Evidence for widespread selection in shaping the genomic landscape during speciation of *Populus*

Jing Wang^{1*}, Nathaniel R. Street², Eung-Jun Park³, Jianquan Liu¹, Pär K. Ingvarsson⁴

¹ Key Laboratory for Bio-resources and Eco-environment, College of Life Science, Sichuan University, Chengdu, China

² Umeå Plant Science Centre, Department of Plant Physiology, Umeå University, 90187 Umeå, Sweden

³ Department of Bioresources, National Institute of Forest Science, Suwon 16631, Republic of Korea

⁴ Department of Plant Biology, Uppsala BioCenter, Swedish University of Agricultural Sciences, PO Box 7080, 75007, Uppsala, Sweden

*Correspondence: wangjing2019@scu.edu.cn

Running title: Genomic impact of selection in *Populus*

1 **Abstract**

2

3 Increasing our understanding of how various evolutionary processes drive the genomic
4 landscape of variation is fundamental to a better understanding of the genomic
5 consequences of speciation. However, the genome-wide patterns of within- and
6 between- species variation have not been fully investigated in most forest tree species
7 despite their global ecological and economic importance. Here, we use whole-genome
8 resequencing data from four *Populus* species spanning the speciation continuum to
9 reconstruct their demographic histories, investigate patterns of diversity and divergence,
10 infer their genealogical relationships and estimate the extent of ancient introgression
11 across the genome. Our results show substantial variation in these patterns along the
12 genomes although this variation is not randomly distributed but is strongly predicted by
13 the local recombination rates and the density of functional elements. This implies that
14 the interaction between recurrent selection and intrinsic genomic features has
15 dramatically sculpted the genomic landscape over long periods of time. In addition, our
16 findings provide evidence that, apart from background selection, recent positive
17 selection and long-term balancing selection are also crucial components in shaping
18 patterns of genome-wide variation during the speciation process.

19

20 **Keywords:** Linked selection, Recombination, Incomplete lineage sorting, Phylogenetic
21 relationship, Ancient introgression, *Populus*

22 **Introduction**

23

24 Determining the evolutionary forces affecting patterns of genome-wide variation has
 25 been a central goal in evolutionary biology over the past several decades (Seehausen et
 26 al., 2014). Furthermore, studying variation in levels of differentiation within and
 27 between closely related species has the potential to yield important insights into the
 28 process of speciation (Ravinet et al., 2017; Wolf & Ellegren, 2017). Studies in a broad
 29 range of taxonomic groups have revealed a picture of a highly heterogeneous genomic
 30 landscape with peaks and valleys of diversity and differentiation (Han et al., 2017;
 31 Nadeau et al., 2012; Stankowski et al., 2019; Turner, Hahn, & Nuzhdin, 2005). Local
 32 peaks of elevated divergence are usually referred to as ‘speciation islands’ and are
 33 thought to represent regions that drive the reproductive isolation between incipient
 34 species (Abbott et al., 2013; Wu, 2001). Between these islands, gene flow acts to
 35 homogenize the remainder of genome and hence acts to limit differentiation (Feder,
 36 Egan, & Nosil, 2012; Nosil, Funk, & Ortiz - Barrientos, 2009). However, a plethora of
 37 recent studies highlight that the heterogeneous patterns of differentiation can evolve
 38 through processes that are unrelated to speciation *per se* (Burri et al., 2015; Cruickshank
 39 & Hahn, 2014). For example, even in the absence of gene flow, natural selection, in the
 40 form of either a selective sweep or background selection, can cause reduced genetic
 41 diversity not only at the target sites under selection but also at linked neutral sites (Han
 42 et al., 2017; Phung, Huber, & Lohmueller, 2016). Such selection could accelerate
 43 lineage sorting and will hence inevitably result in increased genetic differentiation
 44 between species in these regions (Burri, 2017). Furthermore, the long-term action of
 45 linked selection in ancestral as opposed to extant lineages can also affect the amount
 46 and distribution of ancestral polymorphisms (Ma et al., 2018; Munch, Nam, Schierup,

47 & Mailund, 2016; Scally et al., 2012), which can further result in heterogeneous
 48 patterns of genealogical relationships among closely related species (Mailund, Munch,
 49 & Schierup, 2014; Pease & Hahn, 2013). Despite widespread interest in speciation
 50 genomics, there remains little consensus as to how various evolutionary processes have
 51 shaped the genomic landscape during the speciation process that eventually gives rise to
 52 new species (Ravinet et al., 2017).

53 Empirical studies suggest that the formation of the genomic landscape of
 54 diversity during speciation is highly influenced by the demographic histories of the
 55 species, the types of selection acting on different genomic regions and also several other
 56 intrinsic genomic features (Burri, 2017; Ellegren & Galtier, 2016). Disentangling the
 57 effects of speciation (i.e. species split time, strict isolation or divergence with gene flow)
 58 is important for interpreting the patterns of genome-wide variation, because without a
 59 clear picture of the demographic history of the descendant species, it is challenging to
 60 distinguish whether heterogeneous genomic differentiation arose due to genetic drift,
 61 local adaptation or introgression (Nadachowska-Brzyska et al., 2013; Ravinet et al.,
 62 2018). Furthermore, as the speciation process advances, the evolution of genome-wide
 63 patterns of variation can be influenced by different forms of selection (Cutter & Payseur,
 64 2013). Under a background selection model, purifying selection continuously eliminates
 65 deleterious mutations, resulting in reduced levels of genetic diversity at linked loci and
 66 increased levels of F_{ST} (a relative measure of genetic divergence) (Charlesworth, 2012;
 67 Charlesworth, Morgan, & Charlesworth, 1993; Hudson & Kaplan, 1995). Under a
 68 selective sweep model, genetic variants linked to beneficial mutations acted upon by
 69 positive selection hitchhike along and reach high frequency (Kaplan, Hudson, &
 70 Langley, 1989; Smith & Haigh, 1974). Accordingly, even in the absence of gene flow,
 71 selection due to, for instance local ecological adaptation, can result in reduced diversity

72 and increased F_{ST} (Cruickshank & Hahn, 2014). In comparison to purifying and positive
 73 selection, long-term balancing selection favors the maintenance of advantageous
 74 polymorphisms for many generations, which instead result in genomic regions with
 75 elevated genetic diversity and reduced F_{ST} (Charlesworth, 2006; Guerrero & Hahn,
 76 2017). As deleterious mutations are assumed to be much more common compared to
 77 beneficial mutations, background selection has been argued to play a major role in the
 78 evolution of diversity (Burri, 2017; Lohmueller et al., 2011; Phung et al., 2016).
 79 However, many recent simulation studies have shown that background selection alone
 80 is far from sufficient for generating the heterogeneous genomic landscapes observed in
 81 empirical studies of recently diverged species, and other evolutionary processes (such as
 82 positive selection) are thus required to explain the observed patterns (Matthey - Doret
 83 & Whitlock 2019; Stankowski et al., 2019).

84 Regardless of the role of demographic processes and selection, genomic
 85 features, such as recombination rate variation and the heterogeneous density of
 86 functional sites, are also expected to play key roles in mediating the efficacy and extent
 87 of selection and gene flow, as well as how these processes interact as the speciation
 88 process proceeds (Flaxman, Wacholder, Feder, & Nosil, 2014; Hurst, Pál, & Lercher,
 89 2004; Nachman & Payseur, 2012). Local rates of recombination interacts with natural
 90 selection and are known to have a profound effect on patterns of genomic diversity,
 91 incomplete lineage sorting (ILS) and rates of introgression (Begun & Aquadro, 1992;
 92 Comeron, Williford, & Kliman, 2008; Cutter & Payseur, 2013). Independent of the
 93 recombination rate, the density of functional sites can also influence genome-wide
 94 patterns of diversity since functional regions are more likely to experience either
 95 stronger effects of positive or purifying selection compared to nonfunctional regions
 96 where mutations are assumed to have little effect on fitness (Al-Shahrour et al., 2010;

97 Nordborg et al., 2005). The long-term diversity-reducing effects of selection in
 98 functional regions will reduce locally effective population size (N_e), accelerate lineage
 99 sorting and increase genetic divergence between species (Flowers et al., 2011; Hobolth,
 100 Dutheil, Hawks, Schierup, & Mailund, 2011). As it becomes increasingly feasible to
 101 generate whole genome resequencing data from closely related species, the importance
 102 of conserved genomic features in shaping the topography of the genomic landscape of
 103 speciation has increasingly been highlighted by several studies in a diverse set of taxa
 104 showing highly correlated patterns of differentiation among independently species pairs
 105 (Burri, 2017; Delmore et al., 2018; Van Doren et al., 2017; Vijay et al., 2017).

106 Forest trees provide an excellent system to address the genomic architecture of
 107 adaptation and speciation in natural populations because they are mostly
 108 undomesticated without much anthropogenic influence, ecologically important across a
 109 wide variety of habitats and harbour abundant genetic and phenotypic variation (Neale
 110 & Ingvarsson, 2008; Neale & Kremer, 2011). In this study, we focus on four *Populus*
 111 species (*Populus tremula*, *P. davidiana*, *P. tremuloides* and *P. trichocarpa*) that span
 112 the speciation continuum. All four species are all deciduous, obligated outcrossing tree
 113 species that have wide geographical distributions throughout the Northern Hemisphere
 114 (Figure 1A). Among them, *P. tremula* (European aspen), *P. davidiana* (Chinese aspen)
 115 and *P. tremuloides* (American aspen) are sibling aspen species belonging to the same
 116 section of the genus *Populus* (section *Populus*) (Eckenwalder, 1996; Hamzeh &
 117 Dayanandan, 2004). Earlier phylogenetic studies have revealed that *P. tremuloides*
 118 diverged from the other two species following the break-up of the Bering Land bridge,
 119 whereas the uplift of the Qinghai-Tibetan Plateau and the associated climate oscillations
 120 may have driven the divergence between *P. tremula* and *P. davidiana* (Du et al., 2015).
 121 In addition, these aspen species can readily hybridize and their artificial hybrids show

heterosis for many growth and wood characteristics (Hart, De Araujo, Thomas, & Mansfield, 2013), suggesting that the speciation process has not gone to completion among the three aspen species. In comparison, *P. trichocarpa* belongs to a different section of the genus *Populus* (section *Tacamahaca*), and it is reproductively isolated from all aspen species (Jansson & Douglas, 2007). Facilitated by the availability of a high-quality reference genome of *P. trichocarpa* (Tuskan et al., 2006), the four *Populus* species represent a promising model system to investigate how various evolutionary forces have shaped the evolution of the genomic landscape of differentiation across the speciation continuum in forest trees.

