universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Population genomics of speciation in two recently diverged Chinese *Populus* species"

verfasst von / submitted by

Francesca Beclin BSc

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

Master of Science (MSc)

Wien, 2021 / Vienna 2021

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 833
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Ecology and Ecosystems
Betreut von / Supervisor:	Dr. Thibault Leroy

Table of contents

Acknowledgements

First, I want to thank my supervisor Dr. Thibault Leroy for his great support and guidance throughout the work on my Master's Project and this thesis. Furthermore, I want to thank Dr. Huiying Shang for her advice and providing access to the sequencing data, and Dr. Christelle Fraïsse, for advice and answering questions regarding the software DILS. I also want to thank Assoz. Prof. Dipl.-Ing. Dr. Ovidiu Paun for great organizational help. Finally, I want to thank Prof. Dr. Christian Lexer, who amplified my interest in evolutionary biology (through a lecture series), which led me to apply for this Master's Project.

Abstract

Speciation, the continuous evolutionary process leading to the formation of new distinct species, depends on the accumulation of reproductive isolation between two lineages, until they become completely reproductively isolated. During this long process of divergence, specific reproductive barrier loci to gene flow emerge and amass within the genome. This can be due to the establishment of differential adaptations (ecological barriers) between the incipient species, but also due to the fixation of some more intrinsic genetic incompatibilities (non-ecological barriers). To better understand speciation and identify both kinds of barriers to gene flow in genomes, only semi-isolated populations or species, that are still exchanging genes at neutral loci, can be studied.

In this thesis, I studied this process in detail for the two recently diverged (1.3 Mya) species *Populus davidiana* and *P. rotundifolia*. They occupy different ecological niches in Asia (*P. davidiana* is widely distributed in the lowland in central and northeastern China, while *P. rotundifolia* has a narrower range at higher altitudes in the Himalaya and adjacent areas), but their ranges overlap in southwestern China. This thesis has two major objectives: the first is to infer the past demographic history of these two diverging species, in particular regarding gene flow between them, and the second is to identify specific regions in the genome contributing to reproductive isolation between these two species, which could therefore potentially be linked to these different ecological preferences or to some other intrinsic incompatibilities.

Demographic inferences were performed, using the software DILS (Demographic Inferences with Linked Selection), an ABC-based software which performs simulations to infer the most likely demographic scenario. The demographic analyses supported ongoing migration between the two species, and a heterogeneous migration rate along the genome, suggesting a mosaic between regions exchanging genes between the two species and regions without gene flow due to reproductive barriers.

Then, genome scanning for elevated relative (F_{ST}) and absolute divergence levels ($\pi_{Between}$), were performed to identify regions that are strongly differentiated between the two species. We then investigated if these regions might contain genes contributing to reproductive isolation. Manual literature searches were performed to recover the potential function of these genes. Among a short list of 12 genes, two already characterized genes were identified, including *rhd3*, which was described as essential to the formation of root hairs and *wrk3*, which is involved in plant defense to herbivory.

Zusammenfassung

Speziation (Artbildung), der kontinuierliche evolutionäre Prozess, welcher zur Entstehung neuer distinkter Arten führt, basiert auf der Akkumulierung von reproduktiver Isolation zwischen zwei Abstammungslinien, bis diese vollkommen reproduktiv voneinander

isoliert sind. Im Laufe dieser langwierigen Auseinanderentwicklung entstehen an bestimmten Stellen im Genom Barrieren, die den Genfluss zwischen den beiden Arten verhindern. Dies kann entweder durch die Etablierung von differentiellen Anpassungen der beiden entstehenden Arten an ihren jeweiligen Lebensraum bedingt sein, oder aber durch die Fixierung von einigen intrinsischen genetischen Inkompatibilitäten zwischen den Arten. Um den Prozess der Speziation besser zu verstehen und diese beiden Formen von Genfluss-Barrieren in Genomen zu identifizieren, können nur halb-isolierte Populationen oder Arten als Untersuchungsobjekte herangezogen werden, welche noch immer Gene an neutralen Stellen im Genom austauschen.

Im Rahmen dieser Arbeit habe ich diesen Prozess im Detail anhand der zwei Arten *Populus davidiana* und *Populus rotundifolia* untersucht, welche erst vor evolutionär kurzer Zeit (vor etwa 1,3 Millionen Jahren) begonnen haben sich auseinanderzuentwickeln. Diese zwei Arten besetzen unterschiedliche ökologische Nischen in Asien (*P. davidiana* ist im Tiefland im Zentrum und Nordosten Chinas weit verbreitet, während *P. rotundifolia* ein kleineres Verbreitungsgebiet in größeren Höhenlagen im Himalaya und angrenzenden Gebieten hat). Die Verbreitungsgebiete überlappen in Südostchina. Diese Arbeit hat zwei Hauptziele: ersteres ist, die demographische Geschichte dieser beiden divergierenden Arten zu inferieren, mit besonderem Augenmerk auf den Genfluss zwischen ihnen, und das zweite Ziel ist es, spezifische Regionen im Genom zu identifizieren, welche zu reproduktiver Isolation zwischen den beiden Arten beitragen, und welche daher möglicherweise im Zusammenhang mit den unterschiedlichen ökologischen Präferenzen oder mit anderen intrinsischen Inkompatibilitäten der beiden Arten stehen.

