Hard, soft and just right: variations in linked selection and recombination drive genomic divergence during speciation of aspens

Jing Wang¹, Nathaniel R. Street², Douglas G. Scofield¹³⁴, Pär K. Ingvarsson¹

¹ Department of Ecology and Environmental Science, Umeå University, SE-90187, Umeå, Sweden
² Umeå Plant Science Centre, Department of Plant Physiology, Umeå University, SE-90187, Umeå, Sweden
³ Department of Ecology and Genetics: Evolutionary Biology, Uppsala University, Uppsala, Sweden
⁴ Uppsala Multidisciplinary Center for Advanced Computational Science, Uppsala University, Uppsala, Sweden
Corresponding author:

Dr Pär K. Ingvarsson, Department of Ecology and Environmental Science, Umeå University, Umeå, SE 90187, Sweden. Phone: +46907867414; Fax: +46-(0)-90-786-6705; E-mail: par.ingvarsson@emg.umu.se

Keywords: *Populus tremula, P. tremuloides*, Whole-genome re-sequencing, demographic histories, heterogeneous genomic differentiation, linked selection, recombination
Abstract

Despite the global economic and ecological importance of forest trees, the genomic basis of differential adaptation and speciation in tree species is still poorly understood. *Populus tremula* and *P. tremuloides* are two of the most widespread tree species in Northern Hemisphere. Using whole-genome re-sequencing data from 24 *P. tremula* and 22 *P. tremuloides* individuals, we find that the two species diverged ~2.2-3.1 million years ago. The approximately allopatric speciation of the two species was likely the results of the severing of the Bering land bridge combined with the onset of dramatic climatic oscillations throughout the Pleistocene. We detected moderate but also considerable heterogeneous genomic differentiation between species. Rather than being physically clustered into just a few large, discrete genomic ‘islands’ as may be expected when species diverges in the presence of gene flow, we found that the regions of differentiation were particularly steep, narrowly defined and located in regions with substantially suppressed recombination. It appears that species-specific adaptation, mainly involving standing genetic variation via soft selective sweeps, was likely the predominant proximate cause in generating the differentiation islands between species and not local differences in permeability of gene flow. In addition, we identified multiple signatures of long-term balancing selection predating speciation in regions containing immunity and defense-related genes in both species.
Introduction

Understanding how genomes diverge during the process of speciation is a central goal in evolutionary genetics (Nosil, et al. 2009; Strasburg, et al. 2012; Seehausen, et al. 2014). Under neutrality, differentiation is expected to accumulate as a result of the stochastic fixation of polymorphisms by genetic drift (Coyne and Orr 2004). Historical demographic processes can accelerate or decelerate the rate of differentiation through changes in the effective population sizes of nascent daughter species (Avise 2000). In general, both random genetic drift and demographic processes are expected to affect the entire genome (Luikart, et al. 2003). Natural selection, however, is expected to only influence those loci involved in ecological specialization and/or reproductive isolation, resulting in patterns of polymorphisms and divergence that deviate from neutral predictions in genomic regions under selection (Luikart, et al. 2003; Via 2009). The functional architectures of genomes, e.g. mutation and recombination rates, are also important factors in determining genomic landscape of differentiation (Noor and Bennett 2009; Nachman and Payseur 2012; Renaut, et al. 2013; Burri, et al. 2015). For instance, suppressed recombination can lead to increased differentiation by two mechanisms: preventing gene flow between species to avoid the break-up of co-adapted alleles, and the diversity-reducing effects of linked selection (Noor and Bennett 2009). A longstanding challenge in speciation genetics has been to quantify the relative contributions of various evolutionary forces in generating and shaping patterns of genomic divergence during speciation.

With the advance of next generation sequencing (NGS) technologies, a growing number of studies have found highly heterogeneous patterns of genomic

A common explanation is that levels of gene flow between species differ across the genome, where increased genetic divergence in ‘differentiation islands’ is observed in a small number of regions supposed to contain loci involved in reproductive isolation (‘speciation islands’), while the remainder of the genome is still permeable to ongoing gene flow and therefore shows lower levels of differentiation (Nosil, et al. 2009; Sousa and Hey 2013). However, some recent studies have argued that highly differentiated regions should represent ‘incidental islands’ that are not tied to the speciation processes, but result from the diversity-reduced effects of linked selection (either positive or purifying selection) (Noor and Bennett 2009; Turner and Hahn 2010; Nachman and Payseur 2012; Cruickshank and Hahn 2014). In contrast, long-term balancing selection is supposed to maintain stable trans-species polymorphisms and leave signatures of unusually low genetic differentiation between species (Charlesworth 2006). Under this scenario, heterogeneous selection alone is sufficient to generate patterns of heterogeneous genomic differentiation even in complete allopatry (Noor and Bennett 2009; Turner and Hahn 2010; White, et al. 2010).

Exhaustive examination of the above two models is needed in more species and should preferably include details concerning the speciation processes, such as time of divergence, prevalence and rates of gene flow, well-characterized demographic histories and selective and recombination details (Nosil and Feder 2012).

Although largely understudied compared to other model species, forest trees represent a promising system to understand the genomic basis of species divergence and adaptive evolution; as a group they have developed diverse strategies to adapt and thrive across a wide range of climates and environments (Neale and Kremer 2011).
Populus tremula (European aspen) and P. tremuloides (American aspen) are two of the most ecologically important and geographically widespread tree species of the Northern Hemisphere (Figure 1a). Both are keystone species, display rapid growth, high tolerance to environmental stresses, and long-distance pollen and seed dispersal via wind (Eckenwalder 1996; Müller, et al. 2012). Based on their morphological similarity and close phylogenetic relationships, they are considered sister species, or less commonly, conspecific subspecies (Eckenwalder 1996; Wang, et al. 2013). They can readily cross and artificial hybrids usually show high heterosis (Hamzeh and Dayanandan 2004; Tullus, et al. 2012; Wang, et al. 2013). A recent study based on several nuclear and chloroplast loci suggests that the first opening of the Bering land bridge may have driven the allopatric speciation of the two species (Du, et al. 2015).

In accordance with their continent-wide distributions and broad ecological ranges, P. tremula and P. tremuloides harbor among the highest levels of genetic diversity found in plant species thus far (Wang, et al. unpublished data). The extraordinary levels of genetic diversity in both species and the availability of a high-quality reference genome in the congener, P. trichocarpa (Tuskan, et al. 2006), also provide ideal opportunities to identify the relative roles that new mutations (hard selective sweeps) versus pre-existing standing variations (soft selective sweeps) have played during adaptive differentiation and speciation at the genome scale. In hard-sweep models of adaptation, novel beneficial mutations arise and are rapidly fixed in a species, and this process is expected to leave a signature of severely reduced polymorphism in the vicinity of the beneficial mutation (Smith and Haigh 1974). Models of adaptation via soft sweeps assume that beneficial alleles originate from standing genetic variation or by recurrent independent novel mutations (Hermisson and Pennings 2005). Given that the background variation of selectively advantageous
alleles is more heterogeneous in a soft sweep scenario, less severe reductions in levels
of polymorphism are expected for soft sweeps compared to hard sweeps (Herisson
and Pennings 2005; Pennings and Herisson 2006). Based on whole-genome re-
sequencing data in both \textit{P. tremula} and \textit{P. tremuloides}, our goals were to: (1) estimate
the species’ divergence time and reconstruct demographic histories of the two species;
(2) infer and distinguish the relative roles of different evolutionary forces in
generating and shaping patterns of genomic differentiation between species; (3)
evaluate the relative importance of new mutations or standing genetic variation during
adaptive divergence in these widespread forest trees; and (4) identify genomic regions
and genes that may have evolved in response to directional and balancing selection
during speciation.