We use whole-genome re-sequencing in the four *Populus* species to (i) determine their speciation history and characterize whether there is historical gene flow between the now-allopatric species; (ii) examine the fine-scale genomic landscapes of diversity and divergence across species at different stages of divergence; (iii) quantify the extent of genome-wide genealogical discordance and ancient hybridization among the three closely related aspen species; (iv) identify the signatures of positive selection and long-term balancing selection along the genome, and uncover how they impact levels of variation during speciation. Overall, our main aim is to disentangle and understand how the multitude of evolutionary processes have shaped the genomic architecture during speciation.

2. Materials and Methods

2.1 Sample collection, whole-genome resequencing and genotype calling

We used whole genome resequencing data from eight individuals each of *Populus tremula*, *P. tremuloides* and *P. trichocarpa*, as described in Wang et al. (2016a), and

147 additional eight individuals of *P. davidiana* that are first reported in this study (Table
148 S1). The sampling was from a single geographic region for each species (Figure 1A).
149 Briefly, sequencing of all samples was carried out on the Illumina HiSeq 2000 platform.
150 Prior to read mapping, we used Trimmomatic (Lohse et al., 2012) to remove adapter
151 sequences and to trim low quality bases from the start or the end of reads (base quality $\leq$
152 20). If the processed reads were shorter than 36 bases after trimming, the entire reads
153 were discarded. After quality control, we mapped the remaining reads from each
154 individual to the *P. trichocarpa* reference genome (v3.0) (Tuskan et al., 2006) using
155 BWA-MEM algorithm with default parameters, as implemented in bwa-0.7.10 (Li,
156 2013).

157 To minimize the influence of mapping bias, several further filtering steps were
158 employed before genotype calling. First, we used RealignerTargetCreator and
159 IndelRealigner in GATK v3.8.0 (DePristo et al., 2011) to correct for the mis-alignment
160 of bases in regions around insertions and/or deletions (indels). Second, to account for
161 the artifacts due to PCR duplication introduced during library construction, we used the
162 MarkDuplicates method from Picard packages (<http://broadinstitute.github.io/picard/>) to
163 only retain the read or read-pair with the highest summed base quality among those with
164 identical external coordinates and same insert lengths. Additionally, we further
165 discarded site types that likely cause mapping bias based on three criteria: (1) those with
166 extreme read coverage (less than 4 $\times$ or higher than twice of the mean coverage); (2)
167 covered by more than two reads of mapping score equaling zero per individual; (3)
168 overlapping known repetitive elements as identified by RepeatMasker (Tarailo -
169 Graovac & Chen, 2009). Finally, sites that passed all these filtering criteria were used in
170 downstream analyses. This left a total of 168,950,389 sites for further analysis (42.8%
171 of collinear genomic sequences of the *P. trichocarpa* genome assembly).

After filtering, we implemented two complementary approaches for genotype calling. First, to account for the bias inherent in genotype calling approach from next generation sequencing (NGS) data (Nielsen, Korneliussen, Albrechtsen, Li, & Wang, 2012), the population genetic estimates that relied on site frequency spectrum (SFS) were calculated using ANGSD v0.917 (Korneliussen, Albrechtsen, & Nielsen, 2014). Second, for the analyses that require accurate single nucleotide polymorphism (SNP) calls, genotype calling in each individual was performed using HaplotypeCaller of the GATK v3.8.0, and GenotypeGVCFs was then used to merge multi-sample records from the four species together for re-genotyping and re-annotation of the newly merged VCF (DePristo et al., 2011). To minimize genotype calling bias and to retain high-quality SNPs, we further performed several filtering steps: (1) SNPs that overlapped with sites not passing all previous filtering criteria were removed; (2) only bi-allelic SNPs with a distance of at least 5 bp away from any indels were retained; (3) genotypes with read depth (DP) < 5 and/or with genotype quality score (GQ) < 10 were treated as missing, and we then removed all SNPs with a genotype missing rate > 10%. After all these steps of filtering, a total of 8,568,990 SNPs were retained across the four *Populus* species. For the analyses that required imputed and phased dataset, BEAGLE v4.1 (Browning & Browning, 2009) was used to infer haplotypes of individuals within each species.

190

191 **2.2 Phylogenetic relationships and population structure analysis**

192 *Chloroplast phylogeny*

To infer the phylogenetic relationship of the four *Populus* species based on chloroplast data, we first mapped the filtered reads from our resequencing data against the *P. trichocarpa* chloroplast genome using bwa-aln 0.7.10 (Li & Durbin, 2009). Then, UnifiedGenotyper in GATK v3.8.0 was used to call SNPs at all sites (--output_mode

197 EMIT_ALL_SITES). Since chloroplasts are haploid and SNPs are thus expected to be
198 homozygous, the haploid option (-ploidy 1) in UnifiedGenotyper was used. After
199 treating sites with GQ < 30 as missing data, only bi-allelic SNPs with quality by depth
200 (QD) ≥ 10 and with a missing rate $\leq 20\%$ were retained. Finally, a consensus tree was
201 constructed based on 1,292 chloroplast SNPs using maximum likelihood method
202 implemented in SNPhylo (Lee, Guo, Wang, Kim, & Paterson, 2014).

203

204 *Principle component analysis (PCA)*

205 To account for the uncertainty in genotype calls, PCA was performed using ANGSD
206 v0.917 and ngsTools v1.0.1 (Fumagalli, Vieira, Linderth, & Nielsen, 2014). We first
207 used the SAMTools model (Li et al., 2009) in ANGSD to estimate genotype likelihoods
208 from BAM files using only reads with a minimal base quality score of 20 and a minimal
209 mapping quality score of 30 across all individuals. ngsTools was then used to compute
210 the expected covariance matrix across pairs of individuals for the four species based on
211 the genotype posterior probabilities across all filtered sites. Eigenvectors and
212 eigenvalues were generated with the R function eigen from the covariance matrix, and
213 the significance level was determined using the Tracy-Widom test as implemented in
214 EIGENSOFT version 6.1.4 (Patterson, Price, & Reich, 2006).

215

216 *Identity-by-descent (IBD) blocks analysis*

217 To determine the extent to which individuals across the four species shared DNA
218 segments, the identity-by-descent block analysis was performed for the four species
219 using BEAGLE v4.1 (Browning & Browning, 2013) with the following parameters:
220 window=100,000; overlap=10,000; ibdtrim=100; ibdlod=5.

221

222 **2.3 Demographic history reconstruction**

223 *MSMC*

224 We used Multiple Sequentially Markovian Coalescent approach (MSMC v2) (Schiffels
225 & Durbin, 2014) to infer patterns of historical patterns of effective population sizes
226 changes through time for all four *Populus* species. Only sites passing all above filtering
227 criteria were included in analyses. Because different number of individuals and
228 haplotypes provides different resolution for recent and distant population histories, we
229 applied MSMC to phased whole-genome sequences from one (two haplotypes, which
230 can infer more distant size changes), two (four haplotypes, which infer size changes at
231 intermediate time scales) and four (eight haplotypes, which infer the most recent size
232 changes) individuals for each species, respectively. We did not include more haplotypes
233 due to the computational cost of using larger haplotype sets. In total, we have 8, 28 and
234 70 different individual configurations for two-, four-, and eight- haplotype analyses in
235 each species. We ran MSMC on all individual configurations and estimated medians
236 and standard deviations of effective population sizes changes across time. To convert
237 the coalescent scaled time to absolute time in years, we used a mutation rate of 2×10^{-9}
238 per site per year (Koch, Haubold, & Mitchell-Olds, 2000) and a generation time of 15
239 years.

240

241 *Fastsimcoal2*

242 Given the long divergence time and the low number of polymorphic sites shared
243 between aspens and *P. trichocarpa* (Wang, Street, Scofield, & Ingvarsson, 2016a), we
244 used a coalescent simulation-based method implemented in *fastsimcoal2.6* (Excoffier,
245 Dupanloup, Huerta-Sánchez, Sousa, & Foll, 2013) to only infer the demographic and
246 speciation histories only for the three aspen species. For all possible pairs of the three

species, the two-dimensional joint SFS (2D-SFS) was constructed from the posterior probabilities of sample allele frequencies using ANGSD v0.917 (Figure S1). A total of twenty-nine models were evaluated and all models began with the split of an ancestral population into the Eurasia and the North America lineage (*P. tremuloides*) followed by the split of the Eurasian lineage into *P. tremula* and *P. davidiana*. The models differed in terms of (1) whether post-divergence gene flow was present or not, (2) time, level and pattern of gene flow between the three aspen species, and (3) the occurrence and pattern of population expansion in *P. tremuloides* given a genome-wide excess of rare frequency alleles that we observed in this species in our previous study (Wang et al., 2016a) and also in this study (Figure S2). Alternative demographic models were fitted to the joint SFS data. The global maximum likelihood estimates for all demographic parameters under each model were obtained from 50 independent runs, with 100,000 coalescent simulations per likelihood estimates (-n 100000, -N 100000) and 40 cycles of the likelihood maximization algorithm. The models were compared based on the maximum value of likelihood over the 50 independent runs using the Akaike's weight calculated following Excoffier et al. (2013). The model with the maximum Akaike's weight value was chosen as the optimal one. Confidence intervals were generated by performing parametric bootstrapping with 100 bootstrap replicates, and with 50 independent runs in each bootstrap. As for MSMC, we assumed a mutation rate of 2×10^{-9} per site per year and a generation time of 15 years (Koch et al., 2000) when converting estimates to units of years and individuals.

268

269 **2.4 Intra- and inter- species summary statistics**

270 *Intra-species genomic diversity*

271 For each species, based on the SAMTools genotype likelihood model (Li et al., 2009),
 272 we used ANGSD v0.917 (Korneliussen et al., 2014) to estimate allele frequency
 273 likelihoods, obtain a maximum likelihood estimate of the folded site frequency
 274 spectrum and used this to calculate nucleotide diversity (π) in non-overlapping sliding
 275 windows of 10 Kbp and 100 Kbp across the entire genome. Only sites with a minimum
 276 mapping quality of 30 and minimum base quality of 20 were used in the estimation.
 277 Windows were discarded if there were less than 10 % sites left after all of the filtering
 278 steps described above. Since our previous study showed that linkage disequilibrium
 279 (LD) decays within 10 Kbp in different species of *Populus* (Wang et al., 2016a), in the
 280 following we focused more on estimates derived from 10 Kbp windows.