Demografische Inferenzen wurden mithilfe der Software DILS („Demographic Inferences with Linked Selection“) durchgeführt, einer ABC-basierten Software, welche Simulationen ausführt, um das wahrscheinlichste demographische Szenario zu ermitteln. Die durchgeführten demografischen Analysen unterstützen aktuell erfolgenden Genfluss zwischen den beiden Arten, und eine heterogene Migrationsrate entlang des Genoms, was auf ein Mosaik zwischen Regionen hindeutet, die Gene zwischen den beiden Arten austauschen, und Regionen, in welchen aufgrund von reproduktiven Barrieren kein Genfluss erfolgt.Po

Dann wurden Genomscans für erhöhte relative (F_{ST}) und absolute Differenzierung ($\pi_{Between}$) zwischen den beiden Arten durchgeführt, um Regionen zu identifizieren, die sich stark zwischen den beiden Arten unterscheiden. Es wurde untersucht, ob diese Regionen Gene enthalten, welche zu reproduktiver Isolation beitragen könnten. Eine manuelle Literaturrecherche wurde durchgeführt, um die potentielle Funktion dieser Gene zu ermitteln. Unter einer kurzen Liste von 12 identifizierten Genen, waren zwei bereits charakterisierte Gene: *rhd3*, welches als essentiell für die Ausbildung von Wurzelhaaren beschrieben wurde, und *wrk3*, welches an der Verteidigung von Pflanzen gegen Herbivorie beteiligt ist.

1. Introduction

1.1 Speciation-with-gene-flow and selection

Species are the fundamental unit of biodiversity. Despite their central role in biology, many different species definitions exist in the literature, often referred to as "species concepts". In this thesis, I will consider the most widely used species concept among evolutionary biologists (Stankowski & Ravinet, 2021), the "Biological Species Concept" (Mayr, 1942), where species are groups of actually or potentially interbreeding populations that are reproductively isolated from some other groups. Following this view, the understanding of speciation, the evolutionary process leading to new distinct species, is dependent upon the progressive accumulation of reproductive isolation, until the complete isolation of two (or more) species. This process, constantly repeated over millions and billions of years, contributed to the diversification of life on earth.

Darwin first suggested a crucial role of natural (positive) selection in speciation (Darwin, 1859). **Positive selection** is a process that can be important for the progress of speciation: it means that a gene variant that is beneficial to the fitness of a population gains in frequency in the population, due to an increased reproductive success of individuals carrying this allele (i.e. gene variant). Later on, Mayr (1942) rather suggested that speciation mainly occurs due to complete spatial isolation (i.e. allopatry), in other words, in the absence of gene flow between the two diverging groups. More recently, more and more empirical evidence supports that speciation is frequently accompanied by gene flow between the two diverging species (e.g. Nosil, 2008). **Migration** of individuals between sexually reproducing populations leads to a gene exchange (i.e. **gene flow**) of these populations: Individuals from one population procreate with individuals from the other population.

Adaptation to different environments (via divergent positive selection) is thought to be the main driver of speciation under a divergence-with-gene-flow model (isolation with migration or secondary contact) (Feder *et al.*, 2012). This leads to the emergence of reproductive isolation at some loci in the genome that are under strong divergent selection and where gene flow is impeded, therefore representing barriers to gene flow. During the course of speciation, several of such adaptive barrier loci could be accumulated. It should, however, be noted that this process does not necessarily require ecological barriers, since various kinds of intrinsic barriers (which may be non-ecological) can also contribute massively to reproductive isolation and sometimes even initiate speciation (Kulmuni & Westram, 2017). This way, a complex network of interacting reproductive barriers emerges (Ravinet *et al.*, 2017; Feder *et al.*, 2012).

Several other processes have an effect on the rate of accumulation of reproductive barrier loci. Migration and **recombination** (i.e. the exchange of genetic material between homologous chromosomes) counteract positive selection, as they may disintegrate the favorable combinations of adaptive barrier loci to gene flow, in such a way that the extent of

physical linkage and recombination between loci is weighed against the strength of selection. Neutrally evolving regions of the genome or regions where selection is too weak are homogenized by gene flow, until species divergence has progressed to a point where gene flow becomes impossible along the whole genome (Feder *et al.*, 2012).

The fixation of an allele at a barrier locus can have been facilitated by local adaptation, resulting in a so-called **"selective sweep"**, which occurs if strong positive selection acts on a beneficial allele(the selected variant will be swept to a high frequency in one of the populations), reducing the genetic diversity at surrounding neutral loci that are physically linked to the selected variant ("genetic hitchhiking") (Feder *et al.*, 2012). This phenomenon is one of the two sources of "**linked selection**". The extent of the sweep depends on a balance between selection and recombination, increasing with the strength of selection, and decreasing with the recombination rate.

However, there are other processes independent of reproductive isolation which may locally reduce diversity as well. For example, selective sweeps may also target universally adaptive alleles, and in this case, they will not lead to the emergence of barrier loci to gene flow between the diverging species. Another process which is independent of reproductive isolation is **background selection**. This other source of linked selection is due to negative (purifying) selection acting on a deleterious allele aleading to it being purged from the population. Similarly to a "selective sweep", diversity will be reduced at the locus under negative selection, as well as its surrounding region due to the physical linkage, explaining why this source of linked selection is called background selection. Regions of functionally conserved genes are for instance more prone to this effect.

The **effective population size N_e** of a population, which is defined as the number of individuals of a theoretical panmictic population that would experience the same amount of drift, meaning random changes in allele frequency due to stochastic processes, as the real population experiences. Drift affects the whole genome equally. However, N_e will be locally decreased in regions affected by linked selection (described above) Under the influence of positive and background selection, N_e will therefore show a heterogeneous landscape along the genome. As *Ne* is informative about the intensity of genetic drift, the genetic diversity of a population depends to a large extent on N_e, in other words, on the ability to maintain this variation within the population, explaining why N_e is a crucial parameter to explain the within-genome variation in nucleotide diversity (π).

