DISSERTATION / DOCTORAL THESIS

Titel der Dissertation /Title of the Doctoral Thesis

„Genomics of the speciation continuum in

Eurasian *Populus* species “

verfasst von / submitted by

Huiying Shang

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

Wien, 2021 / Vienna 2021

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 794 685 437
Dissertationsgebiet lt. Studienblatt / field of study as it appears on the student record sheet:	Biologie
Betreut von / Supervisor:	Assoz. Prof. Dipl.-Ing. Dr. Ovidiu Paun
Mitbetreut von / Co-Supervisor:	Thibault Leroy, PhD

This thesis is dedicated to the initiator of this research, my PhD mentor and initial supervisor, Prof. Christian Lexer (1971-2019).

I thank him for his enthusiasm for this project, for his guidance, support and encouragement throughout this research. What I have learnt from him will always inspire me to move forward.

Table of contents

Abstract

The study of the mechanisms contributing to the formation of new and distinct species (*i.e.*, speciation) is crucial to understanding the origin of biodiversity. Speciation is a dynamic and continuous process, during which multiple evolutionary forces can be at play, modulating the accumulation of genome-wide divergence and allowing the progressive establishment of reproductive isolation. However, it is still a great challenge to determine which evolutionary factors trigger speciation and the heterogeneous patterns of genomic divergence. This dissertation combines phylogenomics and population genomics tools to investigate the evolution of reproductive isolation among multiple *Populus* species pairs from the early to the late stages of speciation. Starting from these results, I discussed the role of background selection, positive selection, balancing selection and gene flow in shaping genomic patterns of differentiation across the speciation continuum. Based on new empirical data and a literature review, the first chapter gives an overview of (*1*) the phylogenomic relationships of several Eurasian *Populus* species; (*2*) the genome-wide heterogeneity in phylogenetic tree topologies and its correlation with recombination rate; and (*3*) how the genomic architecture of reproductive isolation and the levels of introgressive gene flow vary across the stages of speciation. Our literature survey revealed a variation in reproductive barrier complexity, and a negative correlation between barrier number and intensity of gene flow. Genome-wide topology analysis in *Populus* points to a complex genomic architecture of reproductive isolation. In the second chapter, we investigated the fine-scale patterns of genomic diversity and divergence in *Populus*, and discussed the evolutionary factors shaping the heterogeneous landscape of differentiation across the speciation continuum. We uncover a strong interspecific structure, but also extensive introgression between sympatric or parapatric species pairs. Over the whole continuum of divergence, we recovered a negative correlation between nucleotide diversity and relative divergence across all species pairs, which is consistent with expectations under linked selection. However, the positive correlations between nucleotide diversity and absolute divergence became weaker as the overall divergence level (d_a) increased, suggesting that other forces apart from background selection are also at play. Indeed, the negative correlations between introgression (f_{d}) and F_{ST} in some species pairs indicates the contribution of gene flow in shaping genomic landscapes of differentiation. Besides, strong signals of positive or balancing selection have been found along the genome. In spite of this, our landscape genomics analyses confirmed reduced recombination and linked selection as major factors facilitating the heterogeneous genomic divergence in *Populus*. Overall, the study on several *Populus* species across speciation continuum provides general insights about the formation of the heterogeneous landscape of differentiation.

Zusammenfassung

Artbildung ist ein dynamischer Prozess, bei dem mehrere evolutionäre Faktoren wie natürliche Selektion, genetischer Drift, Genfluss und Mutation zur genomweiten Differenzierung beitragen. Ein genaues Verständnis des Speziationsmechanismus ist essentiell, da die Speziation ein entscheidender Prozess ist, der die Artenvielfalt erzeugt. Es ist jedoch immer noch eine große Herausforderung zu bestimmen, welche evolutionären Faktoren eine relativ wichtige Rolle bei der Speziation spielen. In dieser Arbeit habe ich qualitativ hochwertige Ganzgenom-Resequenzierungsdaten verwendet, um die Entwicklung der reproduktiven Isolation bei mehreren *Populus*-Artenpaaren von der frü henbis zur späten Phase der Divergenz zu untersuchen und die Rolle von Hintergrundselektion, positiver Selektion, ausgleichender Selektion und Genfluss bei der Gestaltung der genomischen Muster der Differenzierung auf Populationsgenom-Ebene ü ber das Speziationskontinuum hinweg zu diskutieren. Das erste Kapitel gibt einen Überblick ü ber(1) die phylogenomischen Beziehungen verschiedener *Populus*-Arten in ganz Eurasien; (2) genomweite phylogenetische Baumtopologien und deren Korrelation mit der Rekombinationsrate; (3) wie die genomische Architektur der reproduktiven Isolation und das Ausmaß des introgressiven Genflusses über die Stadien der Speziation variieren. Unsere Literaturrecherche ergab eine Variation in der Komplexität der Barrieren und eine negative Korrelation zwischen der Anzahl der Barrieren und dem Ausmaß des Genflusses. Eine genomweite Topologie-Analyse der *Populus*-Arten weist auf die komplexe genomische Architektur der reproduktiven Isolation hin. Im zweiten Kapitel werfen wir unter Verwendung populationsgenomischer Daten einen Blick auf feinskalige Muster genomischer Diversität und Divergenz ü ber das gesamte Genom und diskutierten die evolutionären Faktoren bei der Gestaltung der heterogenen Landschaft der Differenzierung und wie sich die genomischen Muster entlang des Speziationskontinuums akkumulieren. Analysen der Populationsstruktur und der Identität durch Abstammung zeigen eine starke interspezifische Struktur, aber auch umfangreiche Introgression zwischen einigen Artenpaaren, insbesondere solchen mit parapatrischer Verbreitung. Vergleiche, die aus den Landschaften der genetischen

Diversität und der Rekombinationsrate für jede der Arten gezogen wurden, oder aus der Verteilung der relativen und absoluten Divergenzniveaus für mehrere Artenpaare, die entlang des Speziationskontinuums verteilt waren, zeigen signifikant konservierte Muster. Über das gesamte Kontinuum der Divergenz konnten wir feststellen, dass die Korrelationen zwischen Nukleotiddiversität und Divergenzlandschaften mit zunehmendem Divergenzniveau (da) schwächer werden. Hinsichtlich der Rekombinationslandschaft wurde die Korrelation mit Fst entlang des Divergenzkontinuums nicht stärker, was auf eine wichtige Rolle der Selektion bei der Erzeugung der heterogenen Divergenzlandschaft hindeutet, jedoch ohne Verstärkung während des Prozesses der Speziation. Schließlich weisen die negativen Korrelationen zwischen Introgression (fd) und Fst bei den Artenpaaren *P. tremuloides* - *P. grandidentata* und *P. tremula* - *P. alba* auf die Rolle des Genflusses bei der Gestaltung der genomischen Landschaft der Divergenz hin. Insgesamt wurden in dieser Dissertation phylogenomische und populationsgenomische Werkzeuge kombiniert, um die Evolution von Barrieren der reproduktiven Isolation bei acht eng verwandten *Populus*-Arten zu diskutieren und die Muster der genomischen Diversität und Differenzierung bei mehreren Artenpaaren über das Speziationskontinuum hinweg zu analysieren. Die landschaftsgenomische Analyse bestätigt, dass die reduzierte Rekombination der Hauptfaktor sein kann, der die heterogene Divergenz in *Populus* erleichtert.

1 Introduction

1.1 Speciation and the evolution of reproductive isolation

Speciation, the continuous process leading to the formation of new reproductively isolated species, is one of the most important fields of research in evolutionary biology. Speciation increases biological diversification and opposes extinction. The balance between the number of speciation and extinction events explains the macroevolutionary net diversification rates. Understanding the process of speciation therefore provides important insights into the evolution of biological diversity. Some of the great advances in the field such as the building of the theoretical population genetics (Wright 1931; Fisher 1950; Nei et al. 1983; Ewens 2012), the formulation of theories of speciation (Dobzhansky 1982; Barton & Charlesworth 1984; Wu 1985; Mayr 1999), and more recently, the shift to massive sequence data (*i.e.,* population genomics) thanks to the rapid development of sequencing technologies. The focus of speciation research has gradually shifted from investigating the spatio-temporal conditions in which speciation is possible (*e.g.,* allopatric vs. sympatric, Endler 1977; Felsenstein 1981; Rice and Hostert 1993; Dieckmann and Doebeli 1999; Coyne & Orr 2004; Mallet, et al. 2009) towards the uncovering of the contribution of the neutral and selective forces, including ecological and non-ecological selection (Hoekstra, et al. 2001; Rieseberg, et al. 2002; Lexer and Fay 2005; Rundell and Price 2009; Nosil 2012). More recently, an increasing number of studies have been devoted to the exploration of evolutionary factors that contribute to the genomic landscape of differentiation across multiple species pairs (Irwin, et al. 2018; Ravinet, et al. 2018; Martin, et al. 2019; Stankowski, et al. 2019).

The gradual establishment of reproductive isolation (Lexer, et al. 2010; Rieseberg and Blackman 2010) is crucial to the process of speciation (Wu 2001; Orr, et al. 2004; Noor and Feder 2006; Feder, et al. 2012; Burri 2017b; Stankowski et al. 2019). One key aspect is therefore to understand the conditions contributing to the emergence of reproductive isolation barriers that promote the formation of new species. Reproductive isolation can be enforced by prezygotic and/or postzygotic barriers. In general, prezygotic barriers prevent populations mating with each other or impede fertilization success, while postzygotic barriers include different types of selection against hybrids, like reducing the fitness of hybrids or intrinsic genetic incompatibilities leading to hybrid sterility (Coyne & Orr 2004). Studies in the last several decades have confirmed that many types of prezygotic barriers contribute to speciation, such as ecological or geographic isolation (Grant et al. 2008; Yassin et al. 2016), mating preference or breeding seasons (Jones, et al. 2006; Jones and Ratterman 2009), divergent flowering phenology and pollinator preference (Bradshaw & Schemske 2003; Valente et al. 2012; Armbruster 2014; Chapurlat et al. 2020).

However, whether there are differences in the types of reproductive isolation barriers that usually evolve at different stages of speciation is still an open question. One expectation is that prezygotic, more economic, isolation barriers evolve earlier than postzygotic isolation barriers and play a more important role in reducing gene flow between species (Ramsey et al. 2003; Dopman et al. 2010; Dell'olivo *et al.* 2011). One of the simplest examples is that sexual selection drives population divergence in jumping spiders (Masta & Maddison 2002). In some cases, postzygotic isolation barriers emerge earlier than prezygotic barriers (Pinheiro, et al. 2013; Johnson, et al. 2015). For example, the crossing experiment between two populations of euryhaline killifish *Lucania parva* in fresh water and salt water found no evidence for prezygotic isolation but reduced survival rate of hybrids, which indicates that postzygotic isolation barriers evolve earlier than prezygotic isolation between these two ecological divergent populations and play an important role in reducing gene flow between species (Kozak et al. 2012). The evolution of barriers also depends on the geography (Coyne and Orr 1997). A study on allopatric, parapatric, and sympatric populations of the butterflies *Heliconius elevatus* and *H. pardalinus* supports this conclusion. Nevertheless, strong reproductive isolation between species pairs is not caused by a single isolation barrier, but by a series of different prezygotic and postzygotic reproductive isolation barriers and their potentially complex interactions (Butlin & Smadja 2018).

Recurrent background selection, a form of linked selection due to negative selection against alleles linked to deleterious variants,, or selective sweeps, another form of linked selection due to positive selection, locally reduce genetic diversity and increase differentiation between species, leading to heterogeneous nucleotide diversity estimates along the genome (the so-called “genomic landscape”). During divergence with gene flow, high divergence regions often contain barrier loci that promote speciation. At early stages of speciation, only few barrier loci exist that contribute to differential adaptation or reproductive isolation, while the rest of the genome can be homogenized by gene flow. As divergence increases, there will be more loci involved in differentiation due to linked selection (Feder, et al. 2012). Under the influence of linked selection, we would expect a positive correlation between genetic diversity and recombination, and a negative correlation between genetic divergence and recombination. This phenomenon has already been widely observed in many organisms, such as butterfly (Martin et al. 2019), maize (Tenaillon *et al.* 2002), and humans (Hellmann *et al.* 2003), suggesting the interplay of selection, genetic diversity and recombination.

Recently, fascinating insights have been gained into the genetic basis of reproductive isolation barriers (Widmer, et al. 2009; Baack, et al. 2015). Yet, general insights about speciation through the identification of the genetic basis of reproductive isolation on a single pair in a given model is limiting. More and more studies perform multiple comparisons of incipient species with contrasted levels of genomic divergence (i.e. population/species pairs across the speciation continuum) to investigate speciation (e.g. *Ficedula* flycatchers, Darwin's finches, *Populus*, hummingbirds, and *Heliconius* butterflies), including the variation in the levels of interspecific gene flow (*i.e.* introgression) across the stages of speciation (Burri, et al. 2015; Han, et al. 2017; Ma et al. 2019) and the genomic architecture of reproductive isolation (Supple, et al. 2015; Henderson and Brelsford 2020). However, more related research is needed to understand the interaction between prezygotic and postzygotic isolation barriers and reveal the genetic architecture of reproductive isolation.

1.2 Phylogenomics

Resolving the evolutionary relationships among species is a critical theme in speciation and systematic study. Until recently it has been difficult to impossible to use a large enough number of independent loci to infer species trees, but this is now changing due to the availability of advanced sequencing technology (Foster et al. 2009; Sims et al. 2009; Fontaine et al. 2015; Nater et al. 2015; Árnason et al. 2018). For decades, the most widely used method to construct phylogenetic relationship of species was based on concatenating sequence data to generate a 'supergene tree' (Qiu et al. 1999; Soltis et al. 1999; Olmstead et al. 2001; Jansen et al. 2007; Wang et al. 2009; Lee et al. 2011). By using this method, thousands of phylogenetic analyses were conducted in all fields of biological research (Brandley et al. 2015; Vargas et al. 2017). Although theoretically, accuracy of the species tree should increase with increasing amounts of data, many studies using different datasets have shown that concatenation methods can yield misleading results if species exhibit very large *Ne* and long generation times, or other features resulting in evolutionary heterogeneity among different genomic regions (Kubatko & Degnan 2007; Liu & Edwards 2009; Liu et al. 2015). Thus, owing to the limits of concatenation methods to infer species trees, multi-species coalescent approaches have drawn much attention for phylogenetic estimation in the presence of high incomplete lineage sorting (ILS) among species (Liu et al. 2010; Song et al. 2012; Zhong et al. 2013; Mirarab & Warnow 2015; Mallo & Posada 2016). Due to widespread interspecific gene flow and vegetative propagation, the wind pollinated *Populus* species have relatively large effective population size. In addition, *Populus* trees are expected to have a long generation time (~20 years). We expect this to be a particularly prominent issue for resolving phylogenetic relationships with concatenation methods. In conclusion, coalescent methods make it possible to reconstruct a species tree under an ILS model and can better reflect the evolutionary relationships of species (Liu *et al.* 2010; Song *et al.* 2012).

Phylogenomic approaches based on tree topology variation, can not only be used to uncover the time and order of branching among the lineages but also to investigate the heterogeneity of the genome and estimate the evolution of reproductive isolation. The signal of true relationships according to the species tree may be more common at low recombination regions, whereas the phylogenetic trees that are discordant with the species tree may be caused by gene flow or incomplete lineage sorting. Thus, at high recombination regions where more introgression occured between species, we would expect higher frequency of introgressed tree topologies that are discordant with the species tree. For example, a genome-wide study in *Heliconius* butterflies identified extensive introgression between parapatric distributed species, which dramatically alters the phylogenetic relationships among species (Martin *et al.* 2019). This pattern provides evidence for barriers to gene flow in shaping genomic landscapes of differentiation. In our studies in *Populus*, we examined genome-wide phylogenetic tree topologies and recombination rate during the early and late stage of speciation to test how the phylogenetic tree topologies patterns may be related to the genomic architecture of reproductive isolation. We expected that under linked selection, the species tree topology would have a higher frequency at regions with low recombination, especially for species in a late stage of speciation, while introgressed tree topologies would have a higher frequency at high recombination regions.

1.3 Genomic landscape of differentiation

Under the view of the biological species concept, speciation is the process of the evolution of reproductive isolation. The process is usually not influenced by a single barrier but by the interaction of multiple isolation barriers (Feder et al. 2012; Nosil, et al. 2017; Butlin and Smadja 2018). Understanding the genetic basis of reproductive isolation barriers is still a major task in the field. In the last decades, advances in sequencing technology have provided an excellent opportunity to disentangle the genomic architecture of reproductive isolation barriers and understand the evolutionary factors that contribute to the heterogeneous landscape of genomic differentiation. Genome-wide scans for high differentiation regions are therefore useful to identify loci involved in reproductive isolation, including those associated with adaptation (Talla, et al. 2017; Wolf and Ellegren 2017).

Genome-wide differentiation can be measured using the fixation index F_{ST}, which is a relative divergence measure between two or more populations, influenced by the levels of within population diversity (Wright 1943; Charlesworth 1998; Jakobsson, et al. 2013). Regions with high F_{ST} and low genetic diversity are thought to be candidates for barriers that contribute to reproductive isolation. Yet recent studies in a wide variety of organisms have confirmed the highly heterogeneous nature of the genome-wide landscape of differentiation (Nosil et al. 2009; Martin, et al. 2013; Lamichhaney, et al. 2015; Vijay, et al. 2016). The high divergence regions do not arise solely due to selection on the barriers to gene flow that contribute to reproductive isolation, but often reflect intrinsic genomic features (e.g., variation of recombination rate or gene density along the genome) (Harrison & Larson 2016). It remains difficult to disentangle which evolutionary factors contribute to the variation of genomic diversity and divergence. Further, another important complementary statistic, D_{XY}, which is the absolute divergence and sensitive to ancestral variation (Nei and Li 1979; Charlesworth 1998) has been promoted (Cruickshank & Hahn, 2014). Based on genetic diversity, D_{XY} and F_{ST}, several models have been summarized (Han et al. 2017; Irwin et al. 2018) to explain the formation of genomic landscapes of differentiation (Fig.1).

The first model (a) is ‘divergence with ongoing gene flow’, selection at loci which contribute to reproductive isolation restricts gene exchange between populations, elevating genomic differentiation (i.e., leading to higher F_{ST} and D_{XY}) and reducing genetic diversity. This model can be used to explain genomic features at the early stage of speciation when extensive gene flow between species may still be frequent. The second model (b) is ‘selection in allopatry’, selection on distinct regions of the genome after a species split into two populations, leading to lower π and higher F_{ST}. As D_{XY} is sensitive to ancestral polymorphism, D_{XY} values are expected to remain stable in this model. The next model (c) is ‘recurrent selection’, when background selection or selective sweeps at certain regions of the genome reduce genetic diversity in the common ancestor, and further selection in the two daughter populations leads to lower D_{XY} and π, but higher F_{ST}. The last model (d) is ‘balancing selection’, when ancestral polymorphisms are maintained at selected sites, resulting in increased D_{XY} and low F_{ST} between species. Therefore, the number of different models on the genome can be counted to explain which models may play key roles in shaping the heterogeneous landscape of differentiation.

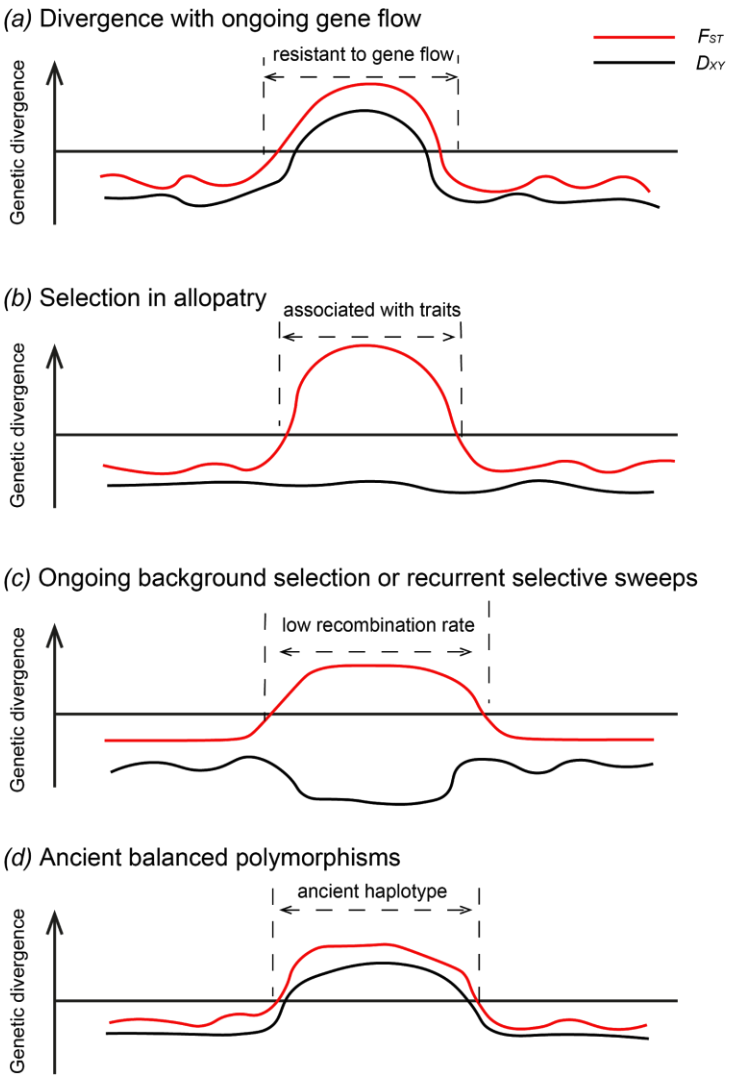


Fig. 1. Expected patterns of genomic islands of divergence under different speciation models. Model *(a)* ‘divergence with ongoing gene flow’ - regions with reproductive barriers prevent gene flow between populations, while the rest of the genome is homogenized by gene flow. Model *(b)* ‘selection in allopatry’ - in allopatric speciation, selection promotes divergence at distinct regions of the genome, leaving low within population genetic diversity, while interspecies diversity does not change. Model *(c)* ‘ongoing background selection or recurrent selective sweeps’ - recurrent selection accumulates divergence at regions of low recombination rate. Model *(d)* ‘ancient balanced polymorphisms’ - highly divergent regions generated before speciation. The figure was redrawn according to Fig. S1 in Han *et al*. 2017.

1.4 Study system

Populus species are widely distributed across the Northern Hemisphere from subtropical to boreal forests, and exhibit strong adaptations to diverse environments. According to the most commonly used classification, *Populus* genus comprises six sections and 29 species, which are traditionally recognized based on morphological traits (Viart 1979; Dickmann & Stuart 1983; Eckenwalder 1996; Heilman 1999). However, obligate outcrossing, abundant wind-pollination, and mixed sexual and vegetative reproductive strategies in *Populus* has led to extensive introgession among species and relatively large effective population size, which complicates phylogenetic inference (Wang, et al. 2016). For example, several studies reported incongruence between phylogenetic trees based on nuclear and chloroplast markers (Liu, et al. 2016; Zhang et al., 2018).

Populus trees have always been favored by mankind because they grow relatively fast to a large size, they can be easily multiplied vegetatively, and have many uses, e.g. fuel, paper industry, furniture or greenery. The genus has also been well studies as a model plant by researchers in various fields because of its favorable genetic attributes such as small genome size (<500 Mb; 2C = 1.1pg in the case of *P. trichocarpa*), diploidy throughout the genus (2*n* = 38), 'porous' species barriers (Meikle 1984; Jansson & Douglas 2007). All these reasons explain why *Populus trichocarpa* was the first tree species sequenced (Tuskan et al. 2006; 2018; Schiffthaler et al. 2019). After more than 15 years of continuous progress, the chromosome-level genome assembly of this species is well curated and annotated, therefore representing one of the best genomic resources available for plants.

In this thesis, we focused on several closely related *Populus* species from section *Populus*, among which *P. alba* and *P. tremula* are two of the most widely distributed species over Eurasia and hybridize naturally and frequently at their hybrid zones (Lexer & Fay 2005; Lexer *et al.* 2010; Stölting et al. 2015), despite the strong postzygotic reproductive isolation barriers due to genomic incompatibilities and variable prezygotic barriers. *Populus davidiana* is also distributed over a wide area, from northeastern to the central part of China, while *P. rotundifolia* (a sister species to *P. davidiana*) inhabits high-altitude regions of Qinghai-Tibetan plateau. The latter two species were thought to have recently undergone parapatric speciation with ongoing gene flow, owing to their divergent ecological environment. A subdivision of *P. davidiana* into northeast and central groups was supported by several pieces of evidence (Zheng et al. 2017b; Song et al. 2021). *Populus adenopoda* widely grows in warm and moist subtropical areas of south and east China (Fan et al. 2018a). The endangered species *P. qiongdaoensis* only occurs on Hainan Islands (Luo & Hong 1987; Liang & Fang 2012). In addition, we also investigated two North American aspens, *P. tremuloides* and *P. grandidentata*, which have distinct morphologies but hybridize in areas of distributions' overlap (Deacon et al. 2019). The divergence of *P. tremuloides*, *P. tremula* and *P. davidiana* is thought to have been triggered by the emergence of Bering Land Bridge and the uplift of Qinghai-Tibetan plateau. Molecular dating analysis has proved *P. tremuloides* splitted earlier than the other two aspens (Du et al. 2015; Wang *et al.* 2020).

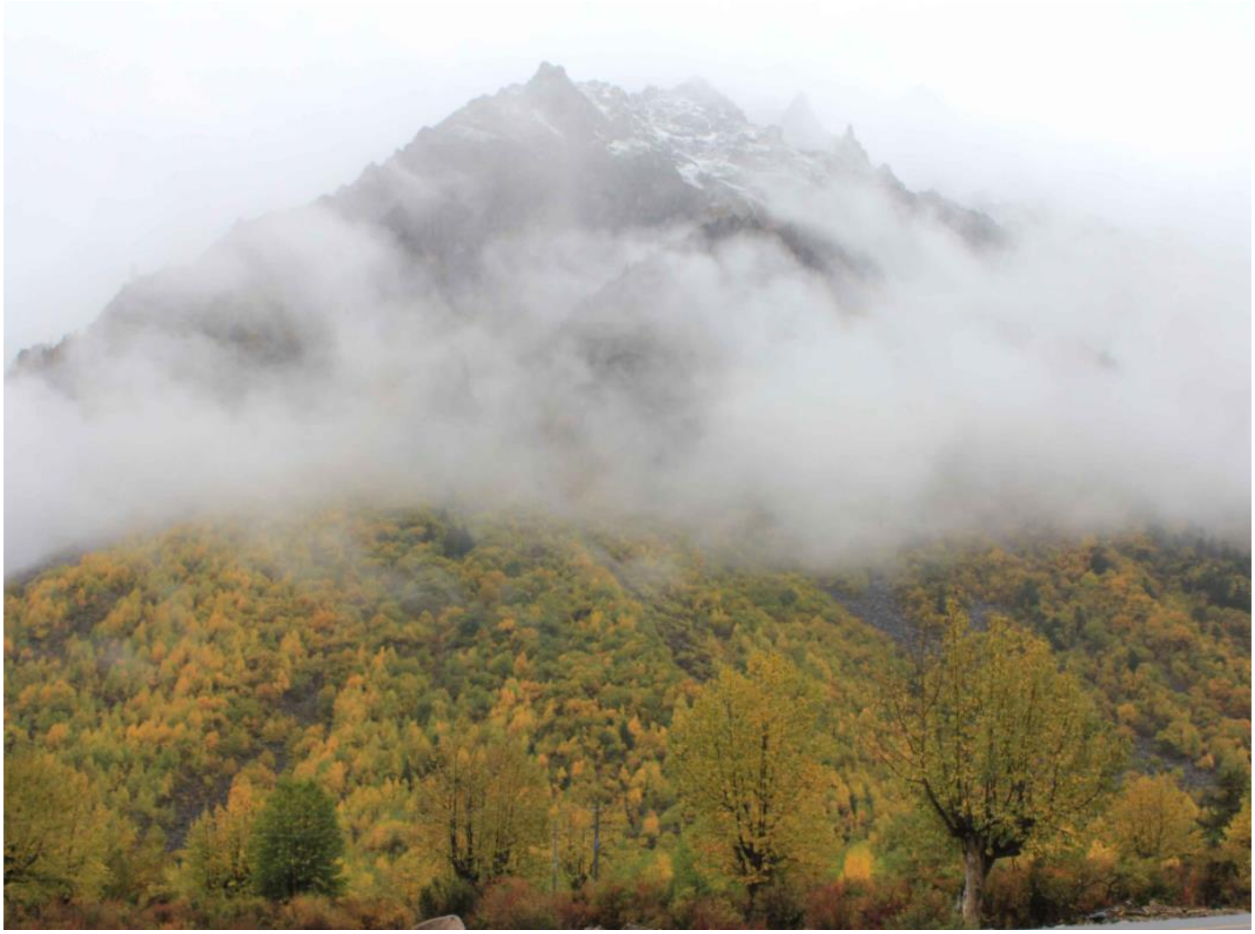


Fig. 2. The photo shows a mixed forest with representative species of *Populus rotundifolia*, *P. mainlingensis*, *Betula platyphylla* and *Picea brachytyla* var. *complanate*. It was taken by Kangshan Mao (reproduced with permission) at Nyingchi, Tibet, China on 7th of October, 2013.

1.5 Research objectives

In the first chapter, we collected the samples and generated whole genome resequencing data of 36 individuals from seven ingroup *Populus* species and two outgroup species (*P. balsamifera* and *P. trichocarpa*) to explore genome-wide patterns of interspecific divergence and gene flow in section *Populus*. In particular, our objectives were the following: (1) to explore phylogenomic relationships in this Eurasian species complex of the model forest tree genus *Populus* with unprecedented depth, making using of both concatenation and coalescent approaches, (2) assess the extent of variation in tree topologies and gene genealogies along the genome, (3) estimate the influence of ILS and gene flow on gene tree topologies, (4) determine how genome-wide phylogenomic patterns are mediated by the genomic architecture of reproductive isolation and recombination rate.

In the second chapter, we resequenced whole genomes for 201 individuals from eight *Populus* species to address the following questions: (1) what are the population structure and demographic histories of each *Populus* species? (2) what are the characteristics of the genomic landscape of differentiation for each species pair across the speciation continuum? (3) are the differentiation patterns repeatable among independent divergence events? (4) what evolutionary factors or processes drive the heterogeneity of genomic landscapes of diversity and differentiation?

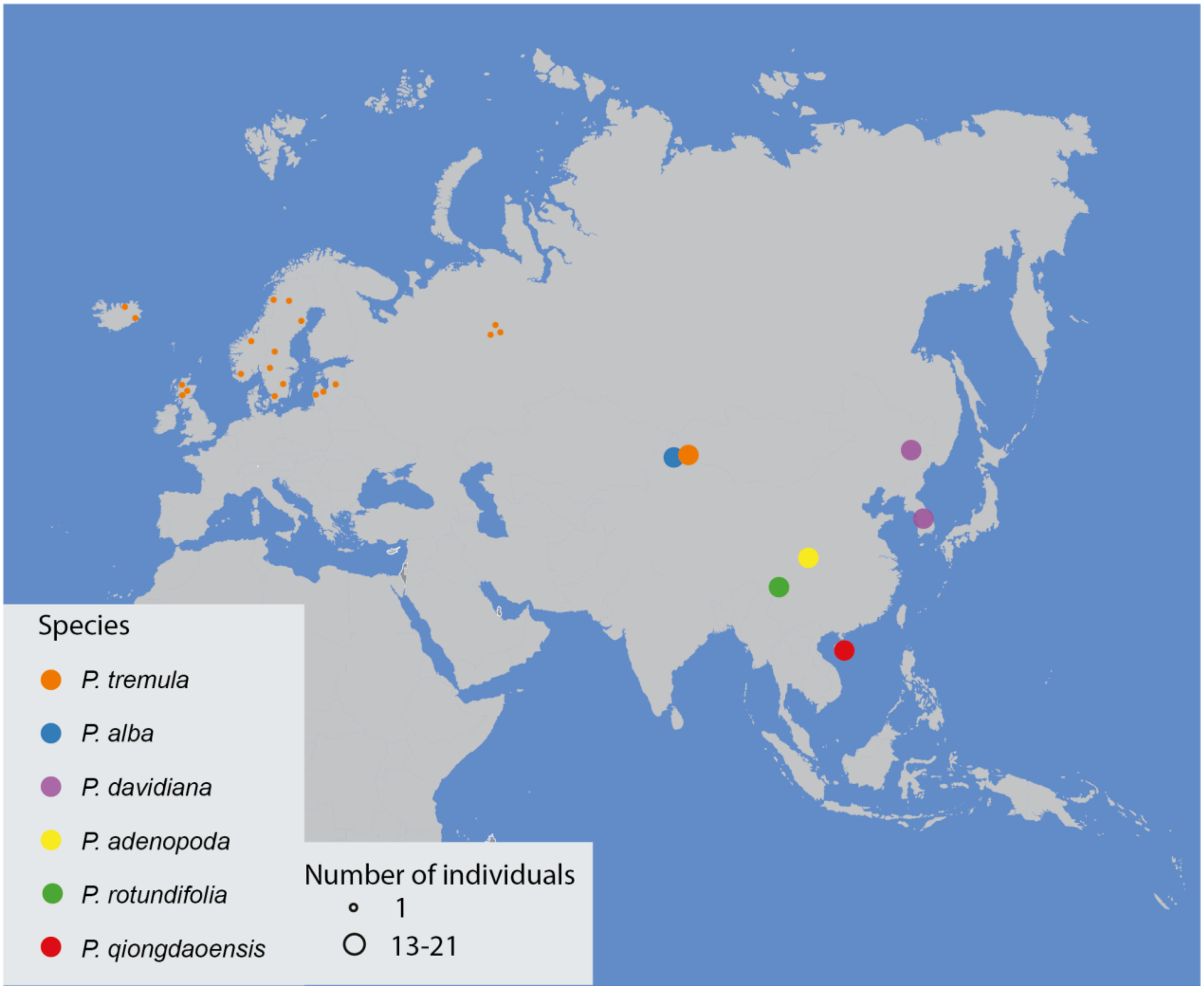


Fig. 3. Map of the *Populus* samples collected in Eurasia. In the map different colors represent different species, while the size of the circle represents the number of samples collected in that place.

1.6 References

1. Akbari A., Vitti J.J., Iranmehr A., Bakhtiari M., Sabeti P.C., Mirarab S. & Bafna V. (2018) Identifying the favored mutation in a positive selective sweep. Nat Methods 15, 279-82.
2. Alexander D.H. & Lange K. (2011) Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics 12, 246.
3. Armbruster W.S. (2014) Floral specialization and angiosperm diversity: phenotypic divergence, fitness trade-offs and realized pollination accuracy. AoB Plants 6.
4. Árnason Ú., Lammers F., Kumar V., Nilsson M.A. & Janke A. (2018) Whole-genome sequencing of the blue whale and other rorquals finds signatures for introgressive gene flow. Sci Adv 4, eaap9873.
5. Bank C., Bürger R. & Hermisson J. (2012) The limits to parapatric speciation: Dobzhansky-Muller incompatibilities in a continent-island model. Genetics 191, 845-63.
6. Barnes B.V. (1959) Natural variation and clonal development of *Populus tremuloides* and *P. grandidentata* in northern lower Michigan.
7. Barton N.H. & Charlesworth B. (1984) Genetic revolutions, founder effects, and speciation. Annual Review of Ecology and Systematics 15, 133-64.
8. Behrmann-Godel J. & Gerlach G. (2008) First evidence for postzygotic reproductive isolation between two populations of Eurasian perch (*Perca fluviatilis* L.) within Lake Constance. Front Zool 5, 3.
9. Bradshaw H.D. & Schemske D.W. (2003) Allele substitution at a flower colour locus produces a pollinator shift in monkeyflowers. Nature 426, 176-8.
10. Brandley M.C., Bragg J.G., Singhal S., Chapple D.G., Jennings C.K., Lemmon A.R., Lemmon E.M., Thompson M.B. & Moritz C. (2015) Evaluating the performance of anchored hybrid enrichment at the tips of the tree of life: a phylogenetic analysis of Australian Eugongylus group scincid lizards. BMC Evol Biol 15, 62.
11. Browning B.L. & Browning S.R. (2013) Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics 194, 459-71.
12. Burri R. (2017a) Dissecting differentiation landscapes: a linked selection's perspective. J Evol Biol 30, 1501-5.
13. Burri R. (2017b) Linked selection, demography and the evolution of correlated genomic landscapes in birds and beyond. Mol Ecol 26, 3853-6.
14. Burri R., Nater A., Kawakami T., Mugal C.F., Olason P.I., Smeds L., Suh A., Dutoit L., Bureš S., Garamszegi L.Z., Hogner S., Moreno J., Qvarnström A., Ružić M., Sæther S.A., Sætre G.P.,

Török J. & Ellegren H. (2015) Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula* flycatchers. Genome Res 25, 1656-65.

15. Butlin R., Debelle A., Kerth C., Snook R.R., Beukeboom L.W., Castillo Cajas R.F., Diao W., Maan M.E., Paolucci S., Weissing F.J., van de Zande L., Hoikkala A., Geuverink E., Jennings J., Kankare M., Knott K.E., Tyukmaeva V.I., Zoumadakis C., Ritchie M.G., Barker D., Immonen E., Kirkpatrick M., Noor M., Macias Garcia C., Schmitt T. & Schilthuizen M. (2012) What do we need to know about speciation? Trends Ecol Evol 27, 27-39.

16. Butlin R.K. & Smadja C.M. (2018) Coupling, Reinforcement, and Speciation. Am Nat 191, 155-72.

17. Campagna L., Gronau I., Silveira L.F., Siepel A. & Lovette I.J. (2015) Distinguishing noise from signal in patterns of genomic divergence in a highly polymorphic avian radiation. Mol Ecol 24, 4238-51.

18. Chapurlat E., Le Roncé I., Ågren J. & Sletvold N. (2020) Divergent selection on flowering phenology but not on floral morphology between two closely related orchids. Ecol Evol 10, 5737-47.

19. Charlesworth B. (1998) Measures of divergence between populations and the effect of forces that reduce variability. Mol Biol Evol 15, 538-43.

20. Charlesworth B., Morgan M.T. & Charlesworth D. (1993) The effect of deleterious mutations on neutral molecular variation. Genetics 134, 1289-303.

21. Coyne J.A. & Orr H.A. (2004) *Speciation*. Sinauer Associates Sunderland, MA.

22. Cruickshank T.E. & Hahn M.W. (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Mol Ecol 23, 3133-57.

23. Deacon N.J., Grossman J.J. & Cavender-Bares J. (2019) Drought and freezing vulnerability of the isolated hybrid aspen *Populus x smithii* relative to its parental species, *P. tremuloides* and *P. grandidentata*. Ecol Evol 9, 8062-74.

24. Dell'olivo A., Hoballah M.E., Gübitz T. & Kuhlemeier C. (2011) Isolation barriers between petunia axillaris and *Petunia integrifolia* (Solanaceae). Evolution 65, 1979-91.

25. DePristo M.A., Banks E., Poplin R., Garimella K.V., Maguire J.R., Hartl C., Philippakis A.A., del Angel G., Rivas M.A., Hanna M., McKenna A., Fennell T.J., Kernytsky A.M., Sivachenko A.Y., Cibulskis K., Gabriel S.B., Altshuler D. & Daly M.J. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43, 491-8.

26. Dickmann D.I. & Stuart K.W. (1983) The culture of poplars in Eastern North America. The culture of poplars in eastern North America.

27. Dieckmann U., Doebeli M., Metz J.A. & Tautz D. (2004) Adaptive speciation. Cambridge University Press.

28. Dobzhansky T. (1982) Genetics and the Origin of Species. Columbia university press.

29. Dopman E.B., Robbins P.S. & Seaman A. (2010) Components of reproductive isolation between North American pheromone strains of the European corn borer. Evolution 64, 881-902.