281

282 *Inter-species genomic divergence*

283 For each species pair, we estimated two divergence metrics across the 10 Kbp and 100
 284 Kbp non-overlapping windows: genetic differentiation (F_{ST}) and sequence divergence
 285 (d_{xy}). Without relying on SNP or genotype calling (Fumagalli et al., 2013), we first used
 286 ANGSD to calculate posterior probabilities of sample allele frequency for each species.
 287 Then, the program ngsFST from the ngsTools package was used to estimate F_{ST}
 288 between species using a method-of-moments estimator, and the program ngsStat was
 289 used to calculate d_{xy} between species at each site. Finally, we averaged these divergence
 290 values across all sites within each window.

291

292 *Population-scaled recombination rate*

293 For each species we used LDhelmet v1.9 (Chan, Jenkins, & Song, 2012), a
 294 coalescent-based, reversible-jump Markov chain Monte Carlo (rjMCM) simulation
 295 method, to estimate the population-scaled recombination rate, ρ . First, VCFtools

(Danecek et al., 2011) and custom shell script were used to tailor the phased genotype data of each chromosome to the necessary input sequence file (fasta format). Then, we used ‘find_confs’ in LDhelmet to concatenate all the input sequences files and generate a haplotype configuration file per species. Thereafter, ‘table_gen’ was used to compute the likelihood lookup table for each species, where we assume the approximate genome-wide neutral diversity (θ) of 0.01 for the three aspen species and of 0.005 for *P. trichocarpa* (Wang et al., 2016a), and the grid of ρ values was specified as $-r$ 0.0 0.1 10.0 1.0 100.0 for all species. In addition, the optional ‘pade’ component of LDhelmet was included in the analysis, which computes the Padé coefficients ($-x$ 11) from the haplotype configuration file. Finally, we ran LDhelmet with window size of 50 SNPs and block penalty of 50 for a total of 1,000,000 iterations, discarding the first 100,000 as burn-in. We then calculated weighted average of the estimated ρ in 10 Kbp and 100 Kbp windows, respectively. Windows with less than 50 SNPs (for 10 Kbp windows) and 200 SNPs (for 100 Kbp windows) left from previous filtering steps were discarded.

2.5 Window-based phylogenomic analysis

Topology weighting

As expected for a clade with rapid radiation, genealogies may vary widely across different genomic regions (Lamichhaney et al., 2015). Given that *P. trichocarpa* is distantly related from the other three aspen species (Hamzeh & Dayanandan, 2004), we used Twisst, a topology weighting method by iterative sampling of subtrees (Martin & Van Belleghem, 2017), to assess and quantify the phylogenetic discordance among the three aspen species along the genome. The genealogical relationships of these species can be defined by three possible topologies: [(*P. tremula*, *P. davidiana*), *P. tremuloides*], [(*P. tremula*, *P. tremuloides*), *P. davidiana*], [(*P. davidiana*, *P.*

321 *tremuloides*), *P. tremula*]. Using *P. trichocarpa* as the outgroup species, local
322 phylogenetic subtrees was inferred in RAxML v8.2.4 (Stamatakis, 2014) with the
323 GTRCATI model over non-overlapping 10 Kbp and 100 Kbp windows. Topology
324 weightings for each window were then computed through determining the number of
325 unique subtrees that match each of the three possible topologies by iteratively sampling
326 a single haplotype from each species (Martin & Van Belleghem, 2017). Windows were
327 discarded in topology weighting estimation if there were < 50 SNPs and < 200 SNPs
328 left from previous filtering steps for 10 Kbp and 100 Kbp windows, respectively.

329

330 *Inference of incomplete lineage sorting*

331 Because the speciation events that resulted in aspen species were close in time (see
332 Results), we expect the lineage sorting process relating these species to be incomplete.
333 Given the three aspen species and the outgroup poplar species (*P. trichocarpa*) with the
334 relationship as (((*P. tremula*, *P. davidiana*), *P. tremuloides*), *P. trichocarpa*), we labeled
335 alleles in *P. tremula*, *P. davidiana* and *P. tremuloides* as A (ancestral allele) if they match
336 the reference allele of *P. trichocarpa* genome, and B (derived allele) otherwise. We then
337 considered segregating sites with (((*P. tremula*, *P. davidiana*), *P. tremuloides*), *P.*
338 *trichocarpa*) patterns as AABAs, ABAAAs, BAAAs, ABBAs, BABAs and BBAAAs. The
339 two SNP patterns ABBAs and BABAs can result from incomplete lineage sorting if we
340 assume no gene flow occurred among species (Durand, Patterson, Reich, & Slatkin,
341 2011; Green et al., 2010). We calculated the level of incomplete lineage sorting (ILS) at
342 site *i* in the genome as:

343

$$344 \text{ ILS} = (C_{\text{ABBA}(i)} + C_{\text{BABA}(i)}) / h \quad (1)$$

$$345 \text{ where } h = (C_{\text{BAAA}(i)} + C_{\text{ABAA}(i)} + C_{\text{AABA}(i)} + 2(C_{\text{BBAA}(i)} + C_{\text{BABA}(i)} + C_{\text{ABBA}(i)})) / 3 \quad (2)$$

346

347 Because population samples were used for all species, at each site we used the
348 frequency of the derived allele in each species to effectively weight each segregating
349 site according to its fit to the six segregation patterns for the three aspen species
350 (Durand et al., 2011), with

351

$$352 \quad C_{BAAA(i)} = \hat{p}_{i1}(1-\hat{p}_{i2})(1-\hat{p}_{i3})(1-\hat{p}_{i4}) \quad (3)$$

$$353 \quad C_{ABAA(i)} = (1-\hat{p}_{i1})\hat{p}_{i2}(1-\hat{p}_{i3})(1-\hat{p}_{i4}) \quad (4)$$

$$354 \quad C_{AABA(i)} = (1-\hat{p}_{i1})(1-\hat{p}_{i2})\hat{p}_{i3}(1-\hat{p}_{i4}) \quad (5)$$

$$355 \quad C_{BBAA(i)} = \hat{p}_{i1}\hat{p}_{i2}(1-\hat{p}_{i3})(1-\hat{p}_{i4}) \quad (6)$$

$$356 \quad C_{BABA(i)} = \hat{p}_{i1}(1-\hat{p}_{i2})\hat{p}_{i3}(1-\hat{p}_{i4}) \quad (7)$$

$$357 \quad C_{ABBA(i)} = (1-\hat{p}_{i1})\hat{p}_{i2}\hat{p}_{i3}(1-\hat{p}_{i4}) \quad (8)$$

358

359 where $\hat{p}_{ij}$ is the frequency of the derived allele at site i in species j .

360

361 The calculation of ILS presents the counts of incomplete lineage sorting pattern
362 ($C_{ABBA(i)}$ and $C_{BABA(i)}$) normalized by the total count of segregating sites (h), which is a
363 proxy of the species tree topology height (Sally et al., 2012). We then summarized the
364 proportion of ILS over non-overlapping 10 Kbp and 100 Kbp windows or in bins with
365 varying distances from the nearest exon.

366

367 **2.6 Analyses of introgression**

368 We first tested for introgression between the three aspen species using the D -statistic,
369 also known as the ABBA-BABA tests. These tests evaluate the imbalance frequency of
370 site patterns (Durand et al., 2011; Green et al., 2010). Using *P. trichocarpa* as the

outgroup, we expect equal counts of the two site patterns (ABBA and BABA) when incomplete lineage sorting causes the site pattern discordance. If discordance is caused by introgression, one of the site patterns is expected to be more prevalent than the other. We applied two approaches to perform the D -statistic test. First, we used `-doAbbababa2` implemented in ANGSD v0.917 (Korneliussen et al., 2014) to directly count ABBA and BABA sites and calculate the D statistic without calling genotypes in non-overlapping 10 Kbp and 100 Kbp windows for the whole genome. Then, jackknife bootstrapping was conducted to estimate significance at the chromosome and whole-genome level. Second, based on allele frequencies at each SNP called by GATK, we calculated the D statistic in 10 Kbp and 100 Kbp non-overlapping windows across the genome with the python script *ABBABABAwindows.py* (https://github.com/simonhmartin/genomics_general) (Martin, Davey, & Jiggins, 2014). Following the detection of introgression among individuals at the genome level, we used a modified f -statistic (f_d) (Martin et al., 2014) to estimate the proportion of introgressed sites at the population level using *ABBABABAwindows.py* (https://github.com/simonhmartin/genomics_general) on non-overlapping 10 Kbp and 100 Kbp windows across the genome.

388

389 **2.7 Genome-wide scan for regions under positive and balancing selection in aspens**

To specifically test for the impact of positive and balancing selection on the genomic landscape during speciation of aspens, we first used a composite likelihood ratio (CLR) statistic implemented in SweepFinder2 (DeGiorgio, Huber, Hubisz, Hellmann, & Nielsen, 2016) to detect regions subject to recent positive selection or selective sweeps in each of the three aspen species. The ancestral allelic state was defined by assuming that the alleles that were the same as those found in the *P. trichocarpa* reference

396 genome was the ancestral alleles. By contrasting the likelihood of the null hypothesis,
397 based on the unfolded site frequency spectrum (SFS) calculated across the genome
398 using $-f$ option, with the likelihood of a model where the SFS has been altered by a
399 recent positive selection event, the CLR statistics was calculated in non-overlapping
400 windows of 10 Kbp. Within each species, windows with CLR values higher than the
401 99th percentile of its distribution were identified as candidate region under selection.

402 Moreover, we identified regions under long-term balancing selection by estimating
403 the summary statistics, β (beta score), which detects the clusters of variants with an
404 excess number of intermediate frequency polymorphisms (Siewert & Voight, 2017).
405 Given that the signals of long-term balancing selection usually is localized to a narrow
406 genomic region (Gao, Przeworski, & Sella, 2015), we used 1 Kbp windows to calculate
407 β values for each core SNPs in the three aspen species. We used the unfolded version of
408 β , with the ancestral and derived allelic states inferred based on comparisons with the
409 outgroup species *P. trichocarpa*. To prevent false positives, we filtered out SNPs with a
410 folded frequency lower than 20%, and defined the SNPs with extreme β scores in the
411 top 1% as significant. Furthermore, only SNPs that are significant in all three species
412 were considered as putative targets of long-term balancing selection. Finally, we binned
413 significant SNPs into 10 Kbp windows for downstream comparisons.

414 Lastly, to assess the effects of positive and balancing selection on the genomic
415 architecture of speciation, we compared outlier windows that were identified as being
416 under positive or balancing selection with the remaining genomic regions using a
417 variety of population genetic statistics, including genetic diversity, divergence, lineage
418 sorting and introgression within and between the three closely related aspen species.
419 Differences between outlier windows and the genome-wide averages for all these
420 statistics were tested using Wilcoxon ranked-sum tests. To further examine whether any

functional classes of genes were over-represented in these candidate regions, we performed gene ontology (GO) analyses using the R package topGO 2.36.0 (Alexa & Rahnenführer 2009). Fisher's exact test was used to calculate the statistical significance of enrichment, and GO terms with *P*-value lower than 0.01 were considered to be significantly enriched.