1.2 Understanding the genetic basis of speciation

Studying speciation is difficult since this process is long, occurring over tens or hundreds of thousands of generations. Without gene flow (i.e. under complete allopatry), the two or more diverging species progressively accumulate differences throughout the genome, in such a way that it is impossible to know which specific genomic differences could lead to genomic incompatibilities, if the two species were to meet again secondarily in a specific contact zone. In this respect, taxa diverging following a scenario of divergence-with-gene-flow offer unique opportunities to understand the genetic bases of speciation. In this

scenario, we can expect no gene flow at reproductive barrier loci, while the rest of the genome is expected to be homogenized by the interspecific gene flow. Therefore, to understand the genetic processes underlying speciation, it is beneficial to look at semi-isolated species which are still exchanging genes, in order to find the speciation genes contributing to reproductive isolation, as they will stand out as specific locations in the genome where there is no gene flow (Ravinet *et al.*, 2017). Therefore, hybrid zones are often considered as natural laboratories for the study of speciation (Hewitt, 1988). This heterogeneity in the levels of gene flow along the genome can be harnessed using scans of the genomic landscape of different summary statistics of divergence (meaning: these summary statistics are computed in non-overlapping sliding windows along the genome) to identify regions with elevated differentiation levels where these statistics deviate from neutral expectations. These regions can then be treated as candidate regions possibly containing genes contributing to reproductive isolation.

One of the summary statistics useful for detecting selection is the fixation index (F_{ST}), a measure of relative differentiation between two or more populations. This index is relative since it depends on the levels of within-population diversity, typically represented by the nucleotide diversity π as proposed by Hudson *et al.* 1992, as well as between-population nucleotide divergence, $\pi_{Between}$ (a measure of absolute differentiation):

$$FST = 1 - \frac{\pi}{\pi Between}$$

Divergent selection or selection against hybrids is expected to cause a peak in F_{ST}, since different alleles will be beneficial for the two populations, respectively, increasing $\pi_{Between}$, and also because selection will contribute to the fixation of the advantageous allele for each population in such a way that π will be reduced. If we assume a simple scenario with two fixed alleles, each providing adaptation to one of two different environments, no variation within each population is expected (π=0), but the diversity between the two populations will be high ($\pi_{Between}$>>0), in such a way that the ratio will be 0 and the F_{ST} will be 1. Therefore, scanning for loci exhibiting with elevated F_{ST} between diverging population or species to detect regions of interest has been proposed (Lewontin & Krakauer, 1973) and was then the focus of considerable methodological developments (e.g. Beaumont & Nichols, 1996; Beaumont & Balding, 2004, Foll & Gaggiotti 2008; see also Akey *et al.*, 2002).

Selection acting at a given locus is not the only force that contributes to peaks of F_{ST} in the genome, as proposed by Cruickshank & Hahn (2014) among others. Since F_{ST} is directly influenced by the levels of within-population diversity, regions of low nucleotide diversity will contain markers exhibiting elevated F_{ST}. This is particularly due to linked selection, either in the form of background selection or selective sweeps of universally adaptive alleles (e.g. Cruickshank & Hahn, 2014). If selection acts against hybrids due to reproductive isolation at a locus, $\pi_{Between,}$ a measure of absolute differentiation, exhibits a peak, since gene flow is inhibited around the selected locus, but homogenizes the rest of the genome. Regions around barrier loci to gene flow are therefore expected to show peaks in both F_{ST} and $\pi_{Between}$ (Irwin *et al.*, 2018). However, $\pi_{Between}$, while having the advantage of being independent of the levels of within-species diversity, also has its own handicap, since it is sensitive to ancestral diversity (Charlesworth, 1998). Therefore, several authors argued that

it is beneficial to combine F_{ST} and $\pi_{Between}$ to increase the reliability of the detection of barriers to gene flow (e.g. Irwin *et al.*, 2018).

1.3 Demographic inferences

Demographic events in the past have been shown to leave specific footprints in some summary statistics of the genetic diversity or divergence of present-day populations. For example, if a population experienced a bottleneck or size expansion in the past, this leads to a shift in the site frequency spectrum: after a population contraction, there are fewer rare alleles left, and after an expansion, there are more singleton alleles than in a population of constant size.

To reconstruct the evolutionary history of diverging species, and in particular to understand if species diverged with or without gene flow, a general strategy is to perform simulations assuming different scenarios regarding different important parameters, including the timing and amount of gene flow between the two species. By comparing the summary statistics computed from these simulated datasets to those of real present-day populations, it is possible to infer the demographic events in the past of those real populations by estimating demographic parameters of the real population, which is the concept behind demographic inferences. One approach to demographic inferences, Approximate Bayesian Computation (ABC), simulates genealogies under different demographic scenarios and compares them to the observed data using summary statistics. Then, the posterior distributions of the demographic parameters to be estimated are approximated from the differences between the simulated and observed summary statistics (Beaumont *et al.*, 2002).

1.4 Study System

The globally distributed genus *Populus* is native to the northern hemisphere and consists of almost 30 species (Zheng *et al.*, 2020; Park *et al.*, 2017). *Populus* represents a good study system for the different stages of speciation, since it contains species pairs that diverged several million years ago, as well as very recently diverged species pairs, spanning the whole speciation continuum (Shang, 2021; Shang *et al.*, 2020). It is ecologically important for wildlife-habitats and watersheds in boreal and temperate forests (Zheng *et al.*, 2017). The species belonging to this genus are wind-pollinated and exhibit a high effective population size, since they show extensive interspecific hybridization and vegetative propagation (Shang, 2021). Furthermore, they are highly adaptable to different environments (Zheng *et al.*, 2017).

This thesis focuses on the most recently diverged species pair *Populus davidiana* and *P. rotundifolia*. These two species have parapatric distributions (i.e. their ranges are adjacent to one another), have diverged about 1.3 Mya and are considered as representative of an early stage of speciation (Shang, 2021).

The species *P. davidiana*, which has a wide distribution range, can be found from northeastern to central China, as well as Mongolia, Korea and the far east of Russia (Zheng *et al.*, 2017). It is considered a keystone species in boreal-forest communities: it is known to prevent soil erosion, soil water loss, to regulate climate and retain ecological stability (Zhang, Wu & Li, 2004). Furthermore, it exhibits a high resistance to drought, cold and barren soils (Hou *et al.*, 2018). Since it naturally occurs mainly on slopes, ridges or gullies, natural populations of *P. davidiana* have been largely preserved from human activities, making it a good model system for phylogeographic analyses (Hou *et al.*, 2018). Growing in dry as well as wet climate, it has been shown to encompass different ecotypes that differ in their drought tolerance and exhibit different morphological and physiological adaptations to wet or dry climate, which are likely under genetic control: the wet climate ecotype shows faster growth rates than the dry climate ecotype (Zhang, Wu & Li, 2004).