30. Du S., Wang Z., Ingvarsson P.K., Wang D., Wang J., Wu Z., Tembrock L.R. & Zhang J. (2015) Multilocus analysis of nucleotide variation and speciation in three closely related *Populus* (Salicaceae) species. Mol Ecol 24, 4994-5005.

31. Eckenwalder J. (1996) Systematics and evolution of *Populus*. U: Stettler, RF, Bradshaw, HD, Heilman, PE, Hinckley, TM, eds.(1996): Biology of *Populus* and Its Implications for Management and Conservation. NRC Research Press, Ottawa, Ontario, Canada.

32. Ellegren H., Smeds L., Burri R., Olason P.I., Backström N., Kawakami T., Künstner A., Mäkinen H., Nadachowska-Brzyska K., Qvarnström A., Uebbing S. & Wolf J.B. (2012) The genomic landscape of species divergence in *Ficedula* flycatchers. Nature 491, 756-60.

33. Ellegren H. & Wolf J.B.W. (2017) Parallelism in genomic landscapes of differentiation, conserved genomic features and the role of linked selection. J Evol Biol 30, 1516-8.

34. Endler J.A. (1977) Geographic variation, speciation, and clines. Princeton University Press.

35. Fan L., Zheng H., Milne R.I., Zhang L. & Mao K. (2018a) Strong population bottleneck and repeated demographic expansions of *Populus adenopoda* (Salicaceae) in subtropical China. Ann Bot 121, 665-79.

36. Fan L., Zheng H., Milne R.I., Zhang L. & Mao K. (2018b) Strong population bottleneck and repeated demographic expansions of *Populus adenopoda* (Salicaceae) in subtropical China. Ann Bot 121, 665-79.

37. Fang Z., Zhao S. & Skvortsov A. (1999) Salicaceae. Flora of China 4, 139-274.

38. Feder J.L., Egan S.P. & Nosil P. (2012) The genomics of speciation-with-gene-flow. Trends Genet 28, 342-50.

39. Felsenstein J. (1981) Skepticism towards Santa Rosalia, or why are there so few kinds of animals? Evolution, 124-38.

40. Fisher R.A. (1950) Gene frequencies in a cline determined by selection and diffusion. Biometrics 6, 353-61.

41. Fitzpatrick B.M., Fordyce J.A. & Gavrilets S. (2008) What, if anything, is sympatric speciation? J Evol Biol 21, 1452-9.

42. Fontaine M.C., Pease J.B., Steele A., Waterhouse R.M., Neafsey D.E., Sharakhov I.V., Jiang X., Hall A.B., Catteruccia F., Kakani E., et al. 2015. Mosquito genomics. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science 347:1258524.

43. Foster J.T., Beckstrom-Sternberg S.M., Pearson T., Beckstrom-Sternberg J.S., Chain P.S., Roberto F.F., Hnath J., Brettin T. & Keim P. (2009) Whole-genome-based phylogeny and divergence of the genus *Brucella*. J Bacteriol 191, 2864-70.

44. Fredrickson R.J., Siminski P., Woolf M. & Hedrick P.W. (2007) Genetic rescue and inbreeding depression in Mexican wolves. Proc Biol Sci 274, 2365-71.

45. Gagnaire P.A., Lamy J.B., Cornette F., Heurtebise S., Dégremont L., Flahauw E., Boudry P., Bierne N. & Lapègue S. (2018) Analysis of genome-wide differentiation between native and introduced populations of the cupped oysters *Crassostrea gigas* and *Crassostrea angulata*. Genome Biol Evol 10, 2518-34.

46. Gao F., Ming C., Hu W. & Li H. (2016) New Software for the Fast Estimation of Population Recombination Rates (FastEPRR) in the Genomic Era. G3 (Bethesda) 6, 1563-71.

47. Gavrilets S. (2004) Fitness landscapes and the origin of species *(MPB-41)*. Princeton University Press.

48. Gebiola M., Kelly S.E., Hammerstein P., Giorgini M. & Hunter M.S. (2016) "Darwin's corollary" and cytoplasmic incompatibility induced by Cardinium may contribute to speciation in Encarsia wasps (Hymenoptera: Aphelinidae). Evolution 70, 2447-58.

49. Grant P.R., Grant B.R. & Petren K. (2008) The allopatric phase of speciation: the sharp-beaked ground finch (*Geospiza difficilis*) on the Galápagos islands. Biol J Linn Soc 69, 287-317.

50. Han F., Lamichhaney S., Grant B.R., Grant P.R., Andersson L. & Webster M.T. (2017) Gene flow, ancient polymorphism, and ecological adaptation shape the genomic landscape of divergence among Darwin's finches. Genome Res 27, 1004-15.

51. Harrison R.G. & Larson E.L. (2016) Heterogeneous genome divergence, differential introgression, and the origin and structure of hybrid zones. Mol Ecol 25, 2454-66.

52. Heilman P.E. (1999) Planted forests: poplars. New Forests 17, 89-93.

53. Henderson E.C. & Brelsford A. (2020) Genomic differentiation across the speciation continuum in three hummingbird species pairs. BMC Evol Biol 20, 113.

54. Hohenlohe P.A., Bassham S., Etter P.D., Stiffler N., Johnson E.A. & Cresko W.A. (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet 6, e1000862.

55. Hou Z. & Li A. (2020) Population genomics reveals demographic history and genomic differentiation of *Populus davidiana* and *Populus tremula*. Front Plant Sci 11, 1103.

56. Irwin D.E., Milá B., Toews D.P.L., Brelsford A., Kenyon H.L., Porter A.N., Grossen C., Delmore K.E., Alcaide M. & Irwin J.H. (2018) A comparison of genomic islands of differentiation across three young avian species pairs. Mol Ecol 27, 4839-55.

57. Jansen R.K., Cai Z., Raubeson L.A., Daniell H., Depamphilis C.W., Leebens-Mack J., Müller K.F., Guisinger-Bellian M., Haberle R.C., Hansen A.K., Chumley T.W., Lee S.B., Peery R., McNeal J.R., Kuehl J.V. & Boore J.L. (2007) Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci U S A 104, 19369-74.

58. Jansson S. & Douglas C.J. (2007) *Populus*: a model system for plant biology. Annu Rev Plant Biol 58, 435-58.

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

60. Kimura M. (1991) The neutral theory of molecular evolution: a review of recent evidence. Jpn J Genet 66, 367-86.

61. Kozak G.M., Rudolph A.B., Colon B.L. & Fuller R.C. (2012) Postzygotic Isolation Evolves before Prezygotic Isolation between Fresh and Saltwater Populations of the Rainwater Killifish, *Lucania parva*. Int J Evol Biol 2012, 523967.

62. Kubatko L.S. & Degnan J.H. (2007) Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst Biol 56, 17-24.

63. Kulmuni J. & Westram A.M. (2017) Intrinsic incompatibilities evolving as a by-product of divergent ecological selection: Considering them in empirical studies on divergence with gene flow. Mol Ecol 26, 3093-103.

64. Lamichhaney S., Berglund J., Almén M.S., Maqbool K., Grabherr M., Martinez-Barrio A., Promerová M., Rubin C.J., Wang C., Zamani N., et al. 2015. Evolution of Darwin's finches and their beaks revealed by genome sequencing. Nature 518:371-375.

65. Lee E.K., Cibrian-Jaramillo A., Kolokotronis S.O., Katari M.S., Stamatakis A., Ott M., Chiu J.C., Little D.P., Stevenson D.W., McCombie W.R., et al. 2011. A functional phylogenomic view of the seed plants. PLoS Genet 7:e1002411.

66. Lehnert S.J., Kess T., Bentzen P., Clément M. & Bradbury I.R. (2020) Divergent and linked selection shape patterns of genomic differentiation between European and North American Atlantic salmon (*Salmo salar*). Mol Ecol 29, 2160-75.

67. Lexer C., Buerkle C.A., Joseph J.A., Heinze B. & Fay M.F. (2007) Admixture in European *Populus* hybrid zones makes feasible the mapping of loci that contribute to reproductive isolation and trait differences. Heredity (Edinb) 98, 74-84.

68. Lexer C. & Fay M.F. (2005) Adaptation to environmental stress: a rare or frequent driver of speciation? J Evol Biol 18, 893-900.

69. Lexer C., Fay M.F., Joseph J.A., Nica M.S. & Heinze B. (2005) Barrier to gene flow between two ecologically divergent *Populus* species, *P. alba* (white poplar) and *P. tremula* (European aspen): the role of ecology and life history in gene introgression. Mol Ecol 14, 1045-57.

70. Lexer C., Joseph J.A., van Loo M., Barbará T., Heinze B., Bartha D., Castiglione S., Fay M.F. & Buerkle C.A. (2010) Genomic admixture analysis in European *Populus* spp. reveals unexpected patterns of reproductive isolation and mating. Genetics 186, 699-712.

71. Li H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997.

72. Liang J. & Fang F. (2012) Survey on wild *Populus qiongdaoensis* T. Hong et P. Luo resource. Tropical Forestry 40, 32-4.

73. Lindtke D., Gompert Z., Lexer C. & Buerkle C.A. (2014) Unexpected ancestry of *Populus* seedlings from a hybrid zone implies a large role for postzygotic selection in the maintenance of species. Mol Ecol 23, 4316-30.

74. Liu L. & Edwards S.V. (2009) Phylogenetic analysis in the anomaly zone. Syst Biol 58, 452-60.

75. Liu L., Xi Z., Wu S., Davis C.C. & Edwards S.V. (2015) Estimating phylogenetic trees from genome-scale data. Ann N Y Acad Sci 1360, 36-53.

76. Liu L., Yu L. & Edwards S.V. (2010) A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol Biol 10, 302.

77. Luo P. & Hong T. (1987) A new species of *Populus* in tropical forests from Hainan. Bull Botanical Res 7, 67-70.

78. Mallet J., Meyer A., Nosil P. & Feder J.L. (2009) Space, sympatry and speciation. J Evol Biol 22, 2332-41.

79. Mallo D. & Posada D. (2016) Multilocus inference of species trees and DNA barcoding. Philos Trans R Soc Lond B Biol Sci 371.

80. Martin S.H., Dasmahapatra K.K., Nadeau N.J., Salazar C., Walters J.R., Simpson F., Blaxter M., Manica A., Mallet J. & Jiggins C.D. (2013) Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. Genome Res 23, 1817-28.

81. Martin S.H., Davey J.W. & Jiggins C.D. (2015) Evaluating the use of ABBA-BABA statistics to locate introgressed loci. Mol Biol Evol 32, 244-57.

82. Martin S.H., Davey J.W., Salazar C. & Jiggins C.D. (2019) Recombination rate variation shapes barriers to introgression across butterfly genomes. PLoS Biol 17, e2006288.

83. Martin S.H. & Jiggins C.D. (2017) Interpreting the genomic landscape of introgression. Curr Opin Genet Dev 47, 69-74.

84. Martinsen G.D., Whitham T.G., Turek R.J. & Keim P. (2001) Hybrid populations selectively filter gene introgression between species. Evolution 55, 1325-35.

85. Masta S.E. & Maddison W.P. (2002) Sexual selection driving diversification in jumping spiders.Proc Natl Acad Sci U S A 99, 4442-7.

86. Mattila T.M., Laenen B., Horvath R., Hämälä T., Savolainen O. & Slotte T. (2019) Impact of demography on linked selection in two outcrossing *Brassicaceae* species. Ecol Evol 9, 9532-45.

87. Mayr E. (1942) Systematics and the origin of species. Columbia Univ. Press, New York. Systematics and the origin of species. Columbia Univ. Press, New York., -.

88. Meikle R.D. (1984) Willows and poplars of Great Britain and Ireland. Botanical Society of the British Isles.

89. Mirarab S. & Warnow T. (2015) ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31, i44-52.

90. Nater A., Burri R., Kawakami T., Smeds L. & Ellegren H. (2015) Resolving evolutionary relationships in closely related species with whole-genome sequencing data. Syst Biol 64, 1000-17.

91. Nei M. (1987) Molecular evolutionary genetics. Columbia university press.

92. Nei M., Maruyama T. & Wu C.I. (1983) Models of evolution of reproductive isolation. Genetics 103, 557-79.

93. Nei M., Suzuki Y. & Nozawa M. (2010) The neutral theory of molecular evolution in the genomic era. Annu Rev Genomics Hum Genet 11, 265-89.

94. Noor M.A. & Bennett S.M. (2009) Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. Heredity (Edinb) 103, 439-44.

95. Nosil P. (2012) Ecological speciation. Oxford University Press.

96. Nosil P. & Feder J.L. (2012) Widespread yet heterogeneous genomic divergence. Mol Ecol 21, 2829-32.

97. Nosil P., Funk D.J. & Ortiz-Barrientos D. (2009) Divergent selection and heterogeneous genomic divergence. Mol Ecol 18, 375-402.

98. Olmstead R.G., Depamphilis C.W., Wolfe A.D., Young N.D., Elisons W.J. & Reeves P.A. (2001) Disintegration of the scrophulariaceae. Am J Bot 88, 348-61.

99. Pabinger S., Dander A., Fischer M., Snajder R., Sperk M., Efremova M., Krabichler B., Speicher M.R., Zschocke J. & Trajanoski Z. (2014) A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform 15, 256-78.

100. Payseur B.A. & Rieseberg L.H. (2016) A genomic perspective on hybridization and speciation. Mol Ecol 25, 2337-60.

101. Pickrell J.K. & Pritchard J.K. (2012) Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet 8, e1002967.

102. Pinho C. & Hey J. (2010) Divergence with gene flow: models and data. Annu Rev Ecol Evol Syst 41, 215-30.

103. Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A., Bender D., Maller J., Sklar P., de Bakker P.I., Daly M.J. & Sham P.C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81, 559-75.

104. Qiu Y.L., Lee J., Bernasconi-Quadroni F., Soltis D.E., Soltis P.S., Zanis M., Zimmer E.A., Chen Z., Savolainen V. & Chase M.W. (1999) The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. Nature 402, 404-7.

105. Rajora O.P. & Dancik B.P. (1992) Genetic characterization and relationships of *Populus alba*, *P. tremula*, and *P. x canescens*, and their clones. Theor Appl Genet 84, 291-8.

106. Ramsey J., Bradshaw H.D., Jr. & Schemske D.W. (2003) Components of reproductive isolation between the monkeyflowers *Mimulus lewisii* and *M. cardinalis* (*Phrymaceae*). Evolution 57, 1520-34.

107. Ravinet M., Faria R., Butlin R.K., Galindo J., Bierne N., Rafajlović M., Noor M.A.F., Mehlig B. & Westram A.M. (2017) Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. J Evol Biol 30, 1450-77.

108. Ravinet M., Yoshida K., Shigenobu S., Toyoda A., Fujiyama A. & Kitano J. (2018) The genomic landscape at a late stage of stickleback speciation: High genomic divergence interspersed by small localized regions of introgression. PLoS Genet 14, e1007358.

109. Renaut S., Grassa C.J., Yeaman S., Moyers B.T., Lai Z., Kane N.C., Bowers J.E., Burke J.M. & Rieseberg L.H. (2013) Genomic islands of divergence are not affected by geography of speciation in sunflowers. Nat Commun 4, 1827.

110. Renaut S., Owens G.L. & Rieseberg L.H. (2014) Shared selective pressure and local genomic landscape lead to repeatable patterns of genomic divergence in sunflowers. Mol Ecol 23, 311-24.

111. Rifkin J.L., Castillo A.S., Liao I.T. & Rausher M.D. (2019) Gene flow, divergent selection and resistance to introgression in two species of morning glories (*Ipomoea*). Mol Ecol 28, 1709-29.

112. Roux C., Fraïsse C., Romiguier J., Anciaux Y., Galtier N. & Bierne N. (2016) Shedding light on the grey zone of speciation along a continuum of genomic divergence. PLoS Biol 14, e2000234.

113. Samuk K., Manzano-Winkler B., Ritz K.R. & Noor M.A.F. (2020) Natural selection shapes variation in genome-wide recombination rate in *Drosophila pseudoobscura*. Curr Biol 30, 1517-28.e6.

114. Samuk K., Owens G.L., Delmore K.E., Miller S.E., Rennison D.J. & Schluter D. (2017) Gene flow and selection interact to promote adaptive divergence in regions of low recombination. Mol Ecol 26, 4378-90.

115. Schiffthaler B., Delhomme N., Bernhardsson C., Jenkins J., Jansson S., Ingvarsson P., Schmutz J. & Street N. (2019) An improved genome assembly of the European aspen *Populus tremula*. bioRxiv, 805614.

116. Schluter D. (2000) The ecology of adaptive radiation. OUP Oxford.

117. Sendell-Price A.T., Ruegg K.C., Anderson E.C., Quilodrán C.S., Van Doren B.M., Underwood V.L., Coulson T. & Clegg S.M. (2020) The genomic landscape of divergence across the speciation continuum in Island-Colonising silvereyes (*Zosterops lateralis*). G3 (Bethesda) 10, 3147-63.

118. Shang H., Hess J., Pickup M., Field D.L., Ingvarsson P.K., Liu J. & Lexer C. (2020) Evolution of strong reproductive isolation in plants: broad-scale patterns and lessons from a perennial model group. Philos Trans R Soc Lond B Biol Sci 375, 20190544.

119. Siewert K.M. & Voight B.F. (2017) Detecting Long-Term Balancing Selection Using Allele Frequency Correlation. Mol Biol Evol 34, 2996-3005.

120. Sims G.E., Jun S.-R., Wu G.A. & Kim S.-H. (2009) Whole-genome phylogeny of mammals: evolutionary information in genic and nongenic regions. Proc Natl Acad Sci U S A 106, 17077-82.

121. Smith J.M. & Haigh J. (1974) The hitch-hiking effect of a favourable gene. Genet Res 23, 23-35.

122. Smýkal P., Nelson M.N., Berger J.D. & Von Wettberg E.J.B. (2018) The Impact of Genetic Changes during Crop Domestication. Agronomy 8, 119.

123. Soltis P.S., Soltis D.E. & Chase M.W. (1999) Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature 402, 402-4.

124. Song S., Liu L., Edwards S.V. & Wu S. (2012) Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc Natl Acad Sci U S A 109, 14942-7.

125. Soria-Carrasco V., Gompert Z., Comeault A.A., Farkas T.E., Parchman T.L., Johnston J.S., Buerkle C.A., Feder J.L., Bast J., Schwander T., Egan S.P., Crespi B.J. & Nosil P. (2014) Stick insect genomes reveal natural selection's role in parallel speciation. Science 344, 738-42.

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

127. Stapley J., Feulner P.G.D., Johnston S.E., Santure A.W. & Smadja C.M. (2017) Variation in recombination frequency and distribution across eukaryotes: patterns and processes. Philos Trans R Soc Lond B Biol Sci 372.

128. Stettler R., Bradshaw T., Heilman P. & Hinckley T. (1996) Biology of *Populus* and its implications for management and conservation. NRC Research Press.

129. Stölting K.N., Paris M., Meier C., Heinze B., Castiglione S., Bartha D. & Lexer C. (2015) Genome-wide patterns of differentiation and spatially varying selection between postglacial recolonization lineages of *Populus alba* (Salicaceae), a widespread forest tree. New Phytol 207, 723-34.

130. Suarez-Gonzalez A., Hefer C.A., Christe C., Corea O., Lexer C., Cronk Q.C. & Douglas C.J. (2016) Genomic and functional approaches reveal a case of adaptive introgression from *Populus balsamifera* (balsam poplar) in *P. trichocarpa* (black cottonwood). Mol Ecol 25, 2427-42.

131. Talla V., Johansson A., Dincă V., Vila R., Friberg M., Wiklund C. & Backström N. (2019) Lack of gene flow: Narrow and dispersed differentiation islands in a triplet of *Leptidea* butterfly species. Mol Ecol 28, 3756-70.

132. Talla V., Kalsoom F., Shipilina D., Marova I. & Backström N. (2017) Heterogeneous Patterns of Genetic Diversity and Differentiation in European and Siberian Chiffchaff (*Phylloscopus collybita abietinus*/*P. tristis*). G3 (Bethesda) 7, 3983-98.

133. Tavares H., Whibley A., Field D.L., Bradley D., Couchman M., Copsey L., Elleouet J., Burrus M., Andalo C., Li M., et al. (2018) Selection and gene flow shape genomic islands that control floral guides. Proc Natl Acad Sci U S A 115, 11006-11.

134. Terhorst J., Kamm J.A. & Song Y.S. (2017) Robust and scalable inference of population history from hundreds of unphased whole genomes. Nat Genet 49, 303-9.

135. Turner T.L. & Hahn M.W. (2010) Genomic islands of speciation or genomic islands and speciation? Mol Ecol 19, 848-50.

136. Tuskan G.A., Difazio S., Jansson S., Bohlmann J., Grigoriev I., Hellsten U., Putnam N., Ralph S., Rombauts S., Salamov A., et al. (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313, 1596-604.

137. Valente L.M., Manning J.C., Goldblatt P. & Vargas P. (2012) Did pollination shifts drive diversification in southern African Gladiolus? Evaluating the model of pollinator-driven speciation. Am Nat 180, 83-98.

138. Vargas O.M., Ortiz E.M. & Simpson B.B. (2017) Conflicting phylogenomic signals reveal a pattern of reticulate evolution in a recent high-Andean diversification (Asteraceae: Astereae: *Diplostephium*). New Phytol 214, 1736-50.

139. Viart M. (1979) Poplars and willows in wood production and land use. Food & Agriculture Org.

140. Vijay N., Bossu C.M., Poelstra J.W., Weissensteiner M.H., Suh A., Kryukov A.P. & Wolf J.B. (2016) Evolution of heterogeneous genome differentiation across multiple contact zones in a crow species complex. Nat Commun 7, 13195.

141. Vijay N., Weissensteiner M., Burri R., Kawakami T., Ellegren H. & Wolf J.B.W. (2017) Genomewide patterns of variation in genetic diversity are shared among populations, species and higher-order taxa. Mol Ecol 26, 4284-95.

142. Wang B., Mojica J.P., Perera N., Lee C.R., Lovell J.T., Sharma A., Adam C., Lipzen A., Barry K., Rokhsar D.S., Schmutz J. & Mitchell-Olds T. (2019) Ancient polymorphisms contribute to genome-wide variation by long-term balancing selection and divergent sorting in *Boechera stricta*. Genome Biol 20, 126.

143. Wang H., Moore M.J., Soltis P.S., Bell C.D., Brockington S.F., Alexandre R., Davis C.C., Latvis M., Manchester S.R. & Soltis D.E. (2009) Rosid radiation and the rapid rise of angiosperm-dominated forests. Proc Natl Acad Sci U S A 106, 3853-8.

144. Wang J., Street N.R., Park E.J., Liu J. & Ingvarsson P.K. (2020) Evidence for widespread selection in shaping the genomic landscape during speciation of *Populus*. Mol Ecol 29, 1120-36.

145. Wang J., Street N.R., Scofield D.G. & Ingvarsson P.K. (2016) Variation in linked selection and recombination drive genomic divergence during allopatric speciation of European and American aspens. Mol Biol Evol 33, 1754-67.

146. Weigel D. & Nordborg M. (2015) Population genomics for understanding adaptation in wild plant species. Annu Rev Genet 49, 315-38.

147. Wright S. (1982) The shifting balance theory and macroevolution. Annu Rev Genet 16, 1-19.

148. Wu C.I. (1985) A stochastic simulation study on speciation by sexual selection. Evolution 39, 66-82.

149. Yang M., He Z., Shi S. & Wu C.I. (2017) Can genomic data alone tell us whether speciation happened with gene flow? Mol Ecol 26, 2845-9.

150. Yassin A., Debat V., Bastide H., Gidaszewski N., David J.R. & Pool J.E. (2016) Recurrent specialization on a toxic fruit in an island *Drosophila* population. Proc Natl Acad Sci U S A 113, 4771-6.

151. Zheng H., Fan L., Milne R.I., Zhang L., Wang Y. & Mao K. (2017) Species delimitation and lineage separation history of a species complex of aspens in China. Front Plant Sci 8, 375.

152. Zhong B., Liu L., Yan Z. & Penny D. (2013) Origin of land plants using the multispecies coalescent model. Trends Plant Sci 18, 492-5.

2 Chapter I

Evolution of strong reproductive isolation in plants: broad-scale patterns and lessons from a perennial model group

Authors: **Huiying Shang[1,2*]**, Jaqueline Hess[1,3], Melinda Pickup[4], David Field[1,5], Pär Ingvarsson[6], Jianquan Liu[7], Christian Lexer[1, †]

Author affiliations:

[1]Department of Botany and Biodiversity Research, University of Vienna, Rennweg 14, 1030 Vienna, Austria

[2]Vienna Graduate School of Population Genetics, Vienna, Austria

[3]Helmholtz Centre for Environmental Research, Halle (Saale), Germany

[4]Institute of Science and Technology (IST), Klosterneuburg, Austria

[5]Edith Cowan University, Perth, Australia

[6]Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden

[7]Key Laboratory for Bio-resources and Eco-environment, College of Life Science, Sichuan University, Chengdu, People's Republic of China

Status: Published, Philos Trans R Soc Lond B Biol Sci 375:20190544

Contribution: Laboratory work, data analysis, interpretation the results, original draft preparation

PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Research

Cite this article: Shang H, Hess J, Pickup M, Field DL, Ingvarsson PK, Liu J, Lexer C. 2020 Evolution of strong reproductive isolation in plants: broad-scale patterns and lessons from a perennial model group. *Phil. Trans. R. Soc. B* **375**: 20190544.
http://dx.doi.org/10.1098/rstb.2019.0544

Accepted: 26 February 2020

One contribution of 19 to a theme issue 'Towards the completion of speciation: the evolution of reproductive isolation beyond the first barriers'.

Subject Areas:
evolution, genetics

Keywords:
speciation, hybridization, reproductive isolation, gene flow, topology discordance, recombination rate

Authors for correspondence:
Huiying Shang
e-mail: huiying.shang@univie.ac.at
David L. Field
e-mail: d.field@ecu.edu.au

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.5018459.

THE ROYAL SOCIETY
PUBLISHING

Evolution of strong reproductive isolation in plants: broad-scale patterns and lessons from a perennial model group

Huiying Shang[1,2], Jaqueline Hess[1,3], Melinda Pickup[4], David L. Field[1,5], Pär K. Ingvarsson[6], Jianquan Liu[7] and Christian Lexer[1,†]

[1]Department of Botany and Biodiversity Research, University of Vienna, Rennweg 14, 1030 Vienna, Austria
[2]Vienna Graduate School of Population Genetics, Vienna, Austria
[3]Helmholtz Centre for Environmental Research, Halle (Saale), Germany
[4]Institute of Science and Technology (IST), Klosterneuburg, Austria
[5]Edith Cowan University, Perth, Australia
[6]Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden
[7]Key Laboratory for Bio-resources and Eco-environment, College of Life Science, Sichuan University, Chengdu, People's Republic of China

HS, 0000-0002-1302-8008; JH, 0000-0003-3281-5434; DLF, 0000-0002-4014-8478; CL, 0000-0002-7221-7482

Many recent studies have addressed the mechanisms operating during the early stages of speciation, but surprisingly few studies have tested theoretical predictions on the evolution of strong reproductive isolation (RI). To help address this gap, we first undertook a quantitative review of the hybrid zone literature for flowering plants in relation to reproductive barriers. Then, using *Populus* as an exemplary model group, we analysed genome-wide variation for phylogenetic tree topologies in both early- and late-stage speciation taxa to determine how these patterns may be related to the genomic architecture of RI. Our plant literature survey revealed variation in barrier complexity and an association between barrier number and introgressive gene flow. Focusing on *Populus,* our genome-wide analysis of tree topologies in speciating poplar taxa points to unusually complex genomic architectures of RI, consistent with earlier genome-wide association studies. These architectures appear to facilitate the 'escape' of introgressed genome segments from polygenic barriers even with strong RI, thus affecting their relationships with recombination rates. Placed within the context of the broader literature, our data illustrate how phylogenomic approaches hold great promise for addressing the evolution and temporary breakdown of RI during late stages of speciation.

This article is part of the theme issue 'Towards the completion of speciation: the evolution of reproductive isolation beyond the first barriers'.

1. Introduction

Current research on speciation genomics strives to tackle two central questions in evolutionary biology: what is the origin and evolution of reproductive barriers in the genomes of diverging populations? And, how do divergent populations or species respond when challenged by hybridization upon secondary contact [1–4]? Theory predicts that speciation may occur in the face of ongoing or episodic gene flow [5]. A rapidly increasing number of speciation genomic studies have started to address divergence with gene flow (DWGF) in a range of different species [6,7], which has been greatly facilitated by advances in second- and third-generation sequencing technologies [8–14]. Hence, speciation genomics has developed into a vibrant research field [3,4,15–17], fuelling debates on topics of fundamental, philosophical and applied interest.

[†]Deceased.

Rapid barrier evolution during DWGF has been predicted by population geneticists for decades and has become widely known as the 'coupling' of individual barrier loci, resulting in mutually strengthened total barriers to gene flow [3,18–21]. It is thought that coupling creates coincidence among the effects of single barrier loci (and thus the traits encoded by them), which may lead to a substantial, but often incomplete, barrier to gene flow [21]. This leads to a 'grey zone' of speciation [22] which may well be responsible for many of the great challenges experienced by taxonomists and systematic biologists in previous decades and centuries. Genetic contact (hybridization) among divergent lineages at these advanced stages of speciation can result in a range of hotly debated outcomes [13,21,23,24]. These may include both heterosis (hybrid vigour) and hybrid breakdown due to genomic incompatibilities including the breakdown of genomic co-adaptation [25–27].

Differentiation between populations and ultimately speciation yields complex patterns of divergence along the genome [28,29]. Theory predicts that individual barrier loci can result in peaks of divergence between species [12,30], but in reality, the interplay of linked selection, variation of recombination rates and density of functional sites results in a complex landscape of peaks and troughs, which may be independent of reproductive isolation (RI). For example, background selection in regions of low recombination and with a high density of functional sites can also lower diversity within species, resulting in divergence peaks between species [31,32]. Nonetheless, several studies have reported a positive correlation between introgression and recombination rate [29,33]. These patterns are consistent with highly polygenic barriers to gene flow and the more efficient removal of introgressed variation in regions of low recombination [34]. However, it remains unclear whether the influence of linked selection on introgressed variation diminishes with time since divergence or whether it holds for organisms with other life histories with high rates of effective recombination.

To this end, hybridizing species have become highly appreciated 'natural labs' for studying speciation [35–37]. This holds true for hybrids formed either during primary divergence or upon secondary contact, and whether the genetic transitions seen in these zones fit with clinal or 'geographic mosaic' evolutionary models [2]. Divergent yet hybridizing taxa can also serve as precious sources of recombinant crosses for studying the genomic architecture of RI and inter-population trait differences [36,38–40]. In addition, hybrid zones enable the impact of introgression on genomic patterns of divergence to be investigated at recent time scales by comparing parapatric populations flanking hybrid zones with allopatric populations [12]. At deeper time scales, studying hybridizing taxa also makes it possible to address important questions regarding the sorting of ancestral variation in young or emerging species, past episodes of gene flow and how this may relate to the evolution of RI [11,33,41]. This is greatly facilitated by recent conceptual developments in merging the analytical toolkits of population genomics and phylogenomics [14,41]. This approach may be particularly useful for organismal groups that maintain leaky reproductive barriers across species complexes for many generations—and thus for millions of years—such as perennial plants with relatively large effective population sizes (N_e), far-ranging pollen and seed dispersal, and the ability to maintain viable genotypes in populations by clonal reproduction [42,43].

Among different study systems for studying speciation in plants, *Populus* has become a perennial model group because of its ecological and economic importance and favourable genetic attributes such as small genome size (less than 500 Mb; $2C = 1.1$ pg in the case of *Populus trichocarpa*), diploidy throughout the genus ($2n = 38$), 'porous' species barriers [44–46] and a well curated and annotated genome assembly [47]. Species of the genus are widespread across the Northern Hemisphere [48]. Several studies have attempted to resolve phylogenetic relationships of species in this genus [49,50], most notably a recent study using resequenced genomes [51]. Obligate outcrossing (dioecy), abundant wind-pollination, and mixed sexual and vegetative reproductive strategies in poplars have led to extensive introgression among species and relatively large effective population size (N_e) [52–55], which complicates phylogenetic inference.

Recent work on speciation genomics in *Populus* has revealed several patterns relevant to understanding the speciation continuum. Firstly, linked selection and recombination rate variation appear to have pervasive effects on genome-wide patterns of genetic diversity and divergence among poplar species, as exemplified by interspecific contrasts involving the two more closely related species *Populus tremula*, *Populus tremuloides* and the more distantly related *P. trichocarpa*. These effects are moderated by important demographic factors and events, such as temporal and interspecific changes in N_e experienced by these temperate tree species in response to climatic cycles [54,55]. At greater levels of divergence (1.73–1.90 Myr), a landmark study by Ma *et al*. [56] revealed the likely determinants of genome-wide patterns of diversity in the two Eurasian desert poplar species, *Populus euphratica* and *Populus pruinosa*, pointing to important roles for the divergent sorting of ancestral polymorphisms and divergent ecological selection. Finally, studies of the highly divergent Eurasian taxa *Populus alba* and *P. tremula* have shown that despite greater than 2.8 Myr of divergence [8], strong post-zygotic barriers due to genomic incompatibilities [57,58] and variable pre-zygotic barriers [58], these taxa still form viable and fertile hybrids within large mosaic hybrid zones in areas of both sym- and parapatry [57,59]. Although these species thus represent a useful showcase example for research on the late stages of speciation, studies that examine genome-wide phylogenomic patterns for taxa pairs representing the early and late stages of speciation are required to better understand how the genomic architecture of RI varies across the stages of speciation.

Beyond particular organismal model groups, categorizing the stage of speciation is dependent on both understanding the level of gene exchange among divergent taxa and identifying the presence of reproductive barriers [60,61]. For plants, the great diversity of mating systems, reproductive strategies and life-history traits may interact to influence the tempo and speed of speciation. Thus, we begin by undertaking a broad analysis including 133 hybridizing species pairs to examine the number of pre- and post-zygotic barriers and how these relate to gene flow in flowering plants. Here, we test the prediction of higher levels of gene flow in species pairs with fewer reproductive isolating barriers. We then 'zoom in' on the genomic footprints of RI and introgressive gene flow in species of the 'model forest tree' genus *Populus* (poplars/aspens/cottonwoods). These include widespread, ecologically divergent Eurasian taxa that provide key examples of the evolutionary mechanisms operating during the late stages of speciation. We analyse 36 re-sequenced genomes from seven species of this Eurasian species complex to examine how the genomic architecture of RI and introgressive gene flow varies across the stages of speciation. Then we analyse the data in a phylogenomic context and examine genome-wide relationships among well sorted versus introgressed tree topologies and recombination rates during both the early and late stages of speciation. Our purpose is to determine how genome-wide phylogenomic patterns (genome-wide tree topologies) are mediated by the genomic architecture of RI and the recombination landscape, and test whether these relations hold across the speciation continuum. Using tree typology weighting and phylogenetic tests for introgression, we compare the amount of gene flow and the relation between introgressed typologies and recombination rate, on five anciently diverged (late-stage speciation) and five recently diverged (early-stage speciation) species. Taken together, this broad to narrow approach provides novel insights into the processes and outcomes of DWGF from the early to late stages of speciation.

2. Material and methods

(a) Plant literature survey

To investigate the interaction between the presence of pre- and post-zygotic reproductive isolating barriers and gene flow, we collated data on hybridization in 133 species pairs, representing 72 genera and 41 plant families (for full description of methods, see Pickup *et al*. [62]). Following Abbott [63], we categorized gene flow into four categories: very low, low, high and variable (different among hybridizing populations) based on criteria and descriptions outlined by Pickup *et al*. [62], which were based on quantitative information on the frequency of hybrids and backcrosses (see also electronic supplementary material, ‘plant literature survey: categorization of gene flow’). For each taxon pair, we identified the presence (1) or absence (0) of each of a set of pre-zygotic and post-zygotic barriers (but we did not attempt to quantify their strength) based on Abbott [63] and descriptions or quantitative assessments from each individual study. Pre-zygotic barriers were: (i) geography (spatial isolation of parental species), (ii) habitat divergence (divergent habitat preference), (iii) divergent flowering phenology, (iv) divergent floral structure, (v) pollinator preference, (vi) mating system and (vii) pollen competition. Mating system (vi) was classified as a pre-zygotic barrier for taxon pairs with divergent mating systems. These include: (i) taxon pairs with a predominantly outcrossing self-compatible species and a highly selfing self-compatible species, (ii) pairs where both taxa are selfing and (iii) pairs including a self-incompatible and self-compatible species (see Pickup *et al*. [62] for details). Post-zygotic barriers were: (i) reduced hybrid viability, (ii) cyto-nuclear interactions, (iii) intrinsic genomic incompatibilities (the interaction between alleles results in lower fitness of individuals), and (iv) extrinsic (ecological context-dependent) incompatibilities, which require divergent ecological environments for the two populations and selection against maladapted hybrids in both environments. A χ^2 contingency test was used to examine if the categories of gene flow (high versus low; combining low, very low and low variable) were associated with the total number of reproductive isolating barriers (combining pre- and post-zygotic barriers) for 123 species pairs where gene flow could be categorized (see Pickup *et al*. [62]). Statistical analysis was conducted in R and tested at $\alpha = 0.05$.

(b) Poplar species and populations sequenced de novo for this study

According to the most commonly used classification of *Populus*, the genus comprises six sections and 29 species [48]. De novo sequence data collection for this study was focused on seven closely related species from section *Populus* (aspens and white poplars) that provide examples of large N_e and large geographical distribution versus small N_e and narrow distributions, sympatric versus parapatric versus allopatric distribution. Among these, *P. alba* (white poplar) and *P. tremula* (Eurasian aspen) are the two most widespread taxa, the former being widely distributed across large parts of southern Eurasia and North Africa, and the latter extending all the way from Scotland to eastern Russia and from northern Scandinavia to the Mediterranean [64]. The two species are at a late stage of speciation, as indicated by partial pre-zygotic and strong post-zygotic reproductive barriers [57,58] and an estimated divergence time of greater than 2.8 Myr [8]. Nevertheless, they still hybridize within large ‘geographical mosaic’ hybrid zones across a broad zone of overlap in Europe and Asia [8,59,65,66]. This species pair serves as a showcase example for the late stage of speciation in this study.

Among the other, more narrowly distributed species, *Populus davidiana* (the Chinese aspen) is distributed from the central to the northeastern part of China, while the Himalayan aspen *Populus rotundifolia* is narrowly endemic to the high-altitude regions of the Qinghai–Tibetan plateau. The two species are thought to have undergone recent parapatric, ecological speciation in the face of gene flow [67]. This species pair thus serves as a showcase example for the early stages of speciation in this study. Among the remaining species sampled and sequenced de novo, *Populus adenopoda* grows in warm and moist subtropical areas of south and east China [68], whereas *Populus qiongdaoensis* is a rare species only known from Hainan island. Publicly available data for the widespread North American trembling aspen *P. tremuloides* were included for comparative purposes.

Our sampling for de novo genome sequencing included 36 accessions from these seven ingroup species and two outgroup taxa from section Tacamahaca, *P. trichocarpa* (black cottonwood) and *Populus balsamifera* (balsam poplar) (electronic supplementary material, table S1). We collected three to five individuals for each species. For species collected in China, genomic DNA was extracted from silica-dried leaves by using the plant DNeasy mini kit (Qiagen, Germany). To increase the quality of total DNA, we used NucleoSpin gDNA clean-up kits for purification of DNA extracts. All libraries were 2× 150 bp paired-end sequenced on an Illumina HiSeq 3000 sequencer at the Institute of Genetics, University of Berne, Switzerland. Illumina HiSeq paired-end reads for *P. tremuloides* and the two outgroup species were downloaded from NCBI using the NCBI SRA toolkit under accession numbers PRJNA299390 and PRJNA276056. Further details about sampling locations and distributions are provided in electronic supplementary material, table S1. The reads of each individual were mapped to the *P. trichocarpa* reference genome using BWA [69]. Details about sequence data processing, variant calling and single nucleotide polymorphism (SNP) quality filtering are provided as electronic supplementary material.