\textbf{Results}

We generated whole-genome sequence data for 24 \textit{P. tremula} and 22 \textit{P. tremuloides},
and more than 88\% of sequenced reads in each sample were aligned to the \textit{P.}
\textit{trichocarpa} reference genome (Table S1). The average coverage of uniquely mapped
reads per site was 25.1 and 22.5 in samples of \textit{P. tremula} and \textit{P. tremuloides},
respectively (Table S1).

\textbf{Population structure}

The genome-wide NGSadmix analysis clearly sub-divided all sampled individuals
into two species-specific groups when the number of clusters was \( K = 2 \) (Figure 1b).
When \( K = 3 \), there was evidence for further population sub-structuring in \textit{P.}
\textit{tremuloides}, where individuals from populations originating in Alberta and Wisconsin
clustered into two subgroups. With $K = 4$, most $P. \text{tremula}$ individuals were inferred to be a mixture of two genetic components, showing slight clinal genetic variation with latitude. No further structure was found when $K = 5$. A principal component analysis (PCA) further supported these results (Figure 1c). Only the first two components were significant (Table S2) based on a Tracy-Widom test, and these explained, respectively, 21.4% and 2.1% of total genetic variance (Figure 1c). Among the total number of polymorphisms between the two species, fixed differences between $P. \text{tremula}$ and $P. \text{tremuloides}$ accounted for 1.1%, whereas 16.7% of polymorphisms were shared between species, with the remaining polymorphic sites being private in either of the two species (Figure 1d).

We examined the extent of population subdivision in $P. \text{tremuloides}$ by measured $F_{ST}$ and $d_{xy}$ between the two $P. \text{tremuloides}$ populations (Alberta and Wisconsin) along individual chromosomes (Table S3). We found low levels of genetic differentiation ($F_{ST}$: $0.0443 \pm 0.0325$) between the populations (Table S3). Total sequence differentiation in the inter-population comparison (mean $d_{xy} = 0.0165 \pm 0.0083$) was similar to mean sequence differences in intra-population comparisons ($\pi_{\text{Alberta}}: 0.0161 \pm 0.0081$; $\pi_{\text{Wisconsin}}: 0.0157 \pm 0.0080$, Table S3), indicating that individuals of the two populations were genetically not more different from each other than individuals within each population. Further tests based on the allele frequency spectrum (Tajima’s D and Fay & Wu’s H) also supported these general patterns (Table S3).

Demographic histories

We used fastsimcoal2 (Excoffier, et al. 2013) to infer the divergence time between $P. \text{tremula}$ and $P. \text{tremuloides}$ and their past demographic histories from the joint site
frequency spectrum. Eighteen divergence models were evaluated (Table S4), and the best fit was provided by a simple isolation-with-migration model, where populations of *P. tremuloides* experienced exponential growth while a stepwise population size change occurred in *P. tremula* after the two species diverged (Figure 2a). The exact parameter estimates of divergence time, migration rates, population sizes and their associated 95% confidence intervals (CI) inferred from 100 parametric bootstraps are given in Table 1. The estimated split time between *P. tremula* and *P. tremuloides* (*T_{DIV}*) was ~2.3 million years ago (Mya) (bootstrap range [BR]: 2.2-3.1 Mya). The contemporary effective population sizes (*N_e*) of *P. tremula* (*N_{P.tremula}* ) and *P. tremuloides* (*N_{P.tremuloides}* ) were 102,814 (BR: 93,688-105,671) and 309,500 (BR: 247,321-310,105) respectively, with both being larger than the effective population size of their common ancestor (*N_{ANC}* = 56,235 [48,012-69,492]). Gene flow (2*N_e*m, where *N_e* is the effective population size and *m* is the migration rate) from *P. tremuloides* to *P. tremula* was higher (0.202 migrants per generation [0.156-0.375]) than in the opposite direction (0.053 [0.052-0.117]). It was most likely due to the higher *N_e* of *P. tremuloides* compared to *P. tremula* (Slatkin 1985), while the overall migration rates in both directions were fairly low given the large *N_e* in both species (Morjan and Rieseberg 2004). The low migration rates are not unexpected given the large geographical distance and disjunct distributions between the two species.

The multiple sequential Markovian coalescent (MSMC)-estimated *N_e* for both *P. tremula* (60,796) and *P. tremuloides* (49,701) at the beginning of species divergence (around 2.3 Mya) were very similar to the fastsimcoal2-based estimates of *N_e* for their ancestral population (Figure 2). Both species experienced similar magnitudes of population decline following their initial divergence, and population expansion in *P. tremuloides* began around 50,000-70,000 years ago and has continued.
up to the present (Figure 2b). *P. tremula* experienced a substantial population expansion following a longer bottleneck than that experienced by *P. tremuloides* (Figure 2b).

**Genome-wide patterns of differentiation and molecular signatures of selection in both high- and low-divergence regions**

We found that the majority of the genome showed moderate genetic differentiation, with average $F_{ST}$ value across the genome being 0.386 (see Figure 3a for the complete distribution). Visual inspection indicates that genetic differentiation was heterogeneous over 10 Kbp non-overlapping windows across the genome (Figure 4).

From the genome-wide empirical $F_{ST}$ distribution, the cutoff for the top 2.5% highly differentiated outlier windows was $F_{ST} > 0.681$ (Figure 3a). After removing windows where $d_{xy}$ values were lower than the genome-wide median (see Materials and Methods), 461 out of 730 highly differentiated windows were retained (Figure 3b).

We based our analysis of genetic signatures of adaptation in these highly divergent regions on our partitioning into categories that were likely to be under either hard or soft selective sweeps based on levels of $\theta_\pi$ within each species (see Materials and Methods; Figure 3b). A minority of outlier windows (5.87%) was consistent with patterns expected for hard selective sweeps in both species (black dots in Figure 3b), while somewhat more outlier windows showed signatures of having been influenced by hard sweeps only in *P. tremula* (12.61%, red dots in Figure 3b) or *P. tremuloides* (8.26%, blue dots in Figure 3b). Soft selective sweeps appear to have affected the remaining 73.3% of divergent outlier regions in one or both species (grey dots in Figure 3b). Compared to the rest of the genome, the divergent outlier regions were characterized by multiple signatures of selection in both species regardless of the
specific category that they belong to (as defined above; \( P < 0.05 \), Mann-Whitney U test). These further signatures include significantly skewed allele frequency spectrum towards rare alleles (more negative Tajima’s D), increased high-frequency derived alleles (more negative Fay & Wu’s H), and stronger signals of linkage disequilibrium (LD) (Table 2 and Figure S4). Divergent regions linked to soft sweeps showed substantially weaker signatures of selection compared to those linked to hard sweeps, with a much subtle excess of low-frequency alleles and weaker levels of LD (Table 2 and Figure S4a,c), but with comparable or even greater excesses of high-frequency derived alleles (Table 2 and Figure S4b). Furthermore, we found that all highly divergent regions showed significantly higher proportion of inter-species fixed differences and lower proportion of inter-species shared polymorphisms compared to the remainder of the genome (Table 2 and Figure S5 a-c). Even after correcting for possible variation in the mutation rate among genomic regions (Feder, et al. 2005), we found that relative node depths (RND) were significantly higher in divergent regions associated with either hard or soft sweeps than the rest of the genome (Table 2 and Figure S5d).