P. rotundifolia occurs in the Himalaya and adjacent areas, with it much narrower range (as compared to *P. davidiana*) spanning the southeastern Qinghai-Tibetan Plateau, the Hengduan Mountains, and the Yunnan-Guizhou Plateau in southeastern China (as well as some parts of Bhutan) (Zheng *et al.*, 2017). Among all *Populus* species, it occurs at the highest elevation, between 2300m and 4500 a.s.l. (Sun, Gong & Liu, 2020).

The ranges of *P. davidiana* and *P. rotundifolia* overlap in the eastern Qinghai-Tibetan Plateau and central China (Zheng *et al.*, 2017). The two species occupy different ecological niches: while *P. davidiana* occurs in the lowland, *P. rotundifolia* occurs at high elevation (ecological niche modeling, Zheng *et al.*, 2017). It might therefore be of interest to identify loci involved in differential adaptation to their contrasting environments, that would represent ecological barriers to gene flow.

Previous studies have found support for admixture and recent introgression between *P. davidiana* and *P. rotundifolia* (Huiying, 2021; Zheng *et al.*, 2017). This suggests that there is ongoing migration between the two species, and that they have either experienced gene flow all along their divergence, as according to the evolutionary model “Isolation with Migration” (IM), or otherwise they went through a period of strict isolation in the past and then at some pint started to exchange genes again, as they experienced “Secondary Contact” (SC).

1.5 Research Objectives

This thesis has two main objectives: the first is to reconstruct the most likely evolutionary divergence scenario of this species pair, particularly regarding gene flow between the two species. Given the importance of gene flow in our understanding of speciation, a particular focus in these demographic inferences has been placed on the temporal variation in the levels of gene flow (i.e. past and ongoing), as well as the heterogeneity of the levels of gene flow along the genome. After this initial description of the history of their divergence, the second objective of this thesis is to scan genomes for regions exhibiting elevated levels of differentiation, and then identify genes falling in these regions to investigate if they could contribute to reproductive isolation.

2. Materials and Methods

2.1 Sequencing data

The data used in this thesis comes from a large recent resequencing project of 13 individuals of the species *Populus davidiana* and 17 individuals of the species *P. rotundifolia* (Shang, 2021; Shang et al. in preparation). For each individual, the sampling site is indicated in table S1. Briefly, for each sample, genomic DNA was extracted with the plant DNeasy mini kit (Qiagen, Germany) and then purified with the NucleoSpin gDNA Clean-up kit (Macherey-Nagel, PA, USA). 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. Among the 13 *P. davidiana* and the 17 *P. rotundifolia* samples, individual sequencing coverages range from 14X to 34X and from 14X to 43X, respectively (Shang *et al.* in preparation, table S1).

2.2 Read mapping and SNP calling

The pipeline used for this project is publicly available on a github repository (https://github.com/ThibaultLeroyFr/PopGenomicsWithDILS). To speed up the computation, these mapping and SNP calling steps were parallelized on several CPUs using two HPC computing clusters (University of Montpellier, France & VSC-4 cluster of the University of Vienna, Austria).

The reads were trimmed with Trimmomatic (v.0.39), using the following parameters: LEADING:3, TRAILING:3, SLIDINGWINDOW:4:15 and MINLEN:50.

The reference genome version 2.2 of *Populus tremula* (Schiffthaler, *et al.* 2019) was then indexed and the trimmed reads were mapped against this reference using BWA (v.0.7.17, Li *et al.* 2013 with the BWA-MEM algorithm (default parameters). Unmapped reads and mapped reads with a low quality (MQ below 20) were discarded. Potential PCR duplicates were then removed using MarkDuplicates (v. 1.140, Picard tools, Broad Institute, 2019, available from http://broadinstitute.github.io/picard/).

In the next step, variants were called using GATK (v. 3.7). First, HaplotypeCaller was used using default parameters on single samples (gVCF) to call SNPs for each species individually. Then, joint genotyping was performed for both species together using the "GenotypeGVCFs" programme. We generated a complete VCF for each species, including both SNPs and non-variant sites, to be able to generate accurate nucleotide diversity estimates for each of the two species (see below). To ensure high quality in our final dataset, we filtered out low-quality SNPs using a series of settings: a quality by depth (QD) < 2.0, a Fisher Strand (FS) bias > 60, a mapping quality (MQ) < 40, a MQranksum < 2 or a

ReadPosRankSum < 2 or a Raw Mapping Quality (Raw_MQ) < 45,000. SNPs satisfying at least one of these conditions were considered as non-variant sites.

2.3 FASTA sequence reconstruction

Given the diploidy of the two focal species, two fasta sequences per individual were reconstructed from the VCF file. Positions with a minimum coverage per individual lower than 3 or higher than 50 were 'hardmasked' with "N" (only base calls with a minimum Illumina quality of 20 were considered). Insertion-deletion mutations (indels) were also excluded in order to keep the sequence length the same than for the reference genome. As a consequence, nucleotide diversity as well as all other summary statistics reported below are only based on SNPs. All positions satisfying these coverage settings were added to the sequence, either by adding the reference allele (non-variant) or by considering the genotype (variant position, i.e. reference or alternate allele depending on the genotypes called at the SNP position for a given individual, see Leroy *et al.*, 2019 for details).

Non-overlapping 10-kbp sequences spanning the whole reconstructed chromosomes were generated using bedtools getfasta (Quinland &Hall, 2010). In addition, CDS were extracted from the reconstructed fasta sequences using the gff of the reference genome v.2.2 of *P. tremula* and the script cutSeqGff.py (Leroy *et al.*, 2019).

