *LFMM2* into q-values with the R package "qvalue" (Storey et al., 2015). Finally, we
266 selected candidate adaptive SNPs with FDR < 1%.

267

268 **Characterization of candidate adaptive SNPs**

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

277 In order to functionally annotate our putative adaptive SNPs, we performed a gene ontology (GO)
278 annotation of the whole *B. pendula* reference gene set provided in Salojärvi et al. (2017), retrieving
279 GO terms associated with the coding regions (Ashburner et al., 2000) using *Omics Box* (Conesa &
280 Götz, 2008). We used *Omics Box* default functional annotation workflow which incorporates BLAST,
281 run on the Viridiplantae non-redundant database, and Interproscan (Zdobnov & Apweiler, 2001). We
282 then identified the genic regions overlapping our candidate adaptive SNPs with *bedtools* v2.28.0
283 (Quinlan & Hall, 2010) and performed a functional enrichment analysis of this adaptive gene set
284 against the fully annotated reference gene set using Fisher's exact test in *Omics Box* (Conesa & Götz,
285 2008). We applied multiple testing corrections (FDR < 5%) and reduced the resulting significantly
286 enriched GOs to the most specific term in the hierarchy for all three ontology levels: biological
287 process, molecular function, and cellular component (Ashburner et al., 2000).

288

289 **Adaption to future climate**

290 We carried out risk of non-adaptedness (RONA) analysis using a slight modification of the method of
291 Rellstab et al. (2016), implemented in the software *pyRona* (Pina-Martins, Baptista, Pappas & Paulo,
292 2018). RONA represents an estimate of the average change in allele frequencies at climate
293 associated SNP loci required in a population to cope with future climatic conditions and is calculated

294 separately for each environmental variable (Rellstab et al., 2016). RONA uses simple linear
295 regressions of the selected current climate environmental variable and the alternative allele
296 frequencies of the candidate loci identified in genome-environment association analyses (Rellstab et
297 al., 2016). The regression coefficients of the environmentally associated loci are then used to predict
298 the allele frequencies of the adaptive loci at a future climate value for the chosen environmental
299 variable (Rellstab et al., 2016). The main improvement in the implementation by Pina-Martins,
300 Baptista, Pappas & Paulo (2018) is that the RONA for an environmental variable is given by the
301 weighted mean RONA of all relevant SNP loci for that environmental variable, weighted by the r^2
302 value of each locus regression, while the original method (Rellstab et al., 2016) uses unweighted
303 means. We computed RONA on an individual sample basis for the seven uncorrelated environmental
304 variables that are expected to change in the future, therefore excluding elevation and its derived
305 measures. We used as future climate the profile ssp370 for 2080-2100 (Figure 4C). For each
306 environmental variable, RONA was calculated by using the candidate adaptive SNP loci identified
307 with *LFMM2* (FDR < 1%). For each individual tree we then calculated weighted mean RONA, giving
308 each environmental variable RONA a weight equivalent to the contribution reported by that variable
309 in the ENM. Furthermore, we reported the maximum RONA for each individual, out of the seven
310 variables for which it was calculated.

311

312 **Results**

313 **SNP discovery**

314 After filtering, the number of reads retained for each sample in the China dataset that we generated
315 ranged between 66,016,635 and 134,129,672 and between 91.5% and 96.7% of these reads per
316 individual were mapped to the reference genome (Table S1, Supporting Information). SNPs filtering
317 of the 83 individuals resulted in 1,497,547 unlinked (r^2< 0.4) SNPs. After filtering, the number of
318 reads retained for each sample in the Salojarvi et al. (2017) dataset ranged between 16,506,088 and

319 237,845,424 and between 82.9% and 97.2% of these reads per individual were mapped to the
320 reference genome. The combined Eurasian dataset resulted in 278,717 SNPs after filtering for
321 missing genotypes. When the China dataset was restricted to pure *B. platyphylla* (71 individuals in
322 SW, CE, and NE China) we were left with 1,387,994 SNPs.

323

324 **Population structure, diversity, and differentiation**

325 For the China dataset *fastSTRUCTURE* analysis with K = 2 separates the samples in north-western
326 China (Xinjiang) from the other populations. At K = 3 south-western and north-eastern populations
327 are further separated and at K = 4 a central population is revealed (Figure 1). The model that
328 maximises the log-marginal likelihood lower bound (LLBO) of the data and that best explains
329 additional weak underlying structure was K = 3, as suggested by the function "*chooseK*" (Figure S4,
330 Supporting Information).

331 For the Eurasian dataset at K = 2, one cluster contains the *B. pendula* individuals from Salojarvi et al.
332 (2017) and the other cluster contains 71 of the 83 individuals that we sequenced from China and the
333 Russian *B. platyphylla* individual from Salojarvi et al. (2019) (Figure 2). The 12 individuals we sampled
334 in north-western China (Xinjiang) are shown as hybrids between *B. pendula* and *B. platyphylla*
335 (Figure 2A). At K = 3, the pattern was similar but a south-western China *B. platyphylla* cluster
336 separates from a north-eastern China *B. platyphylla* cluster (Figure 2B). At K = 4, the only difference
337 was that Irish and two Finnish *B. pendula* samples showed an unknown component (Figure 2C). The
338 *fastSTRUCTURE* model that maximises the log-marginal likelihood lower bound (LLBO) of the data
339 and that best explains additional weak underlying structure was K = 4, as suggested by the function
340 choose K (Figure S5, Supporting Information). The LLBO curve and the cross-validation profile show
341 that the marginal likelihood plateau at K = 2, and so does the prediction error, which stays within 1
342 standard error from K = 2 with added complexity (Figure S5A-B, Supporting Information). These

343 separations suggested by fastSTRUCTURE are reflected in the PCA (Figure S6, Supporting
344 Information).

345 In *fastSTRUCTURE* restricted to the 71 non-hybrid *B. platyphylla* individuals in China, the model that
346 maximises the log-marginal likelihood lower bound (LLBO) of the data was K = 2 and that that best
347 explains additional weak underlying structure was K = 3, as suggested by the function "*chooseK*"
348 (Figure S7A, Supporting Information). The *fastSTRUCTURE* cross-validation score profile shows that
349 the prediction error is lowest at K = 3 (Figure S7B, Supporting Information). At K = 2 the north-
350 eastern population is separated from the south-western population and the individuals in central
351 China appear admixed between these two populations, while at K = 3 the central population is
352 separated (Figure 3).