The cutoff for the bottom 2.5% of the empirical \( F_{ST} \) distribution, denoting outlier windows of exceptionally low levels of interspecies divergence, was \( F_{ST} < 0.169 \) (Figure 3a). After excluding windows with coverage breadth lower than 3 Kbp and retaining windows with high levels of polymorphisms in both species (see Materials and Methods), we identified 49 outlier windows as candidate targets of long-term balancing selection (green dots in Figure 3c). In contrast to the genomic background, these candidate regions showed an excess of intermediate-frequency alleles (higher Tajima’s D and Fay & Wu’s H values), and slightly lower levels of LD (Table 2 and Figure S4). In addition, we found a negligible proportion of fixed
differences and significantly higher proportion of shared polymorphism in these regions (Table 2 and Figure S5a-c). The higher RND values (Table 2 and Figure S5d), however, were likely due to higher levels of ancestral polymorphisms that were maintained by balancing selection before the two species split (Cruickshank and Hahn 2014).

Overall, candidate windows potentially under directional (both hard and soft sweeps) or balancing selection were distributed across the genome (Figure 6). We examined the physical sizes of these selected regions by combining adjacent windows if they were all under selection. We found that the sizes of the selected regions all appeared to be quite small, with the majority of selection likely occurring on a physical scale smaller than 10 Kbp (Figure S6).

Impact of recombination rate on patterns of genetic differentiation

We examined relationships between population-scaled recombination rates (ρ) and levels of inter-species divergence over non-overlapping 10 Kbp windows (Figure S7). We found significant negative correlations between relative divergence $F_{ST}$, which depends on genetic diversity within species, and population recombination rates in both $P. \text{tremula}$ (Spearman’s $\rho=-0.121$, $P$-value<2.2e-16) and $P. \text{tremuloides}$ (Spearman’s $\rho=-0.157$, $P$-value<2.2e-16) (Figure S7a). In contrast to $F_{ST}$, there were significant positive correlations between absolute divergence $d_{xy}$ and recombination rates in both $P. \text{tremula}$ (Spearman’s $\rho=0.199$, $P$-value<2.2e-16) and $P. \text{tremuloides}$ (Spearman’s $\rho=0.140$, $P$-value<2.2e-16) (Figure S7b).

Because $\rho=4N_e c$, where $c$ is the per-generation recombination rate and $N_e$ is the effective population size, reduction of $N_e$ in regions linked to selection will lower local estimates of $\rho$ even if local $c$ is identical. In order to account for such effects, we
compared $\rho/\theta$, between regions with signals of selection and the rest of the genome, because both $\rho$ and $\theta$ are scaled by $N_e$. Relative to genomic background, our results showed significantly suppressed recombination in all selected regions (Figure 5).

**Genes under selection**

The availability of an annotated *P. trichocarpa* genome enabled functional analyses of candidate genes within regions that were putatively under selection. In total, 31 genes were located in divergent regions with signatures of hard sweeps in both species (Table S5); 88 and 48 genes, respectively, were found in divergent regions with signatures of hard sweeps only in *P. tremula* and *P. tremuloides* (Table S5), and 310 genes were located in highly differentiated regions with signatures of soft sweeps in either one or both species (Table S5). Regions containing signatures of long-term balancing selection contained a total of 80 genes (Table S6). Except for divergent regions linked to soft sweeps where a significantly lower concentration of genes was found ($P<0.001$, Mann-Whitney U test), all other selected regions showed comparable gene densities compared to the rest of genome (Figure S8).

We used the Gene Ontology (GO) assignments of those candidate genes putatively under selection to test whether specific GO terms were significantly over-represented. After accounting for multiple comparisons, we did not detect any over-representation for divergent regions with signatures of hard sweeps only in one species or with signatures of soft sweeps. However, among genes with signatures of hard sweeps in both species, we found 16 significantly enriched GO terms, mainly including terms associated with transcription initiation and transcription factor activity (Table S7). Within regions with signatures of long-term balancing selection, 21 significantly overrepresented GO terms were identified among the 80 candidate
genes (Table S8), with most of them being associated with immune response and signal transduction.

Discussion

Species divergence and demographic histories

Our simulation-based analyses indicated that *P. tremula* and *P. tremuloides* diverged around 2.2-3.1 Mya during the Late Pliocene and/or Early Pleistocene. This timing corresponds closely with the first opening of the Bering Strait, which occurred 3.1-5.5 Mya and broke up the overland intercontinental migration route of terrestrial floras between Eurasia and North America (Marincovich and Gladenkov 1999; Gladenkov, et al. 2002). This may have been less of an immediate barrier to wind-dispersed *Populus* than some other tree species, but the severing of the Bering land bridge associated with the onset of dramatic climatic oscillations through the Pleistocene were likely the principal drivers for initial divergence between *P. tremula* and *P. tremuloides* (Comes and Kadereit 1998; Milne and Abbott 2002; Du, et al. 2015). We found evidence of low gene flow between the two species following geographical isolation, most likely due to the repeated opening and closing of the narrow strait resulting from sea level fluctuations during the Quaternary glacial-interglacial cycles (Hu, et al. 2010). Although these features of early divergence rule out a strictly allopatic mode of speciation, given the modern-day large geographic isolation, disjunct distributions and extremely low rates of gene flow, our results support an approximately allopatic model of speciation for these two aspen species (Morjan and Rieseberg 2004).
We found that the estimated effective sizes of current populations in both species were larger than that of their ancestral population. The large contemporary $N_e$ of both species are in agreement with their wide geographic distributions and high levels of genomic diversity (Wang, et al. unpublished data). The coalescent-based, intra-species demographic analyses using MSMC also confirmed this pattern, suggesting that both species have experienced substantial population expansion following long-term population declines after divergence. Population expansion of *P. tremuloides* occurred over the last 50,000-70,000 years, following the retreat of the penultimate glaciation and continuing up to the present (Kaufman and Manley 2004). *P. tremula*, in contrast, experienced a more extended population contraction. Consistent with many other forest trees in Europe, the initiation of the substantial expansion in *P. tremula* coincided with the end of the Last Glacial Maximum (Hewitt 2000; Hewitt 2004).