2.4 Demographic analyses

To infer the history of divergence and gene flow between *P. davidiana* and *P. rotundifolia,* demographic analyses were performed using the software DILS (Demographic Inferences with Linked Selection, Fraïsse *et al.*, 2021). This approach implements an ABC framework, i.e. a likelihood free method. In a nutshell, a large set of summary statistics (see table 1) is calculated from the real data. Independently, a very large number of simulations under various evolutionary scenarios (see below) are performed and for each simulation the same summary statistics as for the real data are computed. One of the main objectives of the ABC procedure is to identify the demographic model that best fits the data. In other words, the strategy here assumes that the best demographic scenario is the one able to reproduce the observed summary statistics. This best demographic scenario is then used to estimate some demographic parameters of the real dataset in order to give a more detailed description of the likely demographic history of the studied population(s).

An important summary statistic used by DILS, which can be applied to infer past changes in effective population size, is Tajima's D, calculated as π, the average number of pairwise differences between nucleotide sequences, and Watterson's θ, the total number of polymorphic sites ($D= \pi - \theta$; Tajima, 1983; Watterson, 1975). Assuming a mutation-drift equilibrium and constant population size, π and θ are expected to be equal (i.e. D=0). π tends to underestimate the number of rare alleles. Therefore, a negative genome-wide Tajima's D indicates an excess of rare alleles, as observed after a recent population expansion (for example following a bottleneck). On the other hand, after a population

contraction, rare alleles tend to have vanished, leading to a deficit of rare alleles and therefore a positive Tajima's D (Tajima, 1989).

The real data we used for these demographic inferences are either sets of 250 non-overlapping 10kb sliding windows sampled by random among all sliding windows spanning the genome or sets of 1,000 CDS sampled by random among all gene models. This random sampling procedure was performed 10 times in order to check the consistency of the demographic analyses with different sets of genomic windows or genes. Once imported in DILS, each dataset was filtered to remove short or low quality sequences as follows: a maximum proportion of unknown nucleotides allowed in the sequence of a locus ≤ 10%; a minimum number of 100 treatable sites allowed in the sequence of each locus and a minimum number of 12 haploid sequences per locus (i.e. 6 individuals) per species.

Four main demographic scenarios of divergence are implemented in DILS. All these scenarios assume the split of an ancestral population with an effective population size (N_{anc}) into two populations or species (pop1 and pop2) at the time T_{SPLIT}. The two species can have different effective population sizes (N_{pop1} and N_{pop2}) (figure 1). Among these four scenarios, two assume no ongoing gene flow, and two assume ongoing gene flow. The first scenario without ongoing gene flow is the Strict Isolation (SI) scenario, where the two species diverged in complete absence of gene flow since the two species split (figure 1). The second model is Ancient Migration (AM), which considers that the species exchanged genes for a period in the beginning of their divergence but stopped at some point in the past (T_{AM}, figure 1). The other two scenarios assume ongoing gene flow: one of them is Isolation with Migration (IM), according to which the two species continuously exchanged genes since T_{SPLIT}. The second model with ongoing gene flow is secondary contact, meaning that there was a period of strict isolation of the two species in the past, but at some point they started exchanging genes again (T_{SC}, figure 1).

DILS hierarchically tests the fit of the different scenario to the real data. First models with or without ongoing migration are compared, and then the software tests among the two models within each of those two categories, respectively (see figure 1). If migration is inferred, the software also infers whether the migration rate is heterogeneous or homogeneous along the genome. Furthermore, the DILS infers (in any case) whether the effective population size is heterogeneous along the genome. Furthermore, based on the simulated datasets under the model that best fits the data, the software will estimate some demographic parameters of the real dataset (e.g. effective population size, migration rate etc.).

Table 1: Summary statistics used for model comparisons in DILS (adapted from the DILS manual available under https://github.com/popgenomics/DILS_web/blob/master/manual.pdf). For each summary statistic, the average and the standard deviation are used.

General
dataset = index
bialsites = number of sites per locus
Summarized SFS
sf = fraction of sites with a xed di

erence between populations
sxA, sxB = fraction of sites with a polymorphism specific to each population
ss = fraction of sites with a polymorphism shared between populations
Polymorphism
piA, piB = pairwise nucleotide diversity (π) for each population (Tajima, 1983)
thetaA, thetaB = Watterson's θ for each population (Watterson, 1975).
Tajima's D
DTajA, DTajB = Tajima's D for each population (Tajima, 1989).
Differentiation and divergence
divAB = raw divergence (D_{xy}) between populations (Nei & Li, 1979)
netdivAB = net divergence (D_a) between populations, measured by D_{xy} - (π_A + π_B)/2 (Nei & Li, 1979)
FST = F_{ST} measured by 1 – π_S/π_T; where π_S is the average nucleotide diversity in each population and π_T is the total nucleotide diversity over the populations (Wright, 1943).
Pearson's correlation between populations/species and across loci
pearson_r_pi = between π_A and π_B
pearson_r_theta = between θ_A and θ_B
pearson_r divAB_netDivAB = between div_{AB} and $netDiv_{AB}$
pearson_r_divAB_FST = between div_{AB} and F_{ST}
pearson_r_netDivAB_FST = between netDivAB and F_{ST}

The following prior parameter distributions were selected (only the ones deviating from the default settings are indicated): Mutation rate = 3.75 x 10^{-8} per site per generation and a generation time of 15 years (Wang et al. 2016, as used in Shang et al. in preparation); Migration rates: M_min=0.4 (default) & M_max=200, model for barriers= bimodal; effective population sizes: N_min=100 (default) & N_max=1,000,000 (default); speciation time (T_split): T_split_min=100 (default) & T_split_max=1,000,000 (default) (DILS manual available under https://github.com/popgenomics/DILS_web/blob/master/manual.pdf).