(c) Phylo- and population genomic data analyses

To assess population structure in our whole-genome dataset, principal component analysis (PCA) was carried out based on biallelic SNPs using PLINK [70]. As an alternative means of depicting genetic relationships, a neighbour-joining (NJ) tree was constructed using PHYLIP v. 3.696 (http://evolution.genetics.washington.edu/phylip.html) and visualized using FigTree v. 1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/).

Owing to the limits of concatenation methods to infer a species tree, especially for species with large effective population size, we constructed a species tree using MP-EST v. 1.5 [71] based on the multi-species coalescent, established statistical support by bootstrapping and estimated divergence times using MCMCTree software in the PAML package [72] as described in the electronic supplementary material. To infer species' demographic histories including N_e changes and the relative timing of species splits, we employed SMC++ v. 1.12.1, which combines a coalescent HMM approach with the computational efficiency of the site frequency spectrum for demographic inference [73]. This approach can use unphased data and has been shown to produce robust results in both the recent and ancient past.

To select poplar taxa at late and early stages of speciation, respectively, we explored the sharing of identity-by-descent (IBD) blocks between pairs of species using BEAGLE v. 4.1 [74] and the parameter settings: window=100 000; overlap = 10 000; ibdtrim = 50; ibdlod = 10, impute = false. To examine variation in genealogies along the genome and identify regions whose evolutionary history deviates from the species tree, we used topology weighting by iterative sampling of subtrees (TWISST) [75] to infer the weights (i.e. frequencies) of all different possible tree topologies for windows along the genome. Data were phased and imputed with BEAGLE v. 4.1 [76] and non-overlapping windows of 50 SNPs were used for inferring trees using PhyML [77]. In order to test the effect of different levels of divergence on tree topologies, we selected five anciently diverged (=late-stage speciation) and five recently diverged (=early-stage speciation) taxa for TWISST analysis. The five late-stage species included the well-studied hybridizing species pair *P. alba* and *P. tremula* introduced earlier, and the five early-stage species included the Chinese aspen *P. davidiana* and the Himalayan aspen *P. rotundifolia*. Based on Zheng *et al.* [67] and our own NJ analysis, we separated *P. davidiana* into two local taxa according to geography, central and northeastern China. All Python scripts used for this analysis can be downloaded at https://github.com/simonhmartin/twisst. Weightings for all topologies were plotted across chromosomes with loess span value set to 0.03. Chromosome-level averages of topology weights were compared with local recombination rates in *P. tremula* [54] in windows of 100 kb.

To gain deeper insights into the ancient and recent admixture events presumably responsible for the observed genome-wide patterns of topology weights for anciently and recently diverged species (above), we examined patterns of IBD tract sharing (above) and inferred ancient and recent admixture using *D*-statistics to test for gene flow [78]. To estimate the extent and direction of gene flow for late-stage speciation taxa, we conducted D_{FOIL} five-taxon tests in 10 kb windows along the genome [78] using *P. trichocarpa* as an outgroup. For early-stage speciation taxa, we quantified gene flow using four-taxon *D*-statistics and *P. alba* as an outgroup; four-taxon tests were deemed sufficient here since our focus was on a single pair of species, *P. davidiana* and *P. rotundifolia*.

3. Results and discussion

(a) Relationships between reproductive barriers and introgressive gene flow in flowering plants

Of the 133 species pairs examined in our literature survey of flowering plants, 105 (78.9%) reported the presence of one or more pre-zygotic reproductive isolating barriers (figure 1*a*). The highest proportion had a single pre-zygotic barrier ($n = 56$, 42.1%) followed by the presence of two barriers ($n = 36$, 27.1%, figure 1*a*). Fewer taxon pairs had three pre-zygotic barriers ($n = 11$, 8.3%), and only two pairs (1.5%) recorded four pre-zygotic barriers. In contrast with the high prevalence of pre-zygotic barriers, fewer than half (42.9%) of the taxon pairs recorded post-zygotic reproductive barriers. Overall, there were also fewer post-zygotic barriers, with most taxon pairs recording only one barrier ($n = 51$, 38.3%; figure 1*b*). Although these analyses only examined the presence or absence of a barrier—rather than its strength—they provide an important overview of how the number of reproductive isolating barriers varies across plant taxa.

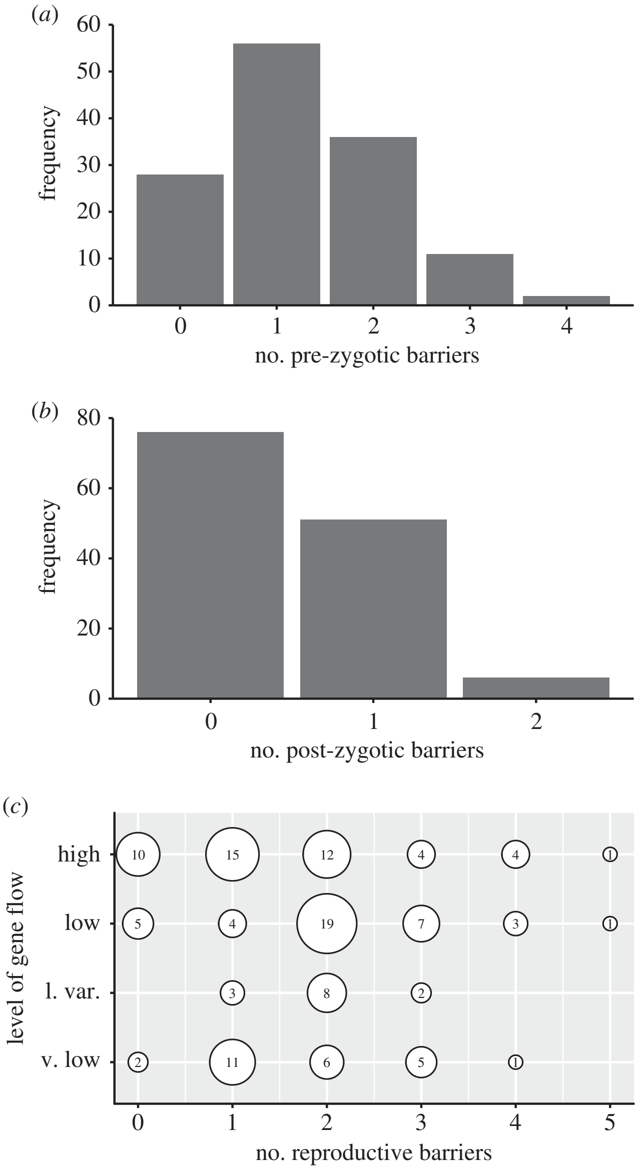


Figure 1. The number of (*a*) pre-zygotic and (*b*) post-zygotic reproductive isolating barriers for 133 angiosperm species pairs. (*c*) The association between the number of reproductive isolating barriers (pre- and post-zygotic) and categories of gene flow for the 133 taxa pairs. l. var., low variable; v. low: very low.

To assess the prediction that reproductive isolating barriers are related to introgressive gene flow [1,2,6,23,63], we examined if there was an association between the total number of barriers (combining pre- and post-zygotic) and the categories of gene flow (high versus low) for the taxon pairs included in our survey. If reproductive isolating barriers are important for the degree of introgressive gene flow, then we would expect higher gene flow for hybridizing taxa with fewer barriers. Indeed, we found a significant negative association between the gene flow categories (high versus low) and the number of reproductive barriers ($\chi^2 = 9.5793$, d.f. = 1, $N = 123$, $p = 0.048$) (figure 1*c*). Although there are

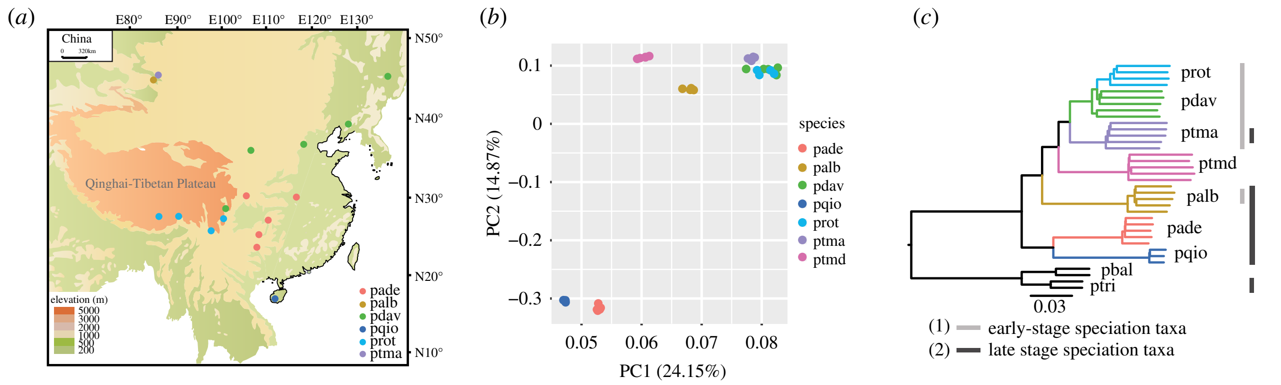


Figure 2. Sample set of individuals from a *Populus* (poplar and aspen) species complex used for whole-genome phylogenomics. (*a*) Sampling locations of six Eurasian *Populus* species. (*b*) PCA of SNP data from resequenced genomes for seven *Populus* species, including the six Eurasian species (*a*) and one North American species, *P. tremuloides*. (*c*) Rooted NJ tree based on the genomic data, with *P. trichocarpa* and *P. balsamifera* as outgroups. (1) and (2) highlight the taxa selected for focused phylogenomic analysis in early stages of speciation (1) and late stages of speciation (2) as determined by IBD tract sharing (figure 3*a*). Species abbreviations: pade, *P. adenopoda*; palb, *P. alba*; pdav, *P. davidiana*; pqio, *P. qiongdaoensis*; prot, *P. rotundifolia*; ptma, *P. tremula*; ptmd, *P. tremuloides*; pbal, *P. balsamifera*; ptri, *P. trichocarpa*.

caveats to this approach (given that presence/absence does not quantify barrier strength [61] and there was not adequate replication to enable a phylogenetically controlled analysis), our results provide some insights into the potential variation in speciation stage across hybridizing plant taxa. Moreover, plants exhibit extensive variation in both life history and mating system [62,79,80]. These differences in life history may mediate the strength of this association between reproductive barriers and gene flow.

Case studies of closely related groups of species can provide further data on the processes underlying RI [6,7]. For example, within our literature survey, there were five hybridizing taxon pairs within the genus *Populus* that are all similar in life history (woody trees) and mating system (dioecious). Of these, *P. alba* and *P. tremula* provide an example of late-stage speciation, with these two taxa exhibiting a number of different reproductive isolating barriers, including habitat divergence, intrinsic incompatibilities and cyto-nuclear incompatibilities [43,57,58,65]. In comparison, *P. davidiana* and *P. rotundifolia* are two recently diverged species that inhabit distinct environments, and which provide an excellent example for the study of RI in the early stage of speciation [67].

(b) Early versus late stages of speciation in *Populus*: novel insights from whole-genome phylogenomics of a poplar species complex

Whole-genome resequencing and reference-mapping of 36 individuals from seven ingroup and two outgroup taxa onto the *P. trichocarpa* genome assembly resulted in an average of 91.7% genomic regions covered, with an average coverage depth of 26.48×, yielding 7 026 036 high-quality SNPs (electronic supplementary material, table S2). *Populus davidiana* (the Chinese aspen) and *P. rotundifolia* (the Himalayan aspen), the two most recently derived species, were not monophyletic in NJ analysis. However, the three sequenced individuals of *P. davidiana* from central China were placed together with *P. rotundifolia* sampled in sym-/parapatry in the same geographical region (figure 2), rather than with conspecific individuals of *P. davidiana* from northeastern China, where its sister taxon *P. rotundifolia* is absent. This is suggestive of hybridization between these species in central China where they co-occur, thus also corroborating recent findings obtained with far more intensive biogeographic sampling but much sparser sampling of the genome [67]. Our two showcase species for late-stage speciation, *P. alba* and *P. tremula*, on the other hand, were clearly separated in PCA and NJ analysis (figure 2).

Our coalescent-based, dated species tree (electronic supplementary material, figure S1 and table S3) broadly reflected genetic relationships seen in the NJ tree and in a recent large-scale phylogenomic study of *Populus* [51], and demographic analysis using the site frequency spectrum and SMC++ complemented this coalescent-based analysis (electronic supplementary material, figure S2). SMC++ indicated an initial reduction in N_e in all species, coincident with the divergence of the major lineages in section *Populus* followed by population recovery to varying degrees (electronic supplementary material, figure S2). The results also reflected species splits seen in our coalescent-species tree, with N_e curves for *P. alba* and *P. tremula* splitting much further back in time than those for *P. davidiana* and *P. rotundifolia*. As expected, the N_e trajectories for *P. alba* and *P. tremula* separated more recently than those for *P. alba* and the North American aspen *P. tremuloides*, consistent with reports of hybridization and introgression between the partially sym-/parapatric Eurasian species *P. alba* and *P. tremula* [8,45,59,65].

Genome-wide patterns of IBD tract sharing (figure 3*a*) allowed us to select groups of both early- and late-stage speciation taxa for subsequent phylogenomic analyses and contrasts. We selected five more recently diverged taxa with weak barriers [67], including *P. davidiana* and *P. rotundifolia*, and five more anciently diverged taxa with strong barriers [57,58], including *P. alba* and *P. tremula*, for genome-wide analyses of tree topologies using TWISST (figure 4). We found a high percentage of discordant tree topologies, especially in early-stage speciation taxa (figure 4), indicating extensive introgression or incomplete lineage sorting (ILS). Of the 15 possible topologies in late-stage speciation taxa, the three most common ones ordered by their frequency were topo6 (green), topo4 (purple) and topo5 (black), and topo6

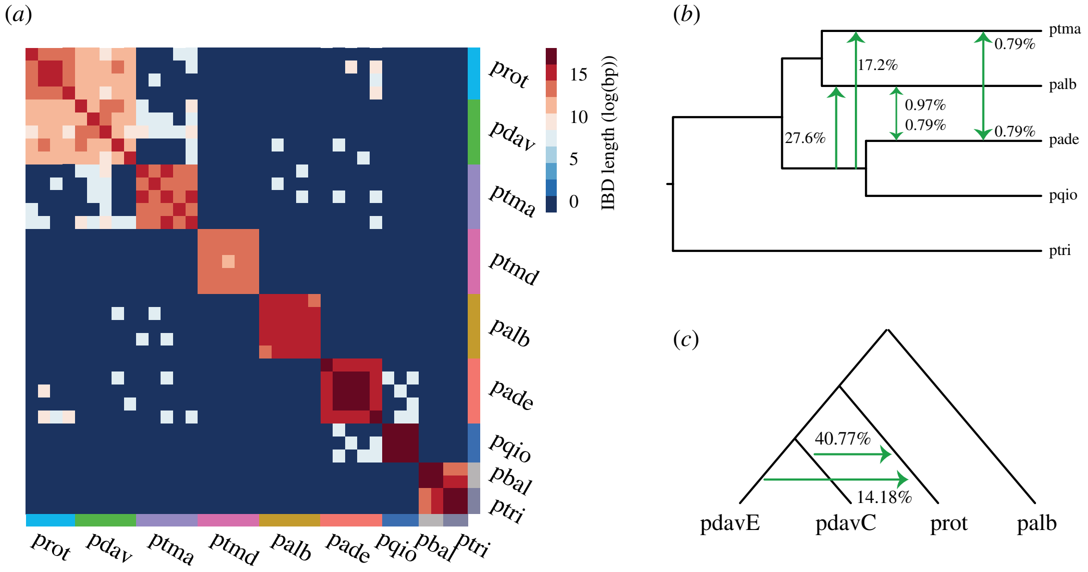


Figure 3. Patterns of haplotype sharing and introgressive gene flow. (*a*) The total length (log) of identical-by-descent (IBD) haplotypes shared between individuals. (*b*) D_{FOIL} five-taxon analysis for five early-stage speciation taxa, and (*c*) four-taxon analysis for four early-stage speciation taxa. Green arrows indicate the direction of gene flow between species; numbers along arrows represent the proportion of genome windows with evidence for gene flow between taxa. The analyses were based on 10 kb windows, retaining values with $p < 0.01$. Taxon abbreviations follow figure 2.

was consistent with the species tree. As expected, more than 10% of genome windows reflecting the species tree (topo6) had completely sorted genealogies in these late-stage speciation taxa (indicated by '% windows with a weighting of 1'). The high weightings of genealogies topo4 and topo5 are indicative of either ILS or ancient introgressive gene flow involving *P. tremula*, *P. alba*, and the ancestor of *P. adenopoda* and *P. qiongdaoensis*. The ancient gene flow hypothesis was supported by D_{FOIL} five-taxon tests (figure 3*b*,*c*), which have been validated to function even with high levels of ILS [78]. This is broadly consistent with widespread interspecific gene flow in *Populus* detected in a recent large-scale phylogenomic study [51].

Under linked selection encompassing both directional selection and background selection against deleterious mutations, we would expect the weights of TWISST species tree topologies to be highest with low recombination rates, while the weights of introgression topologies (or admixture-related parameters more generally) should be released from this constraint or even increase with recombination [33]. This is analogous to the expectation that in the presence of hybrid incompatibilities, introgressed ancestry in populations is more likely to persist in regions of high recombination [81]. In line with this expectation, we observed the expected increase in species tree weights with reduced recombination rates for both early-stage and late-stage speciation taxa (figure 5; electronic supplementary material, table S4; topo6). The weights of putative introgression topologies in late-stage speciation taxa, however, did not show the expected increase for greater recombination rates (figure 5; electronic supplementary material, table S4; topo4 and topo5). Rather, these topologies received appreciable weights across *all* observed recombination rates. This is consistent with a breakdown in correlation between recombination rate and shared haplotype length in deeply divergent *Populus* spp. [51], and suggests that introgressed segments are able to escape barrier loci and linked selection over time, especially in species with high recombination rates.

Outcrossing, wind-pollinated trees such as poplar and aspen species exhibit fairly large N_e (greater than 100 000) and low levels of linkage disequilibrium (LD), consistent with high levels of effective recombination [52]. The decay of LD along chromosomes is even more rapid in species with continuous distributions such as *P. tremula* than in floodplain poplar species with more patchy distributions [53,55,82]. In such high recombination genomes, it should be easier to escape barrier loci [18,83] compared with other organisms with smaller N_e and slower LD decay [14,33]. Also, like other long-lived outcrossing perennial plant species, poplars harbour large amounts of standing genetic variation. This results in complex population genomic signatures of local adaptation, frequently involving subtle allele frequency shifts at many loci [10,66,84]. Importantly, these intraspecific patterns are mirrored by polygenic architectures of fitness-related trait differences between hybridizing species, including our two showcase species for late-stage speciation studied here, *P. alba* and *P. tremula* [39]. In fact, the observed relationships of tree topology weights with recombination rate in strongly divergent species [57,58] are consistent with the polygenic, complex architecture of fitness-related trait differences recently identified by 'admixture mapping' genome-wide association studies in hybrids [39]. Genomic regions supporting the species tree topology in late-stage speciation taxa apparently accumulated owing to linked selection across the genome. Nevertheless, this pattern is also expected to arise as a result of background selection or selective sweeps unrelated to reproductive barriers, effectively lowering N_e for chromosomal regions with low recombination rates [31,32]. Despite these confounding signals, recent simulation studies have shown that background selection alone may not be sufficient to explain recombination rate-dependent divergence landscapes in *Ficedula* flycatchers [85] and monkeyflowers [29], and such modelling approaches using more extensive population genomic data will be useful to further characterize the architecture of RI in deeply divergent poplars.

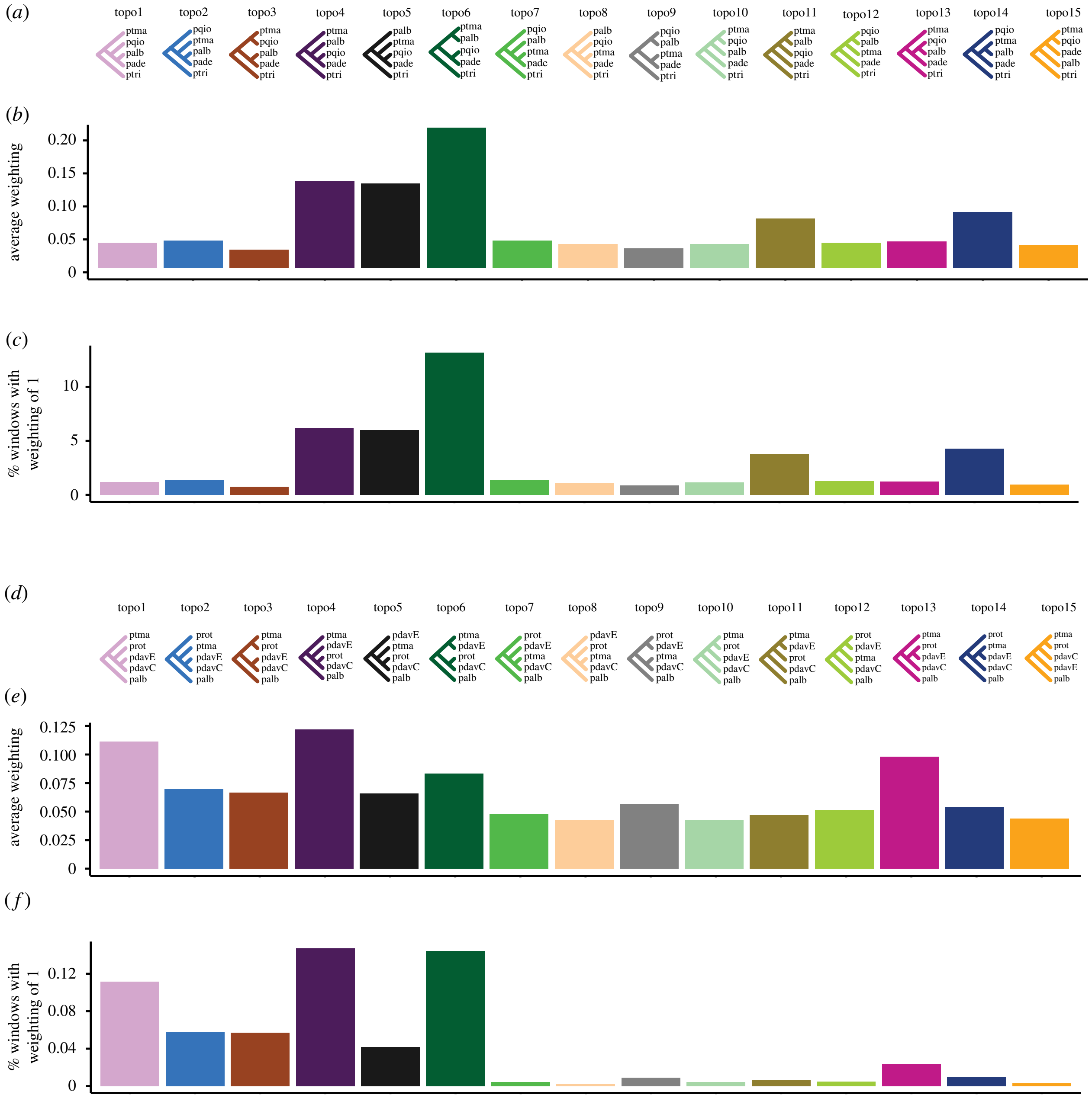


Figure 4. Topology weighting reveals widespread phylogenetic discordance in both early- and late speciation poplar taxa. (*a*) The 15 rooted topologies of late-stage speciation taxa: pade, pqio, ptma, palb and ptri as outgroup. (*b*,*e*) Average weighting of each topology. (*c*,*f*) The percentage of windows exhibiting complete lineage sorting for each topology. (*d*) The 15 rooted topologies of early-stage speciation taxa: pdavC, pdavE, prot, ptma and palb as outgroup. pdavC and pdavE are two populations representing two different phylogeographic lineages of *P. davidiana* from central and northeastern China, respectively. Taxon abbreviations follow figure 2.

In our five selected early-stage speciation taxa, the introgression topology, topo4 (purple)—in which the locally parapatric populations *P. rotundifolia* and central *P. davidiana* were sister taxa—received even higher weightings than the species tree, topo1 (pink) (figure 4). The introgression topology also received consistently higher weightings in well delimited chromosome segments along the genome (electronic supplementary material, figure S3), which is reminiscent of haplotype signatures commonly observed with introgressive gene flow [40,86]. Accordingly, *P. rotundifolia* and *P. davidiana* exhibited extensive sharing of long IBD tracts (figure 3). This might also explain the conspicuous negative correlation between introgressed topology weightings (topo4) and recombination rate seen for these species (figure 5; electronic supplementary material, table S4), with increased weightings at low recombination rates. Increased introgression is not *a priori* expected in low recombination regions [33,81]. The high introgressed topology weightings at low recombination rates (figure 5) can alternatively result from insufficient time to break up long haplotypes stemming from recent introgressive gene flow. A strong positive correlation of the introgression tree and recombination rates as seen in other systems [87] may also be masked by extensive levels of ILS among windows supporting topo4, as suggested by the high frequency of the 'mirrored' topology, grouping together eastern *P. davidiana* and *P. rotundifolia* (topo13, magenta; figure 4). Topo13 also showed a weak negative correlation with recombination rate, highlighting how extensive standing variation in species with large N_e may slow down formation of strong reproductive barriers.

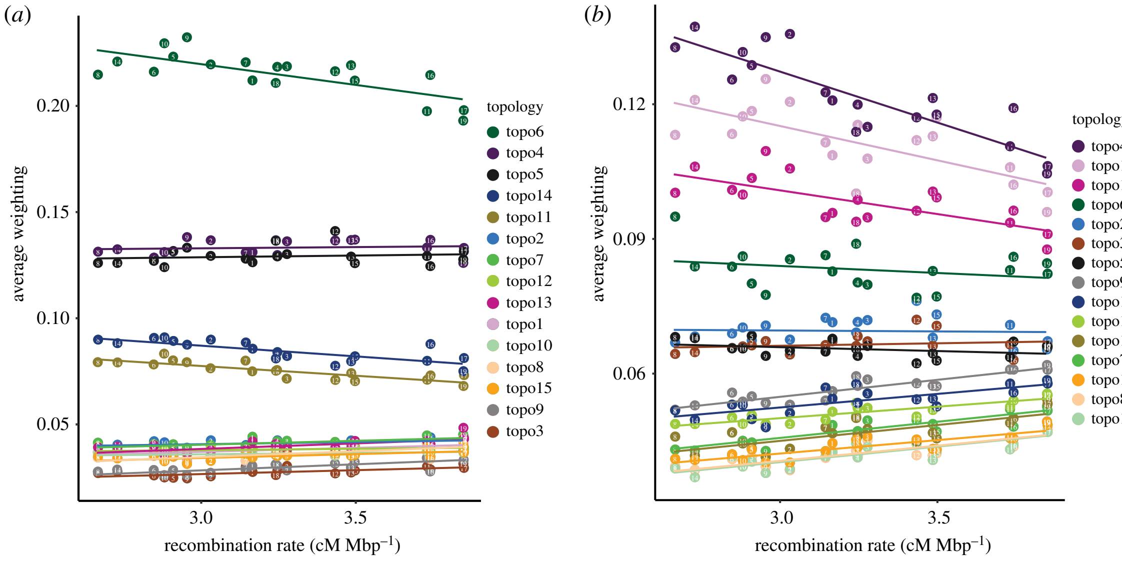


Figure 5. Average weightings per chromosome for all 15 topologies, plotted against average recombination rate (centimorgans per Mbp). Coloured circles represent poplar chromosomes 1–19. (*a,b*) Relationship between weightings of each topology and recombination rate in late-stage and early-stage speciation taxa, respectively, including linear fit. Recombination rate estimates for *P. tremula* were obtained with LDhat based on 100 kb windows [55]. The tree topologies for late- and early-stage speciation taxa are shown in figure 4*a,d*. Correlation and regression statistics are shown in electronic supplementary material, table S4.

4. Conclusion

Our literature survey of hybridizing flowering plant species points to important roles for both pre- and post-zygotic barriers in plant speciation, and indicates that barrier complexity (i.e. the number of different barriers) is linked to an overall reduction in gene flow. Future efforts should explore how different aspects of life-history traits and mating systems (for which plants exhibit extraordinary variation; [62]) mediate the strength of this association, and how plants, animals and fungi differ in this regard. The model tree genus *Populus* offers suitable taxon pairs or groups for addressing the evolution of strong RI during plant speciation; this includes late-stage speciation taxa that are strongly isolated by multiple barriers, but which nevertheless form fertile hybrids. An important future task will be to assess the cumulative action of different pre- and post-zygotic barriers in this group, and how their effects become coupled towards the development of strong RI [4,21]. Each single barrier effect may have a simple or polygenic basis, and some traits may affect multiple barriers [88]. Thus, we anticipate that understanding the evolution of strong RI will benefit greatly from advances in high-throughput phenotyping and the quantitative evolutionary genomics of multivariate trait space.

Our phylogenomic data for a poplar species complex mirrored those from our literature survey, with stronger divergence and greatly reduced IBD tract sharing for late-stage speciation taxa separated by multiple barriers, in contrast with pronounced IBD sharing and topology discordance for early-stage taxa separated mainly by a weak eco-geographic barrier. Genome-wide variation in phylogenetic tree topologies based on 36 sequenced genomes highlights the potential role of both ancient and recent introgressive gene flow for the genomic composition of extant poplar species. This is in addition to ILS, which we must expect to be present at these evolutionary time scales [51]. While the weightings (frequencies) of species tree topologies—and their relationships with recombination rate variation along the genome—were broadly consistent with polygenic barriers and linked selection pinpointed by other studies on *Populus* spp. [54,55], the lack of a strong relationship of putatively introgressed topologies with recombination rates highlights the complexities of barrier formation in this group [39,89]. Complex architectures are expected to arise from a number of factors including (i) high levels of recombination and rapid LD decay along chromosomes in poplars [52,53,55], (ii) long generation times accentuated by the ability of viable genotypes to persist as clones [43], and (iii) large N_e, which enables these completely outcrossing, wind-pollinated tree species to hold extraordinary levels of standing genetic variation. For early-stage speciation taxa, the genome-wide topology/recombination rate relationship pointed to a protracted speciation process and the absence of strong barriers because of the apparent presence of both long introgressed haplotype tracts and high levels of ILS. A similarly protracted process may have been at work for late-stage speciation taxa, supported by an extended period of genetic exchange between the ancestor of *P. adenopoda* and *P. qiongdoaensis* and both *P. tremula* and *P. alba*. Indeed, phylogenomic approaches based on tree topology variation appear to lend themselves to studies of the evolution of strong RI during speciation and the extended time scales this may take. In species complexes of poplars, it appears that despite a polygenic basis of barriers, numbers of barrier loci are still too low (relative to recombination rates and individual selection coefficients) to facilitate strong coupling [86] and thus to prevent the escape of locally adaptive alleles. We hope this work will encourage more studies exploring discordance and concordance between patterns of RI seen through the lenses of different, complementary approaches available to speciation geneticists addressing different time scales.

Data accessibility. Flowering plant literature analysis made use of the database introduced by Pickup *et al.* [62], and an updated data set and R scripts are available from Dryad at https://doi.org/10.5061/dryad.h9w0vt4fw [90]. Extensive biological sample and sequencing statistics are uploaded as electronic supplementary material, tables. All raw read data were uploaded to NCBI and can be found under Bioproject ID PRJNA612655.

Authors' contributions. Study conceived and designed by C.L. together with H.S., J.H., J.L. and D.L.F. Sample collection and laboratory work conducted by H.S. and J.L. Phylogenomic data analysis by

H.S. and J.H. Interpretation of phylogenomic results was undertaken by H.S., J.H., C.L., D.L.F. and P.K.I. M.P. and D.L.F. compiled and analysed the plant literature database. C.L., H.S., J.H., D.L.F. and M.P. drafted the manuscript.

Competing interests. We declare we have no competing interests.

Funding. This work was supported by a fellowship from the China Scholarship Council (CSC) to H.S., Swiss National Science Foundation (SNF) grant no. 31003A_149306 to C.L., doctoral programme grant W1225-B20 to a faculty team including C.L., and the University of Vienna.

Acknowledgements. We thank members of J.L.'s lab for collecting samples, Michael Barfuss and Elfi Grasserbauer for help in the laboratory, the Next Generation Sequencing Platform of the University of Berne for sequencing, the Vienna Scientific Cluster (VSC) for access to computational resources, and Claus Vogel and members of the PopGen Vienna graduate school for helpful discussions.

References

1. Abbott R *et al*. 2013 Hybridization and speciation. *J. Evol. Biol.* **26**, 229–246. (doi:10.1111/j.1420-9101.2012.02599.x)
2. Gompert Z, Mandeville EG, Buerkle CA. 2017 Analysis of population genomic data from hybrid zones. *Annu. Rev. Ecol. Evol. Syst.* **48**, 207–229. (doi:10.1146/annurev-ecolsys-110316-022652)
3. Nosil P, Feder JL, Flaxman SM, Gompert, Z. 2017 Tipping points in the dynamics of speciation. *Nat. Ecol. Evol.* **1**, 1. (doi:10.1038/s41559-016-0001)
4. Ravinet M, Faria R, Butlin RK, Galindo J, Bierne N, Rafajlovic M, Noor MAF, Mehlig B, Westram AM. 2017 Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. *J. Evol. Biol.* **30**, 1450–1477. (doi:10.1111/jeb.13047)
5. Felsenstein, J. 1981 Skepticism toward Santa Rosalia, or why are there so few kinds of animals? *Evolution* **35**, 124–138. (doi:10.1111/j.1558-5646.1981.tb04864.x)
6. Coyne J, Orr H. 2004 *Speciation*. Oxford, UK: Oxford University Press. Sunderland, MA: Sinauer Associates.
7. Nosil P. 2012 *Ecological speciation*. Oxford, UK: Oxford University Press.
8. Christe C, Stolting KN, Paris M, Fraisse C, Bierne N, Lexer C. 2017 Adaptive evolution and segregating load contribute to the genomic landscape of divergence in two tree species connected by episodic gene flow. *Mol. Ecol.* **26**, 59–76. (doi:10.1111/mec.13765)
9. Ellegren H *et al*. 2012 The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature* **491**, 756–760. (doi:10.1038/nature11584)
10. Evans LM *et al*. 2014 Population genomics of *Populus trichocarpa* identifies signatures of selection and adaptive trait associations. *Nat. Genet.* **46**, 1089–1096. (doi:10.1038/ng.3075)
11. Novikova PY *et al*. 2016 Sequencing of the genus *Arabidopsis* identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nat. Genet.* **48**, 1077–1082. (doi:10.1038/ng.3617)
12. Tavares H *et al*. 2018 Selection and gene flow shape genomic islands that control floral guides. *Proc. Natl Acad. Sci. USA* **115**, 11 006–11 011. (doi:10.1073/pnas.1801832115)
13. The Heliconius Genome Consortium. 2012 Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**, 94–98. (doi:10.1038/nature11041)
14. Van Belleghem SM *et al*. 2017 Complex modular architecture around a simple toolkit of wing pattern genes. *Nat. Ecol. Evol.* **1**, 52. (doi:10.1038/s41559-016-0052)
15. Feder JL, Egan SP, Nosil P. 2012 The genomics of speciation-with-gene-flow. *Trends Genet.* **28**, 342–350. (doi:10.1016/j.tig.2012.03.009)
16. Gompert Z, Egan SP, Barrett RD, Feder JL, Nosil P. 2017 Multilocus approaches for the measurement of selection on correlated genetic loci. *Mol. Ecol.* **26**, 365–382. (doi:10.1111/mec.13867)
17. Seehausen O *et al.* 2014 Genomics and the origin of species. *Nat. Rev. Genet.* **15**, 176–192. (doi:10.1038/nrg3644)
18. Barton N, Bengtsson BO. 1986 The barrier to genetic exchange between hybridising populations. *Heredity (Edinb.)* **57**, 357–376. (doi:10.1038/hdy.1986.135)
19. Barton NH, de Cara MA. 2009 The evolution of strong reproductive isolation. *Evolution* **63**, 1171–1190. (doi:10.1111/j.1558-5646.2009.00622.x)
20. Bierne N, Welch J, Loire E, Bonhomme F, David P. 2011 The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Mol. Ecol.* **20**, 2044–2072. (doi:10.1111/j.1365-294X.2011.05080.x)
21. Butlin RK, Smadja CM. 2018 Coupling, reinforcement, and speciation. *Am. Nat.* **191**, 155–172. (doi:10.1086/695136)
22. Roux C, Fraisse C, Romiguier J, Anciaux Y, Galtier N, Bierne N. 2016 Shedding light on the grey zone of speciation along a continuum of genomic divergence. *PLoS Biol.* **14**, e2000234. (doi:10.1371/journal.pbio.2000234)
23. Barton NH, Gale KS. 1993 Genetic analysis of hybrid zones. In *Hybrid zones and the evolutionary process* (ed. RG Harrison), pp. 13–45. New York, NY: Oxford University Press.
24. Martin SH, Jiggins CD. 2017 Interpreting the genomic landscape of introgression. *Curr. Opin. Genet. Dev.* **47**, 69–74. (doi:10.1016/j.gde.2017.08.007)
25. Bar-Zvi D, Lupo O, Levy AA, Barkai N. 2017 Hybrid vigor: the best of both parents, or a genomic clash? *Curr. Opin. Syst. Biol.* **6**, 22–27. (doi:10.1016/j.coisb.2017.08.004)
26. Gavrilets S. 2003 Perspective: models of speciation: what have we learned in 40 years? *Evolution* **57**, 2197–2215. (doi:10.1111/j.0014-3820.2003.tb00233.x)
27. Lindtke D, Buerkle CA. 2015 The genetic architecture of hybrid incompatibilities and their effect on barriers to introgression in secondary contact. *Evolution* **69**, 1987–2004. (doi:10.1111/evo.12725)
28. Han F, Lamichhaney S, Grant BR, Grant PR, Andersson L, Webster MT. 2017 Gene flow, ancient polymorphism, and ecological adaptation shape the genomic landscape of divergence among Darwin's finches. *Genome Res.* **27**, 1004–1015. (doi:10.1101/gr.212522.116)
29. Stankowski S, Chase MA, Fuiten AM, Rodrigues MF, Ralph PL, Streisfeld MA. 2019 Widespread selection and gene flow shape the genomic landscape during a radiation of monkeyflowers. *PLoS Biol.* **17**, e3000391. (doi:10.1371/journal.pbio.3000391)
30. Yeaman S, Aeschbacher S, Burger R. 2016 The evolution of genomic islands by increased establishment probability of linked alleles. *Mol. Ecol.* **25**, 2542–2558. (doi:10.1111/mec.13611)
31. Charlesworth B. 2012 The effects of deleterious mutations on evolution at linked sites. *Genetics* **190**, 5–22. (doi:10.1534/genetics.111.134288)
32. Cruickshank TE, Hahn MW. 2014 Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol. Ecol.* **23**, 3133–3157. (doi:10.1111/mec.12796)
33. Martin SH, Davey JW, Salazar C, Jiggins CD. 2019 Recombination rate variation shapes barriers to introgression across butterfly genomes. *PLoS Biol.* **17**, e2006288. (doi:10.1371/journal.pbio.2006288)
34. Edelman NB *et al*. 2019 Genomic architecture and introgression shape a butterfly radiation. *Science* **366**, 594–599. (doi:10.1126/science.aaw2090)
35. Barton NH, Hewitt GM. 1985 Analysis of hybrid zones. *Annu. Rev. Ecol. Syst.* **16**, 113–148. (doi:10.1146/annurev.es.16.110185.000553)
36. Buerkle CA, Lexer C. 2008 Admixture as the basis for genetic mapping. *Trends Ecol. Evol.* **23**, 686–694. (doi:10.1016/j.tree.2008.07.008)
37. Harrison RG, Larson EL. 2016 Heterogeneous genome divergence, differential introgression, and the origin and structure of hybrid zones. *Mol. Ecol.* **25**, 2454–2466. (doi:10.1111/mec.13582)
38. Brelsford A, Toews DPL, Irwin DE. 2017 Admixture mapping in a hybrid zone reveals loci associated with avian feather coloration. *Proc. R. Soc. B* **284**, 20171106. (doi:10.1098/rspb.2017.1106)
39. Bresadola L, Caseys C, Castiglione S, Buerkle CA, Wegmann D, Lexer C. 2019 Admixture mapping in interspecific *Populus* hybrids identifies classes of genomic architectures for phytochemical,

morphological and growth traits. *New Phytol.* **223**, 2076–2089. (doi:10.1111/nph.15930)