A possible caveat to our demographic inferences is the presence of population subdivision in *P. tremuloides*. However, we found little genetic divergence and similar patterns of genomic variation between the two subpopulations of *P. tremuloides* (see Results), suggesting that subdivision may have occurred too recently to influence our inferences of the speciation processes (Chikhi, et al. 2010). Similarly, sampling could likely be more extensive in both species, to capture a greater extent of the species-wide diversity but this is a perennial concern not restricted to our study. Furthermore, inter-specific hybridization in either species is yet another potential bias. However, there are no other species of *Populus* occurring in the regions from where *P. tremula* were sampled. For *P. tremuloides*, naturally occurring hybridization is only known to occur with *P. grandidentata* in central and eastern North America where the two species co-occur (Pregitzer and Barnes 1980). Therefore any possible
hybridization in our study would be limited to samples from the Wisconsin population of *P. tremuloides*, but as noted above we did not detect any major differences in patterns of genetic variation between the two subpopulations suggesting little or no effect of hybridization.

**Genome-wide patterns of differentiation between two widespread forest tree species**

We detected moderate but also considerable heterogeneous genomic differentiation between *P. tremula* and *P. tremuloides* (Figure 4). Stochastic genetic drift due to geographical isolation has been proposed as the dominant evolutionary force driving the overall patterns of genomic divergence between the two species (Coyne and Orr 2004). Complex demographic histories experienced by the two species also left distinct patterns of genomic variation within and between species. The more dramatic and/or longer period of range expansion in *P. tremuloides* resulted in genome-wide excesses of low frequency alleles (negative Tajima’s D) and the occurrence of more private polymorphisms than in *P. tremula*. In contrast, the longer period of population contraction experienced by *P. tremula* accelerated lineage sorting across the genome due to reductions in effective population size (Fay and Wu 2000), which was supported by the more negative values of Fay & Wu’s H and the greater proportion of derived fixed alleles in *P. tremula* relative to *P. tremuloides* (Table 2). Therefore, our study suggests that neutral processes, e.g. drift and demographic history, were responsible for the majority of genetic differentiation between the two aspen species at a genome-wide scale (Strasburg, et al. 2012).
In addition to the overall pattern generated by these neutral processes, we expected to find regions displaying exceptional differentiation (‘differentiation islands’) that were characterized by multiple independent signatures of positive selection in both species, including excesses of low-frequency alleles, increased high-frequency derived alleles, increased LD, lower proportion of shared polymorphism and/or higher proportion of fixed differences between species compared to the genomic background. Rather than being physically clustered into just a few large, discrete genomic ‘islands’, as expected when species diverge in the presence of gene flow (Turner, et al. 2005), we found differentiation islands to be particularly steep, narrowly defined and located in regions with substantially suppressed recombination throughout the genome. Given the approximately allopatric mode of speciation we envision for the two species, the differentiation islands most likely represent regions harboring loci closely tied to species-specific adaptations rather than those resistant to gene flow (Coyne and Orr 2004; Turner and Hahn 2010; Cruickshank and Hahn 2014). If natural selection is one of the main evolutionary forces shaping patterns of genetic differentiation between these species, regions of low recombination would be expected to show increased $F_{ST}$ values, but not increased $d_{xy}$ values (Noor and Bennett 2009; Cruickshank and Hahn 2014). This occurs because natural selection (through either selective sweeps and/or background selection) removes neutral variation over longer distances in regions of low recombination (Begun and Aquadro 1992). As a consequence, relative measures of divergence (e.g. $F_{ST}$) that rely on within-species diversity are expected to be higher in regions with restricted recombination (Noor and Bennett 2009; Nachman and Payseur 2012). In contrast, increased absolute divergence (e.g. $d_{xy}$) is only expected if reduced gene flow occurred in regions of low recombination (Nachman and Payseur 2012).
accordance with this view, we observed significant negative relationships between population-scaled recombination rates (\(\rho\)) and \(F_{ST}\), but not \(d_{xy}\), in both species (Noor and Bennett 2009; Keinan, et al. 2010).

Taken together, our findings highlight a significant effect of linked selection in generating the heterogeneous differentiation landscape across the genome (Noor and Bennett 2009; Turner and Hahn 2010; Nachman and Payseur 2012; Cruickshank and Hahn 2014).

**Implications for understanding the genetic basis of adaptive evolution**

Patterns of polymorphism and divergence around adaptive sites would allow us to study the degree to which adaptation occurred from new beneficial mutations (hard selective sweeps) or from standing genetic variation (soft selective sweeps) (Pritchard, et al. 2010). Population genetic theory predicts that soft sweeps are common in organisms with large population sizes, because adaptation would not be limited by the availability of beneficial mutations and should proceed primarily from standing genetic variation (Hermisson and Pennings 2005). Accordingly, in forest tree species with large distribution ranges and broad ecological niches, adaptation would seem to be more likely to occur via soft sweeps (Barton and Malik 2010). We found that the large majority (73%) of highly differentiated regions due to adaptive divergence between the two species showed signatures of soft sweeps, but notably 27% of highly differentiated regions showed signatures of hard sweeps in either one or in both aspen species. It has been suggested that regions flanking hard sweeps or regions affected by older hard sweeps could be misidentified as soft sweeps because they may produce spurious population genetic signatures resembling soft sweeps (Schrider, et al. 2015).

Rather than occurring in the “shoulders” of completed hard sweeps, we found that
most regions with signatures of soft sweeps were distributed across the genome without any association to regions linked to hard sweeps (Figure 4). In addition, we found comparable and even more negative values of Fay & Wu’s H in candidate regions linked to soft sweeps compared to those linked to hard sweeps, suggesting that soft selective sweeps are either incomplete or are still ongoing in many of these regions (Fay and Wu 2000). The signals we observed are thus not likely to be a by-product of older hard selective sweeps (Schrider, et al. 2015). Overall, our results suggest that both hard and soft sweeps have been independently involved in divergent adaptation between P. tremula and P. tremuloides (Pritchard and Di Rienzo 2010).

In comparison to divergent selection, long-term balancing selection maintains stable trans-species polymorphisms and leave signatures of unusually low genetic differentiation between species (Charlesworth 2006). We identified a number of genomic regions that are potentially under long-term balancing selection in both P. tremula and P. tremuloides. Apart from low inter-species divergence and high intra-species diversity, these regions are characterized by several other signatures of balancing selection, such as excesses of sites at intermediate frequencies, greater proportions of shared polymorphisms between species and lack of fixed inter-species differences. Due to the long-term effects of recombination on old balanced polymorphisms (Leffler, et al. 2013), we found the signatures and footprints left by balancing selection were generally much narrower and restricted compared to regions under divergent selection, producing comparable or even lower levels of LD than we observed for the rest of the genome. However, after accounting for the influence of local $N_e$ by measuring $\rho/\theta$, we found significantly reduced recombination rates in these regions relative to genome-wide averages, indicating that suppressed
recombination is likely to be critical to the maintenance of beneficial trans-species polymorphisms over long evolutionary time scales (Kamau and Charlesworth 2005).