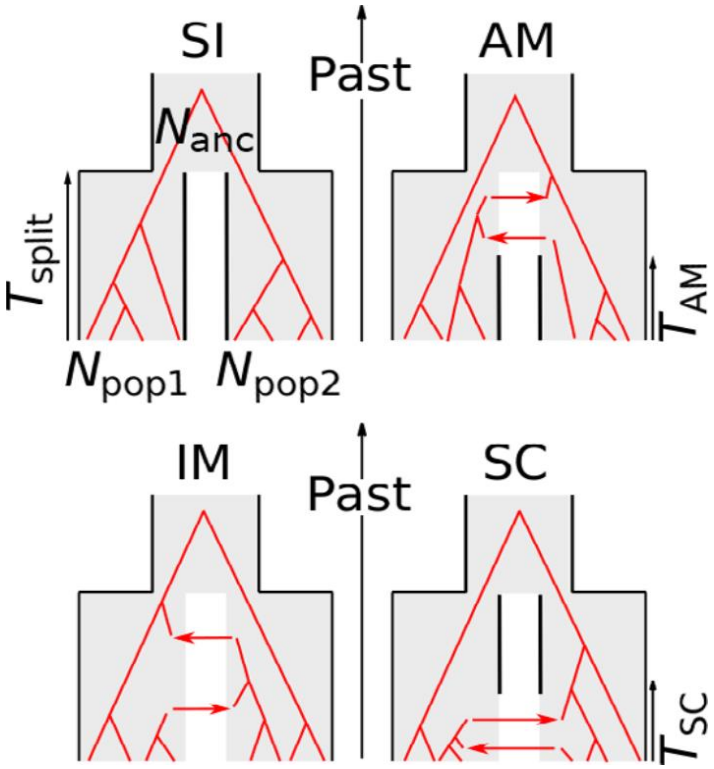


Figure 1: Four main different evolutionary scenarios implemented in DILS. (SI=Strict Isolation; AM=Ancient Migration; IM=Isolation with Migration; SC=Secondary Contact; N_{anc}= ancestral effective population size; N_{pop1} & N_{pop2}= effective population sizes of populations 1 & 2; T_{split}= time of split; T_{AM}= date for the cessation of gene flow; T_{SC}= age of the secondary contact) (after Fraisse et al., 2021).

Particular attention has been paid to model checking. First, the model fit to the real data was checked by performing a series of goodness-of-fit tests (once per summary statistic) to identify the summary statistics that significantly differ between the model and the real data. Second, I visually inspected the joint site frequency spectra (jSFS) of the real data, the model, as well as the residuals between them as implemented in DILS (Fraisse *et al.*, 2021) to verify that the models provide an overall good fit.

2.5 Genome-wide scans for genetic divergence

The rationale was to use genome scans to detect regions exhibiting an excess of relative and/or absolute divergence to identify candidate genes potentially contributing to reproductive isolation between *P. davidiana* and *P. rotundifolia*. We also investigated regions with a deficit of nucleotide diversity and negative Tajima's D (Tajima, 1983; Tajima, 1989) that could be consistent with selective sweeps.

Nucleotide diversity (π, sometimes referred as Tajima's π), the average number of differences between pairs of sequences, was calculated in 10kb genomic sliding windows for each of the two species (Tajima, 1983). Likewise, Tajima's D was calculated as the difference between π and Waterson's θ, another estimator of genetic diversity (Tajima, 1989) computed as the total number of segregating sites, divided by a harmonic number. One important property is that π and θ are assumed to be equal under a scenario of mutation-drift equilibrium and constant population size (then, Tajima's D=π-θ=0). However, since π

tends to underestimate to number of rare alleles, a negative Tajima's D value indicates an excess of rare alleles, which can be expected for a selective sweep.

The median π across the whole genome was calculated for each of the two species, and the difference was considered as an indication of differences between their respective effective population sizes (N_e). Indeed, given the recent divergence of these two *Populus* species, we can assume that they have a relatively similar mutation rates. As a consequence, since the within-species diversity is $\theta = 4Ne\mu$, the observed differences in genetic diversity are expected to translate into differences of long-term N_e. Importantly, N_e can change drastically over relatively short periods of time, for example due to climatic oscillations. The Tajima's D metric is known to capture these recent changes quite well. When Tajima's D is computed over the whole genome and most windows exhibit a deviation from D=0, it is an indication of possible effective population size changes in the recent past, leading to a genome-wide excess or deficit of rare alleles.

For all 10kb sliding windows, we computed the Hudson's F_{ST} (Hudson *et al.*, 1992), as well as $\pi_{Between}$, a measure of absolute differentiation sometimes also referred as D_{XY}, were calculated over sliding 10kb genomic windows, using a pipeline developed by Benoit Nabholz (Univ. Montpellier, France) and made available at https://github.com/benoitnabholz/seq_stat_2pop/releases/tag/v1.1. We excluded windows exhibiting very negative Hudson's F_{ST} values ($F_{ST} < 0.1$). In addition, we computed F_{ST} for each SNP of the dataset. Based on the empirical distribution, we identified the quantile corresponding to the top 1% of F_{ST} values. We then estimated a per-window proportion of highly differentiated SNPs using non-overlapping 10kb genomic sliding windows spanning the whole genome.

F_{ST} and $\pi_{Between}$ are both expected to exhibit peaks in regions contributing to reproductive isolation as compared to the genomic background (Irwin *et al.*, 2018). We therefore used the following two criteria to detect the regions of interests: windows in the top 1% for $\pi_{Between}$ and with a proportion of at least 50% of highly differentiated SNPs. In addition to these metrics, Tajima's D and π were also analyzed to also investigate if the potential barrier could be due a recent sweep in one of the two species.