40. Pallares LF, Harr B, Turner LM, Tautz D. 2014 Use of a natural hybrid zone for genomewide association mapping of craniofacial traits in the house mouse. *Mol. Ecol.* **23**, 5756–5770. (doi:10.1111/mec.12968)
41. Pease JB, Haak DC, Hahn MW, Moyle LC. 2016 Phylogenomics reveals three sources of adaptive variation during a rapid radiation. *PLoS Biol.* **14**, e1002379. (doi:10.1371/journal.pbio.1002379)
42. Eckenwalder JE. 1984 Natural intersectional hybridization between North American species of *Populus* (Salicaceae) in sections Aigeiros and Tacamahaca. II. Taxonomy. *Can. J. Bot.* **62**, 325–335. (doi:10.1139/b84-051)
43. Macaya-Sanz D, Heuertz M, Lindtke D, Vendramin GG, Lexer C, Gonzalez-Martinez SC. 2016 Causes and consequences of large clonal assemblies in a poplar hybrid zone. *Mol. Ecol.* **25**, 5330–5344. (doi:10.1111/mec.13850)
44. Martinsen GD, Whitham TG, Turek RJ, Keim P. 2001 Hybrid populations selectively filter gene introgression between species. *Evolution* **55**, 1325–1335.
45. Rajora OP, Dancik BP. 1992 Genetic characterization and relationships of *Populus alba, P. tremula*, and *P.* × *canescens*, and their clones. *Theor. Appl. Genet.* **84**, 291–298. (doi:10.1007/BF00229485)
46. Suarez-Gonzalez A, Hefer CA, Christe C, Corea O, Lexer C, Cronk QC, Douglas CJ. 2016 Genomic and functional approaches reveal a case of adaptive introgression from *Populus balsamifera* (balsam poplar) in *P. trichocarpa* (black cottonwood). *Mol. Ecol.* **25**, 2427–2442. (doi:10.1111/mec.13539)
47. Tuskan GA *et al*. 2006 The genome of black cottonwood, *Populus trichocarpa* (Torr, Gray). *Science* **313**, 1596–1604. (doi:10.1126/science.1128691)
48. Stettler RF, Bradshaw Jr HD, Heilman PE, Hinckley TM. 1996 *Biology of Populus and its implications for management and conservation*. Ottawa, Canada: NRC Research Press.
49. Cervera MT, Storme V, Soto A, Ivens B, Van Montagu M, Rajora OP, Boerjan W. 2005 Intraspecific and interspecific genetic and phylogenetic relationships in the genus *Populus* based on AFLP markers. *Theor. Appl. Genet.* **111**, 1440–1456. (doi:10.1007/s00122-005-0076-2)
50. Hamzeh M, Périnet P, Dayanandan S. 2006 Genetic relationships among species of *Populus* (Salicaceae) based on nuclear genomic data. *J. Torrey Bot. Soc.* **133**, 519–527. (doi:10.3159/1095-5674(2006)133519:grasop]2.0.co;2)
51. Wang M *et al*. 2019 Phylogenomics of the genus *Populus* reveals extensive interspecific gene flow and balancing selection. *New Phytol.* **225**, 1370–1382. (doi:10.1111/nph.16215)
52. Ingvarsson PK. 2008 Multilocus patterns of nucleotide polymorphism and the demographic history of *Populus tremula*. *Genetics* **180**, 329–340. (doi:10.1534/genetics.108.090431)
53. Slavov GT *et al*. 2012 Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree *Populus trichocarpa*. *New Phytol.* **196**, 713–725. (doi:10.1111/j.1469-8137.2012.04258.x)
54. Wang J, Street NR, Scofield DG, Ingvarsson PK. 2016 Variation in linked selection and recombination drive genomic divergence during allopatric speciation of European and American aspens. *Mol. Biol. Evol.* **33**, 1754–1767. (doi:10.1093/molbev/msw051)
55. Wang J, Street NR, Scofield DG, Ingvarsson PK. 2016 Natural selection and recombination rate variation shape nucleotide polymorphism across the genomes of three related *Populus* species. *Genetics* **202**, 1185–1200. (doi:10.1534/genetics.115.183152)
56. Ma T *et al*. 2018 Ancient polymorphisms and divergence hitchhiking contribute to genomic islands of divergence within a poplar species complex. *Proc. Natl Acad. Sci. USA* **115**, E236–E243. (doi:10.1073/pnas.1713288114)
57. Christe C, Stolting KN, Bresadola L, Fussi B, Heinze B, Wegmann D, Lexer C. 2016 Selection against recombinant hybrids maintains reproductive isolation in hybridizing *Populus* species despite F1 fertility and recurrent gene flow. *Mol. Ecol.* **25**, 2482–2498. (doi:10.1111/mec.13587)
58. Lindtke D, Gompert Z, Lexer C, Buerkle CA. 2014 Unexpected ancestry of *Populus* seedlings from a hybrid zone implies a large role for postzygotic selection in the maintenance of species. *Mol. Ecol.* **23**, 4316–4330. (doi:10.1111/mec.12759)
59. Zeng YF, Zhang JG, Duan AG, Abuduhamiti B. 2016 Genetic structure of *Populus* hybrid zone along the Irtysh River provides insight into plastid-nuclear incompatibility. *Scient. Rep.* **6**, 28043. (doi:10.1038/srep28043)
60. Kisel Y, Barraclough TG. 2010 Speciation has a spatial scale that depends on levels of gene flow. *Am. Nat.* **175**, 316–334. (doi:10.1086/650369)
61. Lowry DB, Modliszewski JL, Wright KM, Wu CA, Willis JH. 2008 The strength and genetic basis of reproductive isolating barriers in flowering plants. *Phil. Trans. R. Soc. B* **363**, 3009–3021. (doi:10.1098/rstb.2008.0064)
62. Pickup M, Brandvain Y, Fraisse C, Yakimowski S, Barton NH, Dixit T, Lexer C, Cereghetti E, Field DL. 2019 Mating system variation in hybrid zones: facilitation, barriers and asymmetries to gene flow. *New Phytol.* **224**, 1035–1047. (doi:10.1111/nph.16180)
63. Abbott RJ. 2017 Plant speciation across environmental gradients and the occurrence and nature of hybrid zones. *J. Syst. Evol.* **55**, 238–258. (doi:10.1111/jse.12267)
64. Dickmann D, Kuzovkina Y. 2008 Poplars and willows in the world. In *Poplars and willows in the world, meeting the needs of society and the environment. International Poplar Commission Working Paper no. IPC/9-2* (eds JG Isebrands, J Richardson), pp. 9–12. Rome, Italy: Food and Agricultural Organization.
65. Lexer C, Joseph JA, van Loo M, Barbara T, Heinze B, Bartha D, Castiglione S, Fay MF, Buerkle CA. 2010 Genomic admixture analysis in European *Populus* spp. reveals unexpected patterns of reproductive isolation and mating. *Genetics* **186**, 699–712. (doi:10.1534/genetics.110.118828)
66. Stölting KN, Nipper R, Lindtke D, Caseys C, Waeber S, Castiglione S, Lexer C. 2013 Genomic scan for single nucleotide polymorphisms reveals patterns of divergence and gene flow between ecologically divergent species. *Mol. Ecol.* **22**, 842–855. (doi:10.1111/mec.12011)
67. Zheng H, Fan L, Milne RI, Zhang L, Wang Y, Mao K. 2017 Species delimitation and lineage separation history of a species complex of aspens in China. *Front. Plant Sci.* **8**, 375. (doi:10.3389/fpls.2017.00375)
68. Fan L, Zheng H, Milne RI, Zhang L, Mao K. 2018 Strong population bottleneck and repeated demographic expansions of *Populus adenopoda* (Salicaceae) in subtropical China. *Ann. Bot.* **121**, 665–679. (doi:10.1093/aob/mcx198)
69. Li H. 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*, 1303.3997 [q-bio.GN].
70. Purcell S *et al*. 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575. (doi:10.1086/519795)
71. Liu L, Yu L, Edwards SV. 2010 A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evol. Biol.* **10**, 302. (doi:10.1186/1471-2148-10-302)
72. Yang Z. 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556.
73. Terhorst J, Kamm JA, Song YS. 2017 Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat. Genet.* **49**, 303–309. (doi:10.1038/ng.3748)
74. Browning BL, Browning SR. 2013 Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459–471. (doi:10.1534/genetics.113.150029)
75. Martin SH, Van Belleghem SM. 2017 Exploring evolutionary relationships across the genome using topology weighting. *Genetics* **206**, 429–438. (doi:10.1534/genetics.116.194720)
76. Browning SR, Browning BL. 2007 Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097. (doi:10.1086/521987)
77. Guindon S, Lethiec F, Duroux P, Gascuel O. 2005 PHYML Online—a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res.* **33**, W557–W559. (doi:10.1093/nar/gki352)
78. Pease JB, Hahn MW. 2015 Detection and polarization of introgression in a five-taxon phylogeny. *Syst. Biol.* **64**, 651–662. (doi:10.1093/sysbio/syv023)
79. Charlesworth D. 2006 Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet.* **2**, e64. (doi:10.1371/journal.pgen.0020064)
80. Goodwillie C, Kalisz S, Eckert CG. 2005 The evolutionary enigma of mixed mating systems in plants: occurrence,

theoretical explanations, and empirical evidence. *Annu. Rev. Ecol. Evol. Syst.* **36**, 47–79. (doi:10.1146/annurev.ecolsys.36.091704.175539)

81. Schumer M *et al*. 2018 Natural selection interacts with recombination to shape the evolution of hybrid genomes. *Science* **360**, 656–660. (doi:10.1126/science.aar3684)
82. Lexer C, Buerkle CA, Joseph JA, Heinze B, Fay MF. 2007 Admixture in European *Populus* hybrid zones makes feasible the mapping of loci that contribute to reproductive isolation and trait differences. *Heredity* **98**, 74–84. (doi:10.1038/sj.hdy.6800898)
83. Uecker H, Setter D, Hermisson J. 2015 Adaptive gene introgression after secondary contact. *J. Math. Biol.* **70**, 1523–1580. (doi:10.1007/s00285-014-0802-y)
84. De Carvalho D *et al.* 2010 Admixture facilitates adaptation from standing variation in the European aspen (*Populus tremula* L.), a widespread forest tree. *Mol. Ecol.* **19**, 1638–1650. (doi:10.1111/j.1365-294X.2010.04595.x)
85. Rettelbach A, Nater A, Ellegren H. 2019 How linked selection shapes the diversity landscape in *Ficedula* flycatchers. *Genetics* **212**, 277–285. (doi:10.1534/genetics.119.301991)
86. Kruuk LE, Baird SJ, Gale KS, Barton NH. 1999 A comparison of multilocus clines maintained by environmental adaptation or by selection against hybrids. *Genetics* **153**, 1959–1971.
87. Edmands S, Timmerman CC. 2003 Modeling factors affecting the severity of outbreeding depression. *Conserv. Biol.* **17**, 883–892. (doi:10.1046/j.1523-1739.2003.02026.x)
88. Smadja CM, Butlin RK. 2011 A framework for comparing processes of speciation in the presence of gene flow. *Mol. Ecol.* **20**, 5123–5140. (doi:10.1111/j.1365-294X.2011.05350.x)
89. McKown AD, Guy RD, Klapste J, Geraldes A, Friedmann M, Cronk QC, El-Kassaby YA, Mansfield SD, Douglas CJ. 2014 Geographical and environmental gradients shape phenotypic trait variation and genetic structure in *Populus trichocarpa*. *New Phytol.* **201**, 1263–1276. (doi:10.1111/nph.12601)
90. Shang H, Hess J, Pickup M, Field DL, Ingvarsson PK, Liu J, Lexer C. 2020 Data from: Evolution of strong reproductive isolation in plants: broad-scale patterns and lessons from a perennial model group. Dryad Digital Repository. (doi:10.5061/dryad.h9w0vt4fw)

Supporting Information, including Supporting Tables and Supporting Figures

Evolution of strong reproductive isolation in plants: broad-scale patterns and lessons from a perennial model group

Huiying Shang[1,2*], Jaqueline Hess[1,3], Melinda Pickup[4], David Field[1,5], Pär Ingvarsson[6], Jianquan Liu[7], Christian Lexer[1, †]

Author affiliations:

[1]Department of Botany and Biodiversity Research, University of Vienna, Rennweg 14, 1030 Vienna, Austria

[2]Vienna Graduate School of Population Genetics, Vienna, Austria

[3]Helmholtz Centre for Environmental Research, Halle (Saale), Germany

[4]Institute of Science and Technology (IST), Klosterneuburg, Austria

[5]Edith Cowan University, Perth, Australia

[6]Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden

[7]Key Laboratory for Bio-resources and Eco-environment, College of Life Science, Sichuan University, Chengdu, People's Republic of China

Whole genome sequencing and data processing.

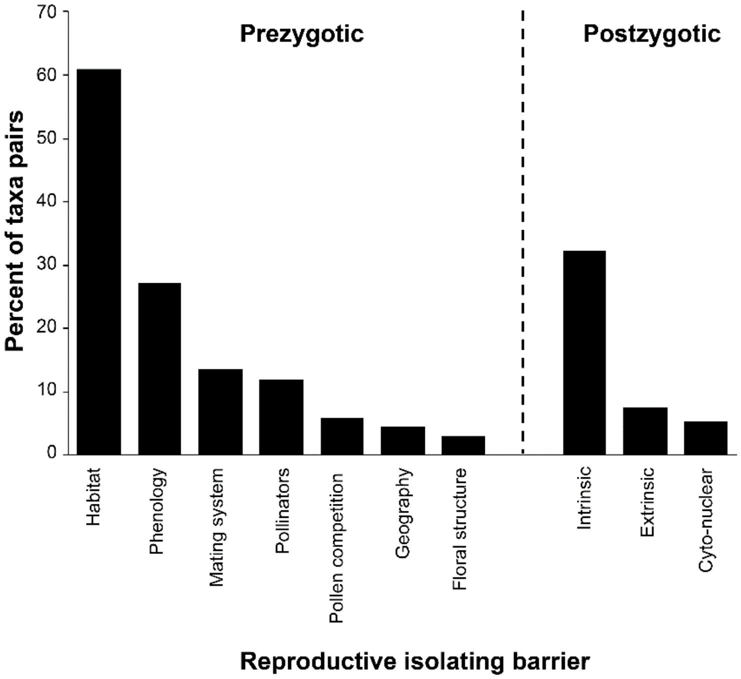
Paired-end reads for all individuals were first analysed using FastQC (http:// www.bioinformatics.babraham.ac.uk/projects/fastqc/) to perform a quality control check of the raw sequence data Then Trimmomatric [1] was used to remove adapters and low-quality reads with the command "TruSeq3-SE. fa: 2:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW: 4:15 MINLEN:36". The reads of each individual were then mapped to the *P. trichocarpa* reference genome [2] using BWA (Version: 0.7.15-r1140) with the end-to-end alignment option [3]. Picard package v2.5 AddOrReplaceReadGroups was used to add group names and MarkDuplicates was used to remove duplicate reads (http://broadinstitute.github.io/picard/). Subsequently, reads in insertion/deletion (indel) regions were identified and realigned using the Genome Analysis Toolkit (GATK v3.6) RealignerTargetCreator and IndelRealigner [4]. Single nucleotide polymorphisms (SNPs) and genotypes were called using the GATK Unified Genotyper and base recalibration using BaseRecalibrator [5]. We used the high-quality SNPs with genotype quality above 20 as reference SNPs. All sites were then filtered with Python script available at Github (https://github.com/Huiying123/phylogenomic_piplines). SNPs were discarded with genotype quality lower than 20, depth lower than 5X or higher than three times the mean depth and sites with a percentage of missing data exceeding 50%. As a result, 7,026,036 SNPs were retained and used for subsequent analysis.

Estimation of a species tree and divergence times from whole-genome phylogenomic data.

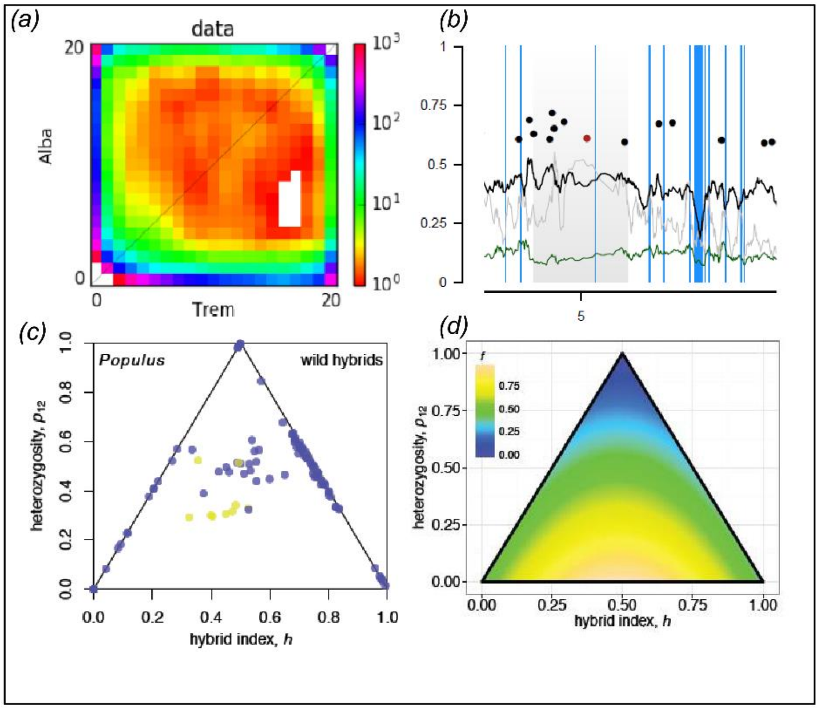
We constructed a species tree using MP-EST version1.5 [6], an approach based on the multi-species coalescent. We first separated the whole genome resequencing data according to the locations of orthologue genes. For each gene, a gene tree was generated in RAxML version 8.0 [7] using the GTRGAMMAI model with 100 bootstrap replicates. The coalescent species tree was then estimated from the 26,041 rooted orthologue gene trees by maximizing a pseudo-likelihood function in MP-EST. Each analysis started with a random number seed and 10 independent tree searches within each run. To evaluate the bootstrap support of the species tree, we randomly picked up half of the gene trees and to use them as input trees to generate the species tree. We repeated this step 100 times.

Divergence times among *Populus* taxa were estimated using MCMCTree software in the PAML package [8]. The program is based on a Bayesian algorithm for species divergence time estimation using fossil constraints. We based our estimation procedure on four-fold degenerate sites and the species tree generated from MP-EST. A molecular clock was assumed according to divergence time estimates for *P. alba* and *P. tremula* (2.8-3.2Ma) available from a previous study [9]. We used 100 million years per unit time for the analysis and the root age was set to < 0.1. The mutation rate of *Populus* was estimated to be 2.5×10^{-9} per site per year and we assumed a generation time of 15 years [10]. Therefore, we set the rgene_gamma = (1, 4) and the rate drift parameter sigma2_gamma = (1, 0.4) under the GTR model. The analysis was run 200,000 times with a burn-in of 10,000 iterations and a sample frequency of 100. We did this twice to check for convergence using TRACER for the posterior distribution. Effective sample size (ESS) exceeded 200 for all parameters.

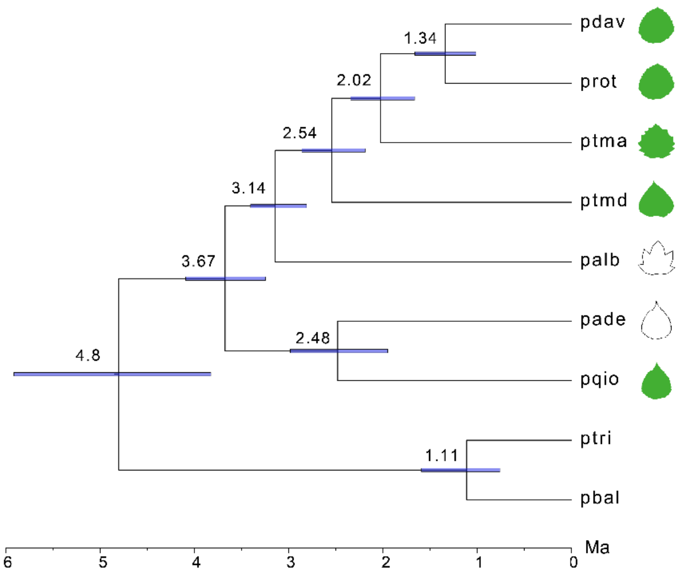
Online supporting figures and tables



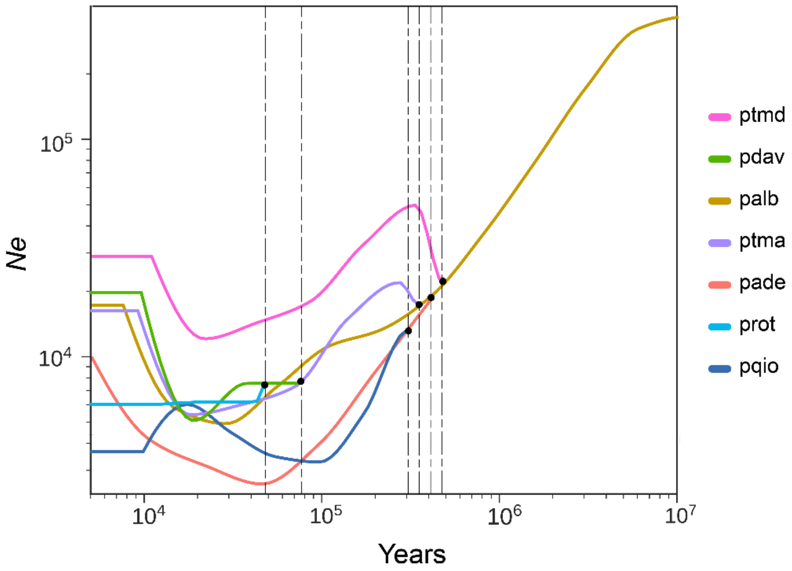
Supporting Information Fig. S1. The percent of taxa pairs with the different pre- and post-zygotic reproductive isolating barriers for 133 angiosperm species pairs. Pre-zygotic barriers are: (i) habitat (divergent habitat preference), (ii) phenology (divergent flowering phenology), (iii) mating system (divergent mating systems – including taxon pairs with a predominantly outcrossing self-compatible species and a highly selfing self-compatible species, pairs where both taxa are selfing, and pairs including a self-incompatible and self-compatible species, (iv) pollinators (pollinator preference), (v) pollen competition, (vi) geography (spatial isolation of parental species) and (vii) floral structure (divergent floral structure). Post-zygotic barriers are: (i) intrinsic incompatibilities (genic incompatibility, Bateson-Dobzhansky-Muller Incompatibilities (BDMIs), reduced hybrid fitness, transmission distortion of diagnostic markers, chromosomal differences, hybrid sterility) (ii) extrinsic incompatibilities (hybrid inviability and reduced hybrid fitness) and (iii) cyto-nuclear interactions (evidence of asymmetries in cross direction and/or outcomes).



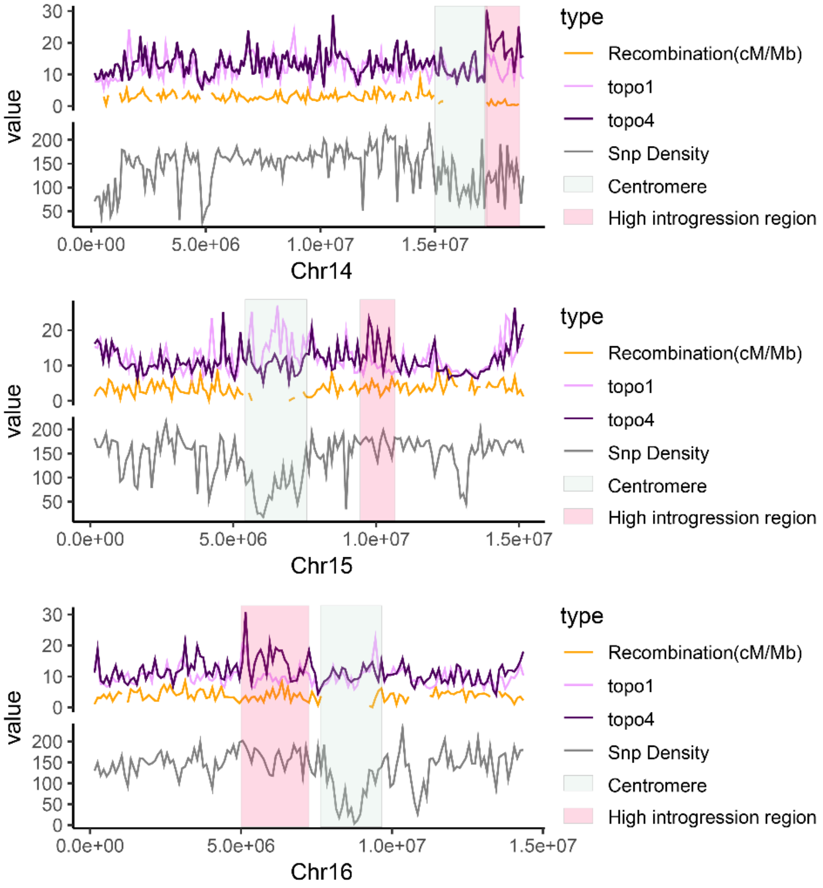
Supporting Information Fig. S2. Genomic patterns and fitness surface estimated for the late-stage speciation taxa *Populus alba* and *P*. *tremula* and their hybrids. **(a)** Site frequency spectrum (SFS) from pooled whole-genome sequence data; **(b)** an example of a genome scanning for genetic differentiation along the chromosome 10 (allele frequency differentials, AFD: black), sequence divergence (Dxy, green), and genomic features (fraction of repetitive DNA, grey); rectangular box (grey) indicates approximate centromere position, blue shades indicate 8kb genomic windows free of fixation; black and red dots indicate AFD and AFD / reduced diversity outlier windows at ≥2SD, respectively. **(c)** Observed survivorship / mortality of seedlings (dots) characterized by their hybrid index (horizontal axis; h, 0= *P. tremula*, 1=*P. alba*) and inter-species heterozygosity (vertical axis; p12, 0 = minimum and 1=maximum); blue and yellow dots denote seedlings alive and dead after an intense selection episode in a common garden trial (several yellow dots are covered by the numerous blue dots = survivors for high values of h and intermediate p12). **(d)** Hybrid breakdown score (f) from Fisher´s geometric model fitted to survivorship data shown in **(c)**, coloured as low (blue) and high (yellow) breakdown, respectively. **(a)** and **(b)** redrawn and adapted from [9], **(c)** and **(d)** redrawn and adapted from [11] and [12].



Supporting Information Fig. S3. Species tree of all nine *Populus* taxa sampled and sequenced for this study (including outgroups) inferred using the coalescent-based method implemented in MP-EST. Divergence time was estimated with MCMCtree using four-fold degenerate sites. The blue bars along nodes indicate the 95% confidence intervals of divergence time. Species abbreviations follow **Fig. 2** of the main paper.



Supporting Information Fig. S4. The dynamic effective population size (*Ne*) and divergence times of each species inferred by SMC++. Both *Ne* (vertical axis) and divergence time in years (horizontal axis) are shown on a log scale. Bifurcation sites indicated by dashed vertical lines (separation points of coloured lines) indicate the divergence times of species pairs.



Supporting Information Fig. S5. Putatively introgressed regions along chromosomes of early-stage speciation taxa. Weightings for species topology (topo1, pink) and introgression topology (topo4, purple) (**Fig. 4d**) are shown along exemplary chromosomes. The population-scaled recombination rate of *P. tremula* (orange) and SNP density (grey) are shown in 100kb windows. Grey boxes indicate approximate centromeric regions. Pink boxes exemplify regions with consistently increased weightings for the introgression topology (topo4), potentially pointing to locally introgressed chromosome segments.

Supporting Table S1. Sampling locations and sample IDs for all sequenced individuals.

ID	Species Name	Location (N, E)		Elevation(m)	Number
MaoKS-CX-2014-083A	*Populus rotundifolia Griff. var. duclouxiana (Dode)*	29.8194	102.2435	2148.66	1
MaoKS-CX-2014-177	*Populus rotundifolia Griff. var. duclouxiana (Dode)*	29.7356	96.0565	3068.15	1
MaoKS-CX-2014-261A	*Populus rotundifolia Griff. var. duclouxiana (Dode)*	27.1423	99.3916	2670.89	1
LiuJQ-QTP-2013-123	*Populus rotundifolia Griff. var. duclouxiana (Dode)*	29.5170	94.8716	2963.13	1
MaoKS-CX-2014-056	*Populus davidiana Dode*	31.5549	102.4176	3362.57	1
LiuJQ-MZL-2013-221	*Populus davidiana Dode*	41.0005	123.1636	321.20	1
LiuJQ-MZL-2013-302	*Populus davidiana Dode*	45.4693	130.9263	407.00	1
LiuJQ-MZL-2013-425	*Populus davidiana Dode*	38.7521	105.9355	1899.92	1
LiuJQ-MZL-2013-167	*Populus davidiana Dode*	39.2277	114.7393	1445.70	1
MaoKS-CX-2014-311	*Populus adenopoda Maxim.*	25.8204	107.3589	918.42	1
MaoKS-CX-2014-320	*Populus adenopoda Maxim.*	27.6683	107.2082	846.08	1
LiuJQ-MZL-2013-055	*Populus adenopoda Maxim.*	29.3404	109.5691	830.00	1
LiuJQ-MZL-2013-063	*Populus adenopoda Maxim.*	32.3771	113.3022	905.92	1
LiuJQ-F-2015-01	*Populus adenopoda Maxim.*	32.7559	105.2528	916.02	1
LiuJQ-Tian-2015-001	*Populus qiongdaoensis* T.Hong et P.Luo	19.1167	109.0925	212.66	1
LiuJQ-Tian-2015-002	*Populus qiongdaoensis* T.Hong et P.Luo	19.1162	109.0923	218.45	2
pop2014-3	*Populus alba*	47.4614	87.8041	500.00	1
pop2014-5	*Populus alba*	47.3817	87.8045	500.00	1
pop2014-15	*Populus alba*	47.3484	87.8669	500.00	1
pop2014-24	*Populus alba*	47.7175	86.8830	482.00	1
pop2014-26	*Populus alba*	47.0117	86.2693	485.00	1
pop2014-45	*Populus tremula*	47.9115	88.1265	993.00	1
pop2014-48	*Populus tremula*	47.9623	88.1792	1236.00	1
pop2014-50	*Populus tremula*	47.9669	88.1836	1272.00	1
pop2014-52	*Populus tremula*	47.9679	88.1852	1305.00	1
pop2014-53	*Populus tremula*	47.9768	88.2001	1333.00	1

Supporting Table S2. Sequencing statistics for all sequenced individuals.

Species name	ID	Raw reads	Cleaned reads	Mapping rate	Average depth	SNP
P. adenopoda	pade0121	51641058	47807522	92.27%	22.18	5082714
	pade31109	69700010	64542562	91.64%	27.11	5873211
	pade32018	76076568	70859083	87.30%	34.34	5984982
	pade5508	62788985	58356536	91.70%	29.76	5538752
	pade6307	85708024	79607881	92.35%	32.61	6310055
P. alba	palb01	43330912	38009829	92.35%	15.55	7298529
	palb02	56211103	48897168	91.03%	19.64	7497659
	palb03	41657930	35537631	90.63%	14.67	6921131
	palb04	49828854	43107169	91.09%	17.69	8094748
	palb05	40179026	32073180	89.82%	12.63	6885215
P. balsamifera	pbal01	61933611	47410704	95.58%	20.39	3316707
	pbal02	66532895	62076405	96.92%	26.54	3402179
P. davidiana	pdav16709	38789412	33915862	92.29%	32.90	7545093
	pdav22110	79929667	74313895	92.71%	38.46	7898376
	pdav30211	92659451	86296246	92.88%	42.87	7734377
	pdav42521	102786028	95584067	92.12%	27.81	8202549
	pdav5607	67161988	62385748	91.11%	15.43	7341735
P. qiongdaoensis	pqioT0103	100321484	92067364	91.47%	42.60	7324558
	pqioT0202	76744734	70556396	91.24%	32.44	6612612
	pqioT0205	82550486	76492429	91.48%	35.56	6704599
P. rotundifolia	prot083A04	47365375	40546868	91.81%	18.44	7046880
	prot12319	78975717	73288740	92.22%	33.01	6973289
	prot17718	60670864	56256163	91.48%	25.19	6891117
	prot261A13	53047205	46089672	86.66%	19.34	5908876
P. tremula	ptma01	37910238	33716847	91.15%	14.29	7627370
	ptma02	63839360	56861868	90.79%	23.24	8815487
	ptma03	25993725	23350515	92.26%	10.43	6980171
	ptma04	49584091	42905174	90.28%	17.68	8327853
	ptma05	57261669	51099511	91.97%	21.92	8863568
P. tremuloides	ptmd01	99361734	92255121	91.70%	28.51	9785394
	ptmd02	96230649	88156718	91.09%	28.35	9890484
	ptmd03	96387182	88430839	91.98%	29.22	9944656

	ptmd04	108934535	100107534	91.23%	31.81	9951773
	ptmd05	130684419	118019244	91.26%	37.69	10145944
P. trichocarpa	ptri01	99715650	95227609	94.03%	38.16	2088232
	ptri02	98048861	91057729	94.72%	34.96	2126432

Supporting Table S3. Divergence time estimates in millions of years (Ma) obtained with MCMC tree within the PAML software package.

Splits	Posterior mean (Ma)	95% HPD CI (Ma)	HPD-CI-width (Ma)
(ptri,pbal)-((pade,pqio),(palb, (ptmd, (ptma, (pdav, prot)))))	4.8	(3.62, 6.06)	2.44
(pade,pqio)-(palb, (ptmd, (ptma, (pdav, prot))))	3.67	(3.17, 4.15)	0.98
palb- (ptmd, (ptma, (pdav, prot)))	3.14	(2.81, 3.41)	0.6
ptmd- (ptma, (pdav, prot))	2.54	(2.16, 2.88)	0.71
ptma-(pdav, prot)	2.02	(1.66, 2.36)	0.7
pdav-prot	1.34	(0.98, 1.66)	0.68
pade-pqio	2.48	(1.85, 3.09)	1.24
ptri-pbal	1.11	(0.70, 1.62)	0.93

Supporting Table S4. Spearman rank correlation and linear regression statistics for relationships between average TWISST tree topology weightings for each chromosome and average recombination rate or gene density in windows of 100kb along each chromosome. Results for the topologies discussed in the main text are indicated by bold type.

Spearman's correlation analysis between average weighting of each topology and recombination or gene density				
	Ancient introgression		**Recent introgression**	
Topology	Average Weighting and recombination	Average Weighting and gene density	Average Weighting and recombination	Average Weighting and gene density
topo1	r=0.5; p=0.0310	r=-0.297; p=0.217	**r=-0.688; p=0.001546**	**r=0.735; p=0.0003408**
topo2	r=0.516; p=0.0255	r=-0.250; p=0.301	r=0; p=1	r=-0.22; p=0.3662
topo3	r=0.660; p=0.00273	r=-0.507; p=0.0267	r=0.289; p=0.2286	r=-0.3418; p=0.152
topo4	**r=0.258 p=0.2852**	**r=0.225; p=0.3545**	**r=-0.854; p=2.2e-16**	**r=0.793; p=5.029e-05**
topo5	**r=0.163; p=0.503**	**r=-0.228; p=0.3487**	r=-0.342; p=0.1518	r=0.283; p=0.2405
topo6	**r=-0.653; p=0.00312**	**r=0.728; p=0.00041**	r=-0.291; p=0.2257	r=0.0413; p=0.8667
topo7	r=0.711; p=0.000927	r=-0.643; p=0.00297	r=0.812; p=2.485e-05	r=-0.842; p=6.213e-06
topo8	r=0.656; p=0.00292	r=-0.774; p=0.00010	r=0.849; p=2.2e-16	r=-0.718; p=0.000525
topo9	r=0.681; p=0.00179	r=-0.887; p=4.29e-07	r=0.884; p=2.2e-16	r=-0.731; p=0.000376
topo10	r=0.311; p=0.1953	r=-0.146; p=0.5513	r=0.796; p=6.04e-05	r=-0.793; p=5.029e-05
topo11	r=-0.844; p=2.2-16	r=0.605; p=0.00611	r=0.658; p=0.00282	r=-0.687; p=0.00115
topo12	r=0.153; p=0.5313	r=-0.266; p=0.2705	r=0.867; p=2.2e-16	r=-0.698; p=0.000896
topo13	r=0.598; p=0.00794	r=-0.682; p=0.0013	r=-0.705; p=0.00105	r=0.783; p=7.16e-05
topo14	r=-0.788; p=9.61e-05	r=0.695; p=0.00096	r=0.739; p=0.000455	r=-0.784; p=7.16e-05
topo15	r=0.646; p=0.00355	r=-0.787; p=6.31e-05	r=0.830; p=3.22e-06	r=-0.713; p=0.000617

Linear regression analysis between average weighting of each topology and recombination or gene density																				
	Ancient introgression										**Recent introgression**									
	Average Weighting and recombination					Average Weighting and gene density					Average Weighting and recombination					Average Weighting and gene density				
Topology	slope	intercept	R2	R2.Ad	P	slope	intercept	R2	R2.Ad	P	slope	intercept	R2	R2.Ad	P	slope	intercept	R2	R2.Ad	P
topo1	0.004	0.026	0.322	0.282	0.011	-0.001	0.050	0.253	0.209	0.028	**-0.015**	**0.161**	**0.517**	**0.488**	**0.001**	**0.005**	**0.056**	**0.495**	**0.465**	**0.001**
topo2	0.002	0.035	0.174	0.125	0.076	-0.001	0.046	0.083	0.029	0.231	0.000	0.071	0.003	-0.056	0.825	0.000	0.071	0.005	-0.054	0.776
topo3	0.004	0.016	0.451	0.419	0.002	-0.001	0.039	0.325	0.285	0.011	0.001	0.063	0.039	-0.017	0.417	0.000	0.071	0.042	-0.015	0.401
topo4	**0.001**	**0.129**	**0.015**	**-0.043**	**0.613**	**0.001**	**0.124**	**0.072**	**0.017**	**0.267**	**-0.023**	**0.195**	**0.761**	**0.747**	**0.000**	**0.007**	**0.049**	**0.560**	**0.535**	**0.000**
topo5	**0.002**	**0.124**	**0.021**	**-0.037**	**0.554**	**-0.001**	**0.136**	**0.028**	**-0.029**	**0.491**	-0.002	0.071	0.132	0.081	0.126	0.000	0.064	0.008	-0.051	0.722
topo6	**-0.020**	**0.278**	**0.530**	**0.502**	**0.000**	**0.007**	**0.144**	**0.502**	**0.472**	**0.001**	-0.003	0.093	0.072	0.017	0.268	0.000	0.082	0.001	-0.057	0.878

topo7	0.004	0.029	0.498	0.469	0.001	-0.001	0.054	0.388	0.352	0.004	0.007	0.024	0.711	0.694	0.000	-0.002	0.073	0.628	0.607	0.000
topo8	0.006	0.018	0.501	0.472	0.001	-0.002	0.057	0.517	0.489	0.001	0.007	0.020	0.724	0.708	0.000	-0.002	0.061	0.407	0.372	0.003
topo9	0.006	0.011	0.548	0.522	0.000	-0.002	0.051	0.563	0.538	0.000	0.008	0.032	0.831	0.821	0.000	-0.002	0.077	0.453	0.421	0.002
topo10	0.001	0.032	0.216	0.169	0.045	0.000	0.039	0.053	-0.002	0.342	0.007	0.019	0.695	0.677	0.000	-0.002	0.064	0.533	0.505	0.000
topo11	-0.009	0.105	0.697	0.679	0.000	0.002	0.050	0.377	0.341	0.005	0.007	0.024	0.574	0.549	0.000	-0.002	0.069	0.420	0.386	0.003
topo12	0.001	0.034	0.082	0.028	0.233	-0.001	0.045	0.170	0.122	0.079	0.005	0.035	0.758	0.744	0.000	-0.001	0.066	0.439	0.406	0.002
topo13	0.005	0.023	0.457	0.425	0.001	-0.002	0.058	0.411	0.377	0.003	-0.011	0.132	0.525	0.497	0.000	0.004	0.057	0.579	0.554	0.000
topo14	-0.010	0.117	0.599	0.576	0.000	0.003	0.055	0.366	0.328	0.006	0.006	0.034	0.557	0.530	0.000	-0.002	0.077	0.574	0.549	0.000
topo15	0.004	0.023	0.467	0.436	0.001	-0.002	0.050	0.594	0.570	0.000	0.006	0.024	0.714	0.697	0.000	-0.002	0.063	0.522	0.494	0.000

Slope: The slope of a regression line represents the rate of change in y as x changes.