353 In PCA on the 71 pure *B. platyphylla* individuals, PC1 explained ~33% of the total variance and
354 separated the north-eastern, central, and south-western populations (Figure S8, Supporting
355 Information). PC2 (15% of total variance) also separates the three populations while PC3 (8% of total
356 variance) scatters the individuals of the north-eastern population. The explained variance levelled-
357 off after PC3 (Figure S8D, Supporting Information).

358 Due to the differentiation between the individuals located in central China and those in the North-
359 East shown by the PCA (Figure S8, Supporting Information) and the recommendation by
360 *fastSTRUCTURE* cross-validation scheme, we assigned individuals to populations according to the
361 admixture coefficients computed with *fastSTRUCTURE* at K = 3. In total, ten individuals were
362 assigned to the central population (CE), seven to the south-western population (SW), 46 to the
363 north-eastern population (NE) and the remaining eight individuals were classified as admixed (Figure
364 3B-C). Based on this assignment, *snmf* identified 17,218 outlier SNPs with FDR < 1% (Figure S9,
365 Supporting Information).

366 Linkage disequilibrium decays rapidly in *B. platyphylla* populations and reaches background levels at
367 approximately 50 kb (Figure S10, Supporting Information). Assessment of pairwise F_{st} between

368 populations showed that the highest levels of differentiation are between the north-eastern and
369 south-western population or the central population and south-western populations (Figure S11,
370 Supporting Information). On the other hand, the central and the north-eastern population reported
371 significantly lower mean pairwise F_{st} (Figure S11, Supporting Information).

372 Nucleotide diversity π in windows of 5kb shows a slight increase in populations from south to north
373 geographically, reporting mean of 0.0057 (~0.6%) in the south-western population, 0.0075 (~0.7%) in
374 the central population and 0.0083 (~0.8%) in the north-eastern population (Figure S12-13,
375 Supporting Information).

376

377 **Species population size and separation history**

378 Assuming that there are three populations of *B. platyphylla* in NE, CE and SW China, and the NE
379 population was the first to diverge, the best-supported fastsimcoal2 model (Figure 3D; Table S2,
380 Supporting Information) suggested that the NE population diverged from the common ancestral
381 population of CE and SW at 2.804 Mya (95% CI: 2.892–3.506 Mya) (Figure 3D). The NE population
382 initially had a small population size. The CE and SW populations diverged 2.797 Mya (95% CI: 2.294–
383 2.768 Mya; Figure 3D). The SW population underwent a slight bottleneck at 2.787 Mya (95% CI:
384 2.129–2.489 Mya) and experienced a slight population expansion at 0.812 Mya (95% CI: 0.477–0.706
385 Mya) (Figure 3D). The estimated present day effective population sizes (Ne) are 3.3×10^5 (95% CI:
386 $3.0–3.4 \times 10^5$) for NE, 1.2×10^5 (95% CI: $1.6–2.2 \times 10^5$) for CE and 0.7×10^5 (95% CI: $1.6–2.2 \times 10^5$) for
387 SW. The best-supported model detected low levels of ancient gene flow between the populations
388 soon after their divergence, no gene flow between 2.797 Mya and 0.812 Mya, and recent gene flow
389 between NE and CE and between SW and CE (Figure 3D).

390

391 MSMC results suggested that the three populations experienced severe bottlenecks (2.51-2.89
392 million years ago) followed by population expansion during the period of 26.1-42.6 thousand years
393 ago (Figure 3E). The NE lineage experienced the least population reduction (Ne, 11235.88 ± 204.75)
394 whereas the SW lineage underwent the most severe population decline (Ne, 6152.70 ± 265.12)
395 (Figure 3E).

396

397 **Environmental niche modelling**

398 Environmental niche models were built based on eleven uncorrelated environmental variables
399 (Figure S1-2, Table S3, Supporting Information) and were based on our *B. platyphylla* samples'
400 locations and GBIF observations of this species (Figure 4).

401 Jack-knife tests suggested that the variables "aspect" and "water-flow direction" did not play a
402 relevant contribution either singly or in combination, so they were excluded (Figure S14-16,
403 Supporting Information). The final Maxent model with the nine remaining variables reported high
404 mean test AUC (0.913 ± 0.017) and low mean test omission rate (0.13, $p < .001$) at a maximum
405 training sensitivity plus specificity logistic threshold of 0.2 (Figure 4A-B; Figure S17-18, Supporting
406 Information). Two variables, "annual precipitation" and "mean temperature of wettest quarter",
407 together contributed to > 50% of the predicting performance of the model, defined as the increase
408 in regularized gain added (or subtracted) to the contribution of the corresponding variable, over
409 each model iteration (Table S4, Supporting Information).

410 Response curves for the final Maxent model are available in Supporting Information. Future
411 projections show an overall contraction of the *B. platyphylla* habitat throughout China, particularly
412 noticeable in central China and in the north-east (Figure 4C). Currently suitable environments
413 located at higher elevations seem to contract less in the future (Figure 4C). The predicted habitat
414 reduces between 19 and 30 % in 2040-2060, and between 17 and 34 % in 2080-2100 (Figure 4C),

415 compared to the present. In the most adverse scenario (ssp370), *B. platyphylla* habitat in China may
416 be reduced by approximately a third of its current extent by 2080-2100 (Figure 4C) and the NE and
417 SW populations will be spatially separated by a large uninhabitable region.

418

419

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

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

425 In total, 7,609 SNPs showed significant associations with one or more environmental variables in
426 *LFMM2* (K = 3) using our criteria (Figure S19, Supporting Information), for a total of 11,304
427 associations. In more details: 4,643 SNP were associated to one environmental variable, 2,424 SNPs
428 to two environmental variables, 384 SNPs to three environmental variables, 133 SNPs to four
429 environmental variables, 21 SNPs to five environmental variables and four SNPs to six environmental
430 variables. Of the 7,609 putative adaptive SNPs identified *LFMM2*, 3,767 were shown to be more
431 differentiated among the NE, CE and SW populations than was expected due to drift in our *snmf* F_{st}
432 outlier test (FDR < 1%).