3. Results

3.1 Phylogenetic relationships, population structure and demographic history

The genome alignment resulted in an average depth of 24.6× across all individuals after quality control (Table S1). The PCA results revealed a clear distinction among the four *Populus* species (Figure S4). Based on the Tracy-Widom test, only the first three components were significant (Table S2). The first principal component (PC1; variance explained=28.79%) separated *P. trichocarpa* from the three aspen species, while the second component (PC2; variance explained=7.52%) separated *P. tremuloides* from *P. tremula* and *P. davidiana*. Finally, the third component (PC3, variance explained=5.33%) separated *P. tremula* and *P. davidiana*. The clustering and genetic relationships of the four species were also confirmed by the phylogenetic tree constructed based on the entire chloroplast genomes (Figure 1B). Moreover, we measured the number and length of shared IBD haplotypes within and between species (Figure S5, S6; Table S3, S4). Compared to between-species comparisons, we found much more extensively shared IBD haplotypes for within-species comparisons (Figure S5), although haplotypes shared within *P. tremuloides* were shorter than the other three species (Figure S6A; Table S4). This is likely owing to the higher recombination rate and more rapid decay of linkage disequilibrium (LD) in *P. tremuloides* than other species (Wang et al., 2016a). For the between-species comparisons, we did not observe

any haplotype sharing between the three aspen species and *P. trichocarpa*, confirming the distant relationship between aspens and poplars in the genus *Populus* (Figure S5, Table S3). Within aspens, *P. tremula* and *P. davidiana* shared more and longer haplotypes than either of them shared with *P. tremuloides* (Figure S5, S6B; Table S3, S4), which also supports a closer relationship between these two species, as identified in the chloroplast phylogeny.

To investigate the demographic and speciation histories of the four *Populus* species, we first used MSMC to examine historical fluctuations in the effective population size (N_e) for each species. The results showed that all species experienced a period of population decline during the early Pleistocene cooling (2.5-0.9 million years ago, Mya) (Figure 1C). Compared with the three aspen species, *P. trichocarpa* experienced a more dramatic population decline during this period (Figure 1C, Figure S7), which likely explain the much lower genetic diversity observed in this species relative to others (Table S5). The two North American species, *P. tremuloides* and *P. trichocarpa*, experienced a population expansion from the start of the last ice age (110 thousand years ago, Kya) until the last glacial maximum (LGM, 23-18 Kya) whereas the European species *P. tremula* remained relatively stable. On the other hand, the eastern Asian species, *P. davidiana*, showed pattern of population decline during the entire period (Figure 1C, Figure S7). Our results therefore revealed that before the LGM, forest trees distributed in different continents experienced asynchronous demographic responses to Pleistocene climate changes (Bai et al., 2018). During and following the LGM, all four species experienced a population decline followed by a subsequent rapid population expansion (Figure 1C).

Because of the distant phylogenetic relationship and low levels of shared polymorphism between *P. trichocarpa* and the three aspen species (Wang et al., 2016a),

we therefore explicitly focused on inferring the demographic parameters of the speciation history for the three aspens. After evaluating a total of twenty-nine models (Figure S8), the best-supported model (Figure S8; Table S6) suggests that the Eurasian lineage (the common ancestor of *P. tremula* and *P. davidiana*) split from the North American lineage (*P. tremuloides*) at ~ 2.4 Mya (bootstrap range [BP]: 2.1-3.2 Mya), which is in accordance with our earlier estimates on the divergence time between *P. tremula* and *P. tremuloides* (Wang, Street, Scofield, & Ingvarsson, 2016b). The European lineage (*P. tremula*) and the East-Asian lineage (*P. davidiana*) diverged ~1.7 Mya (BP: 1.5-2.1 Mya) (Figure 1D, Table S7). Our results detected low levels of ancient gene flow between *P. tremula* and *P. davidiana*, and between *P. tremula* and *P. tremuloides* following speciation until around 847 Kya (BP: 539Kya-1.0 Mya) (Figure 1D). After this period the species have remained isolated which is also reflected by their current disjunct geographic distributions (Figure 1A). Compared to the Eurasian lineage of aspens, *P. tremuloides* has experienced a notable population expansion in the recent past (~772 Kya, BP: 440-887 Kya), which is consistent with its genome-wide excess of rare alleles (Figure S2).

487

3.2 General patterns of genome-wide diversity and differentiation

We further characterized genome-wide patterns of nucleotide diversity (π), population recombination rate (ρ) and divergence (F_{ST} and d_{xy}) for the four *Populus* species (Figure 2A; Table S5, Table S8-S10). At the species level, π varied markedly between species, ranging from 0.0063 in *P. trichocarpa* to 0.0148 in *P. tremuloides*, but the average genomic diversity was very similar across the three aspen species (Table S5). In contrast to the patterns observed for π , the population-scaled recombination rate, ρ , was much higher in *P. tremuloides* (0.0273 bp⁻¹) than in the other three species

(0.0096 bp⁻¹-0.0139 bp⁻¹) (Table S8). Variation in genetic divergence (F_{ST} and d_{xy}) among the six species pairs reveals the continuous nature of differentiation along the speciation continuum, with *P. tremula* and *P. davidiana* showing the lowest levels of divergence and with the highest divergence observed between aspens and *P. trichocarpa* (Figure 2A, Table S9, S10).

At the genome level, patterns of genetic diversity and divergence show high levels of parallelism in all pairwise comparisons. The genome-wide profiles of π (average Spearman's $\rho=0.71$) and ρ (average Spearman's $\rho=0.18$) were positively correlated in all possible species pairs (Figure 2B, Table S11). We found little evidence for an association between either π or ρ and the local mutation rate (μ , approximated by the four-fold synonymous substitution rate) (Figure 2B, Table S11). Hence, the broad-scale variation in genetic diversity is conserved across the diverging lineages, which likely arise from a common genomic architecture where linked selection has played a major role in shaping local genomic diversity (Burri, 2017). This is further highlighted by the conserved landscape of recombination rate variation across the genomes of the species and the strong degree of genome synteny that we observed between the genomes of aspens and poplars (Lin et al., 2018). Second, we found that the differentiation landscapes were highly correlated among all species pairs both for the relative (F_{ST}) and the absolute (d_{xy}) measures of genetic differentiation (Figure 2B, Figure S9, Table S11). The highly similar landscape of differentiation among different species pairs could imply phylogenetically conserved genomic features, e.g. conserved landscapes of functional densities and recombination (Burri, 2017; Vijay et al., 2017). Moreover, significantly negative correlations between F_{ST} and π were found in all pairwise comparisons (Figure 2B, Figure S9, Table S11), which is in line with the observation that F_{ST} is sensitive to intra-specific nucleotide diversity (Charlesworth, 1998). In

contrast, only weak correlations were observed between d_{xy} and π . Because d_{xy} largely reflects diversity in a common ancestor (Cruickshank & Hahn, 2014), a weak correlation between d_{xy} and π implies that ancestral diversity might have little impact on extant diversity in the different *Populus* species. In addition to extant diversity, d_{xy} was only weakly correlated with F_{ST} across all comparisons (Figure 3B, Figure S9, Table S11), which further implies that ancestral polymorphisms have had limited contribution to the genomic divergence of extant species (Cruickshank & Hahn, 2014).

3.3 Topology weighting reveals phylogenetic discordance and ancient introgression between *P. tremula* and *P. tremuloides*

Even if the analyses of current population structure and genomic divergence support a clear species relationship for the four *Populus* species, (((*P. tremula*, *P. davidiana*), *P. tremuloides*), *P. trichocarpa*), we used a topology weighting approach to explore to what degree the ‘species tree’ was congruent across the entire genome. Using *P. trichocarpa* as an outgroup, our results reveal widespread incongruence in local genealogies in either 10 Kbp or 100 Kbp non-overlapping windows across the genome (Figure 3, Figure S10). The most prevalent topology, ((*P. tremula*, *P. davidiana*), *P. tremuloides*), which reflects the likely ‘species topology’, has an average weighting of 54.7% and 76.5% across the genome in 10 Kbp and 100 Kbp windows, respectively. Of the other two topologies, the ((*P. tremula*, *P. tremuloides*), *P. davidiana*) topology was much more common (27.0% and 17.6% for 10 Kbp and 100 Kbp windows) compared to ((*P. davidiana*, *P. tremuloides*), *P. tremula*) (18.3% and 5.9% for 10 Kbp and 100 Kbp windows) (Figure 3, Table S12). In general, we observed that larger windows (100 Kbp) produced higher rates of monophyly (windows with a weighting of 1) and a

greater fraction of resolved trees compared to the smaller windows (10 Kbp) (Figure S10, Table S13).

Interestingly, in contrast to all other chromosomes where all three topologies were observed, chromosome 19, which is known to harbor the sex determination region in *Populus* (Yin et al., 2008), showed only a single monophyletic grouping of the ‘species topology’ (Figure 3). Such a pattern is consistent with the expectation that lineage sorting is faster on sex chromosomes compared to autosomes because of its smaller effective population size (Meisel & Connallon, 2013; Vicoso & Charlesworth, 2006). Overall, both incomplete lineage sorting (ILS) and introgression can result in discordance between the local topology and the species tree for recently diverged species. Given that ILS is expected to generate equal frequencies of the alternative topologies (Durand et al., 2011; Mailund et al., 2014), the more frequent topology of ((*P. tremula*, *P. tremuloides*), *P. davidiana*) is likely explained by the occurrence of introgression between *P. tremula* and *P. tremuloides*. We therefore compared the distribution of the branch lengths separating each pair of aspen species among all topology types. Compared to the expectation that species with recent introgression tend to be separated by short branches (Fontaine et al., 2015; Martin & Van Belleghem, 2017), the branch distances between *P. tremula* and *P. tremuloides* were not obviously different from other species-pairs across topology comparisons (Figure S11). This pattern is most likely caused by ancient hybridization between these two species where genetic drift has eradicated most signatures of gene flow after an ancient introgression event (Schumer, Cui, Powell, Rosenthal, & Andolfatto, 2016).