Candidate genes and functions

We were interested in exploring the functional commonalities of candidate genes within regions putatively under either directional or long-term balancing selection in *P. tremula* and *P. tremuloides*. For candidate genes located in highly differentiated regions, no functional over-representation was found in genes linked to either soft sweeps or species-specific hard sweeps. This indicates that a wide range of genes and functional categories are likely to be involved in rapid adaptation in these widespread species (Wolf, et al. 2010). We only found significant enrichment of GO for genes where hard selective sweeps may have occurred in both species. These genes are mainly associated with transcription initiation and transcription factor activity, suggesting that beneficial mutations affecting transcriptional regulation are likely to be a common source of adaptive evolutionary change between recently diverged species (Wittkopp and Kalay 2012). Regions carrying signatures of long-term balancing selection were enriched for genes involved in signal transduction, immune and defense response. It highlights the influence of co-evolutionary arms races between hosts and natural enemies on the persistence of functional genetic diversity in immunity and defense-related genes (Tiffin and Moeller 2006; Salvaudon, et al. 2008). Future studies of these candidate genes are needed to better assess the adaptive genetic potential of these two widespread forest tree species, and to predict how they might respond to current and future climate.
Conclusion

We have provided insights into the recent evolutionary histories and speciation process separating the two closely related forest tree species, *P. tremula* and *P. tremuloides*. Consistent with the approximately allopatric mode of speciation, we detected moderate levels of genomic differentiation between the two species, and genomic regions of pronounced differentiation were found distributed throughout the genome at many small, independent locations, rather than being clustered into a few large genomic “islands”, as is expected under a model of speciation-with-gene flow. Stochastic genetic drift and historical demographic processes have shaped patterns of polymorphism and differentiation at a genome-wide scale in both species. In addition, we found that species-specific adaptation, mainly involving standing genetic variation via soft selective sweeps, was likely the predominant proximate cause generating the differentiation islands between species, rather than local differences in permeability to gene flow. We must note that this adaptation may largely be unrelated to the speciation process. We also identified multiple signatures of long-term balancing selection in regions of exceptionally low differentiation that appear to have predated speciation. Our study thus highlights that future work should integrate more information on the natural histories of speciation, such as divergence time, geographical context, gene flow magnitudes, demographic histories and sources of adaptation when interpreting the meaning of observed genomic patterns of divergence between closely related species.

Materials and Methods
Population samples, sequencing, quality control and mapping

The analysis workflow of this study is shown in Figure S1. We extracted genomic DNA from leaf samples of 24 genotypes in *P. tremula* and 22 genotypes in *P. tremuloides* (Figure 1a and Table S1). We then constructed 2×100 bp paired-end sequencing libraries with target insert sizes of 650bp for all genotypes that were sequenced on the Illumina HiSeq 2000 platform at the Science for Life Laboratory in Stockholm, Sweden. All samples were sequenced to a target coverage of 25X. The sequencing data have been deposited in the Short Read Archive (SRA) at NCBI under accession IDs ranging from XXXXXX-XXXXXX. The same dataset was also used in (Wang, et al. unpublished data).

For all raw sequencing reads (Wang, et al. 2015), we used Trimmomatic (Lohse, et al. 2012) to remove adapter sequences and cut off bases from either the start or the end of reads when the base quality was lower than 20. Reads were completely discarded if there were fewer than 36 bases remaining after trimming. We then mapped all reads to the *P. trichocarpa* reference genome (v3.0) (Tuskan, et al. 2006), with default parameters implemented in bwa-0.7.10 using the BWA-MEM algorithm (Li 2013). Local realignment was performed to correct for the mis-alignment of bases in regions around insertions and/or deletions (indels) using RealignerTargetCreator and IndelRealigner in GATK v3.2.2 (DePristo, et al. 2011).

In order to account for the occurrence of PCR duplicates introduced during library construction, we used MarkDuplicates in Picard (http://picard.sourceforge.net) to remove reads with identical external coordinates and insert lengths. Only the read with the highest summed base qualities was kept for downstream analyses.

Data filtering and genotype calling
Prior to variant and genotype calling, we employed several filtering steps to exclude potential errors caused by paralogous or repetitive DNA sequences. First, after investigating the empirical distribution, we removed sites showing extremely low (<100 reads across all samples per species) or high (>1200 reads across all samples per species) read coverage. Second, as a mapping quality score of zero is assigned for reads that could be equally mapped to multiple genomic locations, we removed sites containing more than 20 such reads among all samples in each species. Third, we removed sites that overlapped with known repeat elements as identified by RepeatMasker (Tarailo-Graovac and Chen 2009). After all filtering steps, there were 42.8% of sites across the genome left for downstream analyses. Among them, 54.9% were found within gene boundaries, and the remainder (45.1%) was located in intergenic regions.

Two alternative bioinformatics approaches were then used (Figure S1): (1) For those population genetic statistics that relied on the inferred site-frequency-spectrum (SFS), estimation was performed directly from genotype likelihoods without calling genotypes (Nielsen, et al. 2011), as implemented in ANGSD (Korneliussen, et al. 2014). Only reads with a minimal mapping quality of 30 and bases with a minimal quality score of 20 were considered. For all filtered sites in both species, we defined the alleles that were the same as those found in the *P. trichocarpa* reference genome as the ancestral allelic state. We used the -doSaf implementation to calculate the site allele frequency likelihood based on the SAMTools genotype likelihood model in all sites (Li, et al. 2009), and then used the –realSFS implementation to obtain a maximum likelihood estimate of the unfolded SFS using the Expectation Maximization (EM) algorithm (Kim, et al. 2011). Several population genetic statistics were then calculated based on the global SFS (Figure S1). (2) For those estimations
that required accurate genotype calls, single nucleotide polymorphisms (SNPs) and
genotypes were called with HaplotypeCaller in GATK v3.2.2 (Figure S1). A number
of filtering steps were performed to reduce false positives from SNP and genotype
calling: (1) Removed SNPs overlapping sites that did not pass all previous filtering
criteria; (2) Removed SNPs with more than 2 alleles in both species; (3) Removed
SNPs at or within 5bp from any indels; (4) Assigned genotypes as missing if their
quality scores (GQ) were lower than 10, and then removed SNPs with more than two
missing genotypes in each species. (5) Removed SNPs showing significant deviation
from Hardy-Weinberg Equilibrium ($P<0.001$) in each species. In total, we identified
5,894,205 and 6,281,924 SNPs passing these criteria across the 24 $P. \text{tremula}$ samples
and 22 $P. \text{tremuloides}$ samples, respectively.

Population structure

Population genetic structure was inferred using the program NGSadmix (Skotte, et al.
2013), which takes the uncertainly of genotype calling into account and works
directly with genotype likelihoods. Only sites with lower than 10% missing data were
used. We first used the SAMTools model (Li, et al. 2009) in ANGSD to estimate
genotype likelihoods and then generated a beagle file for the subset of the genome
that was determined as being variable using a likelihood ratio test ($P$-value $<10^{-6}$)
(Kim, et al. 2011). We predefined the number of genetic clusters $K$ from 2-5, and the
maximum iteration of the EM algorithm was set to 10,000.

As another method to visualize the genetic relationships among individuals,
we performed principal component analysis (PCA) using ngsTools, which accounts
for sequencing errors and uncertainty in genotype calls (Fumagalli, et al. 2014). The
expected covariance matrix across pairs of individuals from both species was
computed based on the genotype posterior probabilities across all filtered sites. Eigenvectors and eigenvalues from the covariance matrix were generated with the R function eigen, and significance levels were determined using the Tracy-Widom test as implemented in EIGENSOFT version 4.2 (Patterson, et al. 2006).