433 The number of the putative environmentally associated SNPs detected with LFMM2 differed among
434 the variables tested (Figure S20, supporting information). "Isothermality" and "mean diurnal range"
435 had a much larger number of candidates than the other variables, with 5,679 and 2,794 putative
436 SNP-environment associations identified for these two variables, respectively. These were followed
437 by "annual precipitation" (1,059 SNPs hits), "precipitation of driest quarter" (735 SNPs hits), "annual
438 mean temperature" (391 SNPs hits), "flow direction" (263 SNPs hits), "mean temperature of wettest
439 quarter" (227 SNPs hits), "TPI" (104 SNPs hits), "slope" (49 SNPs hits) and "precipitation seasonality"

440 (3 SNPs hits). The variable “aspect” did not have any SNP association with an FDR < 1% in *LFMM2*.
441 There was no significant correlation between the number of identified SNP association per variable
442 and the percentage of contribution of each variable to the ENM (Pearson's test $p > 0.05$).

443 The distribution of the candidate adaptive SNPs in the *B. pendula* reference genome shows that they
444 are spread across the entire genome and present on all linkage groups (Figure S20, Supporting
445 Information). The chromosome that reported the largest number of candidates is chromosome 2
446 (738 SNPs), while the lowest number was recorded for chromosome 6 (344 SNPs). The number of
447 SNPs identified per chromosome does not appear correlated with chromosome size (Pearson's test p
448 > 0.05). A hotspot is visible in the second half of chromosome 7 (Figure S20, Supporting
449 Information).

450 PCA based only on the 7,609 adaptive SNPs showed a steep discrepancy in the variance explained by
451 PC1 compared to that of the other PCs (Figure S21, Supporting Information). Furthermore, PCA
452 based on only adaptive SNPs does not show a clear separation between the central and north-
453 eastern population in the first three PCs, while it separates the south-western population (Figure
454 S21, Supporting Information). When using only 7,500 putatively “neutral” SNPs in PCA, the
455 separation between populations is slightly weaker compared to that observed with the whole data
456 set, and individuals in each population are more scattered along the PCA axes (Figure S22,
457 Supporting Information). The pairwise F_{st} patterns of differentiation between populations based only
458 the 7,609 adaptive SNPs and only on neutral loci reflected those observed with the whole SNPs set,
459 however loci under selection showed substantially higher F_{st} compared to those recorded for neutral
460 markers and for the whole SNP set (Figure S23-24, Supporting Information).

461 We further searched for the mRNA regions in the *B. pendula* reference genome (Salojärvi et al.,
462 2017) containing at least one candidate adaptive SNP out of the 7,609 candidates, and we identified
463 1,633 genic regions. A functional enrichment analysis of the identified genic regions reported

464 significant hits (FDR < 5%) at all three gene ontology levels: biological process (3 hits), cellular
465 component (1 hits) and molecular function (3 hits) (Figure S25, Supporting Information).

466

467 **Risk of non-adaptedness to future conditions (RONA)**

468 We calculated RONA for seven environmental variables projected under the future climate profile
469 ssp370, identified as the worst-case scenario for *B. platyphylla* for the years 2080 – 2100 in terms of
470 habitat reduction (Figure 5, Table S5, Supporting Information). The most represented environmental
471 variables in terms of number of associations in the RONA computation were "isothermality" (5,679
472 SNPs, mean r^2 = 0.459), "mean diurnal range" (2,794 SNPs, mean r^2 = 0.0157) and "annual
473 precipitation" (1,059 SNPs, mean r^2 = 0.0976) (Table S5, Supporting Information). The variables that
474 reported higher average r^2 of adaptive SNPs were "isothermality" (5,679 SNPs, mean r^2 = 0.459),
475 "annual mean temperature" (391 SNPs, mean r^2 = 0.195), and "mean temperature of wettest
476 quarter" (227 SNPs, mean r^2 = 0.185) (Table S5, Supporting Information).

477 There was large variability in the expected allele frequency changes required to match future
478 conditions across environmental variables. "Annual mean temperature" and "mean temperature of
479 wettest quarter" reported notably higher mean individual RONA compared to all the other variables,
480 being 0.1773 and 0.2314 respectively (Table S5, Supporting Information).

481 Weighted mean RONA scores were overall relatively low across individuals, with mean of 0.08 (sd =
482 0.03) and maximum of 0.1455, that was reported for an individual in the north-eastern population
483 (HTY15) (Figure 6A, Table S6, Supporting Information). The maximum RONA scores show greater
484 variability across individuals, with mean of 0.25 (sd = 0.09) and maximum of 0.3956, that was
485 recorded for an individual in the north-eastern population (JY23, "mean temperature of wettest
486 quarter") (Figure 6B, Table S6, Supporting Information).

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 S7, 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 S7, 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 (< .1) mean RONA for this specific environmental variable (Table S7, 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 hybridization between these two species in northwest China, and to a lesser extent in central Russia. Populations in northwest China (Xinjiang) 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 & 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 hybridization in these populations. Instead, they attributed their high differentiation and diversity to a northern glacial refugium for white birch in the Altay Mountains. We might have done the same

512 if we had not co-analysed our data with that of Salojärvi et al. (2017). We removed individuals from
513 NW China from our downstream analyses on *B. platyphylla*.

514

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

516 With the NW individuals excluded, *B. platyphylla* in China has a linear distribution from north-
517 eastern to south-western China, along the edge of the inland mountainous region. Occurrence data
518 and environmental niche models suggest that there is currently a single contiguous region currently
519 habitable by *B. platyphylla*. Within our *B. platyphylla* samples, we found three genetic clusters with
520 some admixed individuals between them. Our fastsimcoal results suggest that that the three
521 populations exchanged genetic information soon after their divergence and also recently, with a
522 long period of isolation in between (Figure 3D). The absolute date estimated from fastsimcoal
523 suggests that this period of isolation between the *B. platyphylla* populations was from 2.8 Mya to 0.8
524 Mya. This is hard to fit with estimates of the time of speciation between *B. platyphylla* and *B.*
525 *pendula* which different studies place between 36,000 years (Tsuda et al., 2017) and 2.6 million years
526 (Chen et al., 2021) ago. We note, however, that our demographic simulation results are provisional
527 as it is hard to distinguish among the many factors that affect patterns genetic variation (Johri et al.,
528 2022), and our assumptions about mutation rate and generation time may be incorrect. We cannot
529 therefore exclude the possibility that our simulations are picking up signals of *B. platyphylla*
530 population isolation due to separate refugia in recent glaciations.