Intercept: The intercept is the expected mean value of Y when all X=0.

The R-squared (R2) statistic: provides a measure of how well the model is fitting the actual data, it is the percentage of the response variable variation that is explained by a linear model. The value is a biased estimate based on the sample size.

R2. Ad: unbiased R2

P-value: the ability to reject the null hypothesis. If the p-value is less than 0.05 or 0.01, corresponding respectively to a 5% or 1% chance of rejecting the null hypothesis when it is true. Calculated by F-statistics

Online supporting references

[1] Bolger, A. M., Lohse, M. & Usadel, B. 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120. (DOI:10.1093/bioinformatics/btu170).

[2] Tuskan, G. A. & Difazio, S. & Jansson, S. & Bohlmann, J. & Grigoriev, I. & Hellsten, U. & Putnam, N. & Ralph, S. & Rombauts, S. & Salamov, A., et al. 2006 The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science (New York, N.Y.)* **313**, 1596-1604. (DOI:10.1126/science.1128691).

[3] Li, H. 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. In *arXiv*:1303.3997v2 (q-bio.GN)

[4] DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., et al. 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491. (DOI:10.1038/ng.806).

[5] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303. (DOI:10.1101/gr.107524.110).

[6] Liu, L., Yu, L. & Edwards, S. V. 2010 A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evol Biol* **10**, 302. (DOI:10.1186/1471-2148-10-302).

[7] Stamatakis, A. 2014 RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313. (DOI:10.1093/bioinformatics/btu033).

[8] Yang, Z. 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer applications in the biosciences : CABIOS* **13**, 555-556.

[9] Christe, C., Stolting, K. N., Paris, M., Frasmall yi, U. C., Bierne, N. & Lexer, C. 2017 Adaptive evolution and segregating load contribute to the genomic landscape of divergence in two tree species connected by episodic gene flow. *Mol Ecol* **26**, 59-76. (DOI:10.1111/mec.13765).

[10] Wang, J., Street, N. R., Scofield, D. G. & Ingvarsson, P. K. 2016 Variation in linked selection and recombination drive genomic divergence during allopatric speciation of European and American aspens. *Mol Biol Evol* **33**, 1754-1767. (DOI:10.1093/molbev/msw051).

[11] Christe, C., Stolting, K. N., Bresadola, L., Fussi, B., Heinze, B., Wegmann, D. & Lexer, C. 2016 Selection against recombinant hybrids maintains reproductive isolation in hybridizing *Populus* species despite F1 fertility and recurrent gene flow. *Mol Ecol* **25**, 2482-2498. (DOI:10.1111/mec.13587).

[12] Simon, A., Bierne, N. & Welch, J. J. 2018 Coadapted genomes and selection on hybrids: Fisher's geometric model explains a variety of empirical patterns. *Evol Lett* **2**, 472-498. (DOI:10.1002/evl3.66).

3 Chapter II

Conserved genomic landscapes of differentiation over the *Populus* speciation continuum

Huiying Shang[1,2], Martha Rendón-Anaya[3], Ovidiu Paun[1], David Field[4], Jaqueline Hess[5], Claus Vogl[6], Jianquan Liu[7], Pär K. Ingvarsson[3,8]*, Thibault Leroy[1,8]*, Christian Lexer[1,8 †]

[1]Department of Botany and Biodiversity Research, University of Vienna, Rennweg 14, 1030 Vienna, Austria.

[2]Vienna Graduate School of Population Genetics, Vienna, Austria.

[3]Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden.

[4]Edith Cowan University, Perth, Australia.

[5]Helmholtz Centre for Environmental Research, Halle (Saale), Germany.

[6]VetMed Vienna

[7]Key Laboratory for Bio-resources and Eco-environment, College of Life Science, Sichuan University, Chengdu, People's Republic of China

[8] These authors contributed equally to this work.

[†]Deceased.

*Corresponding author: Thibault Leroy, Department of Botany and Biodiversity Research, University of Vienna, Rennweg 14, 1030 Vienna, Austria; Email: thibault.leroy@univie.ac.at

Pär K. Ingvarsson, Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden; Email: par.ingvarsson@slu.se

Status: Prepared for submission in Molecular Biology and Evolution

Contribution: Laboratory work, data analysis, interpretation the results, original draft preparation

Article

Conserved genomic landscapes of differentiation across *Populus* speciation continuum

Huiying Shang[1,2], Martha Rendón-Anaya[3], Ovidiu Paun[1], David Field[4], Jaqueline Hess[5], Claus Vogl[6], Jianquan Liu[7], Pär K. Ingvarsson[3,8]*, Thibault Leroy[1,8]*, Christian Lexer[1,8 †]

[1]Department of Botany and Biodiversity Research, University of Vienna, Vienna, Austria.

[2]Vienna Graduate School of Population Genetics, Vienna, Austria.

[3]Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden.

[4]Edith Cowan University, Perth, Australia.

[5]Helmholtz Centre for Environmental Research, Halle (Saale), Germany.

[6]Department of Biomedical Sciences, Vetmeduni Vienna, Vienna, Austria

[7]Key Laboratory for Bio-resources and Eco-environment, College of Life Science, Sichuan University, Chengdu, People's Republic of China

[8] These authors contributed equally to this work.

[†]Deceased.

*Corresponding author: Thibault Leroy, Department of Botany and Biodiversity Research, University of Vienna, Rennweg 14, 1030 Vienna, Austria; Email: thibault.leroy@univie.ac.at

Pär K. Ingvarsson, Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden; Email: par.ingvarsson@slu.se

Abstract

Investigating genome-wide variation patterns along a speciation continuum is of central importance to understand the evolutionary processes contributing to lineage diversification. To identify which forces have shaped the genomic landscapes, we resequenced 201 individuals from eight closely related *Populus* species, representing pairs of species at different stages along the speciation continuum. Using population structure and identity by descent analyses, we first revealed extensive introgression between some species pairs, especially those with parapatric distributions. Inferences of the historical changes in effective population sizes support species-specific demographic trajectories, including recent population expansions in the species characterized by broad present-day distributions. We observed highly conserved genomic landscapes, either focusing on within-species (genetic diversity and recombination rate) or among-species variation (relative and absolute divergence levels). Independent of the stage across the divergence continuum, we recovered negative correlations between nucleotide diversity and relative divergence across all species pairs, which is consistent with a substantial contribution of linked selection in shaping these genomic landscapes. However, the positive correlations between nucleotide diversity and absolute divergence landscapes became weaker as the overall divergence level (d_a) increased, suggesting that background selection is not the only factor at play. Through negative correlations between introgression (f_{d}) and F_{ST} in some species pairs, we also found support for an additional role of gene flow in shaping genomic landscapes of differentiation. Nonetheless, linked selection and recombination rate variation appear as major factors shaping the heterogeneous genomic landscape of divergence in *Populus*.

Keywords: differentiation islands, divergence, introgression, identity-by-descent, linked selection, recombination

Introduction

Understanding the evolutionary forces that shape natural genetic variation is a central goal of evolutionary biology. Over the last decade, population genomic studies have documented highly heterogeneous genomic landscapes of differentiation, identifying both regions of elevated and reduced differentiation between diverging populations (Ellegren, et al. 2012; Martin, et al. 2013; Lamichhaney, et al. 2015; Vijay, et al. 2016; Sendell-Price, et al. 2020). Notwithstanding the specific features of individual groups, the overall drivers of this omnipresent pattern can be multifarious (extended below) and their relative contribution is still debated (Wolf and Ellegren 2017).

Hotspots of elevated genetic differentiation relative to genomic background are often referred to as 'differentiation islands' or 'speciation islands' and are assumed to form around loci underlying local adaptation and reproductive isolation. Thus, delineating differentiation islands has recently become a major topic of research in the field of speciation and adaptation genomics (Burri 2017b; Martin and Jiggins 2017; Ravinet, et al. 2018; Tavares, et al. 2018; Stankowski, et al. 2019). Such investigations are facilitated in groups still experiencing interspecific gene flow, i.e., species diverging under an isolation-with-migration or a secondary contact scenario (Harrison and Larson 2016; Roux, et al. 2016; Wolf and Ellegren 2017; Leroy, et al. 2020; Yamasaki, et al. 2020). For such species pairs, effective migration can be reduced around ‘speciation islands’ due to the effect of selection against hybrids, whereas the remainder of the genome can be homogenized by gene flow. In *Heliconius* butterflies for instance, admixture footprints were found to be locally weaker around loci involved in Mullerian mimicry (Martin, et al. 2013). In many plant species, such as monkey-flowers, snapdragons and morning-glories, differentiation islands have been reported to harbor key adaptive loci controlling floral traits, proving the role of selection and gene flow in shaping a highly heterogeneous genomic landscape of differentiation (Ravinet, et al. 2017; Samuk, et al. 2017; Tavares, et al. 2018; Martin, et al. 2019; Rifkin, et al. 2019).

Other empirical studies reported, however, that heterogeneous differentiation landscapes can emerge due to genomic features not causally linked to speciation and adaptation, such as background selection, recombination rate variation, biased gene conversion, and genomic characteristics influenced by life history traits (Corbett-Detig, et al. 2015; Wolf and Ellegren 2017). Linked selection, which includes genetic hitchhiking (Smith and Haigh 1974) and background selection (Charlesworth, et al. 1993), locally reduces effective population size (*Ne*), leading to decreased diversity levels and elevated relative differentiation. Given its potential to modulate the effect of selection on neighboring genomic regions (Charlesworth and Campos, 2014), variation in recombination rate along the genome has been empirically found to be a driver of the genomic landscape of diversity and differentiation (Renaut, et al. 2013; Burri 2017a; Gagnaire, et al. 2018; Henderson and Brelsford 2020). However, a recent study in a triplet of *Leptidea* butterfly species suggested that divergence landscapes are mainly shaped by directional selection, rather than recombination or introgression (Talla, et al. 2019). Finally, a study in *Boechera stricta* demonstrated that ancestral balanced polymorphism may have contributed to the genomic regions of high divergence (Wang, et al. 2019). Therefore, debates still exist regarding the evolutionary processes that contribute to the heterogeneous landscape of differentiation. Disentangling the relative contribution of these different evolutionary processes to interpret the underlying mechanisms remains challenging with only one to a few pairs of species. New empirical research based on multiple pairs of independent lineages along speciation continuum are therefore of particularly topical importance (Wolf and Ellegren 2017).

To understand the processes behind the formation of 'differentiation islands', we summarize four models according to the divergence models proposed by Han *et al* (2017) and Irwin *et al* (2018). In this context, two measures of divergence, the relative (F_{ST}) and the absolute (D_{XY}), are instrumental in distinguishing between evolutionary scenarios (Charlesworth 1998). Genomic regions with high F_{ST} may occur in the absence of local gene flow due to ecological or non-ecological isolating barriers, or due to locally reduced *Ne* in regions of low recombination (Noor and Bennett 2009; Turner and Hahn 2010; Renaut, et al. 2014; Campagna, et al. 2015; Gagnaire, et al. 2018; Henderson and Brelsford 2020). Disentangling the causal processes using only F_{ST} is impractical, as this is directly influenced by the level of within-groups genetic diversity (e.g. Cruickshank and Hahn, 2014). D_{XY} is a complementary measure, which has the advantage of being independent of the levels of diversity within the groups. However, D_{XY} is sensitive to ancestral variation, with higher values in regions of restricted gene flow and decreased in regions under background selection or selective sweeps (Han, et al. 2017; Irwin, et al. 2018). Scanning the whole genome of pairs of species using F_{ST} and D_{XY}, four models can be hypothesized regarding the extent of local gene flow to identify the main drivers of genomic landscapes of differentiation. The first model (hereafter model a) is 'divergence with ongoing gene flow' where selection at loci which contribute to reproductive isolation restricts gene exchange between populations, locally elevating genomic differentiation (both for F_{ST} and D_{XY}) and reducing genetic diversity. The second model (b) is 'allopatric selection' where natural selection acts on distinct regions of the genome after a species split into two populations, leading to lower nucleotide diversity (π) and higher F_{ST}. As D_{XY} is sensitive to ancestral polymorphism, its values are expected to remain relatively stable under this model. The next model (c) is 'recurrent selection' where background selection or selective sweeps at certain genomic regions reduce genetic diversity in both the common ancestor and the two daughter populations, leading to lower D_{XY} and π, but higher F_{ST}. The last model (d) is 'balancing selection' where ancestral polymorphisms are maintained at selected sites, resulting in increased D_{XY} and low F_{ST} between species. When employed in conjunction, the use of genomic estimates of F_{ST}, D_{XY} and π can thus enhance our understanding of the mechanisms at play that shaped the genomic differentiation landscape.

Populus is represented by perennial woody plants, dioecious, and widely distributed across the Northern Hemisphere (Stettler, et al. 1996). *Populus* genus comprises six sections containing 29 species, among which ten species from *Populus* section *Populus* (Stettler, et al. 1996; Jansson, et al. 2010). The genus *Populus* is well studied species in evolutionary biology not only due to their valuable economic and ecological importance, but also because of their small genome sizes (<500Mb), diploidy through the genus (2*n* = 38), wind pollination, extensive gene flow among species, sexual and vegetative reproductive strategies (Rajora and Dancik 1992; Martinsen, et al. 2001; Suarez-Gonzalez, et al. 2016). Among all woody perennial angiosperm species, the first genome sequenced and published was a *Populus* species (*P. trichocarpa*; Tuskan, et al. 2006). In addition to *Populus trichocarpa*, another well-annotated genome assembly is available (*P. tremula*; Schiffthaler, et al. 2019). In this study, we focused on white poplars and aspens from the section *Populus* which are widely distributed in Eurasia and North America (Supplementary material, Fig. S1 and Table S1).

The divergence time among species from *Populus* section *Populus* varies from 1.3 to 4.8 million years ago, as previously reported based on a species tree estimation (Shang, et al. 2020), therefore representing different stages along the speciation continuum. This taxon therefore provides an excellent system to investigate the genomic architecture of speciation. Here, we use whole genome resequencing data from eight *Populus* species (Supplementary material, Fig. S1 and Table S1) to address the following questions: (1) What is the demographic history of this species complex? (2) Are the genomic landscapes of differentiation across species pairs consistent with expectations regarding the various stages along the speciation continuum? (3) Are differentiation patterns across the genomic landscape repeatable among independent lineages? (4) What are the main evolutionary processes driving these heterogeneous landscapes of diversity and differentiation along the speciation continuum?

Results and Discussions

Strong interspecific structure despite interspecific introgression

As a first step, we estimated the relatedness between individuals using the KING toolset (available at http://people.virginia.edu/~wc9c/KING/) to trace potential clone mates produced by vegetative reproduction. This analysis identified 13 duplicated genotypes out of a total of 32 individuals from the Korean population of *P. davidiana*. In addition, all individuals of *P. qiongdaoensis* have been deemed as clone mates (supplementary material, Fig. S2). Therefore, these two populations had been deleted for subsequent analyses. After filtering and quality control (see Methods), the depth of coverage was relatively homogeneous (supplementary material, Fig. S3) and varied from 21× to 32×, resulting in a dataset of 30,539,136 quality SNPs.

We then explored population genetic structure across all seven *Populus* species. Genetic clustering as suggested by the first two principal components of the PCA (Fig. 1a) are consistent with expected genetic divisions, identifying the species previously described as the most divergent (Shang, et al. 2020) such as *P. adenopoda* or *P. grandidentata*, but not the most recently diverged aspens (*P. davidiana* and *P. rotundifolia*). More specifically, the first PC explains 21.9% of the total inertia and separates most species, with a particularly strong separation *P. adenopoda* on one side, and the other *Populus* species on the other. The second PC explains 19.7% of the variance and contributes to isolate several species, especially *P. grandidentata* from the rest of the sampling (Fig. 1a). To have a clearer picture of the structure of recently diverged species, a PCA analysis with only recently diverged aspens was performed (supplementary material, Fig. S4). The first principal component (26.2% variance explained) separated *P. tremuloides* from the other three species; the second principal component (21.7% variance explained) separated *P. tremula* from *P. davidiana* and *P. rotundifolia*.

Neighbor-joining (Fig. 1b) and Admixture (Alexander and Lange 2011) analyses (Fig. 1c) based on all SNPs identified seven genetic groups, consistent with previously identified species boundaries based on the PCA analysis (Fig. 1a). Additionally, Admixture also indicated a potential introgression between the subtropical species *P. adenopoda* and two recently diverged species, *P. davidiana* and *P. rotundifolia* (Fig. 1c and supplementary material, Fig. S5). The IBD analysis (Fig. 1d) also identified 7 reliable clusters, corresponding to the same species boundaries as before, but also identified some shared haplotypes among aspen species *P. davidiana*, *P. rotundifolia* and *P. tremula*, suggesting recent introgression among these species. Identity-by-descent (IBD) analyses (Fig. 1d) also provide support for extensive introgression between species with overlapping distribution, such as *P. alba* and *P. tremula*, or *P. grandidentata* and *P. tremuloides*. These results are consistent with a scenario of divergence with ongoing gene flow, either due to isolation-with-migration or secondary contact, maintained even after substantial divergence times (d_a: 0.023 for *P. alba* - *P. tremula*; d_a: 0.025 for *P. tremuloides* - *P. grandidentata*). For these two species pairs, these net divergence values are indeed larger than the upper boundary for the 'grey zone of speciation' reported by Roux, et al. (2016) for animals (d_a from 0.005 to 0.02), which suggests that plants, especially open-pollinated, might have a shifted or larger 'grey zones' than animals. Future investigations on the plant grey zone based on a large number of taxa are needed to get more general insights about plant diversification.

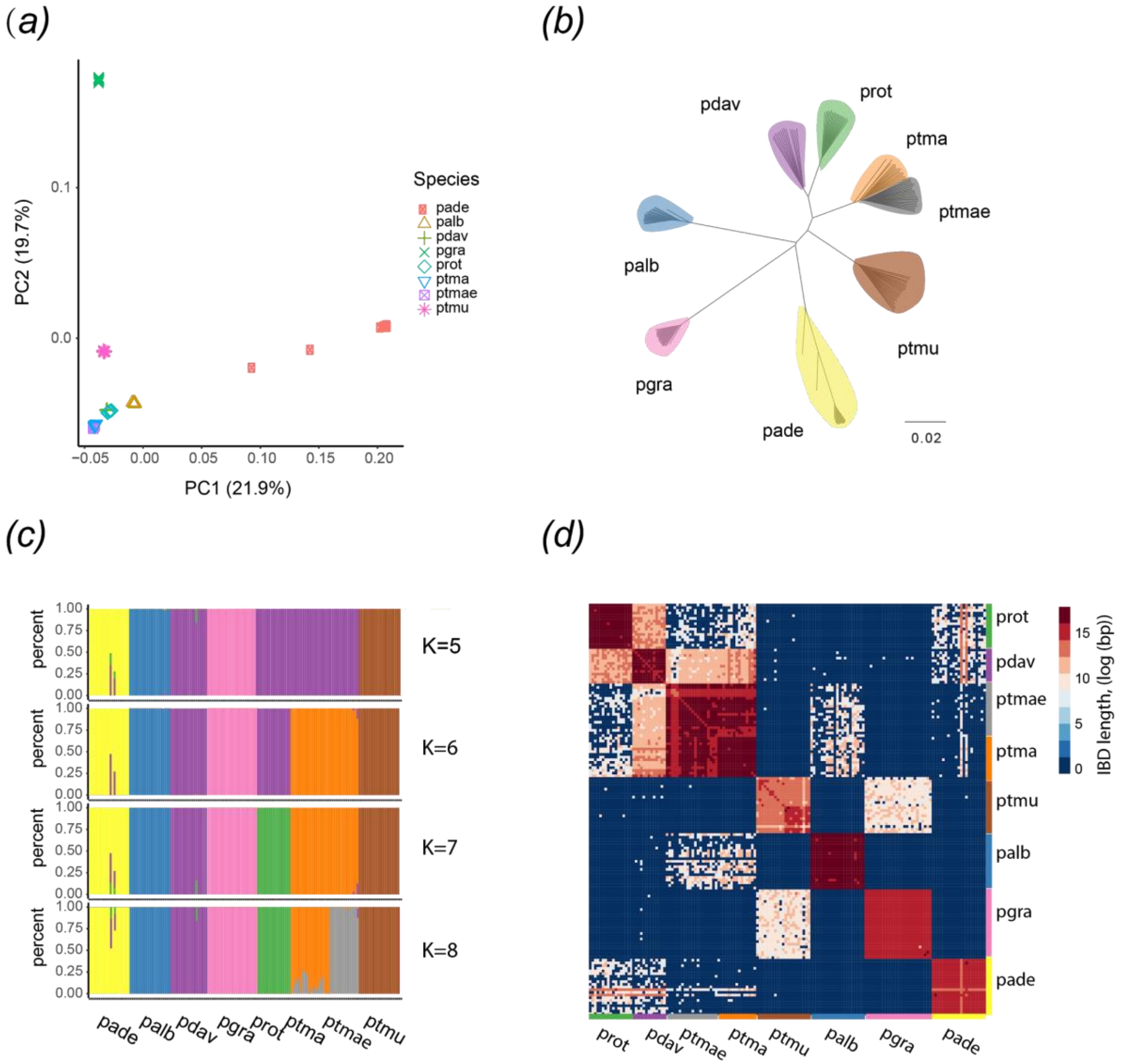


Fig. 1. Genetic structure within *Populus* (poplar and aspen) accessions investigated here. ***(a)*** Principal Component Analysis (PCA) based on all SNPs for seven *Populus* species. Colored symbols represent different species according to legend. The first PC explains 21.9% of the total variance and the second PC explains 19.7% of the variance. ***(b)*** Unrooted neighbour-joining tree of all SNPs. ***(c)*** Population structure analysis based on all SNPs. Estimated membership of each individual's genome for K = 5 to K = 8 as estimated by Admixture (best K = 7). ***(d)*** Identity by descent (IBD) analysis for seven *Populus* species.

Heatmap colours represent the shared haplotype length between species. Species abbreviations: pade, *P. adenopoda*; palb, *P. alba*; prot, *P. rotundifolia*; pgra, *P. grandiantata*; ptmu, *P. tremuloides*; ptma, *P. tremula* collected from China; ptmae, *P. tremula* collected from Europe; pdav, *P. davidiana* collected from China.

Between species variability in demographic trajectories

We then used TreeMix (Pickrell and Pritchard 2012) to recover the phylogenetic relationship among species and infer the admixture events in *Populus* (Fig. 2a). The tree topology was consistent with phylogenetic relationships found in the previous study (Shang, et al. 2020). A drift-only model of divergence (i.e., without migration edges) already explained 95.8% of the total variance. Adding one single migration allowed us to account for 98.9% of the total variance (supplementary material, Fig. S6). This event was inferred between *P. grandidentata* to *P. tremuloides* and is consistent with previous reports of extensive hybridization and introgression between these two species (Deacon, et al. 2019). Adding an additional migration event allowed us to explain 99.6% of the total variance. This second migration edge was inferred from *P. adenopoda* to *P. rotundifolia*. By adding more migration events, the variance explained increases by less than 0.1%, which was considered as too marginal (supplementary material, Fig. S6). Therefore, we considered the bifurcating tree with two migration events as the best scenario explaining the historical relationships among these *Populus* species.

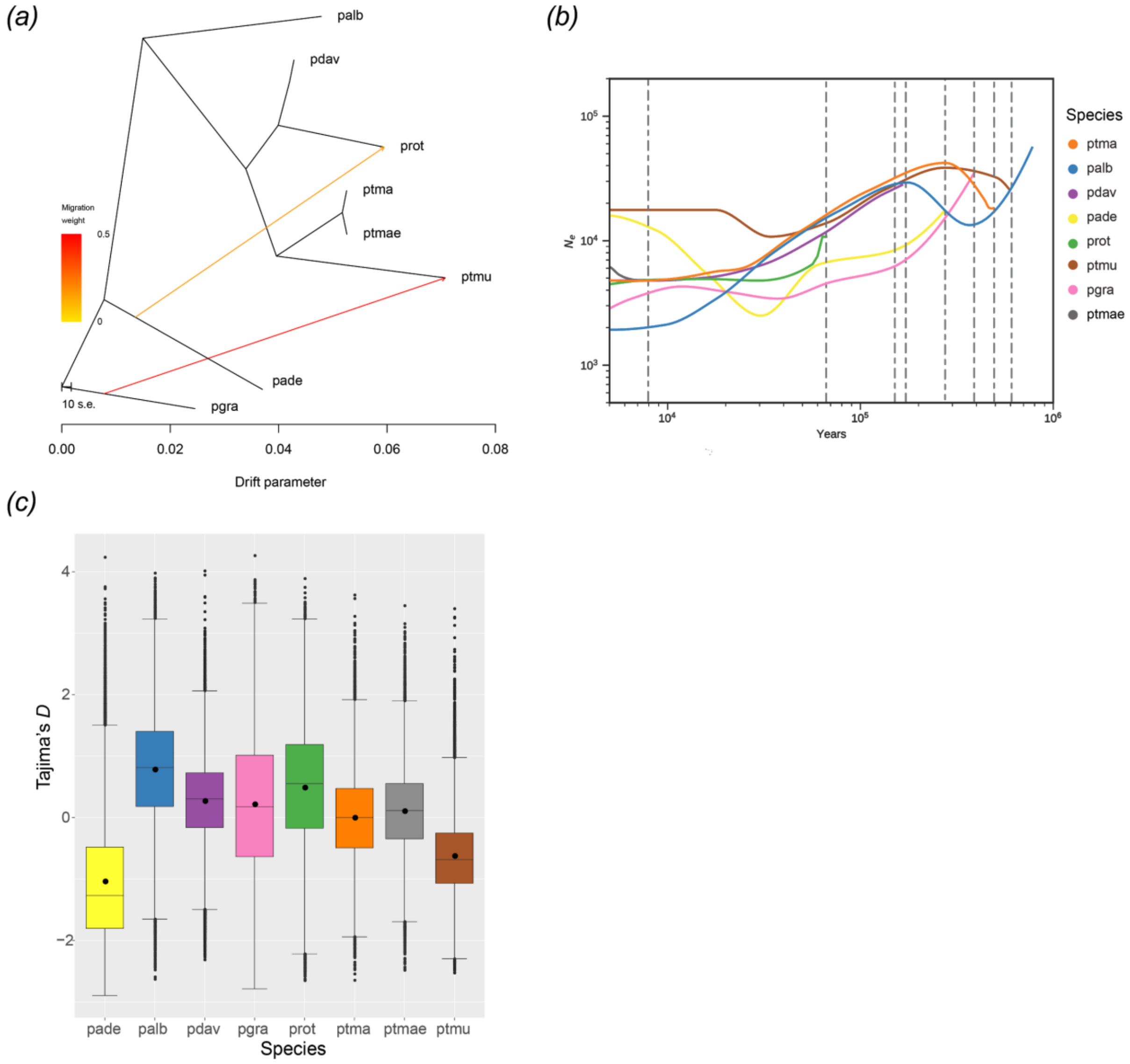


Fig. 2. Demographic history of *Populus* species. ***(a)*** The maximum likelihood tree inferred by TreeMix under a strictly bifurcating model with two migration events. ***(b)*** Changes in effective population size through time inferred by SMC++. ***(c)*** The distribution of Tajima's *D* values calculated over all 10-kb windows.

To infer historical changes in effective population sizes for the investigated *Populus* species we used SMC++ (Terhorst, et al. 2017), a method which only requires unphased sequence data. These inferences (Fig. 2b) support that all species have experienced population size reductions at least over the last 100,000 years, except for the subtropical species *P. adenopoda* and the North

American species *P. tremuloides*. These latter two species have undergone population size expansions from about 30,000 years ago, which seem consistent with their extensive present-day distribution areas in South China and North America, respectively (Eckenwalder 1996). In line with our findings, Fan *et al* (2018) considered the subtropical species *P. adenopoda* may have experienced population contractions during glacial periods, followed by interglacial expansions. Negative mean Tajima's *D* over all non-overlapping 10kb sliding windows spanning the whole genome are consistent with these recent expansion (*P. adenopoda*: -0.97, *P. tremuloides*: -0.61; Fig. 2c). For these two species, the folded-SFS also showed a strong excess of rare variants (*P. adenopoda*: 2.4%, *P. tremuloides*: 5.2%; supplementary material, Fig. S7), which explains these negative Tajima's *D* values (Fig. 2c). Further evidence for different demographic histories among *Populus* species was provided by estimates of genome-wide π and population-scaled recombination rate (ρ) based on 10kb windows. Both π and ρ were found to vary greatly between species (Fig. 3c-d). Since both π and ρ are dependent on *Ne,* the highest π and ρ observed for *P. tremuloides* (Fig. 3c-d) is consistent with larger effective population sizes previously inferred for this species (Fig. 2b). Interestingly, for *P. grandidentata* we found relatively low π but high ρ. In addition, a significantly negative correlation between gene density and π was also evident in all *Populus* species (supplementary material, Fig. S8). The negative correlation between π and ρ in *P. grandidentata* suggests that the effects of linked selection are not confined to low recombination regions (Slotte 2014; Wang, et al. 2016a), but to regions with high gene density. Alternatively, DNA mismatch repair also restricts recombination events, which may also contribute to a negative correlation between π and ρ (Modrich and Lahue 1996).

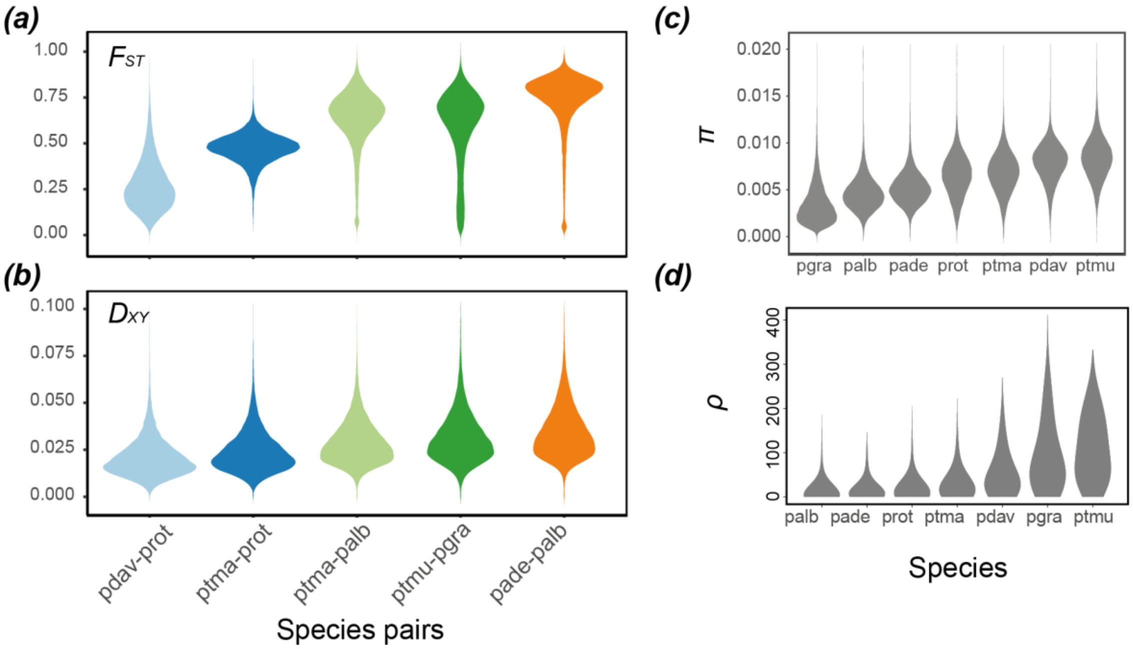


Fig. 3. Summary violin plots of F_{ST}, D_{XY}, π and ρ calculated across 10kb windows. ***(a-b)*** Violin plots of F_{ST} and D_{XY} for five species pairs investigated in this study. ***(c-d)*** Violin plots of π and ρ for seven *Populus* species.

Conserved genomic landscapes across the continuum of divergence

We calculated genome-wide patterns of divergence, nucleotide diversity, gene density and recombination for all *Populus* species across non-overlapping 10kb windows spanning the whole genome. We then investigated the correlations between within-species nucleotide diversity and relative divergence levels, or absolute divergence and recombination rate. Under the influence of linked selection, a lower genetic diversity at high gene density or across low recombination regions is expected. Consistent with this expectation, significantly positive correlations between π and ρ or D_{XY} (Fig. 4b, c, e) and negative correlations between F_{ST} and π (Fig. 4a) were observed across the continuum of divergence represented by the 21 species pairs. As background selection is expected to shape the genomic landscape of differentiation over time, we then used the level of genetic distance between each species pair (d_a) as a proxy for the divergence time and estimated how the correlations between genome-wide diversity or divergence and recombination change across the speciation continuum. We found that the negative relationships between F_{ST} and π or ρ became stronger as d_a increases, which is consistent with the expectation under background selection (Fig. 4a, d). Similar investigations for *π* and D_{XY} showed significantly positive correlations while the trend became weaker as divergence increases (Fig. 4b). As background selection will continue to operate similarly for two species after their split (i.e., the sister species share the same regions with lower *π*), the correlation between D_{XY} and *π* is indeed expected to still be recovered for a long time (Burri 2017a). In other words, the trend is not consistent with the general hypothesis that correlations should still be highly correlated as divergence increases. We also recovered a strong positive correlation between *π* and *ρ* (Fig. 4c), and a similar trend was found as for the investigation of *π* and D_{XY}. Correlations between absolute divergence and the population-scaled recombination rates (Fig. 4e) were significantly positive across the entire speciation continuum, and we do observe that these correlations tend to become stronger as divergence increases across the speciation continuum. The observed patterns (Fig. 4b, c) differ from expectations under a scenario with background selection as the sole factor shaping the heterogeneous landscape of differentiation, indicating that there should be additional evolutionary factors that contribute to it (Burri 2017a).

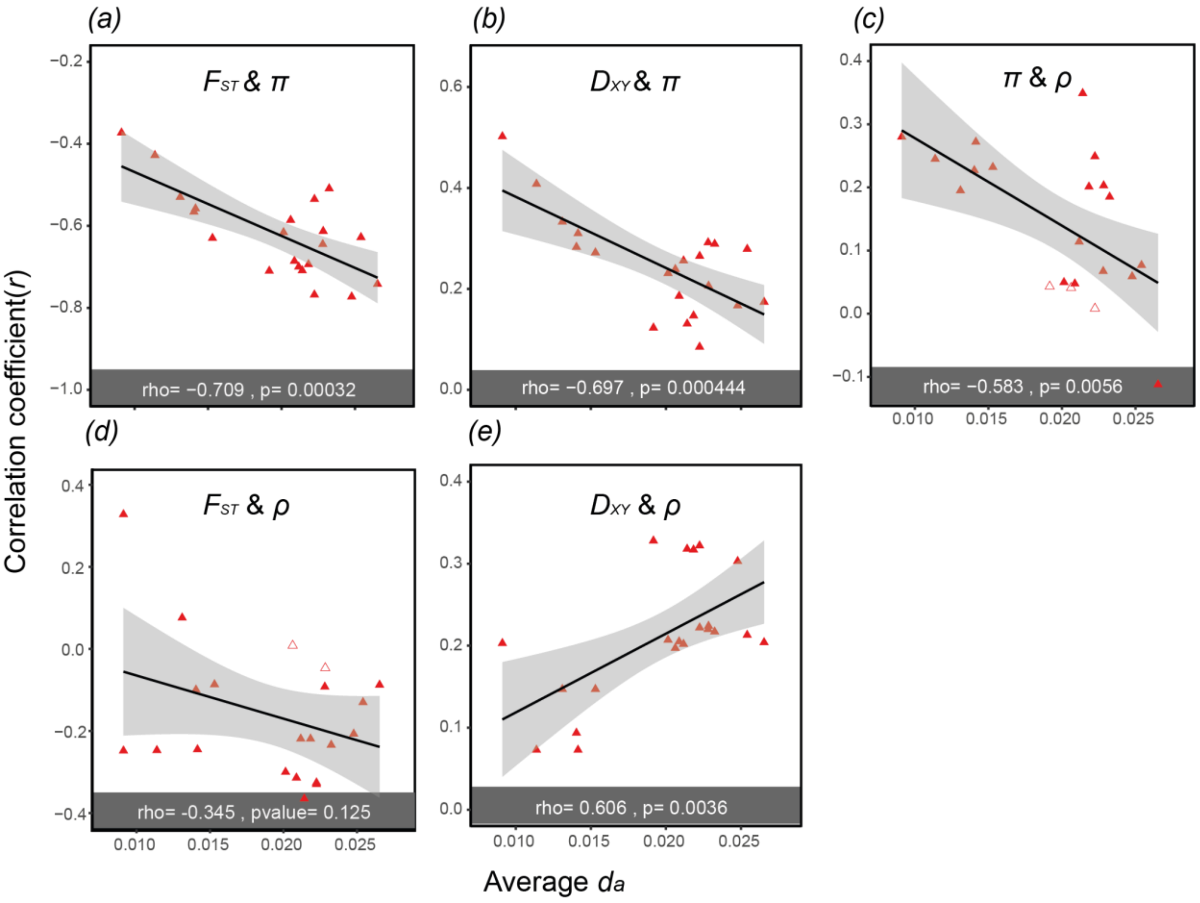


Fig. 4. Correlations between variables for all species comparisons (red filled and unfilled triangles) plotted against the average d_a, used here as a measure of divergence time. The red filled triangles indicate the correlation coefficients are significant ($p < 0.01$) in each panel. The upper panels show how the relationships between average π and ***(a)*** F_{ST}, or ***(b)*** D_{XY} vary for pairs of species with increasing divergence time. ***(c)*** The relationships between average π and ρ for all species pairs investigated. The lower panels ***(d)*** and ***(e)*** show the relationships between ρ and F_{ST} or D_{XY}, respectively.

In-depth pairwise investigations across the speciation continuum

To gain a deeper insight into the factors that have shaped patterns of genome-wide variation, we characterized the genome-wide divergence for five species pairs as representatives of the different stages across the speciation continuum (among all 21 possible species pairs). Overall, we found support for significant interspecific differences for all summary statistics (ANOVA, p<2.2 e^{-16}; Fig. 3). Mean F_{ST} varied from 0.27 between *P. davidiana* - *P. rotundifolia* to 0.73 between *P. adenopoda* - *P. alba* (Fig. 3a) and D_{XY} ranged from 0.021 (*P. davidiana* - *P. rotundifolia*) to 0.035

(*P. adenopoda* - *P. alba*) (Fig. 3b). The average π varied from 0.0035 in *P. grandidentata* to 0.0084 in *P. tremuloides* (Fig. 3c). We inferred lower population-scaled recombination rates in *P. alba* and *P. adenopoda* than in other species, which is consistent with the slower linkage disequilibrium (LD) decays and lower genetic diversity in these two species (Fig. 3d and supplementary material, Fig. S9).