For all detected regions, we performed functional annotations for all genes falling in these regions, or in the immediate vicinity of these detected windows (+5kb on both sides) to take into account that the potential genes could be adjacent to the detected genomic window. Coding sequences of these genes were extracted using the position indicated in the gff. Then, for each gene, a blastn search against the general database NR (NCBI) was performed to find some homologous genes that could be already described and annotated in some other species. Then, extensive manual literature searches were performed to recover the potential function of each gene and if this gene could contribute to adaptation or speciation in the plant literature. This strategy was used as an alternative to Gene Ontology (GO) analysis, in order to provide more precise information regarding the possible roles these genes could play in reproductive isolation, including for the differential adaptation to the different environments that *P. davidiana* and *P. rotundifolia* inhabit.

3. Results

3.1 Nucleotide diversity and Tajima's D

To investigate possible differences in nucleotide diversity between *P. davidiana* and *P. rotundifolia*, a VCF file was generated independently for each species. Despite a lower number of individuals used in *P. davidiana*, I identified more SNPs (16,524,461 SNPs) in the genome of *P. davidiana* than in the genome of *P. rotundifolia* (13,655,815 SNPs). Accordingly, the median per site nucleotide diversity (π) of *P.davidiana*, estimated as 0.0127, was statistically higher than that of *P. rotundifolia*, estimated as 0.0097 over the whole genome. This means that, on average, each *P. davidiana* and *P. rotundifolia* individual exhibits a heterozygous site every 79 bases and 103 bases, respectively. As expected due to the effect of background selection, the nucleotide diversity at synonymous sites (π_s) of all coding regions was slightly lower than on the rest of the genome, but consistent with a substantial difference in their nucleotide diversity levels. On the coding regions, I estimated a median π_s value of 0.0113 for *P. davidiana*, and a median π_s of 0.0084 for *P. rotundifolia* (figure 2).

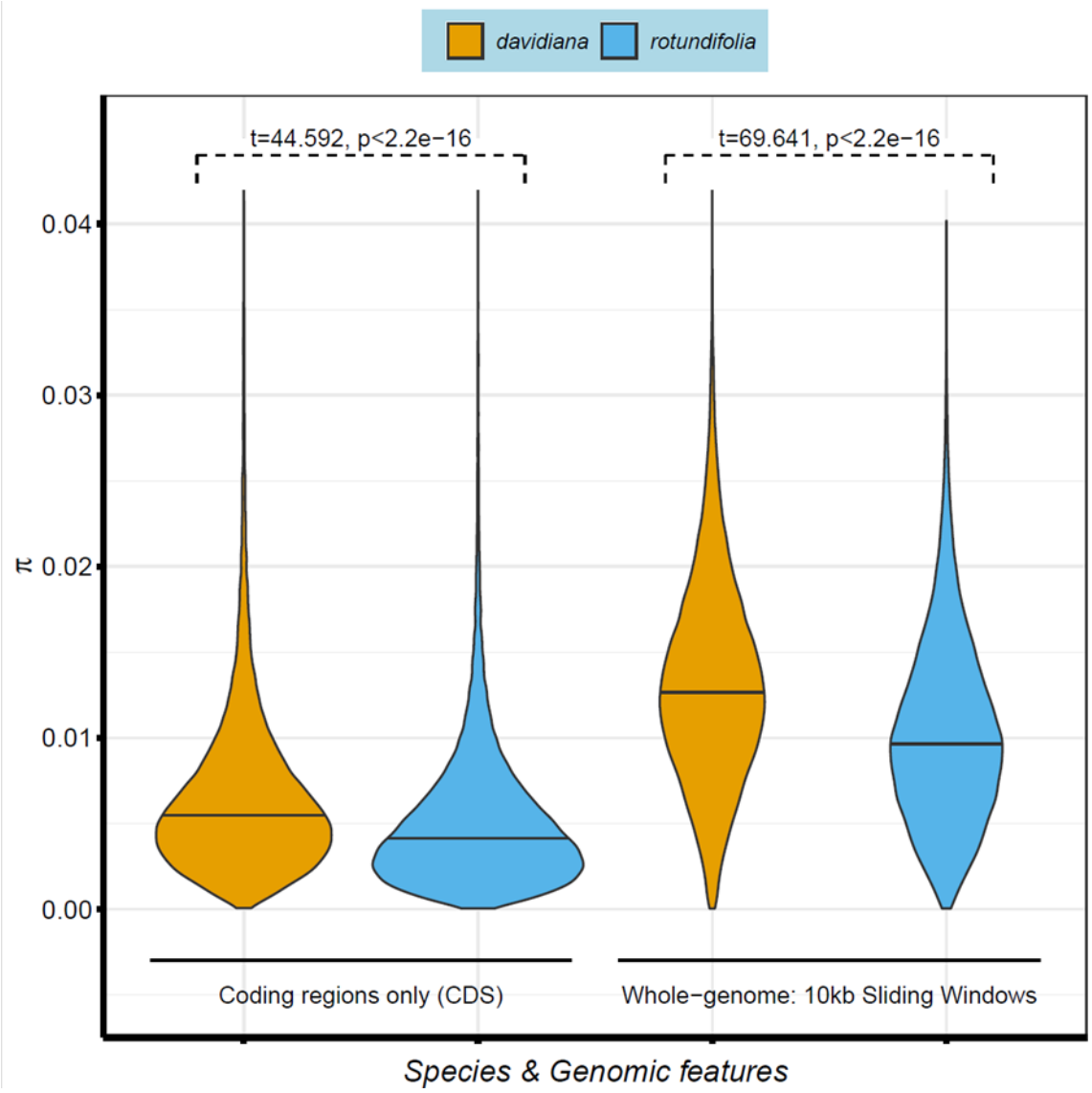


Figure 2: Nucleotide diversity along the genome (π) or at synonymous sites of coding regions (π_S). These estimates were calculated over 10 kb sliding windows (left), or across all coding sequences (right) for both P. davidiana and P. rotundifolia, respectively. Pairwise t-tests were performed to test if the differences between the two species were significant.