531 Consistent with a previous phylogeographic study of Asian white birch showing that genetic diversity
532 of Asian white birch increased along latitude based on SSR (Chen & Lou, 2019), our result confirmed
533 that genomic diversity in the NE is higher than that in CE which is higher than that in SW (Figure S12-
534 13, Supporting Information). It might have been expected that higher diversity in the NE is due to
535 admixture from *B. pendula*, but we find no evidence for this.

536 Ecological niche models indicated that mean temperature of wettest quarter and annual mean
537 precipitation are the most important predictors of *B. platyphylla* distribution and both north-eastern
538 China and the south-western Tibetan plateau currently provide the most suitable environments for
539 this species in China. Future climate projections show an overall decline of the species range and
540 environmental suitability throughout of its current distribution in China, with suitable habitats
541 persisting mostly in areas of higher elevation (Figure 4) and the central China habitable area
542 vulnerable to being lost.

543 Our PCA based only on adaptive SNPs showed that the NE and CE populations formed an
544 overlapping cluster whereas in PCA based on neutral SNPs the two populations separated (Figure
545 S21-22, Supporting Information). This is possibly due to the relatively lower diversifying selection
546 acting on the two populations and the existence of gene flow which can counteract the impact of
547 diversifying selection (Guichoux et al., 2013; Nosil et al., 2009). F_{st} analyses suggested that the level
548 of differentiation is much higher at loci under selection rather than neutral loci, as expected, and the
549 SW populations appears to be highly differentiated from the others at adaptive loci (Figure S23,
550 Supporting Information).

551 Functional enrichment analysis of mRNA intersecting the identified adaptive SNPs suggested that
552 these genes were mostly involved in growth and environmental stress response (Figure S25,
553 Supporting Information), although further studies and functional experiments are needed to confirm
554 these findings. Among the significantly enriched categories, the most interesting in view of climate
555 adaptation were: “regulation of response to stimulus”, “DNA helicase activity” which are enzymes
556 involved in DNA repair and have a crucial role as caretakers of the plant genome against
557 environmental damages (Sami et al., 2021) and “inorganic cation transmembrane transporter
558 activity”, as potassium, the most abundant inorganic cation in plant cells, it is essential for plant
559 growth and development and it has been shown to have a major role in resistance to drought,
560 salinity and fungal infections (Sharma, Dreyer & Ridelsberger, 2013). The 1,633 identified adaptive

561 genes also included several light-response and growth-related genes, according to our OmicsBox
562 annotation, as well as members of all the three GO categories significantly enriched in the putative
563 genes under selection detected by Solojarvi et al. (2017): transmembrane receptor protein tyrosine
564 kinase, peptidyl-histidine phosphorylation, and axis specification. Furthermore, 87 of the 1,633 genic
565 regions detected in this study were identified as putative selective sweeps in *B. pendula* (Salojarvi et
566 al., 2017). Even though it would be interesting to explore all the identified genes singularly, this goes
567 beyond the extent of this work.

568 In line with similar studies in other species (Jordan et al., 2017; Rellstab et al., 2016; Borrell et al.,
569 2019; Pina-Martins et al., 2018), local adaptation appears to be a highly polygenic in *B. platyphylla*,
570 therefore it is likely that maintaining the standing variation and adaptive diversity may be a better
571 solution to aid future climate adaptation, rather than focusing on raising the frequencies of a specific
572 set of adaptive SNPs, as it was proposed previously (Jordan et al. 2017). Two variables, "mean
573 diurnal range" and "isothermality" reported a substantially larger number of significant SNPs
574 association compared to the other climatic variables (Figure S20, Supporting Information), which
575 may have arisen due to the spatial correlation of these environmental variables and genotypes
576 across the sampling range therefore these results should be interpreted with caution and further
577 validation is necessary, however we cannot exclude and highly polygenic control of these variables
578 with many SNPs with low effects involved. We did not identify any significant correlation between
579 the number of SNPs associated with each climate variable, and their percentage contribution in our
580 ENM, unlike some other studies (Borrell et al., 2019; Rellstab et al., 2016). However, we note that it
581 is not a logical necessity as the variables with higher discriminatory power in the ENM could be
582 limiting species ranges precisely because they lack adaptation (Borrell et al., 2019).

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

586 (Figure 5), with the exception of “annual mean temperature” and “mean temperature of wettest
587 quarter”, very likely reflecting the larger relative projected change of temperature derived variables
588 compared with that of environmental variables related to precipitation throughout China. The
589 expected allelic frequencies changes are similar to those reported in other studies (Jordan et al.,
590 2017; Pina-Martins et al., 2019; Rellstab et al., 2016). Weighted mean RONA incorporated the ENM
591 with EAA and generally remains low and it is always below 0.2 (Figure 6A). We also reported the
592 maximum RONA per individual out of the seven variables tested (Figure 6B) because even if a
593 population, in this case an individual, has “low” average RONA for a given future projection, it might
594 still be its highest RONA that will determine how much it will need to shift its allelic frequencies in
595 order respond to future selective pressure (Pina-Martins, Baptista, Pappas & Paulo, 2018). Maximum
596 RONA across individuals showed significantly higher values, as expected. Interestingly, the maximum
597 RONA plot also shows particularly high values for four individuals in the south-western population
598 located at the very southern edge of this population distribution, reported for precipitation
599 seasonality which is expected to have a much larger relative change at these locations (Figure 6B).
600 Overall, both weighted mean and maximum RONA match well with our ENM for ssp370 2080-2100,
601 with lower RONA values reported within the inferred projected suitable environment (Figure 6).

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

612 It is not straightforward to determine when RONA is too high to result in a lag between allele
613 frequencies and adaptation however previous studies on *Fagus sylvatica* have observed frequencies
614 changes of 0.1 – 0.2 per decade in populations (Jump, Hunt, Martinez-Izquierdo, & Peñuelas, 2006),
615 therefore, following suggestions (Jordan et al., 2017) we can hypothesize that frequencies changes <
616 0.1 may be matched by populations naturally, whereas larger changes may require some sort of aid
617 such as assisted gene flow or assisted migration (Borrell et al., 2019). However, the above statement
618 is highly dependent on the population size and generation time (Gain & Francoise, 2021), therefore
619 it should be treated as a rough indication.