The degree of correlation of both the relative and absolute divergence landscapes between pairs of species supports a highly conserved pattern among the five investigated species pairs (Fig. 5a-b). Whereas highly significant for all pairs of species, the correlation coefficients for the pairwise comparisons of the nucleotide diversity landscapes also vary substantially (Fig. 5c), from 0.16 (*P. tremula* versus *P. grandidentata*) to 0.52 (*P. rotundifolia* versus *P. davidiana*). This degree of variation follows the phylogenetic distance, with strongest correlation coefficients for the phylogenetically closest pair of species: *P. rotundifolia* and *P. davidiana* (Fig. 5c). Pairwise comparisons of the local recombination rates inferred independently for all species also revealed only positive correlations (Fig. 5d), suggesting at least a moderate degree of conservation of the recombination landscape between species. The highest positive correlation coefficient of ρ was again observed when comparing the recombination landscapes of the two closest related species, *P. davidiana* and *P. rotundifolia* (0.47), while the lowest correlation was observed for *P. davidiana* and *P. grandidentata* (0.08). Most of the lower values (correlation coefficients < 0.2) were found when comparing *P. grandidentata* with the other species, suggesting an increased uniqueness in the recombination landscape in this species. Overall, landscapes of genetic diversity, divergence and recombination rate remain relatively stable across different species or species pairs (Fig. 5), which implies relatively conserved genomic features across all species. This phenomenon has also been observed in some other plant and animal models (Nosil and Feder 2012; Renaut, et al. 2014; Burri, et al. 2015; Wang, et al. 2020). This main result is consistent with the important role of linked selection, and therefore the recombination landscape, in shaping genome-wide genetic diversity levels.

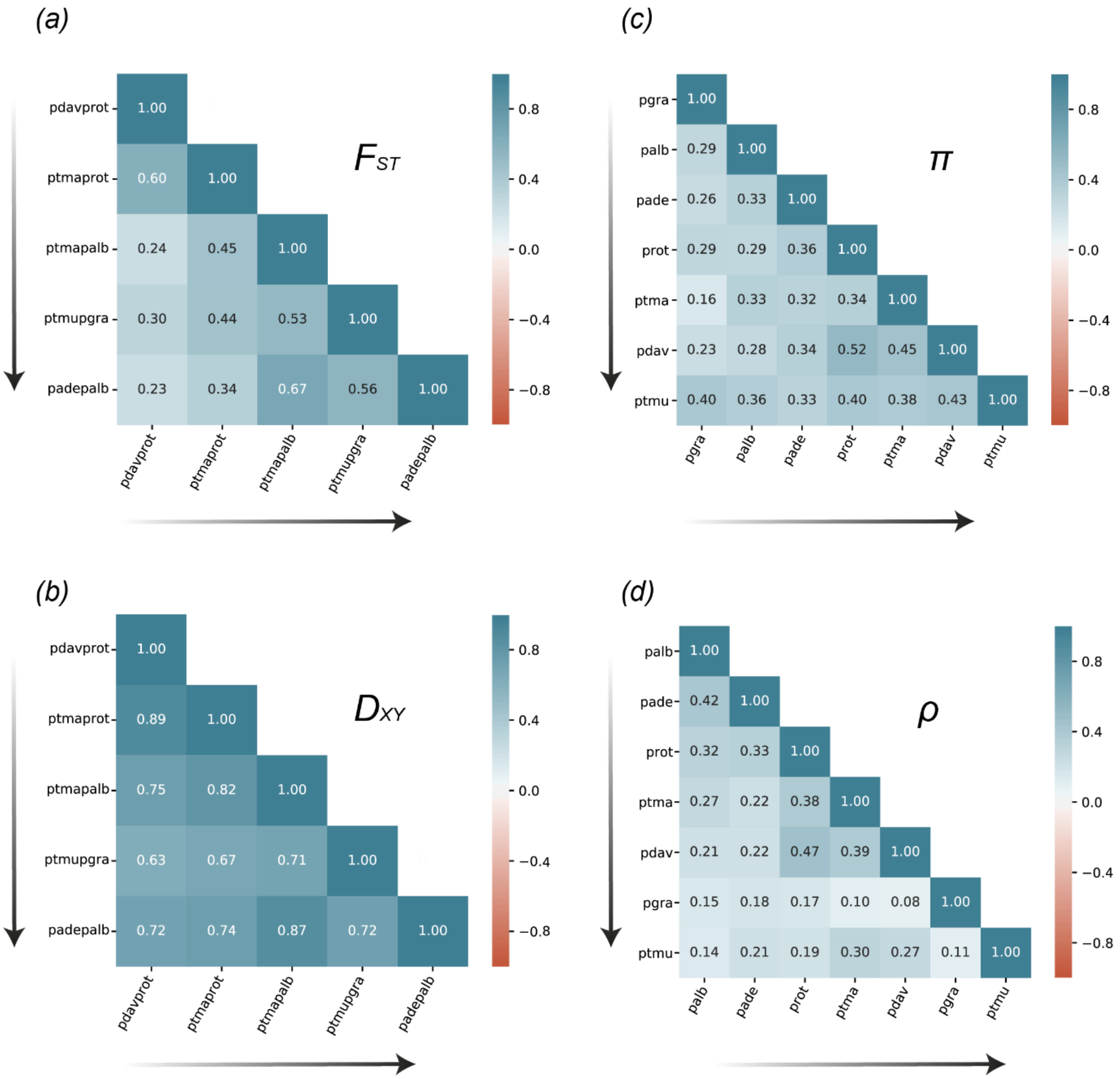


Fig. 5. Correlations analyses of within-species diversity or among-species divergence landscapes ***(a-b)*** Correlation coefficients of F_{ST} or D_{XY} between species pairs. The species pairs are ordered across the speciation continuum (arrows). ***(c-d)*** Correlation coefficients of π or ρ between species. The order of the species is based on the values of π or ρ of each species (from low to high, arrows; see also Fig. 3). All the values are significantly positively correlated (p < 0.001), respectively.

The impact of positive and balancing selection on genomic landscapes of differentiation

To identify selective sweeps for each species, we first used the integrated Selection of Allele Favored by Evolution method (iSAFE; Akbari, et al. 2018) to identify sites that exhibit a likely signature of positive selection (sites with iSAFE score > 0.1). Then, the 10 kb windows significantly enriched in these sites were considered as genomic regions with evidence for positive selection. The numbers of such windows across the seven species ranged from 102 (*P. grandidentata*) to 592 (*P. tremula*) (supplementary material, Fig. S10 and Table S2). For the five investigated species pairs, the windows exhibiting signatures of positive selection had significantly higher F_{ST} (Fig. 6a) or D_{XY} (Fig. 6b) values compared to a random set of windows sampled across the whole genome.

We also used BetaScan (Siewert and Voight 2017) to search for signatures of balancing selection regions across the genome of each species. As above, we investigated non-overlapping 10kb windows to identify those enriched for sites identified as under balancing selection. We observed from 358 genomic regions potentially evolving under balancing selection in *P. adenopoda* and to 2,510 regions in *P. grandidentata* (supplementary material, Fig. S11 and Table S2). We then focused on regions identified in at least two species as good candidates for genomic regions evolving under balancing selection. As expected, regions under balancing selection regions exhibit lower median F_{ST} (Fig. 6a) and higher median D_{XY} (Fig. 6b) values than a set of randomly sampled genomic regions, even if these results are significant for only three species pairs for F_{ST}: *P. davidiana* - *P. rotundifolia, P. tremula* - *P. alba* and *P. tremuloides* - *P. grandidentata*.

To test more explicitly the role of interspecific gene flow on genomic landscapes of divergence, we evaluated genome-wide introgression (f_d; Martin, et al. 2015) in two species pairs known to frequently hybridize in nature (Barnes 1959; Lexer, et al. 2005; Lexer, et al. 2007; Christe, et al. 2017). For both species pairs, we observed $f_d > 0$ for a majority of genomic windows (68% for *P. grandidentata* - *P. tremuloides*; 66% for *P. tremula* - *P. alba*), which confirms frequent introgression in these systems. Significantly negative correlations were observed across the genome for F_{ST} and f_d in both *P. tremula* - *P. alba* and *P. tremuloides* - *P. grandidentata* (Fig. 6c).

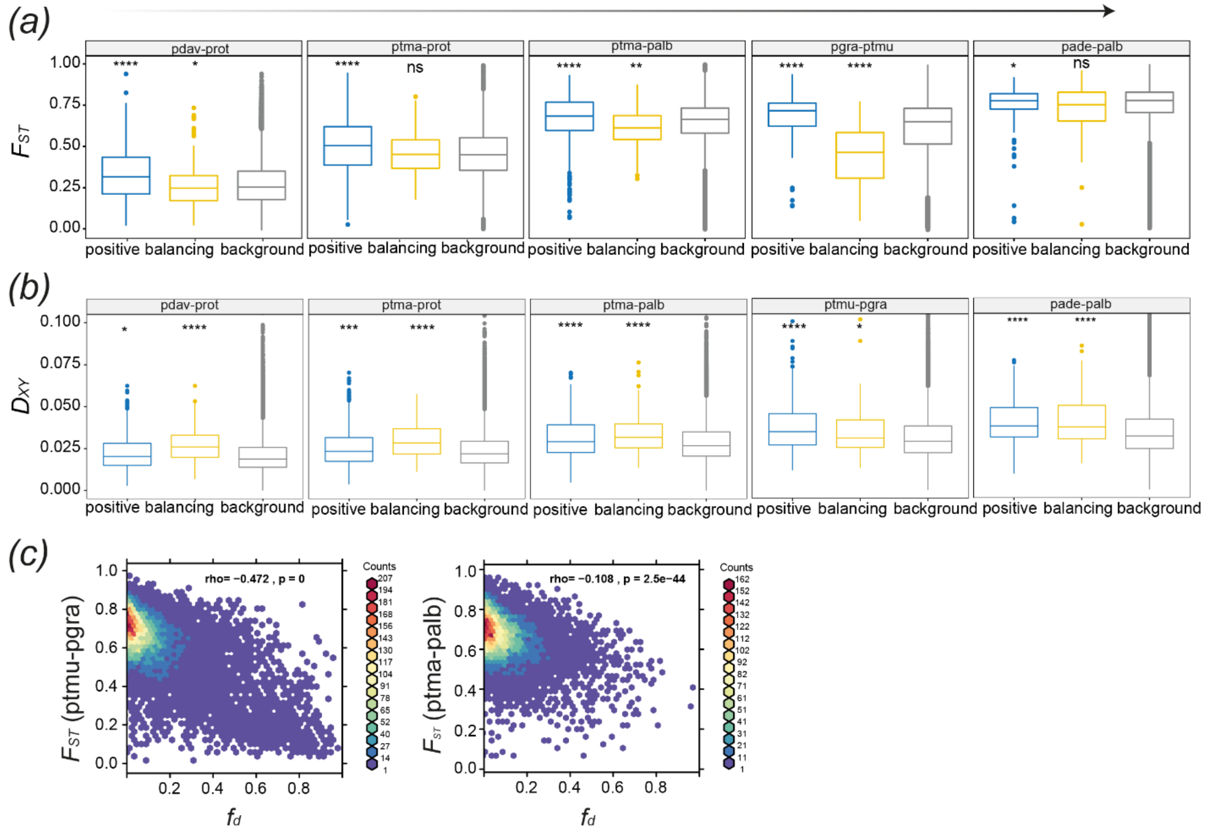


Fig. 6. Comparison of ***(a)*** F_{ST} or ***(b)*** D_{XY} at regions under positive selection (blue boxes), balancing selection (yellow boxes) and genomic background (grey boxes). The box plot represents the distribution within unique selection windows (positive selection) in each species, or shared selected windows (balancing selection) in each species pair. The arrow on the top indicates the divergence increase across the speciation continuum. ***(c)*** Correlation analysis between F_{ST} and f_d reveals introgression between species.

Genomic divergence across a speciation continuum (models of genomic patterns of differentiation)

Based on F_{ST}, D_{XY}, and π, we summarized four models to help explain the genomic divergence patterns observed (supplementary material, Fig. S12) following the strategy developed by Han et al. (2017) and Irwin et al. (2018). The major differences of these models are the role and extent of gene flow, and the type of selection acting to shape genome-wide genetic differentiation. Our results show a heterogeneous distribution of the four models along the genome for all five species pairs (supplementary material, Fig. S13-S17 and Table S3). A great number of regions (74.3%-78.7%) fit a model of "allopatric selection" (model b in supplementary material, Fig. S12 and

Table S3), i.e., regions exhibiting an elevated F_{ST} (tail of the distribution) but a moderate D_{XY} (in the middle of the distribution). Such a signature is expected to be consistent with recent footprints of positive or background selection on genomic differentiation. Genomic regions fitting the model of 'balancing selection' (model d in supplementary material, Fig. S12 and Table S3) are the second most frequent for all investigated species pairs. This model is characterized by an elevated D_{XY} but a low F_{ST} implying the action of balancing selection in shaping the heterogeneous landscape of divergence. In addition, we found support for reproductive barriers (model a in supplementary material, Fig. S12 and Table S3) in all five species pairs, but only for a very limited number of windows, suggesting that genomic heterogeneity in the levels of gene flow due to the species barriers do not play a major role in shaping genomic differentiation landscapes as selection occurring independently in each species does. Interestingly, this result holds true for all species pairs we investigated, *i.e.*, regardless of the stage along the *Populus* speciation continuum. Consistent with our findings, the newest finding in three hummingbird species pairs also confirmed the role of linked selection, recombination rate or gene density in shaping the landscape of genomic divergence (Henderson and Brelsford 2020). Overall, our study estimated evolutionary factors that contribute to the genomic landscape of differentiation across multiple closely related *Populus* species pairs with different levels of gene flow. In the future, more studies on multiple species pairs across the speciation continuum are needed to reveal the drivers of speciation and genomic differentiation.

Conclusions

In this study, we reconstructed the demographic history and investigated the evolution of the genomic landscape of diversity and divergence across a speciation continuum, using eight closely related species of *Populus* section *Populus* as models. By investigating evolution of diversity and differentiation landscapes across this speciation continuum, we provided , to our knowledge, one of the most ambitious case studies in terms of the number of species pairs analyzed (see also Stankowski, et al. 2019). Our analyses are consistent with a prominent contribution of linked selection in shaping these genomic landscapes. Both correlation analysis and model-based inference indeed support this major role of linked selection, recombination rate and gene density in shaping the empirical patterns of genomic differentiation. The study also confirmed the importance of gene flow in this system, through extensive introgression among species with parapatric distributions, despite a high level of divergence among the most divergent hybridizing species (d_a = 0.025). Overall, this work provides a prime example of the strategic importance of investigating these genomic patterns on a large number of species to better access the temporal evolution of the genomic landscapes of diversity and differentiation.

Materials and method

Sampling, sequencing and reads processing

Two hundred and one samples were collected from eight species of *Populus* section *Populus* in Eurasia and North America (supplemental material, Fig. S1 and Table S1). The leaves were dried in silica gel first and were then used for genomic DNA extraction with Plant DNeasy Mini Kit (Qiagen, Germany). To increase the purity of total DNA, we used the NucleoSpin gDNA Clean-up kit (Macherey-Nagel, Germany). Whole genome resequencing was performed with 2 x 150bp paired-end sequencing technology on Illumina HiSeq 3000 sequencer at the Institute of Genetics, University of Bern, Switzerland.

All raw sequencing reads were mapped to *P. tremula* 2.0 reference genome (Schiffthaler, et al. 2019) using BWA-MEM, as implemented in bwa v0.7.10 (Li 2013). Samtools v1.3.1 was used to ignore alignments with mapping quality below 20 (Li, et al. 2019). Read-group information including library, lane, sample identity and duplicates were recorded using Picard v2.5 (http://broadinstitute.github.io/picard/). Sequencing reads around insertions and deletions (i.e., indels) were realigned using RealignerTargetCreator and IndelRealigner in the Genome Analysis Toolkit (GATK v3.6) (DePristo, et al. 2011). We used GATK HaplotypeCaller and then GenotypeGVCFs for the individual SNP calling and for the joint genotyping, respectively, using default parameters among all samples. Finally, we performed several filtering steps using GATK to retain only high-quality SNPs: (1) ‘QD’ < 2.0; (2) 'FS > 60.0'; (3) 'MQ < 40.0'; (4) 'ReadPosRankSum < -8.0'; (5) 'SOR > 4.0'; (6) 'MQRankSum < -12.5'. Besides, we also excluded loci with missing data more than 30% and discarded two individuals which had very low depth of coverage (< 10), as calculated using VCFtools v0.1.15 (http://vcftools.sourceforge.net/man_latest.html). The scripts for snp calling are available at https://github.com/Huiying123/Populus_speciation/tree/main/population_snp_calling.

Family relatedness and population structure analysis

To avoid the influence of clone mates to population genomics estimates, we estimated kinship coefficients using the KING toolset for family relationship inference based on pairwise comparisons of SNP data (http://people.virginia.edu/~wc9c/KING/manual.html). The software classifies pairwise relationships into four categories according to the estimated kinship coefficient: a negative kinship coefficient estimation indicates the lack of a close relationship.Estimated kinship coefficients higher than >0.354 correspond to duplicates, while coefficients ranging from [0.177, 0.354], [0.0884, 0.177] and [0.0442, 0.0884] correspond to 1st-degree, 2nd-degree, and 3rd-degree relationships, respectively.

After discarding individuals with low depth and high inbreeding coefficient (F > 0.9, *P. qiongdaoensis*) as well as duplicates identified with the KING toolset, we then used VCFtools v0.1.15 (http://vcftools.sourceforge.net/man_latest.html) to calculate the mean depth of coverage and heterozygosity for each individual. Based on the results of these analyses, we concluded that all *P. qiongdaoensis* individuals are probably clone mates. As a consequence, only seven species were used for the subsequent analyses.

We used PLINK (Purcell, et al. 2007) to generate a variance-standardized relationship matrix for principal components analysis (PCA) and a distance matrix to build a neighbor joining tree (NJ-tree). The NJ tree was constructed using PHYLIP v.3.696 (https://evolution.genetics .washington.edu/phylip.html). Both PCA and NJ-tree analyses were performed based on the full set of SNPs. In addition, we used ADMIXTURE v1.3 for the maximum-likelihood estimation of individual ancestries (Alexander and Lange 2011). This analysis was run for *K* from 2 to 10, and the estimated parameter standard errors were generated using 200 bootstrap replicates. We also performed an IBD blocks analysis using BEAGLE v5.1 (Browning and Browning 2013) to detect identity-by-descent segments between pairs of species. The parameters we used are: window=100,000; overlap=10,000; ibdtrim=100; ibdlod=10.

Demographic trajectory reconstruction

To reconstruct the demographic history of *Populus* species, we first inferred the history of species splits and mixture based on genome wide allele frequency data using TreeMix v1.13 (Pickrell and

Pritchard 2012). We removed the sites with missing data and performed linkage pruning. We then ran TreeMix implementing a default bootstrap and a block size of 500 SNPs (-k=500). The best migration event was evaluated according to the greatest increase of total variation explained. The plotting R functions of the Treemix suite were then used to visualize the results. In addition, we used SMC++ to estimate historical changes in effective population size (Terhorst, et al. 2017). The split time for species pairs was estimated based on the joint frequency spectrum for both species. Due to widespread vegetative reproduction in some poplar species, we adjusted the generation time to 20 years rather than the generally thought 15 years (Wang, et al. 2016); a mutation rate of 2.5×10^{-9} per site per year was also assumed in order to scale results to real time (Macaya-Sanz, et al. 2012).

Nucleotide diversity and divergence estimates

Nucleotide diversity, as well as relative and absolute divergence estimates were calculated based on genotype likelihoods. We used ANGSD v0.93 (http://www.popgen.dk/angsd/index.php/ANGSD) to estimate statistical parameters from the BAM files for all *Populus* species. First, we used '*dosal 1*' to calculate site allele frequency likelihood and then used '*readSFS*' to estimate folded site frequency spectra (SFS). Genome-wide diversity and Tajima's *D* were calculated with the parameter '*-doThetas 1*' in ANGSD based on the folded SFS of each species. We selected two population genomic statistics to estimate divergence F_{ST} and D_{XY}. We estimated SFS for each population separately and then used it as a prior to generated 2D-SFS for each species pair. F_{ST} of each species pair were estimated with the parameters '*realSFS fst*' based on the 2D-SFS. Finally, we averaged the F_{ST} value of sites over 10kb windows. To estimate D_{XY}, we used ANGSD to calculate minor allele frequencies with the parameters '*-GL 1 -doMaf 1 -only_proper_pairs 1 -uniqueOnly 1 -remove_bads 1 -C 50 -minMapQ 30 -minQ 20 -minInd 4 -SNP_pval 1e-3 -skipTriallelic 1 -doMajorMinor 5*' and then computed D_{XY} as follows: $D_{XY} = A_1 * B_2 + A_2 * B_1$, with A and B being the allele frequencies of A and B, and 1 and 2 being the two populations. We averaged D_{XY} across 10kb windows.

To examine the relationships among diversity, differentiation and recombination landscapes, we estimated Pearson's correlation coefficient between pairs of these statistics. These tests were performed across genomic windows for the 21 possible *Populus* species pairs. Finally, we used d_a (D_{XY} – mean π) as a measure of divergence time.

Population-scale recombination rate and linkage disequilibrium

We estimated population scaled recombination rate (ρ) with FastEPRR (Gao, et al. 2016) for each species separately. To reduce the effect of population size on the estimation of recombination rate, we randomly selected 13 individuals for each species, corresponding to the number of individuals available for *Populus davidiana* (pdav). First, we filtered all missing and non-biallelic sites with VCFtools and then phased the data with the parameters "*impute=true nthreads=20 window=10,000 overlap=1,000 gprobs=false*" in Beagle v5.1 (Browning and Browning 2013). Finally, we ran FastEPRR v2.0 (Gao, et al. 2016) with a window size of 10kb. After getting the results, we estimated the correlation between recombination rate of one species to another. To evaluate LD decay, we used PLINK (Purcell, et al. 2007) to obtain LD statistics for each species. Parameters were set as follows: '*--maf 0.1 --*r^2 *gz --ld-window-kb 500 --ld-window 99999 --ld-window-*r^2 *0*'. LD decay was finally plotted in R.

Divergent regions of exceptional differentiation

We further investigated genomic differentiation landscapes across multiple species pairs along the speciation continuum and identified which evolutionary factors contribute to genomic differentiation. We reported genomic regions showing elevated or decreased values of F_{ST}, D_{XY} and π across 10kb windows. Windows falling above the top 5% or below the bottom 5% of F_{ST} and D_{XY} were considered. For these specific windows, we then classified them following the four models of divergence suggested by Irwin *et al*. 2018 and Han *et al* 2017. These four models differ in the role of gene flow (with or without), or the type of selection (selective sweep, background selection or balancing selection).

Genome-wide scan for regions under positive and balancing selection

We further tested the impact of positive and balancing selection on genomic patterns of differentiation across *Populus* species. We used integrated Selection of Allele Favored by Evolution (iSAFE) to detect signatures of selective sweeps using whole-genome sequence data (Akbari, et al. 2018). First, we phased the data with Beagle v4.1 (Browning and Browning 2013).

Second, we ran iSAFE using *P. trichocarpa* as outgroup. The output is a non-negative iSAFE score for each mutation. We measured the number of selected sites within 10kb windows for each species and selected the top 1% windows.

We also identified ancient balancing selection regions using BetaScan v1.0 (Siewert and Voight 2017), which detects the signature of an excess number of intermediate frequency polymorphisms near a balanced variant. We used 1kb windows to calculate β values for five *Populus* species pairs across the speciation continuum.

Once positive or balancing selection outlier regions were identified, we compared the values of F_{ST}, D_{XY}, π, recombination rate and f_d (see next section) at these regions with the rest of the genome to see if selection also shaped the genomic landscapes of differentiation.

Estimation of introgression

To detect introgression between species at genomic level, we computed *f*-statistics index (f_d) (Martin, et al. 2015) on non-overlapping 10kb windows spanning the genome. To perform this analysis, we used the modified version made available on github (https://github.com/simonhmartin/genomics_general) and computed these statistics for two pairs of species (*P. alba*-*P. tremula* and *P. tremuloides*-*P. grandidentata*) that hybridize frequently at their overlap regions respectively. The test uses three populations and an outgroup with the relationship (((P1, P2), P3), O) to measure an excess of shared variation between P1 and P3 ($D<0$) or P2 and P3 ($D>0$). We selected the subtropical species *P. qiongdaoensis* as an outgroup and used *P. tremuloides* and *P. alba* as the third population for the two species pairs, respectively.

Acknowledgements

We thank members of Jianquan Liu's lab for collecting samples and to Chris Cole for providing samples from *P. grandidentata*. Michael Barfuss and Elfi Grasserbauer provided assistance in the laboratory. Sequencing was performed at the Next Generation Sequencing Platform of the University of Berne and at the National Genomics Infrastructure (NGI) of Science for Life Laboratory, Stockholm. The Vienna Scientific Cluster (VSC) and the Swedish National Infrastructure for Computing (SNIC) at Uppsala Multidisciplinary Centre for Advanced

Computational Science (UPPMAX) provided access to computational resources. We are also grateful to members of the PopGen Vienna graduate school for helpful discussions. This manuscript is dedicated to the memory of our friend and colleague Prof. Christian Lexer.

Data accessibility

All raw read data was uploaded to NCBI and can be found under Bioproject ID PRJNAXXX. All the scripts used for the analysis are available on XXX.

Author contributions

Study conceived and designed by Christian Lexer and Huiying Shang. Laboratory work conducted by Huiying Shang. Population genomic data analysis by Huiying Shang and Martha Rendón-Anaya. Interpretation of the results was undertaken by Huiying Shang, Christian Lexer, Martha Rendón-Anaya, Pär K. Ingvarsson, Thibault Leroy, Jaqueline Hess, David Field, Claus Vogl and Ovidiu Paun. The manuscript was drafted by Huiying Shang, with help from Thibault Leroy and Ovidiu Paun.

Funding

This work was supported by a fellowship from the China Scholarship Council (CSC) to Huiying Shang, Swiss National Science Foundation (SNF) grant no.31003A_149306 to Christian Lexer, doctoral programme grant W1225-B20 to a faculty team including Christian Lexer, and the University of Vienna.

ORCID

Huiying Shang, https://orcid.org/0000-0002-1302-8008

Ovidiu Paun, https://orcid.org/0000-0002-8295-4937

Thibault Leroy, https://orcid.org/0000-0003-2259-9723

References

Akbari A, Vitti JJ, Iranmehr A, Bakhtiari M, Sabeti PC, Mirarab S, Bafna V. 2018. Identifying the favored mutation in a positive selective sweep. Nat Methods 15:279-282.

Alexander DH, Lange K. 2011. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics 12:246.

Barnes BV. 1959. Natural variation and clonal development of *Populus tremuloides* and *P. grandidentata* in northern lower Michigan.

Browning BL, Browning SR. 2013. Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics 194:459-471.

Burri R. 2017a. Dissecting differentiation landscapes: a linked selection's perspective. J Evol Biol 30:1501-1505.

Burri R. 2017b. Linked selection, demography and the evolution of correlated genomic landscapes in birds and beyond. Mol Ecol 26:3853-3856.

Burri R, Nater A, Kawakami T, Mugal CF, Olason PI, Smeds L, Suh A, Dutoit L, Bureš S, Garamszegi LZ, et al. 2015. Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula* flycatchers. Genome Res 25:1656-1665.

Campagna L, Gronau I, Silveira LF, Siepel A, Lovette IJ. 2015. Distinguishing noise from signal in patterns of genomic divergence in a highly polymorphic avian radiation. Mol Ecol 24:4238-4251.

Charlesworth B. 1998. Measures of divergence between populations and the effect of forces that reduce variability. Mol Biol Evol 15:538-543.

Charlesworth B, Campos JL. 2014. The relations between recombination rate and patterns of molecular variation and evolution in *Drosophila*. Annu Rev Genet 48:383-403.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics 134:1289-1303.

Christe C, Stölting KN, Paris M, Fraïsse C, Bierne N, Lexer C. 2017. Adaptive evolution and segregating load contribute to the genomic landscape of divergence in two tree species connected by episodic gene flow. Mol Ecol 26:59-76.

Corbett-Detig RB, Hartl DL, Sackton TB. 2015. Natural selection constrains neutral diversity across a wide range of species. PLoS Biol 13:e1002112.

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

Deacon NJ, Grossman JJ, Cavender-Bares J. 2019. Drought and freezing vulnerability of the isolated hybrid aspen *Populus x smithii* relative to its parental species, *P. tremuloides* and *P. grandidentata*. Ecol Evol 9:8062-8074.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491-498.

Du S, Wang Z, Ingvarsson PK, Wang D, Wang J, Wu Z, Tembrock LR, Zhang J. 2015. Multilocus analysis of nucleotide variation and speciation in three closely related *Populus* (Salicaceae) species. Mol Ecol 24:4994-5005.

Eckenwalder J. 1996. Systematics and evolution of *Populus*. U: Stettler, RF, Bradshaw, HD, Heilman, PE, Hinckley, TM, eds.(1996): Biology of *Populus* and Its Implications for Management and Conservation. In: NRC Research Press, Ottawa, Ontario, Canada.

Ellegren H, Smeds L, Burri R, Olason PI, Backström N, Kawakami T, Künstner A, Mäkinen H, Nadachowska-Brzyska K, Qvarnström A, et al. 2012. The genomic landscape of species divergence in *Ficedula* flycatchers. Nature 491:756-760.

Fan L, Zheng H, Milne RI, Zhang L, Mao K. 2018. Strong population bottleneck and repeated demographic expansions of *Populus adenopoda* (Salicaceae) in subtropical China. Ann Bot 121:665-679.

Fang Z, Zhao S, Skvortsov A. 1999. Salicaceae. Flora of China 4:139-274.

Gagnaire PA, Lamy JB, Cornette F, Heurtebise S, Dégremont L, Flahauw E, Boudry P, Bierne N, Lapègue S. 2018. Analysis of Genome-Wide Differentiation between Native and Introduced Populations of the Cupped Oysters *Crassostrea gigas* and *Crassostrea angulata*. Genome Biol Evol 10:2518-2534.

Gao F, Ming C, Hu W, Li H. 2016. New Software for the Fast Estimation of Population Recombination Rates (FastEPRR) in the Genomic Era. G3 (Bethesda) 6:1563-1571.

Han F, Lamichhaney S, Grant BR, Grant PR, Andersson L, Webster MT. 2017. Gene flow, ancient polymorphism, and ecological adaptation shape the genomic landscape of divergence among Darwin's finches. Genome Res 27:1004-1015.

Harrison RG, Larson EL. 2016. Heterogeneous genome divergence, differential introgression, and the origin and structure of hybrid zones. Mol Ecol 25:2454-2466.

Henderson EC, Brelsford A. 2020. Genomic differentiation across the speciation continuum in three hummingbird species pairs. BMC Evol Biol 20:113.

Irwin DE, Milá B, Toews DPL, Brelsford A, Kenyon HL, Porter AN, Grossen C, Delmore KE, Alcaide M, Irwin JH. 2018. A comparison of genomic islands of differentiation across three young avian species pairs. Mol Ecol 27:4839-4855.

Jansson S, Bhalerao R, Groover A. 2010. Genetics and genomics of *Populus*: Springer.

Lamichhaney S, Berglund J, Almén MS, Maqbool K, Grabherr M, Martinez-Barrio A, Promerová M, Rubin CJ, Wang C, Zamani N, et al. 2015. Evolution of Darwin's finches and their beaks revealed by genome sequencing. Nature 518:371-375.

Leroy T, Rougemont Q, Dupouey JL, Bodénès C, Lalanne C, Belser C, Labadie K, Le Provost G, Aury JM, Kremer A, et al. 2020. Massive postglacial gene flow between European white oaks uncovered genes underlying species barriers. New Phytol 226:1183-1197.

Lexer C, Buerkle CA, Joseph JA, Heinze B, Fay MF. 2007. Admixture in European *Populus* hybrid zones makes feasible the mapping of loci that contribute to reproductive isolation and trait differences. Heredity (Edinb) 98:74-84.

Lexer C, Fay MF, Joseph JA, Nica MS, Heinze B. 2005. Barrier to gene flow between two ecologically divergent *Populus* species, *P. alba* (white poplar) and *P. tremula* (European aspen): the role of ecology and life history in gene introgression. Mol Ecol 14:1045-1057.

Lexer C, Joseph JA, van Loo M, Barbará T, Heinze B, Bartha D, Castiglione S, Fay MF, Buerkle CA. 2010. Genomic admixture analysis in European *Populus* spp. reveals unexpected patterns of reproductive isolation and mating. Genetics 186:699-712.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-2079.

Macaya-Sanz D, Heuertz M, López-de-Heredia U, De-Lucas AI, Hidalgo E, Maestro C, Prada A, Alía R, González-Martínez SC. 2012. The Atlantic-Mediterranean watershed, river basins and glacial history shape the genetic structure of Iberian poplars. Mol Ecol 21:3593-3609.

Martin SH, Dasmahapatra KK, Nadeau NJ, Salazar C, Walters JR, Simpson F, Blaxter M, Manica A, Mallet J, Jiggins CD. 2013. Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. Genome Res 23:1817-1828.

Martin SH, Davey JW, Jiggins CD. 2015. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. Mol Biol Evol 32:244-257.

Martin SH, Davey JW, Salazar C, Jiggins CD. 2019. Recombination rate variation shapes barriers to introgression across butterfly genomes. PLoS Biol 17:e2006288.

Martin SH, Jiggins CD. 2017. Interpreting the genomic landscape of introgression. Curr Opin Genet Dev 47:69-74.

Martinsen GD, Whitham TG, Turek RJ, Keim P. 2001. Hybrid populations selectively filter gene introgression between species. Evolution 55:1325-1335.

Modrich P, Lahue R. 1996. Mismatch repair in replication fidelity, genetic recombination, and cancer biology. Annu Rev Biochem 65:101-133.

Noor MA, Bennett SM. 2009. Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. Heredity (Edinb) 103:439-444.

Nosil P, Feder JL. 2012. Widespread yet heterogeneous genomic divergence. Mol Ecol 21:2829-2832.

Pickrell JK, Pritchard JK. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet 8:e1002967.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559-575.

Rajora OP, Dancik BP. 1992. Genetic characterization and relationships of *Populus alba*, *P. tremula*, and *P. x canescens*, and their clones. Theor Appl Genet 84:291-298.

Ravinet M, Faria R, Butlin RK, Galindo J, Bierne N, Rafajlović M, Noor MAF, Mehlig B, Westram AM. 2017. Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. J Evol Biol 30:1450-1477.

Ravinet M, Yoshida K, Shigenobu S, Toyoda A, Fujiyama A, Kitano J. 2018. The genomic landscape at a late stage of stickleback speciation: High genomic divergence interspersed by small localized regions of introgression. PLoS Genet 14:e1007358.

Renaut S, Grassa CJ, Yeaman S, Moyers BT, Lai Z, Kane NC, Bowers JE, Burke JM, Rieseberg LH. 2013. Genomic islands of divergence are not affected by geography of speciation in sunflowers. Nat Commun 4:1827.

Renaut S, Owens GL, Rieseberg LH. 2014. Shared selective pressure and local genomic landscape lead to repeatable patterns of genomic divergence in sunflowers. Mol Ecol 23:311-324.

Rifkin JL, Castillo AS, Liao IT, Rausher MD. 2019. Gene flow, divergent selection and resistance to introgression in two species of morning glories (*Ipomoea*). Mol Ecol 28:1709-1729.

Roux C, Fraïsse C, Romiguier J, Anciaux Y, Galtier N, Bierne N. 2016. Shedding Light on the Grey Zone of Speciation along a Continuum of Genomic Divergence. PLoS Biol 14:e2000234.

Samuk K, Owens GL, Delmore KE, Miller SE, Rennison DJ, Schluter D. 2017. Gene flow and selection interact to promote adaptive divergence in regions of low recombination. Mol Ecol 26:4378-4390.

Schiffthaler B, Delhomme N, Bernhardsson C, Jenkins J, Jansson S, Ingvarsson P, Schmutz J, Street N. 2019. An improved genome assembly of the European aspen *Populus tremula*. bioRxiv:805614.

Sendell-Price AT, Ruegg KC, Anderson EC, Quilodrán CS, Van Doren BM, Underwood VL, Coulson T, Clegg SM. 2020. The Genomic Landscape of Divergence Across the Speciation Continuum in Island-Colonising Silvereyes (*Zosterops lateralis*). G3 (Bethesda) 10:3147-3163.

Shang H, Hess J, Pickup M, Field DL, Ingvarsson PK, Liu J, Lexer C. 2020. Evolution of strong reproductive isolation in plants: broad-scale patterns and lessons from a perennial model group. Philos Trans R Soc Lond B Biol Sci 375:20190544.

Siewert KM, Voight BF. 2017. Detecting Long-Term Balancing Selection Using Allele Frequency Correlation. Mol Biol Evol 34:2996-3005.

Slotte T. 2014. The impact of linked selection on plant genomic variation. Brief Funct Genomics 13:268-275.

Smith JM, Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genet Res 23:23-35.

Stankowski S, Chase MA, Fuiten AM, Rodrigues MF, Ralph PL, Streisfeld MA. 2019. Widespread selection and gene flow shape the genomic landscape during a radiation of monkeyflowers. PLoS Biol 17:e3000391.

Stettler R, Bradshaw T, Heilman P, Hinckley T. 1996. Biology of Populus and its implications for management and conservation: NRC Research Press.

Stölting KN, Paris M, Meier C, Heinze B, Castiglione S, Bartha D, Lexer C. 2015. Genome-wide patterns of differentiation and spatially varying selection between postglacial recolonization lineages of *Populus alba* (Salicaceae), a widespread forest tree. New Phytol 207:723-734.

Suarez-Gonzalez A, Hefer CA, Christe C, Corea O, Lexer C, Cronk QC, Douglas CJ. 2016. Genomic and functional approaches reveal a case of adaptive introgression from *Populus balsamifera* (balsam poplar) in *P. trichocarpa* (black cottonwood). Mol Ecol 25:2427-2442.

Talla V, Johansson A, Dincă V, Vila R, Friberg M, Wiklund C, Backström N. 2019. Lack of gene flow: Narrow and dispersed differentiation islands in a triplet of *Leptidea* butterfly species. Mol Ecol 28:3756-3770.

Tavares H, Whibley A, Field DL, Bradley D, Couchman M, Copsey L, Elleouet J, Burrus M, Andalo C, Li M, et al. 2018. Selection and gene flow shape genomic islands that control floral guides. Proc Natl Acad Sci U S A 115:11006-11011.

Terhorst J, Kamm JA, Song YS. 2017. Robust and scalable inference of population history from hundreds of unphased whole genomes. Nat Genet 49:303-309.

Turner TL, Hahn MW. 2010. Genomic islands of speciation or genomic islands and speciation? Mol Ecol 19:848-850.

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, et al. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313:1596-1604.

Vijay N, Bossu CM, Poelstra JW, Weissensteiner MH, Suh A, Kryukov AP, Wolf JB. 2016. Evolution of heterogeneous genome differentiation across multiple contact zones in a crow species complex. Nat Commun 7:13195.

Wang B, Mojica JP, Perera N, Lee CR, Lovell JT, Sharma A, Adam C, Lipzen A, Barry K, Rokhsar DS, et al. 2019. Ancient polymorphisms contribute to genome-wide variation by long-term balancing selection and divergent sorting in *Boechera stricta*. Genome Biol 20:126.

Wang J, Street NR, Scofield DG, Ingvarsson PK. 2016a. Natural Selection and Recombination Rate Variation Shape Nucleotide Polymorphism Across the Genomes of Three Related *Populus* Species. Genetics 202:1185-1200.