To further investigate patterns of ancient introgression between *P. tremula* and *P. tremuloides*, we calculated two statistics associated with the ABBA-BABA test across the genome. The *D*-statistic is used to test for ancient gene flow by comparing the

570 imbalance of ABBAs and BABAs, and the f_d -statistic is used to estimate the fraction of
571 the genome shared through ancient introgression. For the D -statistics, we also
572 implemented two different approaches, which differed in whether the called genotypes
573 was relied or not. We find that the estimates of the two approaches are highly correlated
574 with each other (Figure S12), suggesting that this statistic is robust to identify
575 introgression regardless of which type of data is used. Genome-wide estimates of the
576 D -statistic and f -statistic showed a general pattern of positive values over 10 Kbp and
577 100 Kbp non-overlapping windows (Table S14), confirming that *P. tremuloides* has a
578 closer genetic relationship with *P. tremula* than with *P. davidiana*. Thus, the significant
579 asymmetry in genetic relationship together with the excess of shared sequence
580 polymorphism between *P. tremula* and *P. tremuloides* (Figure S13) all provide evidence
581 for historical gene flow between the currently allopatric Eurasian and North American
582 aspen species.

583

584 ***3.4 Long-term effects of selection in shaping patterns of diversity, divergence,*** 585 ***incomplete lineage sorting (ILS) and levels of introgression in Populus species***

586 To evaluate the impact of natural selection on genetic diversity, divergence, ILS and
587 gene flow in the context of speciation, we examined the correlations between these
588 genetic parameters and factors affecting the extent and efficiency of selection. First,
589 regions with a high density of potential targets for selection are expected to experience
590 stronger linked selection simply because selection occurs more often in such regions
591 (Al-Shahrour et al., 2010; Flowers et al., 2011). We therefore examined the relationship
592 between intraspecific diversity, species divergence and the density of functional
593 elements, defined as the proportion of protein-coding sites within a 10 Kbp or 100 Kbp
594 window (coding density). We hypothesized that if selection contributes to the reduction

595 of diversity at linked neutral sites, its effect is expected to be more pronounced in
596 regions with greater content of functional elements (Ravinet et al., 2017). Consistent
597 with this prediction, we observed a significantly negative relationship between π and
598 functional content (Figure 4A). This correlation was robust to the presence of
599 confounding factors such as GC content, recombination rate and the choice of window
600 size (Table S15).

601 Moreover, if natural selection was acting on the ancestral polymorphisms prior to
602 the divergence of the two descendant lineages, it could also have an effect on the
603 genetic divergence between species (Munch et al., 2016; Scally et al., 2012). We
604 therefore examined the relationship between interspecies divergence (both F_{ST} and d_{xy})
605 and coding density, and observed negative relationships for both F_{ST} and d_{xy} (Figure
606 4B, C), especially between species with longer divergence times (e.g. aspens and *P.*
607 *trichocarpa*) (Table S17, Table S19). Indeed, if a region experiences natural selection
608 during divergence, it should show lower π within species and higher F_{ST} between
609 species because F_{ST} is sensitive to intra-specific genetic variation (Cruickshank &
610 Hahn, 2014). Accordingly, a positive correlation between coding density and F_{ST} is
611 predicted. The opposite pattern we observe here indicates that long-term natural
612 selection, most likely due to background selection in functional regions, has
613 continuously contributed to the reduced ancestral polymorphism and genetic divergence
614 in regions with greater functional content (Phung et al., 2016). In fact, because of the
615 accumulation of the large amount of new mutations since speciation, ancestral
616 polymorphism may only account for a small amount of the overall average divergence
617 between distantly related species (Edwards & Beerli, 2000). However, the variance of
618 ancestral polymorphism, largely affected by natural selection in ancestral populations,
619 can on the other hand make a substantial contribution to the variability of genome-wide

620 patterns of divergence between species (McVicker, Gordon, Davis, & Green, 2009;
621 Phung et al., 2016).

622 To further explore the role of natural selection during the divergence of the three
623 aspen species, we examined the extent of ILS across the genome, which can aid to infer
624 evolutionary process in ancestral populations (Mailund et al., 2014; Pease & Hahn,
625 2013). The pattern of ILS along the genome offers information about the local
626 differences in the ancestral effective population size of the *aspen* ancestor (Pamilo &
627 Nei, 1988). Both purifying and positive selection in the ancestral population are
628 expected to reduce ancestral population size in regions targeted by selection, resulting in
629 increased rates for lineages to coalesce and leaving less available for ILS (Dutheil,
630 Munch, Nam, Mailund, & Schierup, 2015; Munch et al., 2016; Prüfer et al., 2012;
631 Scally et al., 2012). In agreement with this, we found that the fraction of ILS decreases
632 with increasing coding density (Figure 4D), and this relationship remained even after
633 correcting for the confounding variables (Table S21). Within coding exons, ILS is ~19 %
634 lower and the suppression of ILS extends several thousand bps away from coding genes
635 (Figure S14). Similarly, the proportion of the topology reflecting the true species tree
636 increases with coding density (Figure 4E; Table S21).

637 In addition, given that the level of admixture estimated by f_d between *P. tremula*
638 and *P. tremuloides* show considerable heterogeneity across the genome (Figure S15),
639 we examined whether selection may have played a role in shaping genome-wide
640 patterns of introgression. We estimated the relationship between f_d and coding density
641 and found a significantly negative correlation (Figure 4F; Table S21), indicating that
642 there is greater selection against introgressed alleles in regions enriched for genes
643 (Harris & Nielsen, 2016). The occurrence of this pattern is not likely an artefact of
644 reduced power, as regions with a high density of functionally important elements are

645 expected to have experienced stronger long-term selection and exhibit lower levels of
646 ILS. Accordingly, our power to detect introgression is expected to be elevated close to
647 these regions (Sankararaman et al., 2014; Sankararaman, Mallick, Patterson, & Reich,
648 2016). Taken together, it is clear that natural selection has had a strong impact on
649 patterns of phylogenetic discordance across the genome among closely related aspen
650 species. However, it is not yet clear to what extent this heterogeneity might be due to
651 incomplete lineage sorting of ancestral polymorphisms or due to ancient introgression.
652 More explicit experimental designs in future studies are needed to tease apart these
653 different processes and explore how natural selection and hybridization act in
654 combination to shape the genome-wide phylogenetic heterogeneity among recently
655 diverged species.

656 In addition to the local density of functional elements, recombination rates can also
657 interact with natural selection to influence the genomic distribution of genetic diversity
658 and divergence (Figure 4). High recombination can rapidly decouple linked loci and
659 restrict the effect of selection on linked neutral sites (Begun & Aquadro, 1992; Cutter &
660 Payseur, 2013). We found that π and F_{ST} showed positive and negative correlations,
661 respectively, with local recombination rates (Figure 4A-C; Table S16, S18, S20). In
662 contrast to the predicted pattern of speciation with gene flow where reduced d_{xy} is
663 expected in regions of high recombination (Nachman & Payseur, 2012), we did not find
664 any relationship between d_{xy} and recombination rate. These observations are in
665 accordance with the expectation that linked selection is prevalent and has genome-wide
666 effects in shaping patterns of genetic diversity and divergence at linked sites in *Populus*
667 (Nachman & Payseur, 2012; Wang et al., 2016b). On the other hand, we found that the
668 incidence of ILS increases with the recombination rate (Figure 4D), which was still
669 observable even after correcting for the confounding variables of coding density and

GC content (Table S22). Given that variation in ILS across the genome approximately reflects variation in ancestral N_e (Degnan & Salter, 2005; Pamilo & Nei, 1988), the stronger effects of recurrent natural selection in low-recombination regions also reduced N_e in ancestral populations and hence made ILS less likely to occur (Charlesworth et al., 1993; Martin, Davey, Salazar, & Jiggins, 2019; Pease & Hahn, 2013). We did not find obvious correlation between f_d and recombination rate (Figure 4F, Table S22), might because barriers to introgression has been sculpted by long-term selection and genetic drift after the ancient gene flow and cannot be predicted by recombination rate estimated from current populations. Overall, all these results suggest that the patterns of diversity, divergence and genealogical relationships among the three closely related aspen species are not randomly distributed along the genome, but instead are strongly structured by the interaction between widespread natural selection and intrinsic genomic features, as well as their influence on retention of signatures of ancient gene flow.

683

684 *3.5 The impact of positive and balancing selection on genomic architecture of* 685 *speciation*

Although widespread background selection is likely to have had a large effect in shaping the heterogeneous genomic landscape of variation within and between species (Burri, 2017; Charlesworth, 2012), we were interested in assessing whether positive selection or long-term balancing selection have also played important roles in driving these processes. To identify the impact of positive selection, we performed a composite-likelihood based (CLR) test to scan the genomes for signals of positive selection in each of the three aspen species (Figure 5A). For each species, we considered the windows with a CLR value in the top 1 percentile as candidate region under positive selection. In total, we detected 538 outlier windows across the three

species, and only 13 among them were shared by all species (Figure 5B). Our results suggest that most putative sweeps are likely species-specific and may result from relatively recent positive selection that has occurred independently in various lineages after speciation. Compared to genome-wide averages, outlier windows have significantly lower nucleotide diversity, lower recombination rates, higher F_{ST} but similar d_{xy} (Figure 5C). In addition, the outlier windows show significantly higher average weightings of the ‘species topology’ (Topo2) and lower levels of ILS compared to genomic background (Figure 5C,D). The ancestral admixture proportion (f_d) between *P. tremula* and *P. tremuloides* is also significantly reduced in the outlier windows (Figure 5D), suggesting that strong selection in these regions may have contributed to the reproductive barriers isolating closely related species (Martin et al., 2019).

To further study how long-term balancing selection may have driven the evolution of the genomic landscape during speciation, we used a summary statistics, β (beta score), to search for signals of balancing selection across the genome for each aspen species (Siewert & Voight, 2017). As we did for positive selection, we only consider variants with β scores falling in the top 1% as candidate variants. Furthermore, variants simultaneously detected in all three species are considered as potential targets of long-term balancing selection. With this criteria we identified a total of 519 variants putatively under long-term balancing selection across the three aspen species (Figure 6A,B). These variants were unevenly distributed in the genome, and to make them comparable to our previous analyses we clustered them into 32 regions of 10 Kbp windows (Figure 6A). We found significantly higher nucleotide diversity, higher recombination rates, lower F_{ST} , and higher d_{xy} in the regions under balancing selection compared to the genomic background (Figure 6C). Moreover, we found lower weightings of the ‘species topology’, higher ILS, and lower f_d in the candidate balancing

720 selection regions although the results were not significant, likely due to the small
721 number of windows showing evidence for balancing selection (Figure 6C,D). We
722 therefore infer that long-term balancing selection may not only influence the genomic
723 landscape of diversity and divergence but may also play a role in shaping the
724 genealogical relationship and barriers to introgression among closely related species
725 (Charlesworth, 2006; Wang et al., 2019).