620

621 **Concluding remarks**

622 This study provides insights into both the evolution and ecology of Asian white Birch, *B. platyphtylla*,
623 in China. We showed that *B. platyphylla* and *B. pendula* are better considered distinct species, as
624 their differentiation is clear at the genomic level. We identified three distinct lineages of *B.*
625 *platyphylla* in China distributed along a latitudinal gradient from the north-east to the south-west,
626 and a hybrid zone between *B. platyphylla/B. pendula* in north-western China. The three *B.*
627 *platyphylla* lineages appear to have had a long period of isolation from one another, followed by
628 more recent gene flow. Our species distribution model shows that *B. platyphylla* prefers
629 mountainous environments with cool summer and moderate levels of precipitations. Future
630 prediction shows significant reduction on the suitable habitat of this species in China under every
631 future scenario tested, up to a third reduction compared to the current extent. We identified signals
632 of local adaptation throughout the genome of this species, suggesting that climate control is a highly
633 polygenic mechanism. We estimated the degree of maladaptation to inform conservation strategies.
634 Despite the limitation of our method, we showed that rapidly rising temperature, particularly in the
635 summer months, poses a risk to this species and our environmental niche model suggests that
636 current habitats located in central China and in the north-eastern region adjacent to the Changbai

637 mountains will likely become unsuitable by 2080-2100, while current habitats located at higher
638 elevations, such as the south-western Tibetan plateau and the region surrounding the Greater
639 Khingan mountains range in the north-east will remain suitable. It is possible that populations
640 outside the predicted suitable environment may be able to adapt, however climate, particularly
641 temperature, might change to an extent that adaptation will not be possible and assisted migration
642 might be the only option possible to rescue some populations. The estimated required changes in
643 allelic frequencies at adaptive loci do not appear as concerning within the predicted suitable
644 environment, however with the current dataset, based on individuals' allelic frequencies, it is
645 difficult to estimate with confidence an allelic frequency change threshold after which populations
646 will fail to adapt on time naturally and will require assisted gene flow from donor pre-adapted
647 populations.

648

649 **Acknowledgements**

650 This work was funded by the National Natural Science Foundation of China (31770230 and
651 31600295) to Nian Wang. Gabriele Nocchi is on a PhD programme run jointly between Queen Mary
652 University of London and the Royal Botanic Gardens, Kew funded by the UK Government
653 Department for Environment, Food and Rural Affairs (Defra), in association with Action Oak.

654 We thank Stefan Götz, David Seide and the BioBam Bioinformatics team for kindly letting us use
655 their software OmicsBox for the functional annotation and functional enrichment analysis of the
656 identified adaptive genes. We thank Richard Nichols (Queen Mary University of London), Yiye Liang
657 and Baosheng Wang (South China Botanical Garden, Chinese Academy of Sciences) for their expert
658 advice on our analyses.

659 This research utilised Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT:
660 http://doi.org/10.5281/zenodo.438045.

661

662 **References**

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

666 2. Alberto, F., Aitken, S., Alía, R., González-Martínez, S., Hänninen, H., & Kremer, A. et al. (2013).
667 Potential for evolutionary responses to climate change – evidence from tree populations.
668 *Global Change Biology*, *19*(6), 1645-1661.

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

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

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

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

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

680 8. Caye, K., Deist, T., Martins, H., Michel, O., & François, O. (2016). TESS3: fast inference of
681 spatial population structure and genome scans for selection. *Molecular Ecology Resources*,
682 *16*(2), 540-548.

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

686 10. Chang C. C., Chow, CC, Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015)
687 Second-generation PLINK: rising to the challenge of larger and ricer datasets. *GigaScience*,
688 *4*(1)

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

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

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

696 14. Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M., … Durbin, R. and
697 1000 Genomes Project Analysis Group. (2011). The Variant Call Format and VCFtools.
698 *Bioinformatics*, *27*(15), 2156–2158.

15. 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.

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

17. 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.

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

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

20. 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.

21. 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.

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

23. 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.

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

25. 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.

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

27. 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.

28. 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.

29. 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.

30. 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.

31. Jorhi, P., Eyre-Walker, A., Gutenkunst, R. N., Lohmueller, K. E., & Jensen, J. D. (2022). On the prospect of achieving accurate joint estimation of selection with population history. *Genome Biology and Evolution*, evac088.

32. 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.

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

34. Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics*, *15* (1), 356.

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

36. 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.

37. 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.

38. 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.

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

40. 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.

41. 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.

42. 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.

43. 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.

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

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

46. 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.

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

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

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

50. 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.

51. 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.

52. 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.

53. 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.

54. 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.

55. Schiffels, S., & Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, *46* (8), 919-925.

56. 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.

57. 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.

58. 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. http://github.com/jdstorey/qvalue

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

60. 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.

61. 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.

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

63. 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.

64. 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.

65. 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.

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

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

68. 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.

Data accessibility

Sequencing read data are archived in the NCBI through the BioProject accession PRJNA854976.

Benefit sharing statement

This research was led by scientists from the country providing genetic samples. As described above, all data have been shared with the broader public via appropriate biological databases.

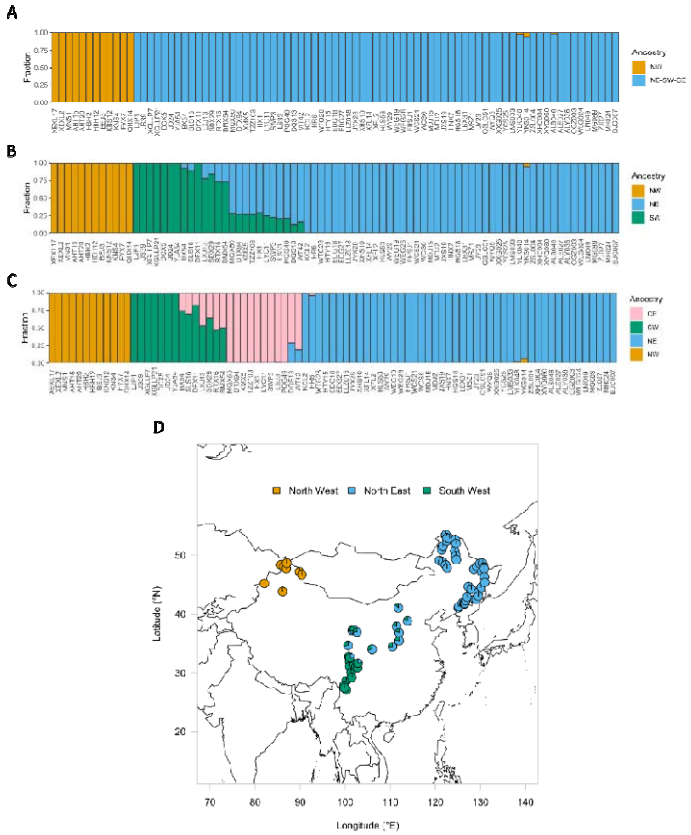
Author contributions

RB and NW conceived the project. NW and JD performed the sampling and extracted DNA. LY, JD, and YG performed mapping and SNPs calling. GN and JW analysed the data and wrote the manuscript. RB oversaw the project and helped write the manuscript.