Wang J, Street NR, Scofield DG, Ingvarsson PK. 2016b. Variation in Linked Selection and Recombination Drive Genomic Divergence during Allopatric Speciation of European and American Aspens. Mol Biol Evol 33:1754-1767.

Wang M, Zhang L, Zhang Z, Li M, Wang D, Zhang X, Xi Z, Keefover-Ring K, Smart LB, DiFazio SP, et al. 2020. Phylogenomics of the genus *Populus* reveals extensive interspecific gene flow and balancing selection. New Phytol 225:1370-1382.

Wolf JB, Ellegren H. 2017. Making sense of genomic islands of differentiation in light of speciation. Nat Rev Genet 18:87-100.

Yamasaki YY, Kakioka R, Takahashi H, Toyoda A, Nagano AJ, Machida Y, Moller PR, Kitano J. 2020. Genome-wide patterns of divergence and introgression after secondary contact between *Pungitius* sticklebacks. Philos Trans R Soc Lond B Biol Sci 375:20190548.

Zheng H, Fan L, Milne RI, Zhang L, Wang Y, Mao K. 2017. Species Delimitation and Lineage Separation History of a Species Complex of Aspens in China. Front Plant Sci 8:375.

Supporting Information, including Supporting Tables and Supporting Figures

Huiying Shang[1,2], Martha Rendón-Anaya[3], Ovidiu Paun[1], David Field[4], Jaqueline Hess[5], Claus Vogl[6], Jianquan Liu[7], Pär K. Ingvarsson[3,8]*, Thibault Leroy[1,8]*, Christian Lexer[1,8 †]

[1]Department of Botany and Biodiversity Research, University of Vienna, Vienna, Austria.

[2]Vienna Graduate School of Population Genetics, Vienna, Austria.

[3]Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden.

[4]Edith Cowan University, Perth, Australia.

[5]Helmholtz Centre for Environmental Research, Halle (Saale), Germany.

[6]Department of Biomedical Sciences, Vetmeduni Vienna, Vienna, Austria.

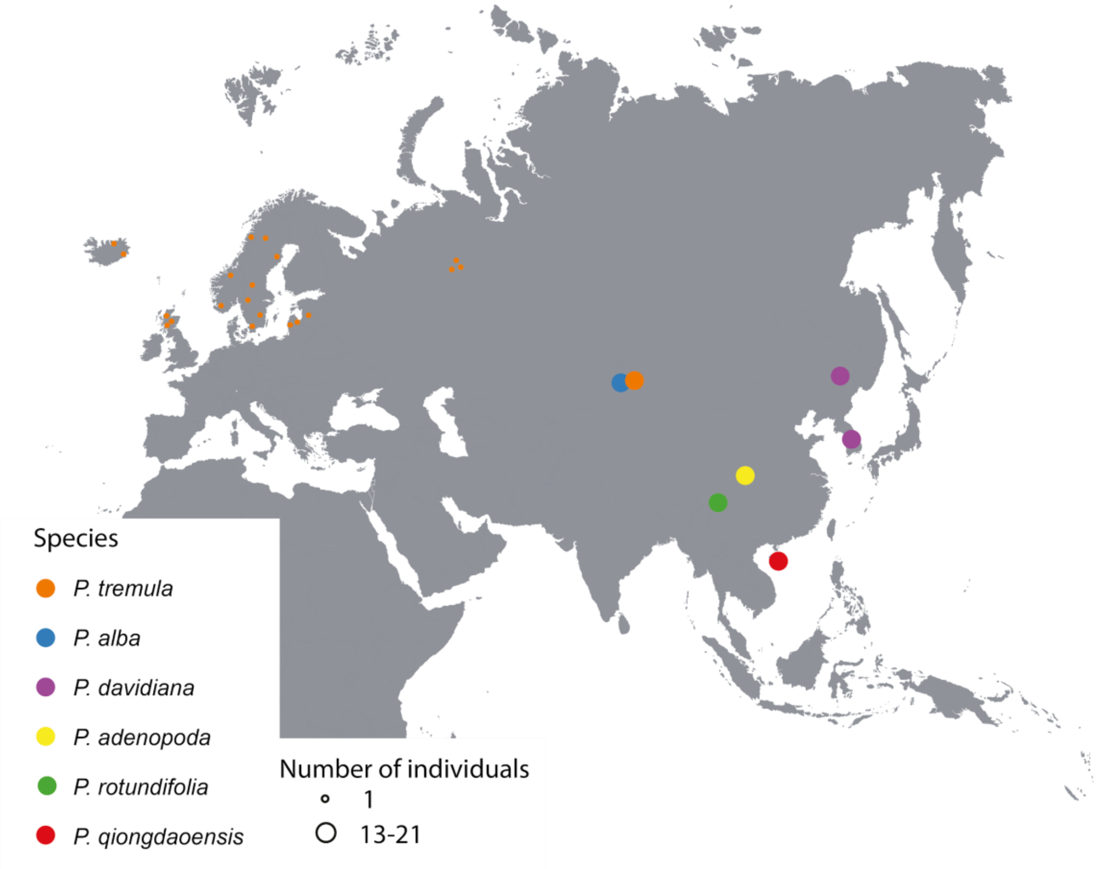
[7]Key Laboratory for Bio-resources and Eco-environment, College of Life Science, Sichuan University, Chengdu, People's Republic of China

[8] These authors contributed equally to this work.

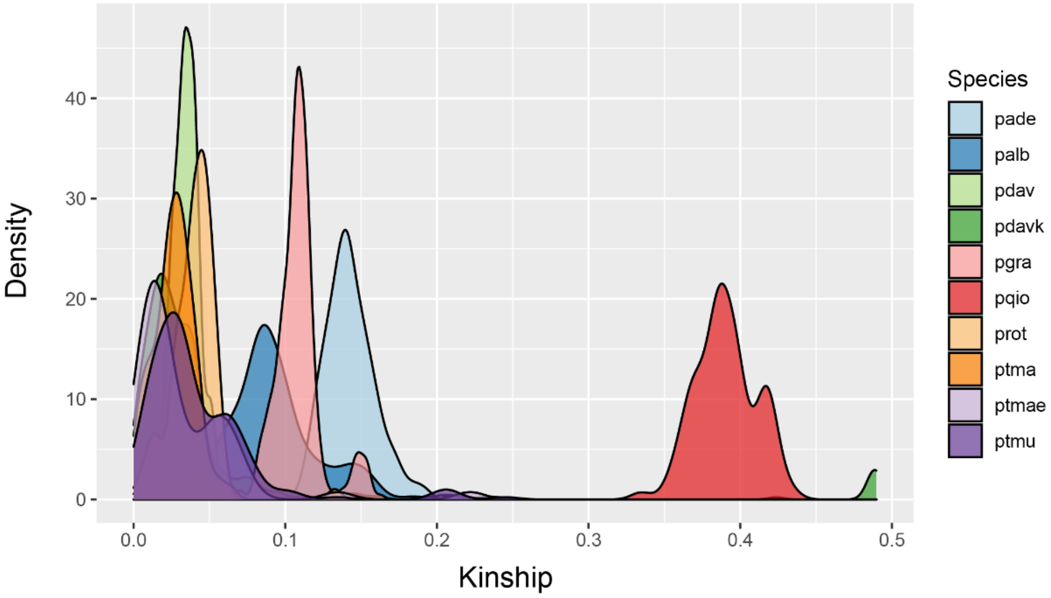
[†]Deceased.

*Corresponding author: Thibault Leroy, Department of Botany and Biodiversity Research, University of Vienna, Rennweg 14, 1030 Vienna, Austria; Email: thibault.leroy@univie.ac.at

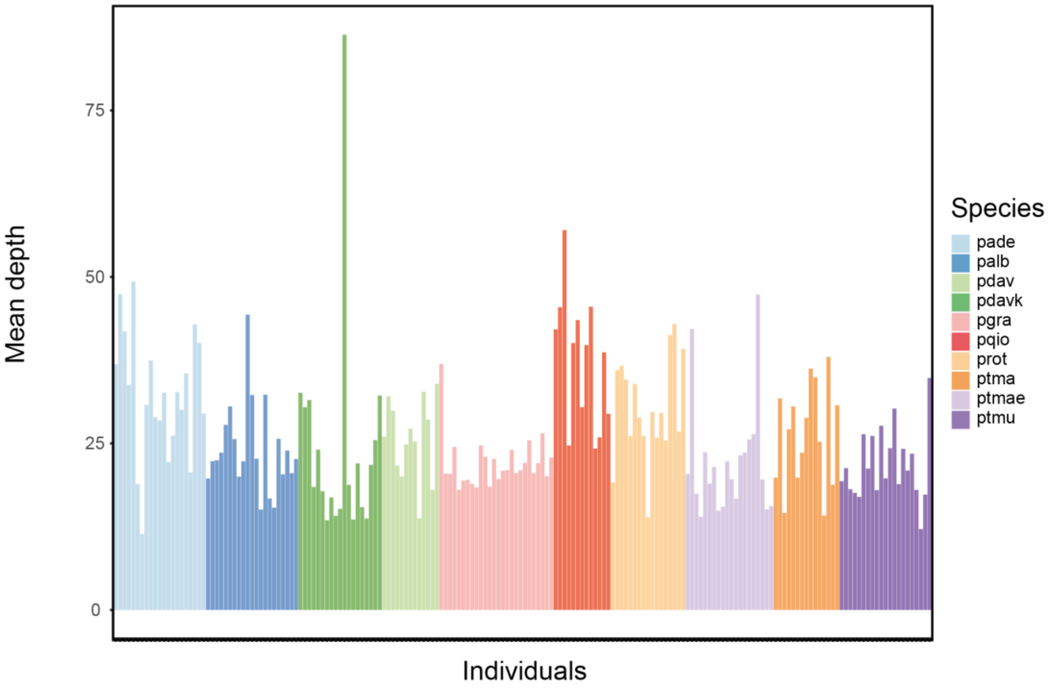
Pär K. Ingvarsson, Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden; Email: par.ingvarsson@slu.se



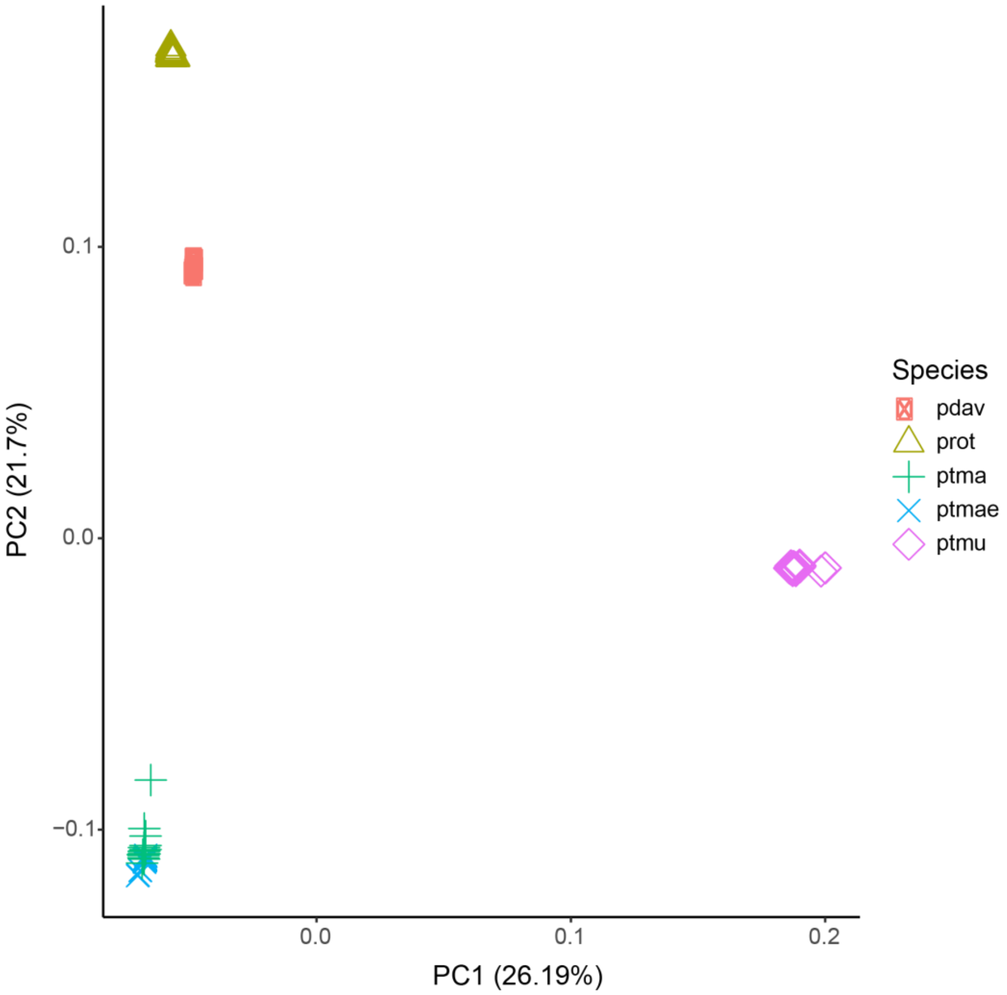
Supporting Fig. S1. Map of the *Populus* samples collected in Eurasia. In the map, the different colors represent the different species, while the size of the circle is relative to the amount of samples collected in that location.



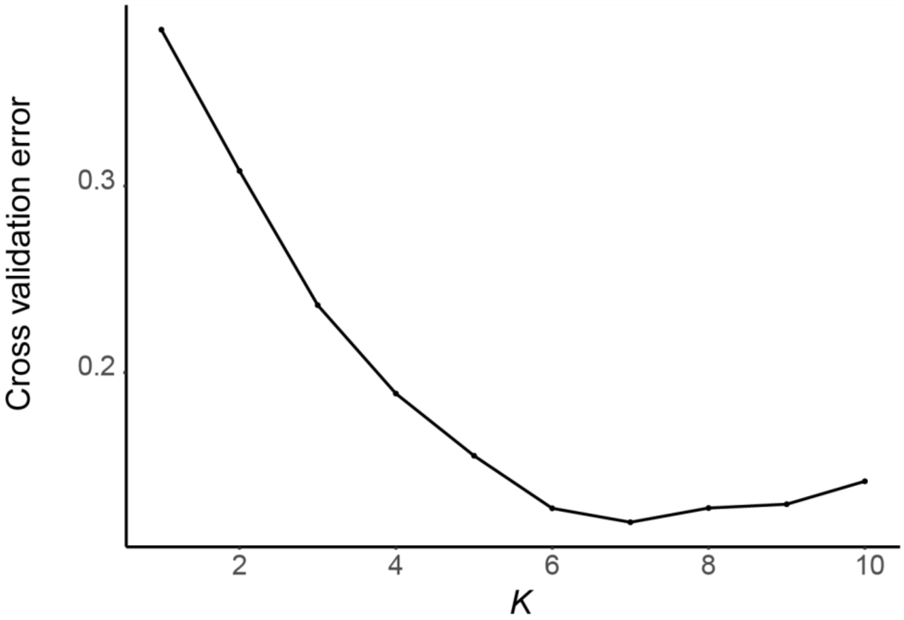
Supporting Fig. S2. Family relationship analysis for all eight *Populus* species. Kinship coefficient ranges: >0.354, [0.177, 0.354], [0.0884, 0.177] and [0.0442, 0.0884] corresponding to duplicate/MZ twin, 1st-degree, 2nd-degree, and 3rd-degree relationships, respectively. Species abbreviations: pade, *P. adenopoda*; palb, *P. alba*; pdav, *P. davidiana* collected in China; pdavk, *P. davidiana* collected in South Korea; pgra, *P. grandidentata*; pqio, *P. qiongdaoensis*; prot, *P. rotundifolia*; ptma, *P. tremula* collected in China; ptmae, *P. tremula* collected in Europe; ptmu, *P.tremuloides*.



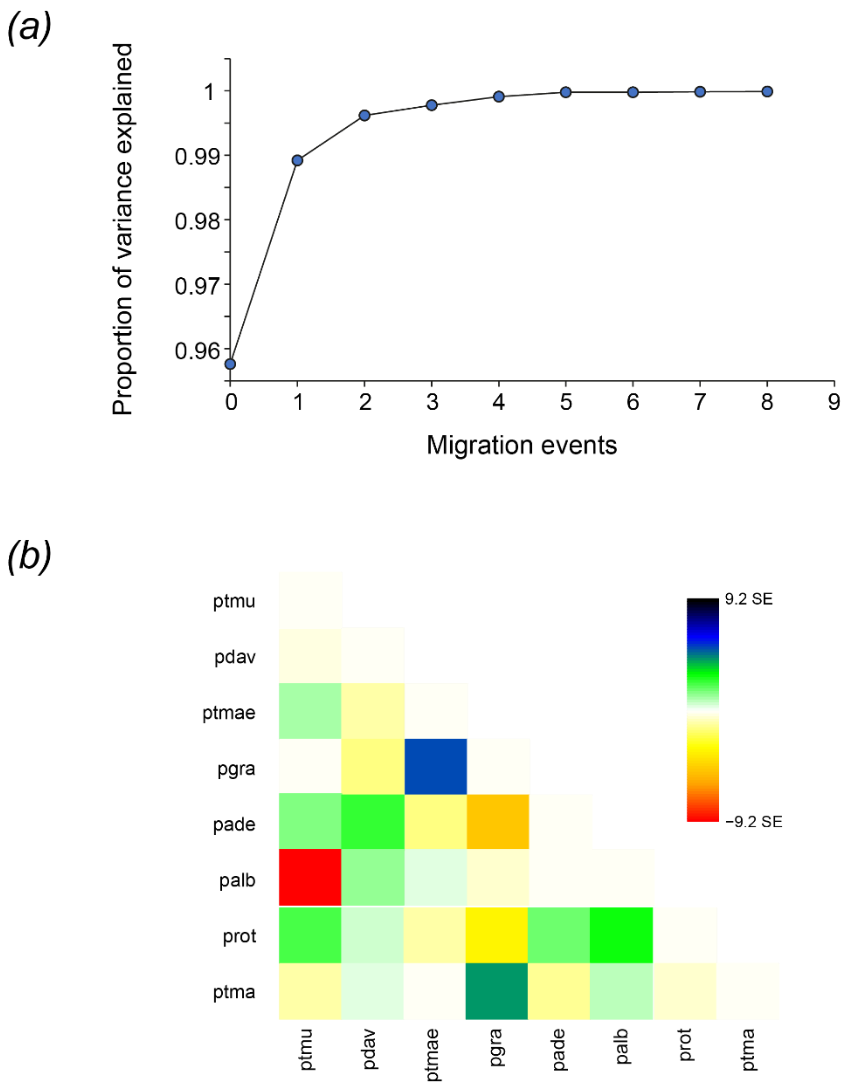
Supporting Fig. S3. Mean depth of each *Populus* accession investigated. Species abbreviations: pade, *P. adenopoda*; palb, *P. alba*; pdav, *P. davidiana* collected in China; pdavk, *P. davidiana* collected in South Korea; pgra, *P. grandidentata*; pqio, *P. qiongdaoensis*; prot, *P. rotundifolia*; ptma, *P. tremula* collected in China; ptmae, *P. tremula* collected in Europe; ptmu, *P. tremuloides*.



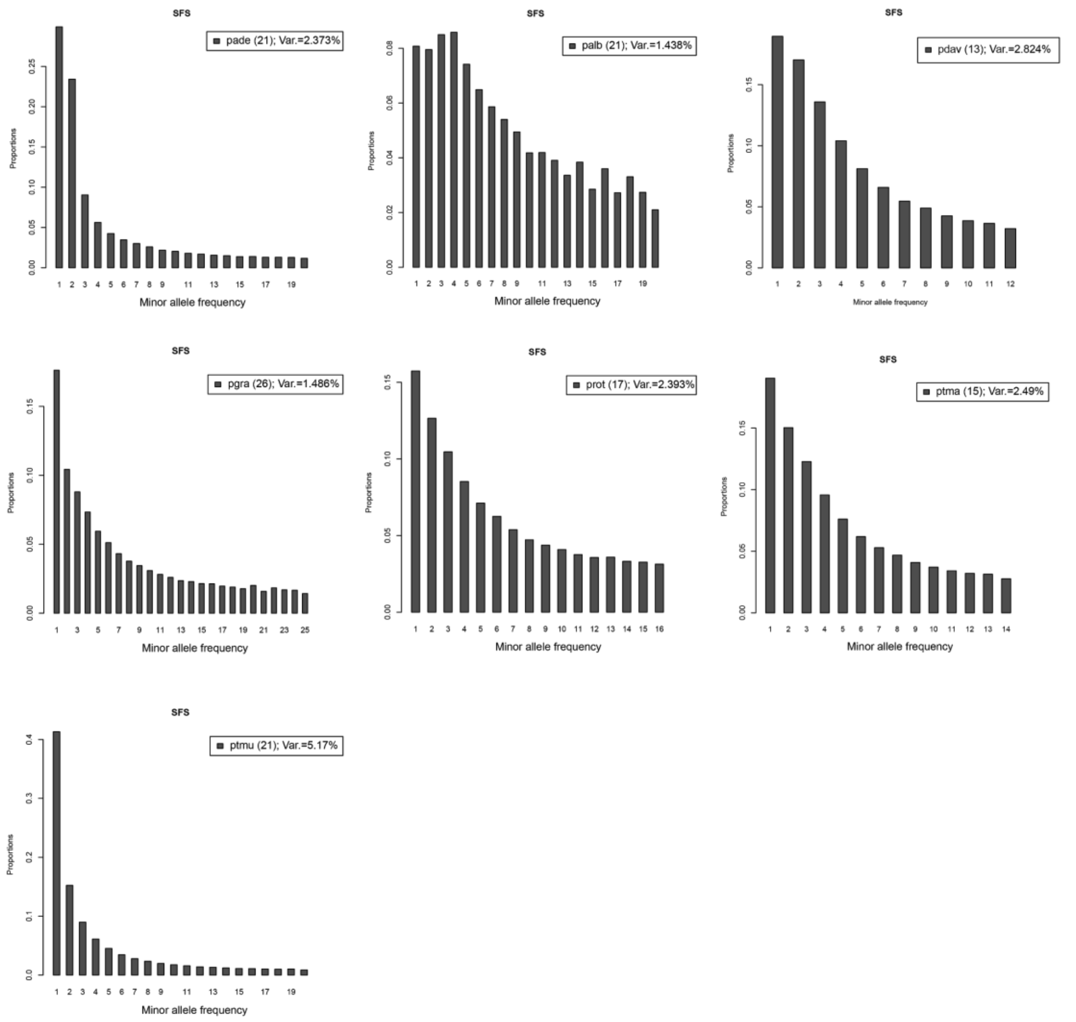
Supporting Fig. S4. Principal component analysis of SNP data from resequenced genomes for four recently diverged *Populus* species. Species abbreviations: pdav, *P. davidiana* collected in China; prot, *P. rotundifolia*; ptma, *P. tremula* collected in China; ptmae, *P. tremula* collected in Europe; ptmu, *P. tremuloides*.



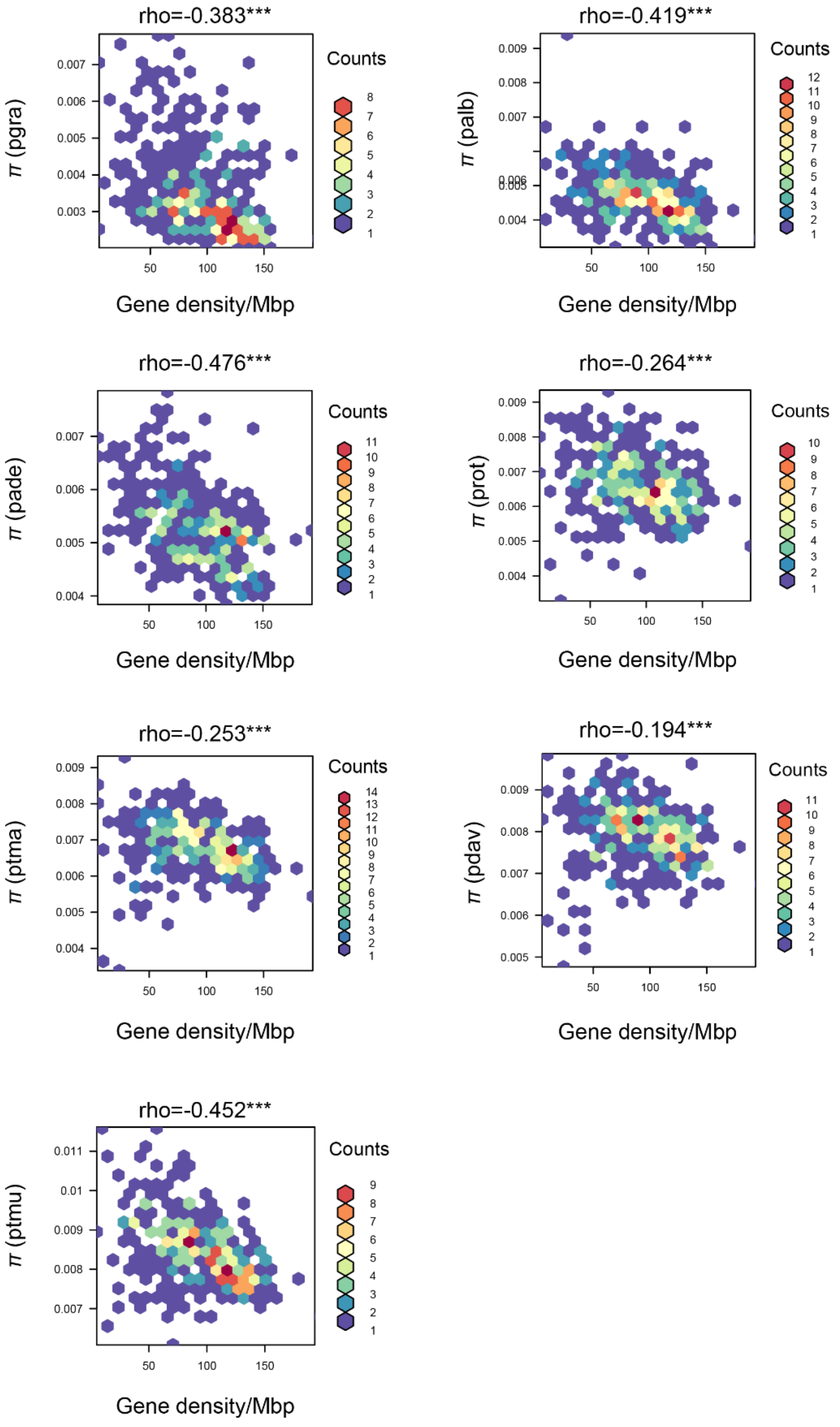
Supporting Fig. S5. Plot of ADMIXTURE cross validation error from *K*=1 through *K*=10. The lowest cross validation error obtained at *K*=7.



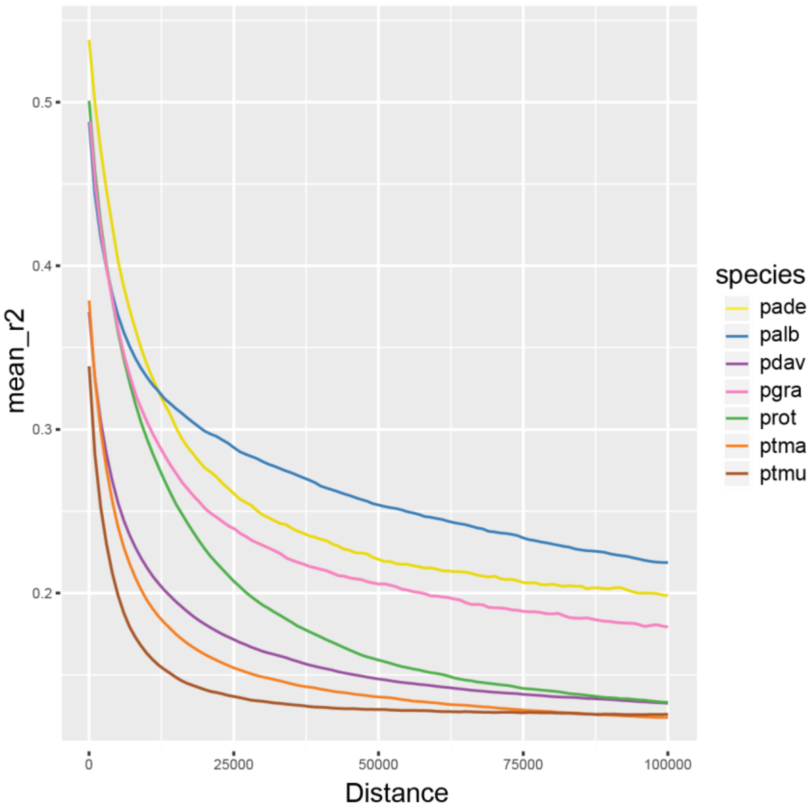
Supporting Fig. S6. Treemix analysis results. ***(a)*** plot of proportion of variance explained by the ten models run in TreeMix analysis using m=0 to m=8 migration events. ***(b)*** Residual heatmap from tree with two migration events.



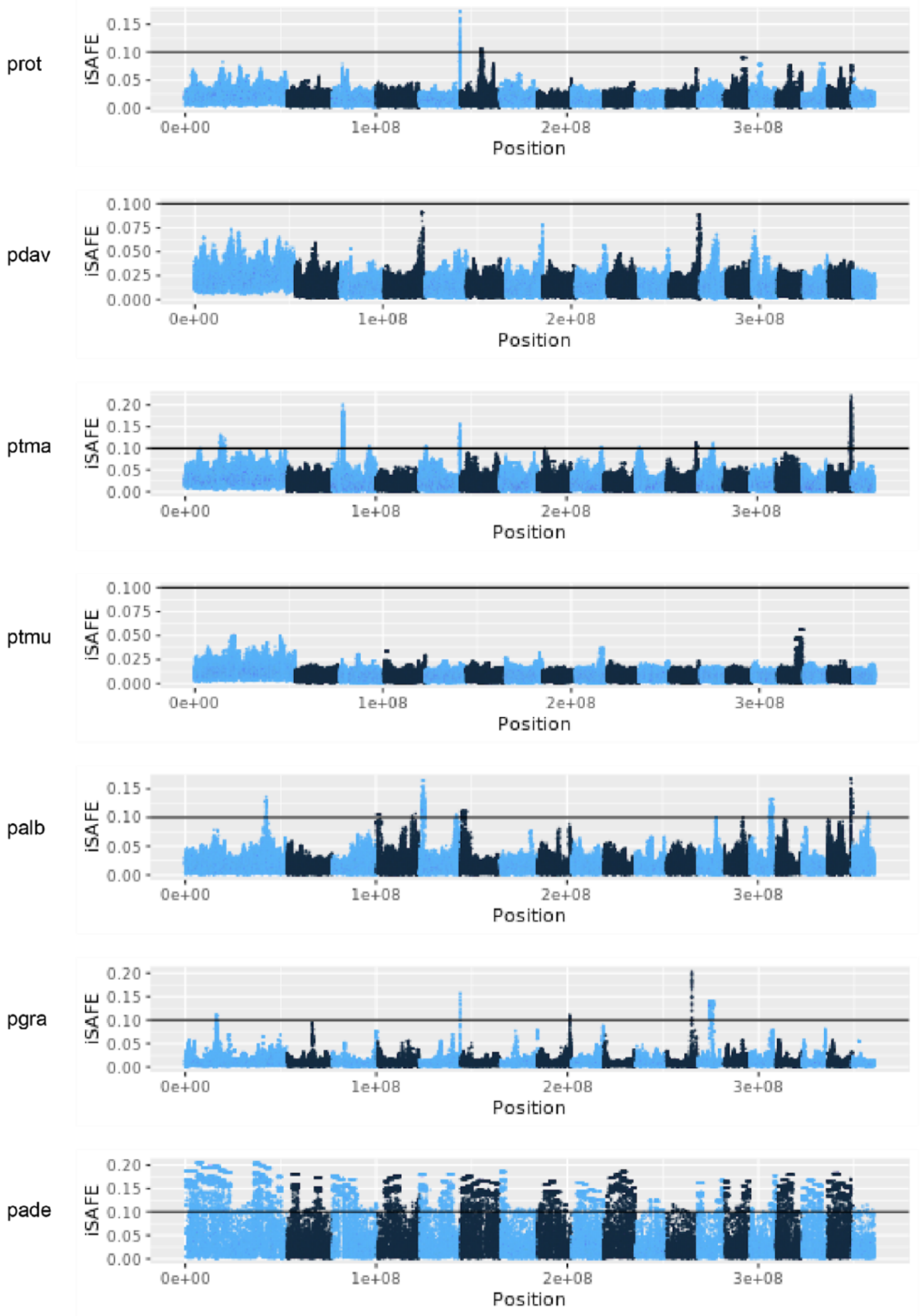
Supporting Fig. S7. The folded site frequency spectrum analysis for seven *Populus* species. X-axes show allele accounts in the population, whereas y-axes show the frequency. Species abbreviations: pade, *P. adenopoda*; palb, *P. alba*; pdav, *P. davidiana*; pgra, *P. grandidentata*; prot, *P. rotundifolia*; ptma, *P. tremula*; ptmu, *P. tremuloides*.



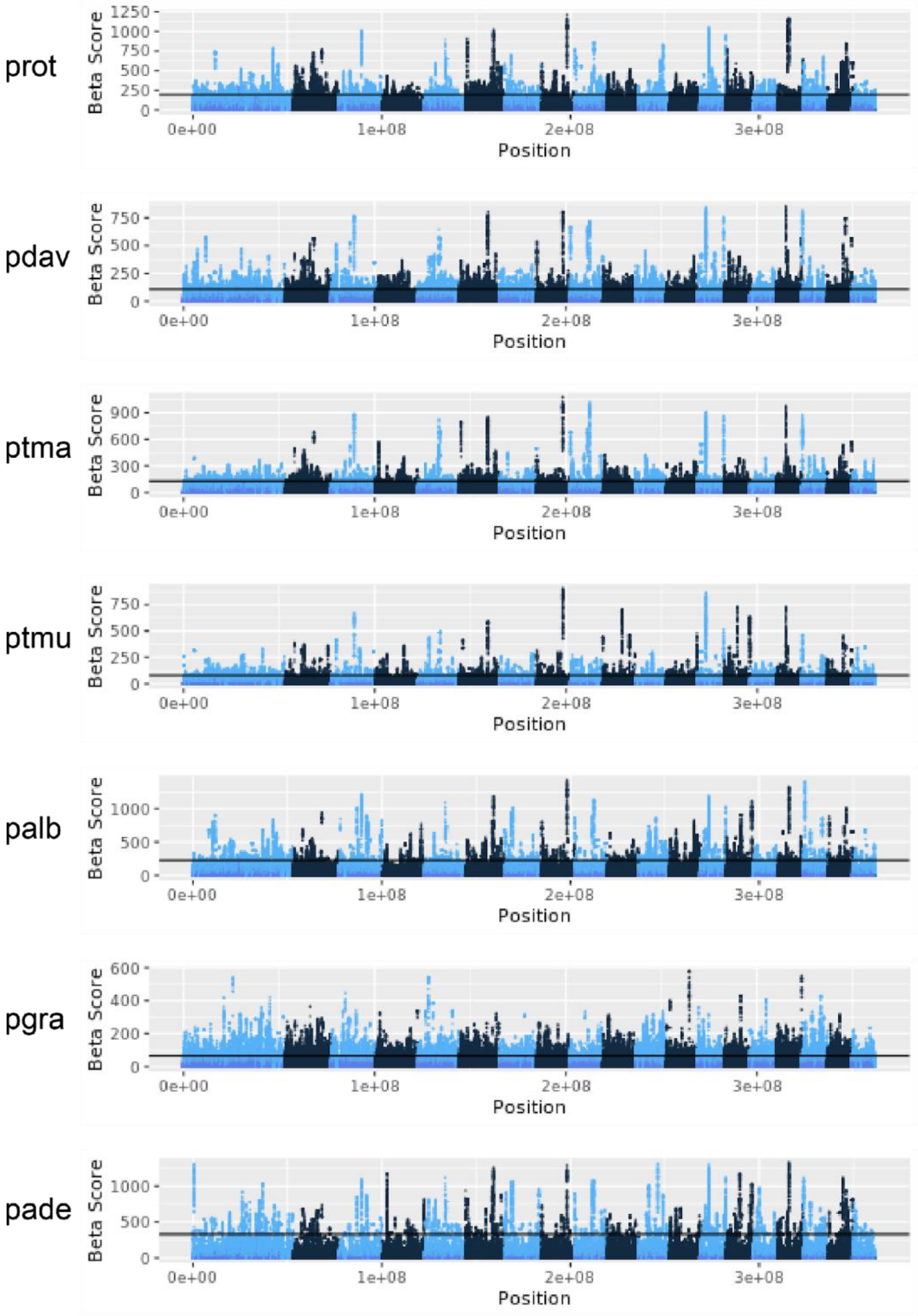
Supporting Fig. S8. Correlation analysis between gene density and genetic diversity for each *Populus* species. Species abbreviations: pade, *P. adenopoda*; palb, *P. alba*; pdav, *P. davidiana*; pgra, *P. grandidentata*; prot, *P. rotundifolia*; ptma, *P. tremula*; ptmu, *P. tremuloides*.



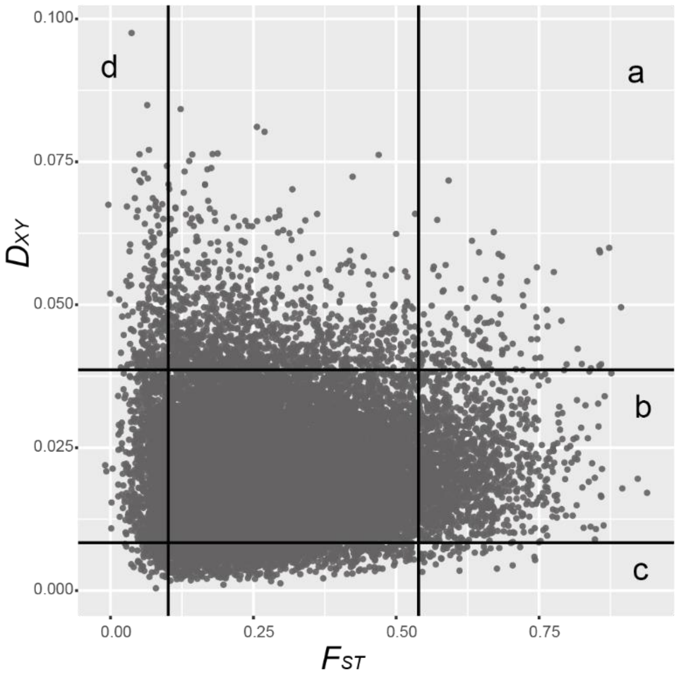
Supporting Fig. S9. Linkage disequilibrium (LD) decay by distance across eight *Populus* species. Species abbreviations: pade, *P. adenopoda*; palb, *P. alba*; pdav, *P. davidiana*; pgra, *P. grandidentata*; prot, *P. rotundifolia*; ptma, *P. tremula*; ptmu, *P. tremuloides*.