726 Finally, to assess whether there were any specific biological functions that were
727 significantly over-represented on genes located in regions identified as undergoing
728 either positive (506 genes) or long-term balancing selection (32 genes), we performed
729 gene ontology (GO) enrichment analysis. We did not detect over-representation for any
730 functional category among the candidate genes under long-term balancing selection. In
731 contrast, we identified 31 significantly enriched GO categories (Fisher's exact test,
732 $P < 0.01$) for genes under positive selection (Table S23). These GO clusters were
733 primarily associated with metabolic processes (DNA, nucleic acid, cellular
734 macromolecule and aromatic compound, molybdopterin cofactor), biosynthetic
735 processes (molybdopterin cofactor, vitamin B6), cell morphogenesis and gene
736 expression regulation. Together these functional clusters are biologically relevant for
737 plant adaptation, because the biosynthesis of a panoply of diverse natural chemicals
738 serve as important adaptive strategies for sessile long-lived trees to adapt to
739 ever-changing abiotic and biotic environments (Weng, 2014).

740

741 **4. Discussion**

742 Much of our knowledge of how genomic landscape builds in the speciation process is
743 drawn from studies focusing on two young species pairs with ongoing gene flow. Very
744 few examples of now-allopatric species pairs along the speciation continuum have been

investigated. Here, we focus our research on four widespread *Populus* species that are allopatrically distributed in northern Hemisphere. After characterizing their speciation and demographic histories, we find that species in different continent exhibited idiosyncratic responses to Pleistocene climate changes. In addition, ancient gene flow was detected between extant Eurasian and North American aspen species (*P. tremula* and *P. tremuloides*). We also investigated the evolutionary forces that have shaped genome-wide patterns of variations within and between species. Our results have found substantial variation in genetic diversity, divergence, species relationships and the extent of introgression along the genome. Variation in these patterns is predictable and can be largely explained by genome-wide variation in the strength and extent of both recent and ancient selection, which depends on the recombination rate and the local density of functional sites. Our findings therefore provide evidence of how recurrent selection interacts with genomic features to shape the genomic landscape during species divergence. We further demonstrate that not only background selection, positive and long-term balancing selection also play crucial roles in shaping genomic variation and phylogenetic relationship among the recently diverged aspen species. Overall, this study highlights the striking impacts of natural selection in shaping within- and between-species genomic variation through speciation.

Acknowledgements

All analyses were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under the projects SNIC2016-7-89 and SNIC 2017/1-499. Financial support was provided by National Natural Science Foundation of China (31971567) and the Fundamental Research Funds for the Central Universities.

770

771 **References**

772

773 Abbott, R., Albach, D., Ansell, S., Arntzen, J. W., Baird, S. J., Bierne, N., . . . Buggs, R.
774 (2013). Hybridization and speciation. *Journal of Evolutionary Biology*, 26(2),
775 229-246.

776 Al-Shahrour, F., Minguez, P., Marqués-Bonet, T., Gazave, E., Navarro, A., & Dopazo,
777 J. (2010). Selection upon genome architecture: conservation of functional
778 neighborhoods with changing genes. *PLoS Computational Biology*, 6(10),
779 e1000953.

780 Alexa, A., & Rahnenführer, J. (2009). Gene set enrichment analysis with topGO.
781 *Bioconductor Improv*, 27.

782 Bai, W. N., Yan, P. C., Zhang, B. W., Woeste, K. E., Lin, K., & Zhang, D. Y. (2018).
783 Demographically idiosyncratic responses to climate change and rapid
784 Pleistocene diversification of the walnut genus *Juglans* (Juglandaceae) revealed
785 by whole-genome sequences. *New Phytologist*, 217(4), 1726-1736.

786 Begun, D. J., & Aquadro, C. F. (1992). Levels of naturally occurring DNA
787 polymorphism correlate with recombination rates in *D. melanogaster*. *Nature*,
788 356(6369), 519.

789 Browning, B. L., & Browning, S. R. (2009). A unified approach to genotype imputation
790 and haplotype-phase inference for large data sets of trios and unrelated
791 individuals. *The American Journal of Human Genetics*, 84(2), 210-223.

792 Browning, B. L., & Browning, S. R. (2013). Improving the accuracy and efficiency of
793 identity-by-descent detection in population data. *Genetics*, 194(2), 459-471.

794 Burri, R. (2017). Interpreting differentiation landscapes in the light of long-term linked
795 selection. *Evolution Letters*, 1(3), 118-131.

796 Burri, R., Nater, A., Kawakami, T., Mugal, C. F., Olason, P. I., Smeds, L., . . .
797 Garamszegi, L. Z. (2015). Linked selection and recombination rate variation
798 drive the evolution of the genomic landscape of differentiation across the
799 speciation continuum of *Ficedula* flycatchers. *Genome Research*, 25(11),
800 1656-1665.

801 Chan, A. H., Jenkins, P. A., & Song, Y. S. (2012). Genome-wide fine-scale
802 recombination rate variation in *Drosophila melanogaster*. *PLoS Genetics*, 8(12),
803 e1003090.

804 Charlesworth, B. (1998). Measures of divergence between populations and the effect of
805 forces that reduce variability. *Molecular Biology and Evolution*, 15(5), 538-543.

806 Charlesworth, B. (2012). The effects of deleterious mutations on evolution at linked
807 sites. *Genetics*, 190(1), 5-22.

808 Charlesworth, B., Morgan, M., & Charlesworth, D. (1993). The effect of deleterious
809 mutations on neutral molecular variation. *Genetics*, 134(4), 1289-1303.

810 Charlesworth, D. (2006). Balancing selection and its effects on sequences in nearby
811 genome regions. *PLoS Genetics*, 2(4), e64.

812 Comeron, J. M., Williford, A., & Kliman, R. (2008). The Hill–Robertson effect:
813 evolutionary consequences of weak selection and linkage in finite populations.
814 *Heredity*, 100(1), 19.

815 Cruickshank, T. E., & Hahn, M. W. (2014). Reanalysis suggests that genomic islands of
816 speciation are due to reduced diversity, not reduced gene flow. *Molecular*
817 *Ecology*, 23(13), 3133-3157.

818 Cutter, A. D., & Payseur, B. A. (2013). Genomic signatures of selection at linked sites:
819 unifying the disparity among species. *Nature Reviews Genetics*, 14(4), 262.

820 Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . .
821 Sherry, S. T. (2011). The variant call format and VCFtools. *Bioinformatics*,
822 27(15), 2156-2158.

823 DeGiorgio, M., Huber, C. D., Hubisz, M. J., Hellmann, I., & Nielsen, R. (2016).
824 SweepFinder2: increased sensitivity, robustness and flexibility. *Bioinformatics*,
825 32(12), 1895-1897.

826 Degnan, J. H., & Salter, L. A. (2005). Gene tree distributions under the coalescent
827 process. *Evolution*, 59(1), 24-37.

828 Delmore, K. E., Lugo Ramos, J. S., Van Doren, B. M., Lundberg, M., Bensch, S., Irwin,
829 D. E., & Liedvogel, M. (2018). Comparative analysis examining patterns of
830 genomic differentiation across multiple episodes of population divergence in
831 birds. *Evolution Letters*, 2(2), 76-87.

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

835 Du, S., Wang, Z., Ingvarsson, P. K., Wang, D., Wang, J., Wu, Z., . . . Zhang, J. (2015).
836 Multilocus analysis of nucleotide variation and speciation in three closely
837 related *Populus* (Salicaceae) species. *Molecular Ecology*, 24(19), 4994-5005.

838 Durand, E. Y., Patterson, N., Reich, D., & Slatkin, M. (2011). Testing for ancient
839 admixture between closely related populations. *Molecular Biology and*
840 *Evolution*, 28(8), 2239-2252.

841 Dutheil, J. Y., Munch, K., Nam, K., Mailund, T., & Schierup, M. H. (2015). Strong
842 selective sweeps on the X chromosome in the human-chimpanzee ancestor
843 explain its low divergence. *PLoS Genetics*, 11(8), e1005451.

844 Eckenwalder, J. E. (1996). Systematics and evolution of *Populus*. *Biology of Populus*
845 *and its Implications for Management and Conservation*, 7, 32.

846 Edwards, S., & Beerli, P. (2000). Perspective: gene divergence, population divergence,
847 and the variance in coalescence time in phylogeographic studies. *Evolution*,
848 54(6), 1839-1854.

849 Ellegren, H., & Galtier, N. (2016). Determinants of genetic diversity. *Nature Reviews*
850 *Genetics*, 17(7), 422.

851 Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., & Foll, M. (2013).
852 Robust demographic inference from genomic and SNP data. *PLoS Genetics*,
853 9(10), e1003905.

854 Feder, J. L., Egan, S. P., & Nosil, P. (2012). The genomics of
855 speciation-with-gene-flow. *Trends in Genetics*, 28(7), 342-350.

856 Flaxman, S. M., Wacholder, A. C., Feder, J. L., & Nosil, P. (2014). Theoretical models
857 of the influence of genomic architecture on the dynamics of speciation.
858 *Molecular Ecology*, 23(16), 4074-4088.

859 Flowers, J. M., Molina, J., Rubinstein, S., Huang, P., Schaal, B. A., & Purugganan, M.
860 D. (2011). Natural selection in gene-dense regions shapes the genomic pattern of
861 polymorphism in wild and domesticated rice. *Molecular Biology and Evolution*,
862 29(2), 675-687.

863 Fontaine, M. C., Pease, J. B., Steele, A., Waterhouse, R. M., Neafsey, D. E., Sharakhov,
864 I. V., . . . Kakani, E. (2015). Extensive introgression in a malaria vector species
865 complex revealed by phylogenomics. *Science*, 347(6217), 1258524.

866 Fumagalli, M., Vieira, F. G., Korneliussen, T. S., Linderroth, T., Huerta-Sánchez, E.,
867 Albrechtsen, A., & Nielsen, R. (2013). Quantifying population genetic
868 differentiation from next-generation sequencing data. *Genetics*, 195(3), 979-992.
869 Fumagalli, M., Vieira, F. G., Linderroth, T., & Nielsen, R. (2014). ngsTools: methods
870 for population genetics analyses from next-generation sequencing data.
871 *Bioinformatics*, 30(10), 1486-1487.

872 Gao, Z., Przeworski, M., & Sella, G. (2015). Footprints of ancient - balanced
873 polymorphisms in genetic variation data from closely related species. *Evolution*,
874 69(2), 431-446.

875 Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., . . . Fritz,
876 M. H.-Y. (2010). A draft sequence of the Neandertal genome. *Science*,
877 328(5979), 710-722.

878 Guerrero, R. F., & Hahn, M. W. (2017). Speciation as a sieve for ancestral
879 polymorphism. *Molecular Ecology*, 26(20), 5362-5368.