847 **Figures and Tables**

848 **Figure 1.** *fastSTRUCTURE* results for the China dataset. **A)** Bar plot of *fastSTRUCTURE* results of 83
849 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)**
850 Map of the individuals showing their genetic ancestry composition according to fastSTRUCTURE at K
851 = 3.

852



853

854

855

856

857

858

859

860

861

862

863

864

Figure 2. *fastSTRUCTURE* results for the Eurasian dataset. **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.

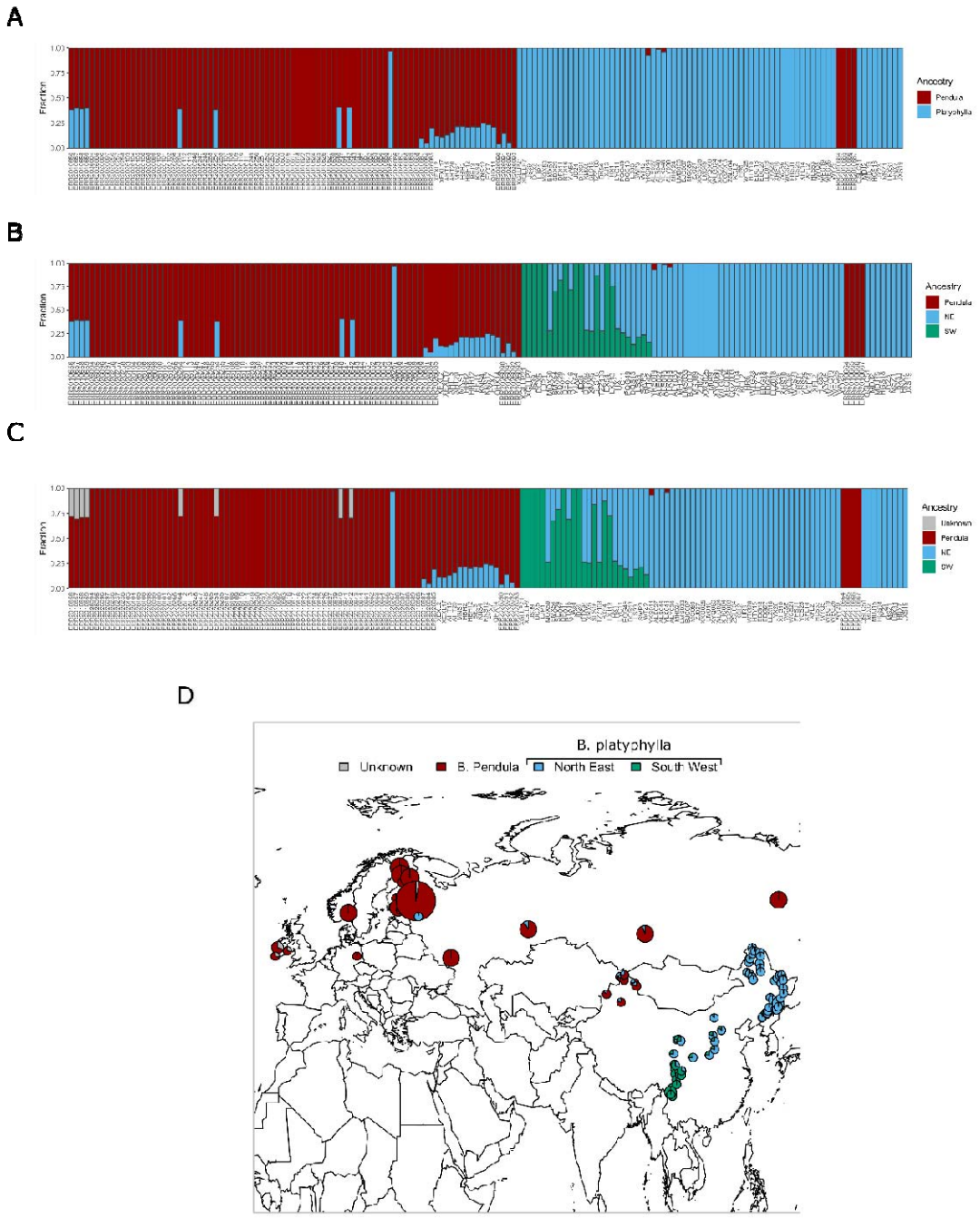


Figure 3. *fastSTRUCTURE*, *MSMC* and *fastsimcoal* results for Chinese *B. platyphylla*. **A)** Bar plot of *fastSTRUCTURE* results of 71 *B. platyphylla* individuals, based on 1,387,994 unlinked SNPs ($r^2 < 0.4$) at K = 2, and **B)** K=3. **C)** Map of the sampled individuals showing their genetic ancestry composition according to *fastSTRUCTURE* at K = 3. **D)** A demographic scenario of the NE, CE and SW lineages modelled using fastsimcoal2. Shown in light grey are the ancestral populations. The arrows show the per generation migration rate (m) between the three lineages. **E)** Historical effective population size of the three populations estimated by MSMC. Solid lines and shading represent medians and ±standard deviation calculated across pairs of haplotypes, respectively.