Demographic history
To infer demographic history associated with speciation of *P. tremula* and *P. tremuloides*, we used a coalescent simulation-based method implemented in fastsimcoal 2.5.1 (Excoffier, et al. 2013). We calculated two-dimensional joint site frequency spectrum (2D-SFS) from posterior probabilities of sample allele frequencies by ngsTools (Fumagalli, et al. 2014). 100,000 coalescent simulations were used for the estimation of the expected 2D-SFS and log-likelihood for a set of demographic parameters in each model. Global maximum likelihood estimates for each model were obtained from 50 independent runs, with 10-40 conditional maximization algorithm cycles, as implemented in fastsimcoal 2.5.1. Eighteen divergence models were examined (Figure S2). All models began with the split of the ancestral population into two sub-populations and differed in terms of (i) whether post-divergence gene flow was present or not, (ii) levels and patterns of gene flow between the two species, and (iii) how population size changes occurred, either at the time of species divergence or afterwards (Figure S2). Model comparison was based on the maximum value of likelihood over the 50 independent runs using the Akaike information criterion (AIC) and Akaike’s weight of evidence (Excoffier, et al. 2013). The model with the maximum Akaike’s weight value was chosen as the optimal one. We assumed a mutation rate of $2.5 \times 10^{-9}$ per site per year in *Populus* (Koch, et al. 2000) and a generation time of 15 years when converting estimates to units of years.
and individuals. Parameter confidence intervals of the best model were obtained by
100 parametric bootstraps, with 50 independent runs in each bootstrap.

We then employed a newly developed multiple sequential Markovian
colescent (MSMC) method (Schiffels and Durbin 2014), which is an extension of a
pairwise sequential Markovian coalescent (PSMC) method (Li and Durbin 2011), to
estimate variation of scaled population sizes ($N_e$) over historical time in both species.
Prior to the analysis, all segregating sites within each species were phased and
imputed using fastPHASE v1.4.0 (Scheet and Stephens 2006). Because MSMC
measures the time to the first coalescence between all pairs of haplotypes, resolution
for recent population size changes can be enhanced if more haplotypes are used
(Schiffels and Durbin 2014). We applied MSMC to phased whole-genome sequences
from one (two haplotypes), two (four haplotypes) and four (eight haplotypes)
individuals in each species, respectively. We did not include more haplotypes because
of the high computational cost of greater samples. A generation time of 15 years and a
rate of $2.5 \times 10^{-9}$ mutations per nucleotide per year (Koch, et al. 2000) were used to
covet the scaled times and population sizes into real times and sizes.

**Genome-wide patterns of differentiation**

Because linkage disequilibrium (LD) decays within 10 kilobases (Kbp) in both *P.
tremula* and *P. tremuloides* (Wang, et al. unpublished data), we divided the genome
into 39,406 non-overlapping windows of 10 Kbp in size to investigate patterns of
genomic differentiation between species. For a window to be included in the
downstream analyses, we required there to be at least 1 Kbp sites left after all above
filtering steps. Levels of genetic differentiation between species at each site were
estimated using method-of-moments $F_{ST}$ estimators implemented in ngsFST from the
ngsTools package (Fumagalli, et al. 2014), which calculates indices of the expected genetic variance between and within species from posterior probabilities of sample allele frequencies, without relying on SNP or genotype calling (Fumagalli, et al. 2013). We then averaged $F_{ST}$ values of all sites within each 10 Kbp non-overlapping window.

We defined outlier windows of exceptionally high interspecies divergence as windows above the top 2.5% of the $F_{ST}$ empirical distribution. As $F_{ST}$ is a relative measure of differentiation and is sensitive to intra-species genetic variation (Charlesworth 1998; Cruickshank and Hahn 2014), we calculated another measure of differentiation, $d_{xy}$, which is the pairwise nucleotide divergence between species and that is independent of within-species diversity (Nei 1987). $d_{xy}$ was calculated from sample allele frequency posterior probabilities at each site using ngsStat from ngsTools software package (Fumagalli, et al. 2014), and was then averaged over non-overlapping 10 Kbp windows. Regions with high $F_{ST}$ but low $d_{xy}$ are more likely to be caused by low ancestral polymorphism at times pre-dating speciation (Cruickshank and Hahn 2014), therefore we retained only those outlier windows with $d_{xy}$ values higher than the genome-wide median value.

We identified windows below the bottom 2.5% of the $F_{ST}$ empirical distribution as outlier windows of exceptionally low levels of interspecies divergence. Through further screening, we found a skewed pattern of low coverage breadth in lowly differentiated windows compared to the genomic background and the highly diverged windows (Figure S3). There is thus the possibility that regions showing low genetic differentiation may contain some artifacts arising from mis-aligned reads due to repetitive sequences or paralogs, despite the stringent quality filters we have
imposed. We thus performed another more stringent filtering on these regions by only retaining windows with at least 3 Kbp sites left from previous quality filtering steps.

**Molecular signature of selection in high- and low-divergence regions**

To assess directional selection in highly divergent regions, we considered both hard- and soft-sweep models. We calculated levels of genetic polymorphism ($\theta_\pi$) using an empirical Bayes approach with the maximum likelihood of unfolded SFS as a prior in ANGSD, within each 10 Kbp non-overlapping window (Kim, et al. 2011). Because hard sweeps would leave a signature of more severe reductions in levels of polymorphism compared to soft sweeps (Hermisson and Pennings 2005; Pennings and Hermisson 2006), a 5% threshold of $\theta_\pi$ was applied for outlier windows of exceptional differentiation to classify divergent regions into four mutually exclusive categories: (1) hard selective sweeps occurred in both species if $\theta_\pi$ dropped below the bottom 5% of empirical distribution in both species; (2) and (3) hard selective sweeps occurred in only one of *P. tremula* or *P. tremuloides* if $\theta_\pi$ dropped below the bottom 5% of empirical distribution only in the respective species; (4) adaptations occurred from standing genetic variation (soft selective sweeps) if $\theta_\pi$ appeared similar to background levels (not below the threshold). In contrast to directional selection, one of the strongest signatures of long-term balancing selection is an excess of polymorphism surrounding the target of selection (Charlesworth 2006). We considered outlier windows with exceptionally low $F_{ST}$ values as potentially being subject to long-term balancing selection if $\theta_\pi$ were in the top 5% of the empirical distribution in both species.