880 Hamzeh, M., & Dayanandan, S. (2004). Phylogeny of *Populus* (Salicaceae) based on
881 nucleotide sequences of chloroplast TRNT-TRNF region and nuclear rDNA.
882 *American Journal of Botany*, 91(9), 1398-1408.

883 Han, F., Lamichhaney, S., Grant, B. R., Grant, P. R., Andersson, L., & Webster, M. T.
884 (2017). Gene flow, ancient polymorphism, and ecological adaptation shape the
885 genomic landscape of divergence among Darwin's finches. *Genome Research*,
886 27(6), 1004-1015.

887 Harris, K., & Nielsen, R. (2016). The genetic cost of Neanderthal introgression.
888 *Genetics*, 203(2), 881-891.

889 Hart, J., De Araujo, F., Thomas, B., & Mansfield, S. (2013). Wood quality and growth
890 characterization across intra-and inter-specific hybrid aspen clones. *Forests*,
891 4(4), 786-807.

892 Hobolth, A., Dutheil, J. Y., Hawks, J., Schierup, M. H., & Mailund, T. (2011).
893 Incomplete lineage sorting patterns among human, chimpanzee, and orangutan
894 suggest recent orangutan speciation and widespread selection. *Genome*
895 *Research*, 21(3), 349-356.

896 Hudson, R. R., & Kaplan, N. L. (1995). Deleterious background selection with
897 recombination. *Genetics*, 141(4), 1605-1617.

898 Hurst, L. D., Pál, C., & Lercher, M. J. (2004). The evolutionary dynamics of eukaryotic
899 gene order. *Nature Reviews Genetics*, 5(4), 299.

900 Jansson, S., & Douglas, C. J. (2007). *Populus*: a model system for plant biology. *Annual*
901 *Review of Plant Biology*, 58, 435-458.

902 Kaplan, N. L., Hudson, R., & Langley, C. (1989). The "hitchhiking effect" revisited.
903 *Genetics*, 123(4), 887-899.

904 Koch, M. A., Haubold, B., & Mitchell-Olds, T. (2000). Comparative evolutionary
905 analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*,
906 *Arabis*, and related genera (Brassicaceae). *Molecular Biology and Evolution*,
907 17(10), 1483-1498.

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

910 Lamichhaney, S., Berglund, J., Almén, M. S., Maqbool, K., Grabherr, M.,
911 Martínez-Barrio, A., . . . Zamani, N. (2015). Evolution of Darwin's finches and
912 their beaks revealed by genome sequencing. *Nature*, 518(7539), 371.

- 913 Lee, T.-H., Guo, H., Wang, X., Kim, C., & Paterson, A. H. (2014). SNPhylo: a pipeline
914 to construct a phylogenetic tree from huge SNP data. *BMC Genomics*, 15(1),
915 162.
- 916 Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with
917 BWA-MEM. *arXiv:1303.3997*.
- 918 Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with
919 Burrows–Wheeler transform. *Bioinformatics*, 25(14), 1754-1760.
- 920 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Durbin, R.
921 (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*,
922 25(16), 2078-2079.
- 923 Lin, Y.-C., Wang, J., Delhomme, N., Schiffthaler, B., Sundström, G., Zuccolo, A., . . .
924 Cossu, R. M. (2018). Functional and evolutionary genomic inferences in
925 Populus through genome and population sequencing of American and European
926 aspen. *Proceedings of the National Academy of Sciences*, 115(46),
927 E10970-E10978.
- 928 Lohmueller, K. E., Albrechtsen, A., Li, Y., Kim, S. Y., Korneliussen, T., Vinckenbosch,
929 N., . . . Grarup, N. (2011). Natural selection affects multiple aspects of genetic
930 variation at putatively neutral sites across the human genome. *PLoS Genetics*,
931 7(10), e1002326.
- 932 Lohse, M., Bolger, A. M., Nagel, A., Fernie, A. R., Lunn, J. E., Stitt, M., & Usadel, B.
933 (2012). R obi NA: A user-friendly, integrated software solution for
934 RNA-Seq-based transcriptomics. *Nucleic Acids Research*, 40(W1),
935 W622-W627.
- 936 Ma, T., Wang, K., Hu, Q., Xi, Z., Wan, D., Wang, Q., . . . Abbott, R. J. (2018). Ancient
937 polymorphisms and divergence hitchhiking contribute to genomic islands of
938 divergence within a poplar species complex. *Proceedings of the National
939 Academy of Sciences*, 115(2), E236-E243.
- 940 Mailund, T., Munch, K., & Schierup, M. H. (2014). Lineage sorting in apes. *Annual
941 Review of Genetics*, 48, 519-535.
- 942 Martin, S. H., Davey, J. W., & Jiggins, C. D. (2014). Evaluating the use of
943 ABBA–BABA statistics to locate introgressed loci. *Molecular Biology and
944 Evolution*, 32(1), 244-257.
- 945 Martin, S. H., Davey, J. W., Salazar, C., & Jiggins, C. D. (2019). Recombination rate
946 variation shapes barriers to introgression across butterfly genomes. *PLoS
947 Biology*, 17(2), e2006288.
- 948 Martin, S. H., & Van Belleghem, S. M. (2017). Exploring evolutionary relationships
949 across the genome using topology weighting. *Genetics*, 206(1), 429-438.
- 950 Matthey-Doret, R., & Whitlock, M. C. (2019). Background selection and FST:
951 consequences for detecting local adaptation. *Molecular Ecology*. doi:
952 10.1111/mec.15197.
- 953 McVicker, G., Gordon, D., Davis, C., & Green, P. (2009). Widespread genomic
954 signatures of natural selection in hominid evolution. *PLoS Genetics*, 5(5),
955 e1000471.
- 956 Meisel, R. P., & Connallon, T. (2013). The faster-X effect: integrating theory and data.
957 *Trends in genetics*, 29(9), 537-544.
- 958 Munch, K., Nam, K., Schierup, M. H., & Mailund, T. (2016). Selective sweeps across
959 twenty millions years of primate evolution. *Molecular Biology and Evolution*,
960 33(12), 3065-3074.
- 961 Nachman, M. W., & Payseur, B. A. (2012). Recombination rate variation and
962 speciation: theoretical predictions and empirical results from rabbits and mice.

- 963 *Philosophical Transactions of the Royal Society B: Biological Sciences*,
964 367(1587), 409-421.
- 965 Nadachowska-Brzyska, K., Burri, R., Olason, P. I., Kawakami, T., Smeds, L., &
966 Ellegren, H. (2013). Demographic divergence history of pied flycatcher and
967 collared flycatcher inferred from whole-genome re-sequencing data. *PLoS*
968 *Genetics*, 9(11), e1003942.
- 969 Nadeau, N. J., Whibley, A., Jones, R. T., Davey, J. W., Dasmahapatra, K. K., Baxter, S.
970 W., . . . Blaxter, M. L. (2012). Genomic islands of divergence in hybridizing
971 *Heliconius* butterflies identified by large-scale targeted sequencing.
972 *Philosophical Transactions of the Royal Society B: Biological Sciences*,
973 367(1587), 343-353.
- 974 Neale, D. B., & Ingvarsson, P. K. (2008). Population, quantitative and comparative
975 genomics of adaptation in forest trees. *Current opinion in plant biology*, 11(2),
976 149-155.
- 977 Neale, D. B., & Kremer, A. (2011). Forest tree genomics: growing resources and
978 applications. *Nature Reviews Genetics*, 12(2), 111.
- 979 Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., & Wang, J. (2012). SNP calling,
980 genotype calling, and sample allele frequency estimation from new-generation
981 sequencing data. *PloS One*, 7(7), e37558.
- 982 Nordborg, M., Hu, T. T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., . . . Goyal, R.
983 (2005). The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biology*,
984 3(7), e196.
- 985 Nosil, P., Funk, D. J., & Ortiz-Barrientos, D. (2009). Divergent selection and
986 heterogeneous genomic divergence. *Molecular Ecology*, 18(3), 375-402.
- 987 Pamilo, P., & Nei, M. (1988). Relationships between gene trees and species trees.
988 *Molecular Biology and Evolution*, 5(5), 568-583.
- 989 Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis.
990 *PLoS Genetics*, 2(12), e190.
- 991 Pease, J. B., & Hahn, M. W. (2013). More accurate phylogenies inferred from low-
992 recombination regions in the presence of incomplete lineage sorting. *Evolution*,
993 67(8), 2376-2384.
- 994 Phung, T. N., Huber, C. D., & Lohmueller, K. E. (2016). Determining the effect of
995 natural selection on linked neutral divergence across species. *PLoS Genetics*,
996 12(8), e1006199.
- 997 Prüfer, K., Munch, K., Hellmann, I., Akagi, K., Miller, J. R., Walenz, B., . . . Winer, R.
998 (2012). The bonobo genome compared with the chimpanzee and human
999 genomes. *Nature*, 486(7404), 527.
- 1000 Ravinet, M., Faria, R., Butlin, R., Galindo, J., Bierne, N., Rafajlović, M., . . . Westram,
1001 A. (2017). Interpreting the genomic landscape of speciation: a road map for
1002 finding barriers to gene flow. *Journal of Evolutionary Biology*, 30(8),
1003 1450-1477.
- 1004 Ravinet, M., Yoshida, K., Shigenobu, S., Toyoda, A., Fujiyama, A., & Kitano, J.
1005 (2018). The genomic landscape at a late stage of stickleback speciation: High
1006 genomic divergence interspersed by small localized regions of introgression.
1007 *PLoS Genetics*, 14(5), e1007358.
- 1008 Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., . . .
1009 Reich, D. (2014). The genomic landscape of Neanderthal ancestry in present-day
1010 humans. *Nature*, 507(7492), 354.

1011 Sankararaman, S., Mallick, S., Patterson, N., & Reich, D. (2016). The combined
1012 landscape of Denisovan and Neanderthal ancestry in present-day humans.
1013 *Current Biology*, 26(9), 1241-1247.

1014 Scally, A., Dutheil, J. Y., Hillier, L. W., Jordan, G. E., Goodhead, I., Herrero, J., . . .
1015 Marques-Bonet, T. (2012). Insights into hominid evolution from the gorilla
1016 genome sequence. *Nature*, 483(7388), 169.

1017 Schiffels, S., & Durbin, R. (2014). Inferring human population size and separation
1018 history from multiple genome sequences. *Nature Genetics*, 46(8), 919.

1019 Schumer, M., Cui, R., Powell, D. L., Rosenthal, G. G., & Andolfatto, P. (2016). Ancient
1020 hybridization and genomic stabilization in a swordtail fish. *Molecular Ecology*,
1021 25(11), 2661-2679.