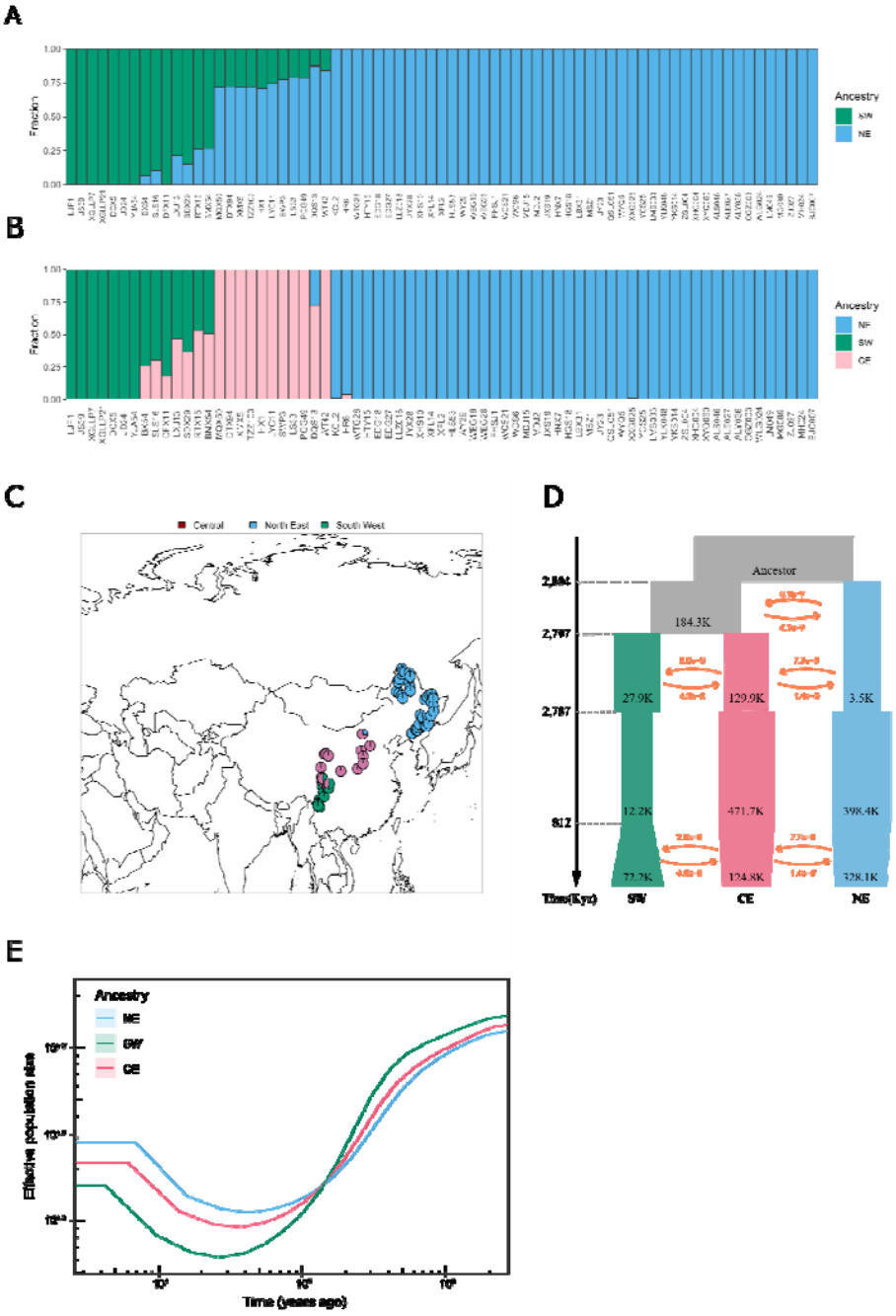


Figure 4. Environmental niche model (ENM) of *B. platyphylla* in China. Red points represent the individuals sampled in this study included in the model, blue points represent observations downloaded from the GBIF (http://www.gbif.org). **A)** Present time ENM showing the HSI throughout China. **B)** Binomial representation of the present time habitat suitability index (HSI), using the maximum training sensitivity plus specificity threshold of 0.2. **C)** Binomial HSI projections (> 0.2) of *B. platyphylla* under four different climate scenarios (ssp126, ssp245, ssp370 and ssp585) at two different time points (2041-2060 and 2080-2100). The percentage in each plot shows the decrease in suitable environment area compared with the current habitat.