We compared different unions of outlier windows to the remaining portion of the genome by a variety of additional population genetic statistics in both species.
First, Tajima’s D (Tajima 1989) and Fay & Wu’s H (Fay and Wu 2000) were calculated from sample allele frequency likelihoods in ANGSD. Second, levels of LD and population-scaled recombination rates ($\rho$) were estimated based on the SNP data created by GATK. To evaluate levels of LD within each 10 Kbp window, the correlation coefficients ($r^2$) between SNPs with pairwise distances larger than 1 Kbp were calculated using VCFtools v0.1.12b (Danecek, et al. 2011). Population-scaled recombination rates $\rho = 4N_e c$ (Where $N_e$ is the effective population size and c is the recombination rate) were estimated using the Interval program of LDhat 2.2 (McVean, et al. 2004) with 1,000,000 MCMC iterations sampling every 2,000 iterations and a block penalty parameter of five. The first 100,000 iterations of the MCMC iterations were discarded as burn-in. Resulting estimates of $r^2$ and $\rho$ were averaged over each 10 Kbp window. In any of the two species, windows were discarded in the estimation of $r^2$ and $\rho$ if there were less than 3 Kbp and/or 10 SNPs left from previous filtering steps. Finally, we used the program ngsStat (Fumagalli, et al. 2014) to calculate another three measures of genetic differentiation in each window: with P. trichocarpa as an outgroup, the proportion of fixed differences that is caused by either fixed derived alleles in P. tremula or P. tremuloides among all segregating sites; the proportion of inter-species shared polymorphisms among all segregating sites; and the relative node depth (RND). RND was calculated by dividing the $d_{xy}$ of the two aspen species by $d_{xy}$ between aspen (represented by 24 samples of P. tremula in this study) and P. trichocarpa (24 samples; see (Wang, et al. unpublished data). Significance of the differences between regions putatively under selection and the genome-wide averages for all above mentioned population genetic statistics were examined using one-sided Wilcoxon ranked-sum tests.
Gene ontology (GO) enrichment

To determine whether any functional classes of genes were overrepresented among regions putatively under selection, we performed functional enrichment analysis of gene ontology (GO) using Fisher’s exact test by agriGO’s Term Enrichment tool (http://bioinfo.cau.edu.cn/agriGO/index.php; (Du, et al. 2010). GO groups with fewer than two outlier genes were excluded from this analysis. P-values of Fisher’s exact test were further corrected for multiple testing with Benjamini-Hochberg false discovery rate (Benjamini and Hochberg 1995). GO terms with a corrected P-value <0.05 were considered to be significantly enriched.

Acknowledgements

We are grateful to Rick Lindroth for providing access to the samples of *P. tremuloides* used in this study. We thank Carin Olofsson for extracting DNA for all samples used in this study. The research has been funded through grants from Vetenskapsrådet and a Young Researcher Award from Umeå University to PKI. JW was supported by a scholarship from the Chinese Scholarship Council (No. 2011618053).

References


Cruickshank TE, Hahn MW. 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Molecular ecology 23:3133-3157.


Figures and Tables

Figure 1. Geographic distribution and genetic structure of 24 *Populus tremula* and 22 *P. tremuloides*. (a) Map showing the current geographic distribution of *P. tremula* (red) and *P. tremuloides* (blue). Yellow circles and triangles indicate the locations where the 24 individuals of *P. tremula* and 22 individuals of *P. tremuloides* were sampled. (b) Genetic structure of the two aspen species inferred using NGSadmix. The y-axis quantifies subgroup membership, and the x-axis shows the sample ID for each individual. (c) Principal component analysis (PCA) plot based on genetic covariance among all individuals of *P. tremula* (red circle) and *P. tremuloides* (green square and blue triangle). The first two principle components (PCs) are shown, with PC1 explaining 21.04% (*P* = 2.51 × 10^{-19}, Tacey-Widom test) of the overall genetic variation and separating the two species and PC2 explaining 2.09% (*P* = 9.65 × 10^{-4}, Tracy-Widom test) of the overall variation and separating the Wisconsin samples (blue triangle) of *P. tremuloides* from Alberta (green square). (d) Pie chart summarizing the proportion of fixed, shared and exclusive polymorphisms of the two aspen species.

Figure 2. Demographic history of *Populus tremula* and *P. tremuloides*. (a) Graphical summary of the most likely inferred demographic scenario of speciation implemented in fastsimcoal2. (b) Multiple sequential Markovian coalescent (MSMC) estimates of the effective population size (*N_e*) changes for *P. tremula* (red line) and *P. tremuloides* (blue line) based on the inference from two (dashed), four (dotted) and eight (solid) phased haplotypes. Time scale on the x-axis is calculated assuming a
neutral mutation rate per generation ($\mu$) = $3.75 \times 10^{-8}$ and generation time (g) = 15 years. The grey bar indicates the speciation time inferred by *fastsimcoal2*.

**Figure 3. Illustration of the strategy for detecting candidate regions under selection.** (a) Frequency histogram of the distribution of $F_{ST}$ values over 10 Kbp non-overlapping windows across the genome. Data points located to the left ($F_{ST}=0.169$) and right ($F_{ST}=0.681$) vertical dashed lines (corresponding to the bottom and top 2.5% of the empirical $F_{ST}$ distribution) were identified as regions with exceptional low and high differentiation between *P. tremula* and *P. tremuloides*. (b) For windows with exceptional high differentiation between species, two steps of filtering were performed. In step 1, we filtered windows where $d_{xy}$ was smaller than the genome-wide median value. In step 2, the bottom 5% thresholds (short dashed lines) of nucleotide diversity ($\theta_\pi$) was applied to classify divergent windows into four mutually exclusive categories: hard selective sweeps occurred in both species (black dots) if $\theta_\pi$ dropped below the bottom 5% of empirical distribution in both species; hard sweeps occurred in only one of *P. tremula* (red dots) and *P. tremuloides* (blue dots) if $\theta_\pi$ dropped below the bottom 5% of empirical distribution only in the respective species; adaptation occurred from standing genetic variation (soft selective sweeps, grey dots) if $\theta_\pi$ appeared similar as background (not below the threshold). In both (b) and (c) the long dashed lines indicate the genome-wide median values of $\theta_\pi$ in *P. tremula* and *P. tremuloides*, respectively. (c) For windows with exceptional low differentiation between species, we performed two filtering steps to identify regions potentially under long-term balancing selection. In step 1, we filtered windows where the coverage breadth was lower than 3 Kbp. In step 2, only windows where $\theta_\pi$ was above the top 5% of empirical distributions (short dashed lines) in both species were
considered as being under long-term balancing selection (green dots). Please refer to Table 2 for the proportion and other genomic features for each category.

**Figure 4. Genome-wide divergence.** Chromosomal distribution of genetic differentiation ($F_{ST}$) between *Populus tremula* and *P. tremuloides*. The small, light blue dots indicate $F_{ST}$ values estimated over 10 Kbp non-overlapping windows. Grey lines indicate $F_{ST}$ values estimated over 100 Kbp non-overlapping windows. Locations for windows showing specific signatures of selection are highlighted with colored bars above the plot. Among them, candidate windows under hard selective sweeps in either one or both species (red, blue and black bars) are located on the topside; candidate windows with signatures of soft sweeps in either one or both species (grey bars) are located in the middle; and candidate windows under long-term balancing selection (green bars) in both species are located at the bottom of all bars.

**Figure 5. Comparisons of recombination rates ($\rho/\theta_\pi$) between selected regions and genomic background in *P. tremula* (a) and *P. tremuloides* (b).** Black, red, blue, grey, green and light blue boxplots represent $\rho/\theta_\pi$ within regions with signatures of hard sweeps in both species, only in *P. tremula*, only in *P. tremuloides*, soft sweeps in either one or both species, long-term balancing selection and the rest of the genome, respectively. Asterisks designate significant differences between candidate regions with signatures of selection and the rest of genomic regions by Mann-Whitney U test (* $P$-value < 0.05; ** $P$-value < 1e-4; ***$P$-value <2.2e-16).