1022 Seehausen, O., Butlin, R. K., Keller, I., Wagner, C. E., Boughman, J. W., Hohenlohe, P.
1023 A., . . . Brännström, Å. (2014). Genomics and the origin of species. *Nature*
1024 *Reviews Genetics*, 15(3), 176.

1025 Siewert, K. M., & Voight, B. F. (2017). Detecting long-term balancing selection using
1026 allele frequency correlation. *Molecular biology and evolution*, 34(11),
1027 2996-3005.

1028 Smith, J. M., & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetics*
1029 *Research*, 23(1), 23-35.

1030 Stamatakis, A. (2014). RAXML version 8: a tool for phylogenetic analysis and
1031 post-analysis of large phylogenies. *Bioinformatics*, 30(9), 1312-1313.

1032 Stankowski, S., Chase, M. A., Fuiten, A. M., Rodrigues, M. F., Ralph, P. L., &
1033 Streisfeld, M. A. (2019). Widespread selection and gene flow shape the genomic
1034 landscape during a radiation of monkeyflowers. *PLoS Biology*, 17(7), e3000391.

1035 Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to identify repetitive
1036 elements in genomic sequences. *Current Protocols in Bioinformatics*, 25(1),
1037 4.10.11-4.10.14.

1038 Turner, T. L., Hahn, M. W., & Nuzhdin, S. V. (2005). Genomic islands of speciation in
1039 *Anopheles gambiae*. *PLoS Biology*, 3(9), e285.

1040 Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., . . .
1041 Salamov, A. (2006). The genome of black cottonwood, *Populus trichocarpa*
1042 (Torr. & Gray). *Science*, 313(5793), 1596-1604.

1043 Van Doren, B. M., Campagna, L., Helm, B., Illera, J. C., Lovette, I. J., & Liedvogel, M.
1044 (2017). Correlated patterns of genetic diversity and differentiation across an
1045 avian family. *Molecular Ecology*, 26(15), 3982-3997.

1046 Vicoso, B., & Charlesworth, B. (2006). Evolution on the X chromosome: unusual
1047 patterns and processes. *Nature Reviews Genetics*, 7(8), 645.

1048 Vijay, N., Weissensteiner, M., Burri, R., Kawakami, T., Ellegren, H., & Wolf, J. B.
1049 (2017). Genomewide patterns of variation in genetic diversity are shared among
1050 populations, species and higher - order taxa. *Molecular Ecology*, 26(16),
1051 4284-4295.

1052 Wang, B., Mojica, J. P., Perera, N., Lee, C.-R., Lovell, J. T., Sharma, A., . . . Rokhsar,
1053 D. S. (2019). Ancient polymorphisms contribute to genome-wide variation by
1054 long-term balancing selection and divergent sorting in *Boechera stricta*. *Genome*
1055 *Biology*, 20(1), 126.

1056 Wang, J., Street, N. R., Scofield, D. G., & Ingvarsson, P. K. (2016a). Natural selection
1057 and recombination rate variation shape nucleotide polymorphism across the
1058 genomes of three related *Populus* species. *Genetics*, 202(3), 1185-1200.

- 1059 Wang, J., Street, N. R., Scofield, D. G., & Ingvarsson, P. K. (2016b). Variation in
1060 linked selection and recombination drive genomic divergence during allopatric
1061 speciation of European and American aspens. *Molecular Biology and Evolution*,
1062 33(7), 1754-1767.
- 1063 Weng, J. K. (2014). The evolutionary paths towards complexity: a metabolic
1064 perspective. *New Phytologist*, 201(4), 1141-1149.
- 1065 Wolf, J. B., & Ellegren, H. (2017). Making sense of genomic islands of differentiation
1066 in light of speciation. *Nature Reviews Genetics*, 18(2), 87.
- 1067 Wu, C. I. (2001). The genic view of the process of speciation. *Journal of Evolutionary*
1068 *Biology*, 14(6), 851-865.
- 1069 Yin, T., DiFazio, S. P., Gunter, L. E., Zhang, X., Sewell, M. M., Woolbright, S. A., . . .
1070 Wang, M. (2008). Genome structure and emerging evidence of an incipient sex
1071 chromosome in *Populus*. *Genome Research*, 18(3), 422-430.

1072 **Data Accessibility Statement**

1073
1074
1075 Raw whole genome resequencing data generated for this study have been deposited in
1076 the NCBI short read archive under accession number PRJNA576115

1077 **Author Contributions**

1078
1079
1080 J.W. conceived the study, analyzed the data and wrote the manuscript. E.J.P. provided
1081 the materials of *P. davidiana* used in this study. N.R.S., J.L., P.K.I. read and
1082 commented on the manuscript. All authors approved the final manuscript.

Figure legends:

Figure 1. Phylogenetic and population genetic analyses of four *Populus* species.

(A) Sampling locations (black circle) of eight individuals from each of the four *Populus* species included in this study. Species ranges for *P. tremula*, *P. davidiana*, *P. tremuloides* and *P. trichocarpa* are indicated by red, green, blue and purple shading, respectively. (B) Maximum-likelihood phylogenetic tree reconstructed based on complete chloroplast sequences. Color scheme for the four species is the same in A-C. (C) Historical effective population size of the four *Populus* species inferred using MSMC v2 based on sets of eight haplotypes, with solid lines representing medians and shading representing $\pm$ standard deviation calculated across pairs of haplotypes. Yellow bar indicates Early Pleistocene cooling; Glacial and interglacial periods of the Late and Middle Pleistocene are indicated by dark and light grey bars, respectively; black bar indicates the period of Last Glacial Maximum (LGM). (D) Schematic of demographic scenarios of the three aspen species modeled using fastsimcoal2. The ancestral population is shown in light and dark grey respectively for different ancestral lineages. *P. tremula* is in red, *P. davidiana* is in green, and *P. tremuloides* is in blue. The arrows indicate the per generation migration rate (m) between species. Estimated divergence time, effective population size, and gene flow are detailed in Supplementary Table S7.

Figure 2. Genome-wide landscape of genetic diversity and divergence within and between species.

(A) Chromosomal landscape of (a) the density of coding sequences (CD); (b) nucleotide diversity (π); (c) recombination rate (ρ); (d) the relative measure of genetic divergence (F_{ST}) and (e) the absolute measure of genetic divergence (d_{xy}). (B) Distribution of correlation coefficients (Spearman's ρ) shown as violin plots for population summary statistics characterizing genomic features (neutral mutation rate μ) and variation within (π , ρ) and between species (F_{ST} , d_{xy}) calculated at 100 Kbp windows. Subscripts 'i, j' symbolize all possible combinations of correlations between two species $i=1 \dots (n-1)$ and $j=(i+1) \dots n$ for within-species measures; Capital letters 'I, J' symbolize inter-species statistics. Correlations exclude pseudo-replicated species comparisons. Detailed information can be found in Supplementary Table S11.

Figure 3. Heterogeneous distribution of phylogenies in three aspen species.

Chromoplots for 19 chromosomes show the distribution of three possible rooted phylogenetic relationships inferred from 100 Kbp genomic regions for *P. tremula* (*P. tra*), *P. davidiana* (*P. dav*) and *P. tremuloides* (*P. trs*), with *P. trichocarpa* as outgroup species. The colored vertical bars represent the windows with complete monophyly of the three alternative topologies as shown in the lower right, where the proportion of the three topologies in 100 Kbp and 10 Kbp (in parenthesis) across the genome are also shown. Across all chromosomes, the D statistic generally tends toward positive values, indicating ancient introgression between *P. tra* and *P. trs* may have been occurring across the genome.

Figure 4. Widespread impact of linked selection. Relationship between

recombination rate (blue), coding density (red) and (A) genetic diversity, (B) F_{ST} , (C) d_{xy} , (D) incomplete lineage sorting (ILS), (E) weighting of the 'species' tree ($[P. tra, P. dav], P. trs$) and (F) the estimated admixture proportion (f_d) between *P. tra* and *P. trs*. Quantile binning is for visualization. The points and error bars indicate the means and $1.96 \times$ standard errors. Statistical tests were performed on the unbinned data and detailed correlation coefficients are shown in supplementary Table S15-Table S22.

1133

1134 **Figure 5. Identification of positive selection.** (A) Positive selection analysis by
1135 SweepFinder2 reveals windows that are candidates for being under positive selection in
1136 the three aspen species: *P. tremula* (*P. tra*), *P. davidiana* (*P. dav*) and *P. tremuloides* (*P.*
1137 *trs*). Horizontal red line indicates the cut-off of composite likelihood ratio (CLR)
1138 statistics. (B) The Venn diagram represents shared and unique selected windows
1139 detected in the three species. (C) Comparison of genetic diversity, recombination rate,
1140 F_{ST} , d_{xy} , and average weightings of the ‘species’ topology between candidate regions
1141 under positive selection (red boxes) and genomic background (grey boxes). (D)
1142 Comparison of incomplete lineage sorting (ILS) and the estimated admixture proportion
1143 (f_d) between candidate regions under positive selection (red boxes) and genomic
1144 background (grey boxes). Asterisks designate significant differences between candidate
1145 positive selected regions and the rest of genomic regions by Mann-Whitney *U*-test (^{n.s.}
1146 Not significant; * *P* value<0.01; ** *P* value<0.001; *** *P* value <1e-4).

1147

1148 **Figure 6. Identification of long-term balancing selection.** (A) Signals of balancing
1149 selection across all chromosomes in the three aspen species: *P. tremula* (*P. tra*), *P.*
1150 *davidiana* (*P. dav*) and *P. tremuloides* (*P. trs*). Horizontal red line indicates the cut-off
1151 of the β statistics. Only the signals detected in all three aspen species (red dots) were
1152 considered as being under long-term balancing selection. (B) The Venn diagram
1153 represents shared and unique selected SNPs detected in the three species. (C)
1154 Comparison of genetic diversity, recombination rate, F_{ST} , d_{xy} , and average weightings
1155 of the ‘species’ topology between candidate regions under long-term balancing
1156 selection (red boxes) and genomic background (grey boxes). (D) Comparison of
1157 incomplete lineage sorting (ILS) and the estimated admixture proportion (f_d) between
1158 candidate regions under long-term balancing selection (red boxes) and genomic
1159 background (grey boxes). Asterisks designate significant differences between candidate
1160 balancing selected regions and the rest of genomic regions by Mann-Whitney *U*-test (^{n.s.}
1161 Not significant; * *P* value<0.01; ** *P* value<0.001; *** *P* value <1e-4).

1162

A **B** **C** **D**

A

B

D

● Coding density ● Recombination rate

A

B

C

D

E

F

Percentile of coding density and recombination rate

A **B** **D** **C**