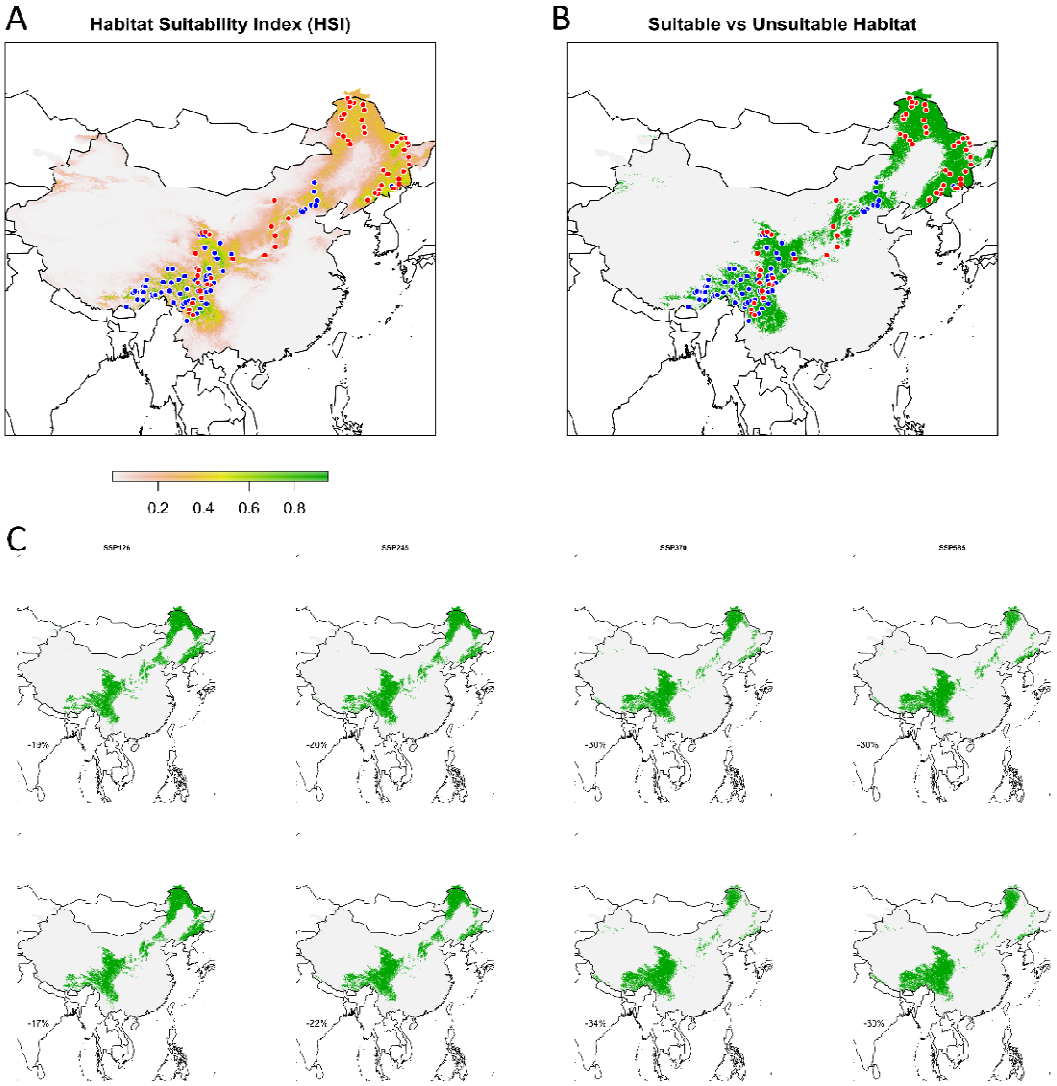


Figure 5. Risk of non-adaptedness (RONA) for the 71 *B. platyphylla* individuals**.** RONA calculated independently for each of the seven climatic variables projected to change in the future. The future climate profile ssp370 at 2080-2100 was used for RONA calculation. AD: admixed individuals. AMT: annual 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.

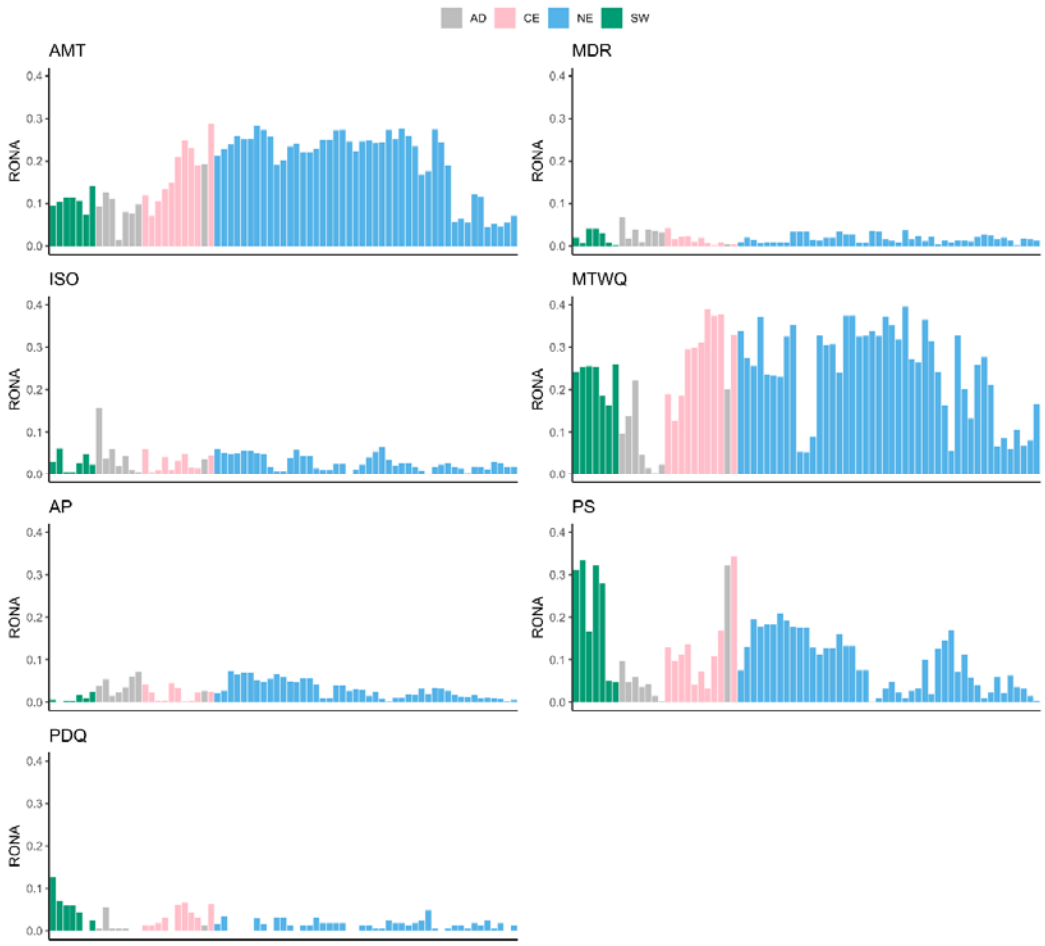
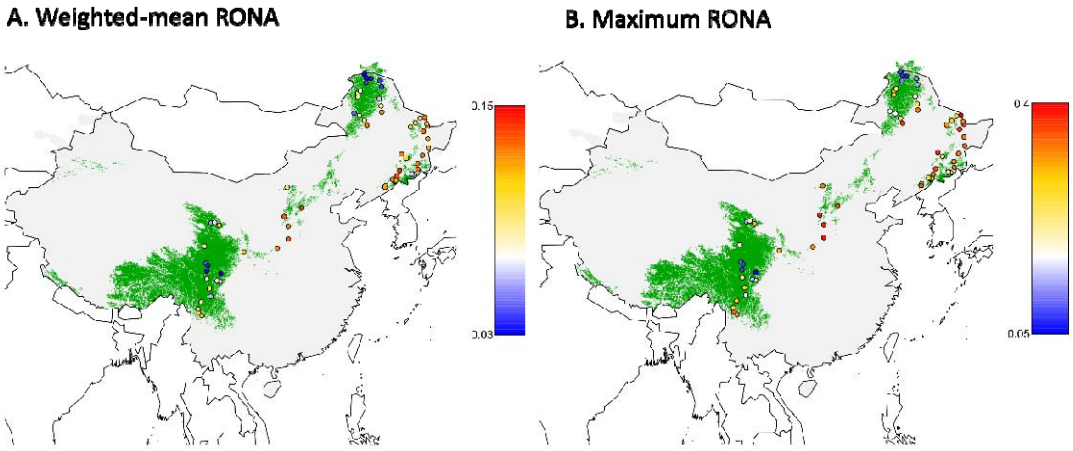



Figure 6. Weighted-mean RONA and MAX RONA. **A)** Map showing weighted mean RONA for the 71 *B. platyphylla* individuals, out of the seven environmental variables included in the analysis. Each variable RONA was given a weight equivalent to its percent contribution in the ENM. **B)** Map showing maximum RONA per individual, out of the seven environmental variables tested. Green shading represents suitable environment in 2080-2100 under profile ssp370 according to our ENM.