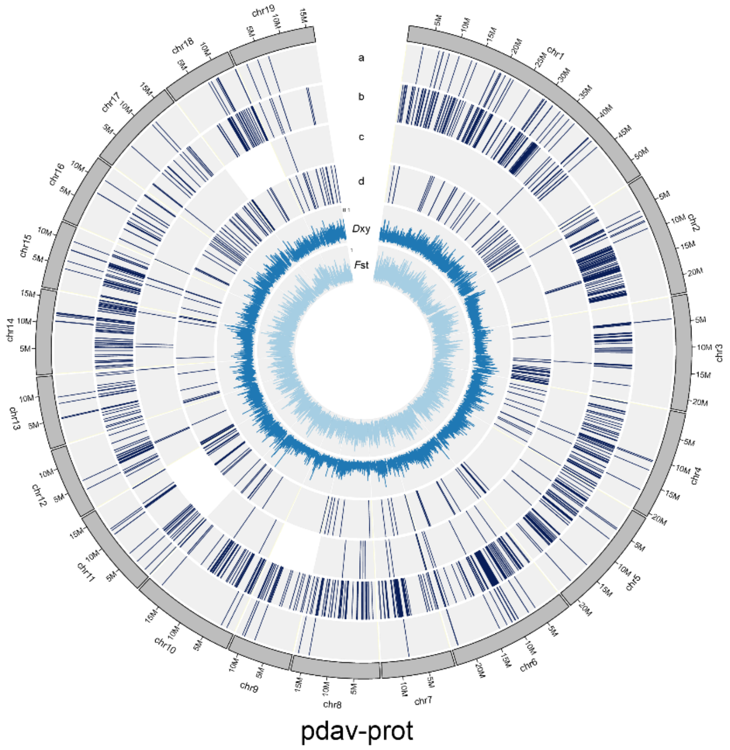
Supporting Fig. S10. Identification of positive selection. (a) Manhattan plot of iSAFE score for seven *Populus* species. Horizontal black line indicates a cutoff based on iSAFE score (> 0.1). Species abbreviations: pade, *P. adenopoda*; palb, *P. alba*; pdav, *P. davidiana* collected in China; pgra, *P. grandidentata*; prot, *P. rotundifolia*; ptma, *P. tremula* collected in China; ptmu, *P. tremuloides*.



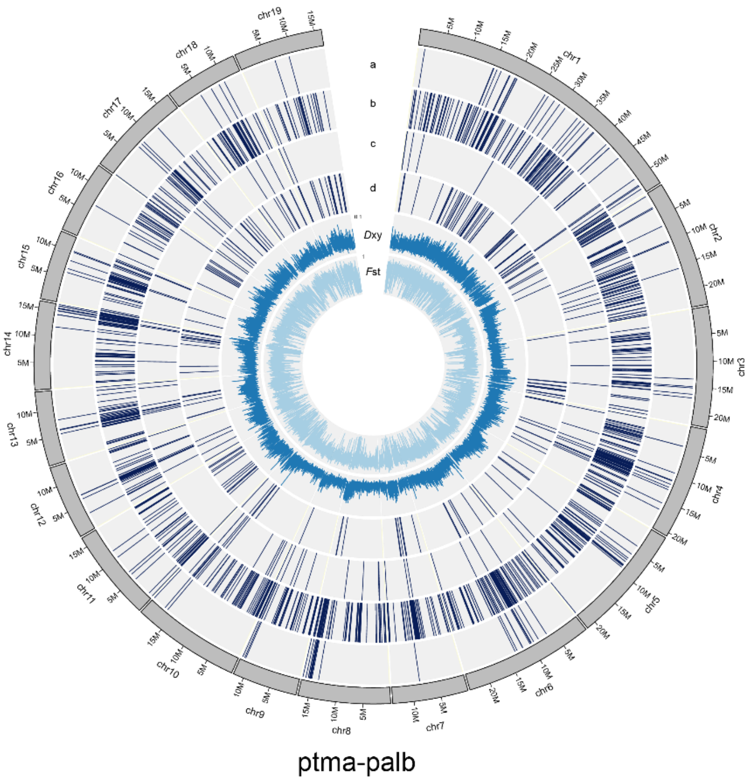
Supporting Fig. S11. Identification of balancing selection. (a) Manhattan plot of Beta score for seven *Populus* species. Horizontal black line indicates a cutoff based on beta score (top 1%). Species abbreviations: pade, *P. adenopoda*; palb, *P. alba*; pdav, *P. davidiana* collected in China; pgra, *P. grandidentata*; prot, *P. rotundifolia*; ptma, *P. tremula* collected in China; ptmu, *P. tremuloides*.



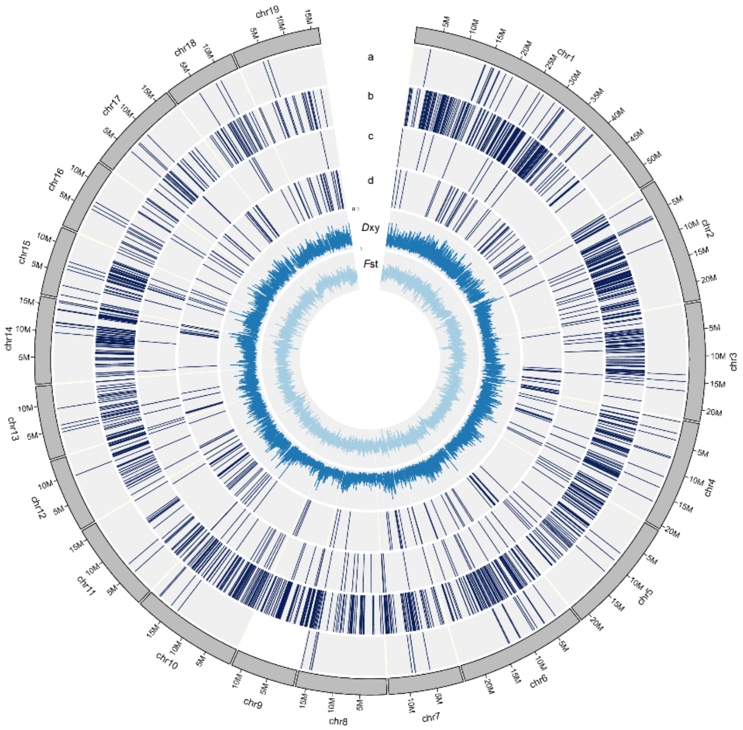
Supporting Fig. S12. Speciation models for the formation of genomic islands of relative differentiation. All the grey points were *F_{ST}* or *D_{XY}* values based on non-overlapping 10kb windows across the whole genome. The lines were 5% threshold for *F_{ST}* and *D_{XY}* values. For example, model (a) explains the regions with top 5% of both *F_{ST}* and *D_{XY}*. In this model, selection at loci which contribute to reproductive isolation restricts gene exchange between populations, elevating genomic differentiation (lead to higher *F_{ST}* and *D_{XY}*) and reducing genetic diversity. The second model (b) is 'allopatric selection', selection on distinct regions of the genome after a species split into two populations, leading to lower π and higher *F_{ST}*. As *D_{XY}* is sensitive to ancestral polymorphism, thus *D_{XY}* values remain stable in this model. The next model (c) is 'recurrent selection', in this model background selection or selective sweeps at certain regions of the genome reduce genetic diversity in the common ancestor and then selection on the same regions of two daughter populations, leading to lower *D_{XY}* and *π*, higher *F_{ST}*. The last model (d) is 'balancing selection', ancestral polymorphisms are maintained at selected sites, resulting in increased *D_{XY}* and low *F_{ST}* between species.



Supporting Fig. S13. Genomic regions associated to the four speciation models (a, b, c, and d) across the whole genome in species pair *P. davidiana* – *P. rotundifolia*. D_{XY} and F_{ST} were calculated in 10kb non-overlapping sliding windows.

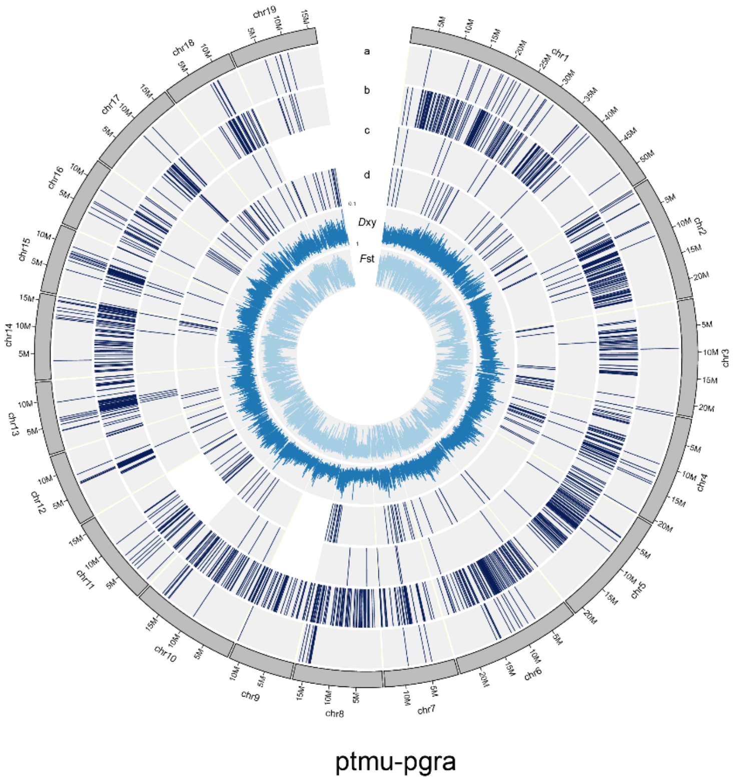


Supporting Fig. S14. Genomic regions associated to the four speciation models (a, b, c, and d) across the whole genome in species pair *P. tremula* – *P. alba*. D_{XY} and F_{ST} were calculated in 10kb non-overlapping sliding windows.

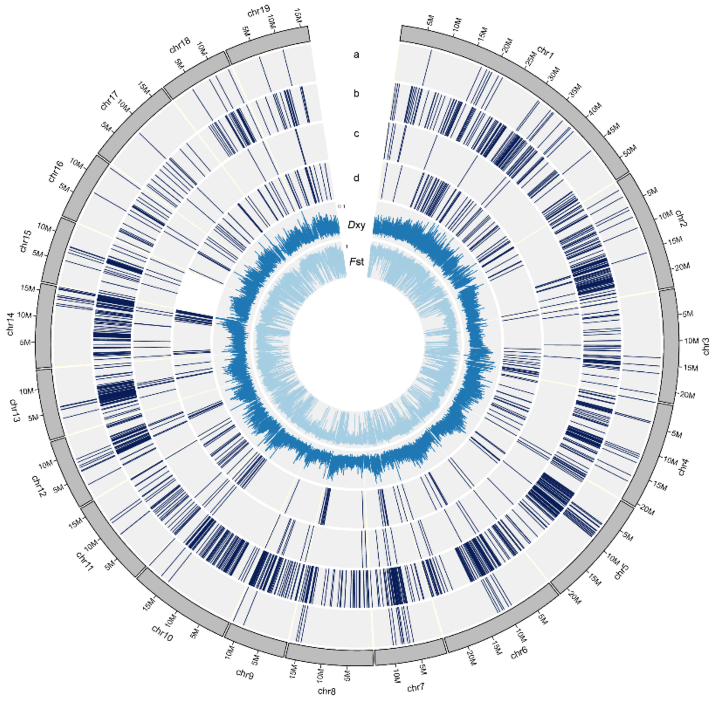


ptma-prot

Supporting Fig. S15. Genomic regions associated to the four speciation models (a, b, c, and d) across the whole genome in species pair *P. tremula – P. rotundifolia*. D_{XY} and F_{ST} were calculated in 10kb non-overlapping sliding windows.



Supporting Fig. S16. Genomic regions associated to the four speciation models (a, b, c, and d) across the whole genome in species pair *P. tremuloides* – *P. grandidentata*. D_{XY} and F_{ST} were calculated in 10kb non-overlapping sliding windows.



pade-palb

Supporting Fig. S17. Genomic regions associated to the four speciation models (a, b, c, and d) across the whole genome in species pair *P. adenopoda – P. alba*. D_{XY} and F_{ST} were calculated in 10kb non-overlapping sliding windows.

Supporting Table S1. Coordination of *Populus* samples used in this study.

Sampled_Trees	Number	Code_number	*Species*	Latitude_(N)	Longitude_(E)	Coverage
pop2014-1	21	palb01	*Populus alba*	47.54	87.90	19.60
pop2014-2		palb02	*Populus alba*	47.54	87.90	22.20
pop2014-3		palb03	*Populus alba*	47.46	87.80	22.33
pop2014-4		palb04	*Populus alba*	47.38	87.80	23.47
pop2014-5		palb05	*Populus alba*	47.38	87.80	27.62
pop2014-6		palb06	*Populus alba*	47.37	87.82	30.37
pop2014-7		palb07	*Populus alba*	47.35	87.86	25.48
pop2014-9		palb09	*Populus alba*	47.35	87.87	19.90
pop2014-10		palb10	*Populus alba*	47.35	87.87	22.15
pop2014-12		palb12	*Populus alba*	47.35	87.87	44.09
pop2014-13		palb13	*Populus alba*	47.35	87.87	32.06
pop2014-15		palb15	*Populus alba*	47.35	87.87	22.53
pop2014-18		palb18	*Populus alba*	47.35	87.89	14.97
pop2014-19		palb19	*Populus alba*	47.35	87.89	32.11
pop2014-20		palb20	*Populus alba*	47.49	87.33	16.61
pop2014-21		palb21	*Populus alba*	47.49	86.33	15.23
pop2014-24		palb24	*Populus alba*	47.72	86.88	25.50
pop2014-26		palb26	*Populus alba*	47.01	86.27	20.20
pop2014-37		palb37	*Populus alba*	47.84	86.66	23.77
pop2014-38		palb38	*Populus alba*	47.83	86.66	20.40
pop2014-41		palb41	*Populus alba*	47.83	86.67	22.51
pop2014-45	15	ptma45	*Populus tremula*	47.91	88.13	19.84
pop2014-48		ptma48	*Populus tremula*	47.96	88.18	31.73
pop2014-50		ptma50	*Populus tremula*	47.97	88.18	14.52
pop2014-52		ptma52	*Populus tremula*	47.97	88.19	27.08
pop2014-53		ptma53	*Populus tremula*	47.98	88.20	30.50
pop2014-55		ptma55	*Populus tremula*	47.98	88.22	19.81

pop2014-56		ptma56	*Populus tremula*	47.98	88.23	23.57
pop2014-57		ptma57	*Populus tremula*	47.98	88.24	28.84
pop2014-58		ptma58	*Populus tremula*	47.99	88.24	36.18
pop2014-59		ptma59	*Populus tremula*	47.99	88.24	34.85
pop2014-62		ptma62	*Populus tremula*	47.99	88.24	25.21
pop2014-67		ptma67	*Populus tremula*	47.99	88.24	14.13
pop2014-68		ptma68	*Populus tremula*	48.00	88.27	37.94
pop2014-70		ptma70	*Populus tremula*	48.00	88.26	18.76
pop2014-71		ptma71	*Populus tremula*	47.99	88.26	30.67
pop2014-73	13	pdav73	*Populus davidiana*	45.32	127.35	25.89
pop2014-74		pdav74	*Populus davidiana*	45.10	128.00	31.94
pop2014-75		pdav75	*Populus davidiana*	45.02	128.18	29.83
pop2014-76		pdav76	*Populus davidiana*	44.95	128.79	21.57
pop2014-77		pdav77	*Populus davidiana*	44.93	128.97	19.95
pop2014-78		pdav78	*Populus davidiana*	44.92	128.99	24.74
pop2014-79		pdav79	*Populus davidiana*	44.53	129.79	27.09
pop2014-80		pdav80	*Populus davidiana*	44.77	129.15	25.18
pop2014-81		pdav81	*Populus davidiana*	44.93	128.73	13.71
pop2014-82		pdav82	*Populus davidiana*	45.25	127.73	32.63
pop2014-83		pdav83	*Populus davidiana*	45.25	127.71	28.47
pop2014-84		pdav84	*Populus davidiana*	45.32	127.32	17.93
pop2014-85		pdav85	*Populus davidiana*	45.33	127.27	33.84
LiuJQ-F-2015-01-01	21	pade01	*Populus adenopoda Maxim.*	32.76	105.25	36.63
LiuJQ-F-2015-01-02		pade02	*Populus adenopoda Maxim.*	32.76	105.25	47.01
LiuJQ-F-2015-01-03		pade03	*Populus adenopoda Maxim.*	32.76	105.25	41.52
LiuJQ-F-2015-01-04		pade04	*Populus adenopoda Maxim.*	32.76	105.25	33.50
LiuJQ-F-2015-01-05		pade05	*Populus adenopoda Maxim.*	32.76	105.25	48.90
LiuJQ-F-2015-01-06		pade06	*Populus adenopoda Maxim.*	32.76	105.25	18.70
LiuJQ-F-2015-01-07		pade07	*Populus adenopoda Maxim.*	32.76	105.25	11.26
LiuJQ-F-2015-01-08		pade08	*Populus adenopoda Maxim.*	32.76	105.25	30.45

LiuJQ-F-2015-01-09		pade09	*Populus adenopoda Maxim.*	32.76	105.25	37.09
LiuJQ-F-2015-01-10		pade10	*Populus adenopoda Maxim.*	32.76	105.25	28.67
LiuJQ-F-2015-01-11		pade11	*Populus adenopoda Maxim.*	32.76	105.25	28.22
LiuJQ-F-2015-01-12		pade12	*Populus adenopoda Maxim.*	32.76	105.25	32.40
LiuJQ-F-2015-01-13		pade13	*Populus adenopoda Maxim.*	32.76	105.25	21.95
LiuJQ-F-2015-01-14		pade14	*Populus adenopoda Maxim.*	32.76	105.25	25.95
LiuJQ-F-2015-01-15		pade15	*Populus adenopoda Maxim.*	32.76	105.25	32.41
LiuJQ-F-2015-01-16		pade16	*Populus adenopoda Maxim.*	32.76	105.25	29.76
LiuJQ-F-2015-01-17		pade17	*Populus adenopoda Maxim.*	32.76	105.25	35.21
LiuJQ-F-2015-01-18		pade18	*Populus adenopoda Maxim.*	32.76	105.25	20.47
LiuJQ-F-2015-01-19		pade19	*Populus adenopoda Maxim.*	32.76	105.25	42.56
LiuJQ-F-2015-01-20		pade20	*Populus adenopoda Maxim.*	32.76	105.25	39.80
LiuJQ-F-2015-01-21		pade21	*Populus adenopoda Maxim.*	32.76	105.25	29.21
MaoKS-CX-2014-261A-01	17	prot261A01	*Populus rotundifolia*	27.14	99.39	19.06
MaoKS-CX-2014-261A-02		prot261A02	*Populus rotundifolia*	27.14	99.39	35.85
MaoKS-CX-2014-261A-03		prot261A03	*Populus rotundifolia*	27.14	99.39	36.47
MaoKS-CX-2014-261A-04		prot261A04	*Populus rotundifolia*	27.14	99.39	34.46
MaoKS-CX-2014-261A-05		prot261A05	*Populus rotundifolia*	27.14	99.39	25.95
MaoKS-CX-2014-261A-06		prot261A06	*Populus rotundifolia*	27.14	99.39	33.79
MaoKS-CX-2014-261A-07		prot261A07	*Populus rotundifolia*	27.14	99.39	28.73
MaoKS-CX-2014-261A-08		prot261A08	*Populus rotundifolia*	27.14	99.39	26.00
MaoKS-CX-2014-261A-09		prot261A09	*Populus rotundifolia*	27.14	99.39	13.86
MaoKS-CX-2014-261A-10		prot261A10	*Populus rotundifolia*	27.14	99.39	29.59
MaoKS-CX-2014-261A-11		prot261A11	*Populus rotundifolia*	27.14	99.39	25.72
MaoKS-CX-2014-261A-12		prot261A12	*Populus rotundifolia*	27.14	99.39	29.44
MaoKS-CX-2014-261A-13		prot261A13	*Populus rotundifolia*	27.14	99.39	25.29
MaoKS-CX-2014-261A-14		prot261A14	*Populus rotundifolia*	27.14	99.39	41.12
MaoKS-CX-2014-261A-15		prot261A15	*Populus rotundifolia*	27.14	99.39	42.80
MaoKS-CX-2014-261A-16		prot261A16	*Populus rotundifolia*	27.14	99.39	26.64
MaoKS-CX-2014-261A-17		prot261A17	*Populus rotundifolia*	27.14	99.39	39.09

LiuJQ-Tian-2015-001-01	14	pqioT0101	*Populus qiongdaoensis T.Hong et P.Luo*	19.12	109.09	41.81
LiuJQ-Tian-2015-001-02		pqioT0102	*Populus qiongdaoensis T.Hong et P.Luo*	19.12	109.09	45.14
LiuJQ-Tian-2015-001-03		pqioT0103	*Populus qiongdaoensis T.Hong et P.Luo*	19.12	109.09	56.67
LiuJQ-Tian-2015-001-04		pqioT0104	*Populus qiongdaoensis T.Hong et P.Luo*	19.12	109.09	24.46
LiuJQ-Tian-2015-002-01		pqioT0201	*Populus qiongdaoensis T.Hong et P.Luo*	19.12	109.09	39.75
LiuJQ-Tian-2015-002-02		pqioT0202	*Populus qiongdaoensis T.Hong et P.Luo*	19.12	109.09	43.20
LiuJQ-Tian-2015-002-03		pqioT0203	*Populus qiongdaoensis T.Hong et P.Luo*	19.12	109.09	30.17
LiuJQ-Tian-2015-002-04		pqioT0204	*Populus qiongdaoensis T.Hong et P.Luo*	19.12	109.09	39.48
LiuJQ-Tian-2015-002-05		pqioT0205	*Populus qiongdaoensis T.Hong et P.Luo*	19.12	109.09	45.17
LiuJQ-Tian-2015-002-06		pqioT0206	*Populus qiongdaoensis T.Hong et P.Luo*	19.12	109.09	23.98
LiuJQ-Tian-2015-002-07		pqioT0207	*Populus qiongdaoensis T.Hong et P.Luo*	19.12	109.09	25.65
LiuJQ-Tian-2015-002-08		pqioT0208	*Populus qiongdaoensis T.Hong et P.Luo*	19.12	109.09	38.38
LiuJQ-Tian-2015-002-09		pqioT0209	*Populus qiongdaoensis T.Hong et P.Luo*	19.12	109.09	4.57
LiuJQ-Tian-2015-002-10		pqioT0210	*Populus qiongdaoensis T.Hong et P.Luo*	19.12	109.09	29.23
Alb10-3	22	ptmu01	*P.tremuloides*	NA	NA	19.18
Alb13-1		ptmu02	*P.tremuloides*	NA	NA	21.13
Alb16-1		ptmu03	*P.tremuloides*	NA	NA	18.01
Alb17-4		ptmu04	*P.tremuloides*	NA	NA	17.45
Alb25-4		ptmu05	*P.tremuloides*	NA	NA	16.87
Alb27-1		ptmu06	*P.tremuloides*	NA	NA	26.23
Alb31-1		ptmu07	*P.tremuloides*	NA	NA	21.09
Alb33-2		ptmu08	*P.tremuloides*	NA	NA	25.99
Alb35-2		ptmu09	*P.tremuloides*	NA	NA	17.87
Alb6-3		ptmu10	*P.tremuloides*	NA	NA	27.43
Albb15-3		ptmu11	*P.tremuloides*	NA	NA	19.61
Dan1-1C13		ptmu12	*P.tremuloides*	NA	NA	24.15
Dan2-1B7		ptmu13	*P.tremuloides*	NA	NA	30.06
PG1-1B4		ptmu14	*P.tremuloides*	NA	NA	18.75
PG2-1B9		ptmu15	*P.tremuloides*	NA	NA	24.03
PG3-1B6		ptmu16	*P.tremuloides*	NA	NA	20.79

PI12-1B14		ptmu17	*P.tremuloides*	NA	NA	3.06
PI3-1B3		ptmu18	*P.tremuloides*	NA	NA	23.28
Sau1-1B10		ptmu19	*P.tremuloides*	NA	NA	17.88
Sau2-1B2		ptmu20	*P.tremuloides*	NA	NA	12.05
Sau3-1B13		ptmu21	*P.tremuloides*	NA	NA	17.20
Wau1-1B5		ptmu22	*P.tremuloides*	NA	NA	34.64
Bonghyeon4_2	32	pdavk01	*P. davidiana*	NA	NA	32.63
Daehwa18-2		pdavk02	*P. davidiana*	NA	NA	30.40
Daehwa6-1		pdavk03	*P. davidiana*	NA	NA	30.46
Dongdu2-1		pdavk04	*P. davidiana*	NA	NA	31.55
KR_BD_5		pdavk05	*P. davidiana*	37.92	128.39	18.37
KR_DW_18_2		pdavk06	*P. davidiana*	NA	NA	23.98
KR_DW_18_3		pdavk07	*P. davidiana*	NA	NA	17.63
KR_DW_6		pdavk08	*P. davidiana*	37.50	28.46	19.13
KR_DW_6-2		pdavk09	*P. davidiana*	37.50	28.46	17.77
KR_KM_1		pdavk10	*P. davidiana*	37.29	129.24	13.41
KR_OD_19-1		pdavk11	*P. davidiana*	NA	NA	16.86
KR_OD_19-2		pdavk12	*P. davidiana*	NA	NA	17.85
KR_OD_19-4		pdavk13	*P. davidiana*	37.80	128.54	14.05
KR_PD_15		pdavk14	*P. davidiana*	38.16	128.37	15.10
KR_PG_4		pdavk15	*P. davidiana*	37.83	128.49	86.26
KR_PG_4-1		pdavk16	*P. davidiana*	37.83	128.49	14.51
KR_PK_1-2		pdavk17	*P. davidiana*	36.02	128.70	18.69
KR_PK_2_2		pdavk18	*P. davidiana*	NA	NA	13.50
KR_PK_3_2		pdavk19	*P. davidiana*	NA	NA	21.94
KR_PU-12		pdavk20	*P. davidiana*	37.78	128.58	15.36
KR_SG_5		pdavk21	*P. davidiana*	37.51	128.60	24.07
KR_SW_6_2		pdavk22	*P. davidiana*	NA	NA	13.69
KR_SW_6_3		pdavk23	*P. davidiana*	NA	NA	24.99

KR_SY_3		pdavk24	*P. davidiana*	38.18	128.30	21.73
KR_SY_4		pdavk25	*P. davidiana*	38.18	128.30	25.43
KR_SY_6		pdavk26	*P. davidiana*	38.18	128.30	24.02
KR_WD_2_1		pdavk27	*P. davidiana*	NA	NA	16.55
KR_WD_2_2		pdavk28	*P. davidiana*	NA	NA	21.26
Palgong1-1		pdavk29	*P. davidiana*	NA	NA	32.23
Palgong2-3		pdavk30	*P. davidiana*	NA	NA	27.11
Palgong3-1		pdavk31	*P. davidiana*	NA	NA	31.31
Sogwang9-1		pdavk32	*P. davidiana*	NA	NA	37.65
Sgardur	20	ptmae01	*P. tremula*	65.86	-17.8929	22.28
SJorvik		ptmae02	*P. tremula*	64.84	-14.37	19.57
LV_VIL_02		ptmae03	*P. tremula*	57.19	27.51	20.39
LV_SAL_22		ptmae04	*P. tremula*	56.65	22.66	42.25
LV_LIE_36		ptmae05	*P. tremula*	56.24	21.40	17.42
NO_ALE_07		ptmae06	*P. tremula*	62.48	6.72	13.93
NO_STV_02		ptmae07	*P. tremula*	58.77	5.91	23.66
NO_MIR_01		ptmae08	*P. tremula*	66.31	14.41	18.94
RU_SYK_01		ptmae09	*P. tremula*	61.68	50.99	21.46
RU_SYK_10		ptmae10	*P. tremula*	61.65	51.08	14.88
RU_SYK_20		ptmae11	*P. tremula*	61.67	51.09	15.48
SwAsp003		ptmae12	*P. tremula*	56.71	13.22	16.31
SwAsp033		ptmae13	*P. tremula*	57.83	15.31	21.71
SwAsp045		ptmae14	*P. tremula*	59.64	12.94	22.08
SwAsp067		ptmae15	*P. tremula*	61.20	13.81	23.58
SwAsp096		ptmae16	*P. tremula*	63.98	20.71	24.49
SwAsp109		ptmae17	*P. tremula*	66.36	18.18	40.91
UK_AAR_23		ptmae18	*P. tremula*	56.19	-5.41	19.55
UK_CGM_90		ptmae19	*P. tremula*	57.14	-3.91	15.06
UK_WRS_112		ptmae20	*P. tremula*	57.87	-5.44	15.55

BT4	26	pgra01	*P. grandidentata*	NA	NA	36.91
P11086_106		pgra02	*P. grandidentata*	NA	NA	20.48
P11086_107		pgra03	*P. grandidentata*	NA	NA	20.42
P11086_108		pgra04	*P. grandidentata*	NA	NA	24.47
P11086_109		pgra05	*P. grandidentata*	NA	NA	18.02
P11086_110		pgra06	*P. grandidentata*	NA	NA	19.40
P11086_111		pgra07	*P. grandidentata*	NA	NA	19.54
P11086_112		pgra08	*P. grandidentata*	NA	NA	18.91
P11086_113		pgra09	*P. grandidentata*	NA	NA	18.39
P11086_114		pgra10	*P. grandidentata*	NA	NA	24.69
P11086_115		pgra11	*P. grandidentata*	NA	NA	22.99
P11086_116		pgra12	*P. grandidentata*	NA	NA	18.48
P11086_117		pgra13	*P. grandidentata*	NA	NA	22.66
P11086_118		pgra14	*P. grandidentata*	NA	NA	19.63
P11086_119		pgra15	*P. grandidentata*	NA	NA	20.87
P11086_120		pgra16	*P. grandidentata*	NA	NA	20.99
P11086_121		pgra17	*P. grandidentata*	NA	NA	24.00
P11086_122		pgra18	*P. grandidentata*	NA	NA	20.52
P11086_123		pgra19	*P. grandidentata*	NA	NA	20.96
P11086_124		pgra20	*P. grandidentata*	NA	NA	22.07
P11086_125		pgra21	*P. grandidentata*	NA	NA	25.46
P11086_126		pgra22	*P. grandidentata*	NA	NA	20.53
P11086_127		pgra23	*P. grandidentata*	NA	NA	22.01
P11086_128		pgra24	*P. grandidentata*	NA	NA	26.51
P11086_129		pgra25	*P. grandidentata*	NA	NA	20.06
P11086_130		pgra26	*P. grandidentata*	NA	NA	22.88

NA, no data.

Supporting Table S2. Per-species number of non-overlapping sliding windows detected under positive and balancing selection regions.

Selection type	pdav	prot	ptma	palb	ptmu	pgra	pade
positive selection	438	351	592	218	486	102	235
balancing selection	1379	927	1002	630	814	2510	358

Species abbreviations: pade, *P. adenopoda*; palb, *P. alba*; pdav, *P. davidiana*; pgra, *P. grandidentata*; prot, *P. rotundifolia*; ptma, *P. tremula*; ptmu, *P.tremuloides*.

Supporting Table S3. Percentage of each model summarized based on F_{ST} and D_{XY} in each species pair.

Species_pair	Model	Relative percentage among the detected windows
pdav-prot	a	7.2% (129)
	b	78.1% (1396)
	c	2.6% (46)
	d	12.1% (216)
ptma-prot	a	5.5% (98)
	b	75.7% (1357)
	c	7% (125)
	d	11.8% (212)
ptma-palb	a	8.1% (142)
	b	75.7% (1318)
	c	5.2% (92)
	d	11.6% (204)
pade-palb	a	6.9% (121)
	b	74.3% (1313)
	c	4.9% (86)
	d	13.9% (246)

4 Conclusion and outlook

In the first chapter of the thesis, we extracted from the literature 133 cases of pairs of flowering plants to investigate the interaction of prezygotic and postzygotic reproductive isolation barriers and gene flow. The results pointed to the important role of both prezygotic and postzygotic barriers during speciation. Future efforts should explore how different aspects of life-history traits and mating systems mediate the strength of the negative association between the number of reproductive isolation barriers and the level of gene flow, and how plants, animals and fungi differ in this regard. Then, using *Populus* as an exemplary model group, we analyzed genome-wide variation for phylogenetic tree topologies in both early and late stage of speciation taxa to determine how these patterns may be related to the genomic architecture of reproductive isolation. Genome wide variation in phylogenetic tree topologies highlights the important role of both ancient and recent gene flow in shaping the heterogeneous genomic landscapes of diversity and differentiation. The negative correlation between signals consistent with species tree and recombination rate is consistent with the correlation between polygenic barriers and linked selection. Even though reproductive isolation barriers evolved from early to the late stage of speciation in *Populus,* the number of barriers is not enough to facilitate strong coupling and thus to prevent the escape of locally adaptive alleles (Kruuk, et al. 1999). Taken together, this broad to narrow approach provides novel insights into the processes and outcomes of divergence with gene flow from the early to late stages of speciation. In the future, it is critical to explore the temporal dynamics of prezygotic and postzygotic barriers accumulation, and their coupling along the speciation continuum. The sequencing of ancient DNA from fossil tree samples (Wagner et al. 2018) will probably offer in a reasonable future a new opportunity to trace back the emergence of the genomic islands of speciation.

In the second part of the thesis, we focused on discussing the evolutionary factors in shaping the heterogeneous landscapes of diversity and differentiation. To achieve this goal, we first constructed the demographic history of eight closely related *Populus* species, including estimating gene flow between species and the effective population size change. Our results showed extensive introgression among three closely related species *P. tremula*, *P. davidiana* and *P. rotundifolia*.

Two of the eight investigated species, *P. adenopoda* and *P. tremuloides* had experienced population expansion from 30,000 years ago, which seems consistent with their extensive present-day distribution areas in South China and North America, respectively. Then we calculated genome wide F_{ST} and D_{XY} for 21species pairs from the early to the late stage of speciation. We observed highly conserved genomic landscapes, either at the intraspecific (genetic diversity and recombination rate) and interspecific levels (relative and absolute divergence levels). This may be specific to a group of species in which large-scale structural variation between species may be less common. Over the whole continuum of divergence, we recovered negative correlations between nucleotide diversity and relative divergence across all species pairs, which is consistent with expected effects of linked selection. However, the positive correlations between nucleotide diversity and absolute divergence landscapes became weaker as the overall divergence level (d_a) increased. This may indicate that background selection is also a contributing factor to the heterogeneity of genomic landscapes of differentiation. In addition, the negative correlations between introgression (f_d) and F_{ST} in some species pairs indicates the role of gene flow in shaping genomic landscapes of differentiation. Finally, compared with the genomic background, the regions under balancing selection showed significantly higher D_{XY} across five representative species pairs across the speciation continuum. This result confirmed the role of ancestral polymorphism in shaping the genomic landscape of differentiation in *Populus*. Nonetheless, linked selection and recombination rate variation appear as major factors which have shaped the heterogeneous genomic landscape of divergence in *Populus*. Overall, the empirical study on several *Populus* species across speciation continuum could explain the heterogeneous landscape of differentiation. A better description and functional annotation of the candidate genes contributing to reproductive isolation is now needed. This future work will allow a better understanding of the prezygotic and postzygotic mechanisms of *Populus* speciation.

References

Kruuk LE, Baird SJ, Gale KS, Barton NH. 1999. A comparison of multilocus clines maintained by environmental adaptation or by selection against hybrids. Genetics 153:1959-1971.

Wagner S, Lagane F, Seguin-Orlando A, Schubert M, Leroy T, Guichoux E, Chancerel E, Bech-Hebelstrup I, Bernard V, Billard C, et al. 2018. High-Throughput DNA sequencing of ancient wood. Mol Ecol 27:1138-1154.

5 Appendix

Conference contributions

Science Talk, March 12th, 2018, University of Vienna (Oral presentation)

Title: **Incomplete lineage sorting and gene flow in *Populus* impact on tree topologies and speciation**

Abstract

The genus *Populus* is well studied because of its ecological and economic importance, and because of its favorable genetic attributes such as small genome size (<500 Mb; 2C =1.1pg in the case of P. trichocarpa), diploidy throughout the genus (2n = 38), 'porous' species barriers, and a well curated and annotated genome assembly. However, due to extensive hybridization and introgression between species, it is a great challenge to reconstruct the phylogenetic relationship between species in section *Populus*. In my study, we used both concatenated and coalescent methods to recover the phylogenetic relationship of seven Populus species and test the impact of incomplete lineage sorting and gene flow on tree topologies and speciation. Our results showed plenty of shared haplotypes between recently divergent species *P. davidiana* and *P. rotundifolia*. Both ILS and gene flow contribute to tree topologies, but whole genome population genetics are required to understand drivers of diversification along the divergence (dis-) continuum.

Retreat meeting of the Vienna Graduate School of Population genetics, October 10-12, 2019 at Altaussee, Austria (Oral presentation)

Title: **Variation in linked selection and recombination rate drive genomic divergence in *Populus*?**

Abstract

Speciation is an important topic in evolutionary biology. During the process of speciation, genetic changes gradually accumulate in the genomes of diverging species. Recent studies have documented genomic differentiation highly variable across the genome and the high differentiation regions are usually regarded as ‘genomic islands of speciation’. However, it’s still unclear how the patterns of differentiation generated and what evolutionary processes drive the evolution of genomic islands? To identify which evolutionary processes driving the genomic landscapes we resequenced 201 individuals, from eight closely related Populus species. Population structure and identity by descent analyses revealed strong interspecific structure, but also extensive introgression between some species pairs, especially those with parapatric distributions. Inferences of the historical changes in effective population sizes suggest population expansions in two species, *P. adenopoda* and *P. tremuloides*, which is consistent with the extensive present-day distribution areas of these species in South China and North America, respectively. Comparisons drawn from the genetic diversity and recombination rate landscapes for each of the species, or from the distribution of relative and absolute divergence levels for five species pairs distributed along the speciation continuum, revealed significantly conserved patterns. Significant correlations among species or species pairs were highly repeatable, which indicates that indirect selection is responsible for the differentiation landscape in *Populus*.

6 Acknowledgements

I sincerely appreciate many people who contributed to my PhD research project. First, I would like to thank my initial supervisor, Christian Lexer, who not only pointed me in the right direction during the research and gave me great patience and kindness through the whole project, but set me an excellent example of how to balance work and life, and how to get along well with people, no matter who she or he is, we should respect and show our kindness to them. I learnt a lot from him, his passion for his research career and his positive attitude on things always inspire me. What I have learnt from him will have a lifelong impact on me. I am also very grateful for the advice and feedback from Jaqueline Hess (Jacky), David Field, and Melinda Pickup, especially for their contributions to the first paper. Jacky is a very lovely and smart person, who is not only my colleague but also a very good friend in my heart. She always gave help and encouragement when I got stuck in my life. My supervisors in the last year Thibault Leroy and Ovidiu Paun gave me support, advice and encouragement. I very much appreciate their kind help. It was an honor to know Pelle Ingvarsson, Martha Rendon and their office mates. Many thanks to their support for SNP calling in their lab. The two weeks I spent there left me good memories. I am also very grateful to Jianquan Liu for offering the Asian *Populus* samples.

I would additionally like to thank my office mates through the years. Discussions with them were particularly helpful during interpretation of my data and solution of problems in my life. They are: Camille Christe, Kai N. Stölting, Luisa Bresadola, Margot Paris, Marylaure de La Harpe, Gil Yardeni, Marie K Brandrud, Juliane Baar, Thomas Wolfe, Jacqueline Heckenhauer, Marcela Van Loo, Clara Groot Crego, Neil McNair, Michael Barfuss, Thelma Barbara, Zhiqing Xue, Mimmi Eriksson, Christina Hedderich, Ahmad Muhammad, Aglaia Szukula, Dana Paun and Thelma Barbara. Besides, I am very grateful to Michael for his help during DNA extraction.

It is an honor and pleasure for me to have been part of the Vienna Graduate School of Population Genetics. Joining in the group made me feel like in a big family with people from the world. As the members in the group are working on different research topics. Discussions with them always help to expand my horizons and bring new ideas for my research. Many thanks to some of them: Christian Schlötterer, Joachim Hermisson, Claus Vogl, Magnus Nordborg, Claire Burny, Sheng-

Kai Hsu, Wei-Yun Lai, Filip Wierzbicki, Krzysztof Stankiewicz, Divya Selvaraju, Florian Schwarz, and the visitors coming for a talk there.

Finally, I would like to express my sincere thanks to my family, especially to my lovely husband Kun, who always gives me support, understanding and makes me happy. My mother, for her help in taking care of my baby. My parents-in-law, for their kind help in life during the period when I was writing the thesis. This work would never have been done without their support and encouragement.