**Table 1.** Inferred demographic parameters of the divergence history between *P. tremula* and *P. tremuloides* for the best model shown in Figure 2a.
<table>
<thead>
<tr>
<th>Parameters</th>
<th>Point estimation</th>
<th>95% CI(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower</td>
<td>Upper</td>
</tr>
<tr>
<td>(N_{\text{ANC}})</td>
<td>56235</td>
<td>48012</td>
</tr>
<tr>
<td>(N_{P.\text{tremula}})</td>
<td>102814</td>
<td>93688</td>
</tr>
<tr>
<td>(N_{P.\text{tremuloides}})</td>
<td>309500</td>
<td>247321</td>
</tr>
<tr>
<td>(2N_m_{P.\text{tremuloides} \rightarrow P.\text{tremula}})</td>
<td>0.202</td>
<td>0.156</td>
</tr>
<tr>
<td>(2N_m_{P.\text{tremula} \rightarrow P.\text{tremuloides}})</td>
<td>0.053</td>
<td>0.052</td>
</tr>
<tr>
<td>(T_{\text{DIV}})</td>
<td>2332410</td>
<td>2186760</td>
</tr>
</tbody>
</table>

Parameters are defined in Figure 2a. \(N\) indicates the effective population size of \(P.\text{tremula}\), \(P.\text{tremuloides}\) or their ancestral population, \(m\) indicates the migration rates between species on either direction, \(T_{\text{DIV}}\) indicates the estimated divergence time between the two species from fastsimcoal 2.

\(^a\)Parametric bootstrap estimates obtained by parameter estimation from 100 data sets simulated according to the overall maximum composite likelihood estimates shown in point estimation columns. Estimation were obtained from 100,000 simulations per likelihood.

**Table 2. Summary statistics charactering selected regions in both \(P.\text{tremula}\) and \(P.\text{tremuloides}\).**
<table>
<thead>
<tr>
<th>Parameters</th>
<th>Species</th>
<th>Both*</th>
<th>P. tremula*</th>
<th>P. tremuloides*</th>
<th>Soft*</th>
<th>Balancing*</th>
<th>Background*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P. tremula</td>
<td>0.0039**</td>
<td>0.0046**</td>
<td>0.0092**</td>
<td>0.0112***</td>
<td>0.0404***</td>
<td>0.0147***</td>
</tr>
<tr>
<td></td>
<td>P. tremuloides</td>
<td>0.0054***</td>
<td>0.0103**</td>
<td>0.0059***</td>
<td>0.0133***</td>
<td>0.0422***</td>
<td>0.0159**</td>
</tr>
<tr>
<td>Tajima’s D</td>
<td>P. tremula</td>
<td>-1.2763***</td>
<td>-1.2731***</td>
<td>-0.6377**</td>
<td>-0.7137***</td>
<td>0.2433**</td>
<td>-0.3016**</td>
</tr>
<tr>
<td></td>
<td>P. tremuloides</td>
<td>-1.9246***</td>
<td>-1.4904**</td>
<td>-1.8498***</td>
<td>-1.4244***</td>
<td>-0.2711***</td>
<td>-1.1606**</td>
</tr>
<tr>
<td>Fay&amp;Wu’s  H</td>
<td>P. tremula</td>
<td>-0.6659**</td>
<td>-0.8390***</td>
<td>-0.6167**</td>
<td>-0.7786***</td>
<td>-0.0986**</td>
<td>-0.4008**</td>
</tr>
<tr>
<td></td>
<td>P. tremuloides</td>
<td>-0.3908*</td>
<td>-0.4265**</td>
<td>-0.5809**</td>
<td>-0.6168***</td>
<td>-0.0954**</td>
<td>-0.3236**</td>
</tr>
<tr>
<td>r²</td>
<td>P. tremula</td>
<td>0.2796*</td>
<td>0.2700**</td>
<td>0.2786**</td>
<td>0.2571**</td>
<td>0.1718**</td>
<td>0.2115**</td>
</tr>
<tr>
<td></td>
<td>P. tremuloides</td>
<td>0.2894**</td>
<td>0.2336**</td>
<td>0.2250**</td>
<td>0.2202**</td>
<td>0.1393**</td>
<td>0.1575**</td>
</tr>
<tr>
<td>Fixed (%)</td>
<td>P. tremula</td>
<td>0.0997***</td>
<td>0.0609***</td>
<td>0.0625***</td>
<td>0.0470***</td>
<td>~0**</td>
<td>0.0055**</td>
</tr>
<tr>
<td></td>
<td>P. tremuloides</td>
<td>0.0767***</td>
<td>0.0349***</td>
<td>0.0644***</td>
<td>0.0372**</td>
<td>~0**</td>
<td>0.0035**</td>
</tr>
<tr>
<td>Shared (%)</td>
<td>P. tremula</td>
<td>0.0487***</td>
<td>0.0688***</td>
<td>0.0727***</td>
<td>0.1001***</td>
<td>0.3718***</td>
<td>0.1662**</td>
</tr>
<tr>
<td></td>
<td>P. tremuloides</td>
<td>0.8280***</td>
<td>0.7447***</td>
<td>0.7544***</td>
<td>0.7258***</td>
<td>0.1330***</td>
<td>0.3836**</td>
</tr>
<tr>
<td>F_{ST}</td>
<td>P. tremula</td>
<td>0.0270*</td>
<td>0.0297**</td>
<td>0.0309**</td>
<td>0.0448***</td>
<td>0.0468***</td>
<td>0.0248**</td>
</tr>
<tr>
<td></td>
<td>P. tremuloides</td>
<td>0.8404**</td>
<td>0.7527***</td>
<td>0.7789**</td>
<td>0.8105**</td>
<td>0.7614***</td>
<td>0.5509**</td>
</tr>
</tbody>
</table>

*Both indicates candidate regions with signatures of hard selective sweeps in both species; P. tremula and P. tremuloides indicate candidate regions with signatures of hard sweeps only in the respective species; Soft indicates candidate regions with signatures of soft sweeps in either one or both species; Balancing indicates candidate regions with signatures of long-standing balancing selection in both species. Background indicates the rest of the genome not showing exceptional differentiation. Asterisks designate significant differences between the regions under selection and the rest of genomic regions by Mann-Whitney U test (* P-value < 0.05; ** P-value < 1e-4; ***P-value <2.2e-16).
(a) The diagram illustrates the population sizes of two species, *P. tremula* and *P. tremuloides*. The ancestral population size (ANC) is 56235. The population size of *P. tremula* is 102814, and the population size of *P. tremuloides* is 309500.

(b) The graph shows the effective population size over time for both species, with different haplotypes. The x-axis represents time in years ago, and the y-axis represents effective population size on a logarithmic scale.
Step 1. Filter windows where coverage breadth < 3000bp

Step 2.
Hard sweeps in both species
Hard sweeps only in *P. tremula*
Hard sweeps only in *P. tremuloides*
Soft sweeps in either one or both species
Balancing selection in both species