The median Tajima's D values of two species are both slightly positive. The median Tajima's D of *P. davidiana* is 0.470, the median Tajima's D of *P. rotundifolia* is 0.814 (figure 3). The coding sequences showed slightly lower median Tajima's D for both species, but still consistent with slight deviations from 0: for *P. davidiana* it was estimated to be 0.118736, and for *P. rotundifolia 0.391704*, which is statistically higher than the Tajima's D of *P. davidiana* (figure 3).

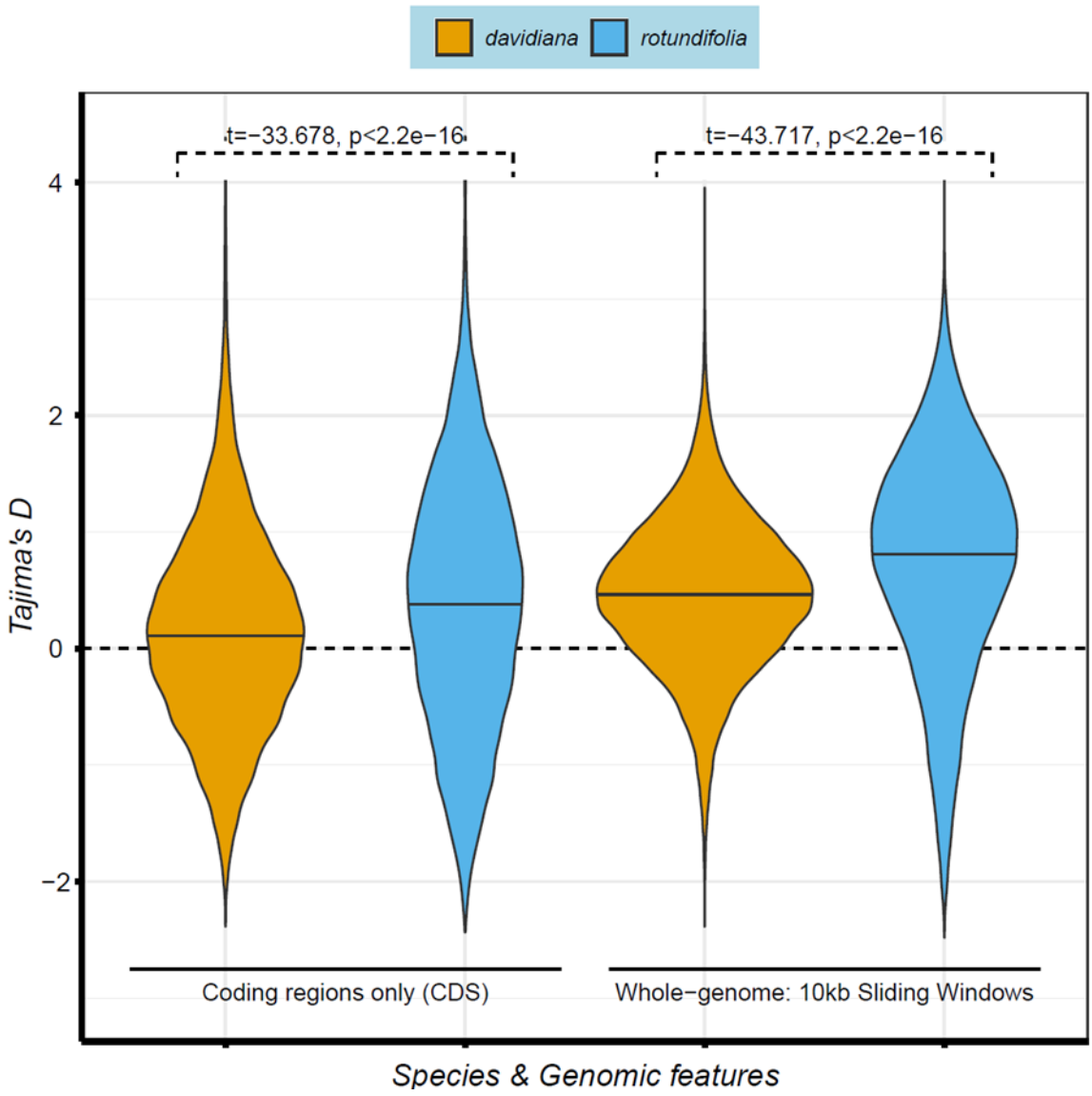


Figure 3: Genomic variation calculated over 10 kb sliding windows (left) and variation over the coding sequences (right) of Tajima's D values for P. davidiana and P. rotundifolia, respectively.

In addition to all these species-specific investigations, a joint genotyping was performed using the samples of both species. Our subsequent analyses will focus on this dataset containing both species. This VCF contains 19,062,189 SNPs. Using this dataset, a median genome-wide F_{ST} (Hudson) value of 0.0893, and a median genome-wide $\pi_{Between}$ value of 0.0112 were estimated, suggesting an overall low level of differentiation between these species.

3.2 Demographic analyses

To learn more about the demographic history of these species, I then used the previously described summary statistics, as well as some others (39 in total, see the

Materials and Methods section for a full list), in order to infer the demographic history of these two species using an ABC framework.

a) Model choice:

These tests were performed for ten randomly selected sets of 250 10kb genomic windows, as well as 1000 randomly selected coding sequences, respectively (inferred models and posterior probabilities for all tests are shown in tables S2 & S3). All inferred scenarios in the tests on 10kb genomic windows supported ongoing gene flow (and most in the tests on CDS), either isolation with migration or Secondary contact.

Assuming constant effective population sizes since the divergence, the model "Isolation with Migration" (IM) was always inferred as the best model for all sets of 10kb genomic windows (with posterior probabilities for migration ranging from 0,72 – 0,92, and for IM from 0,75 – 0,88). Assuming the same models but allowing for variation in effective population sizes during the divergence of *P. davidiana* and *P. rotundifolia*, seven sets supported SC (post. prob.: 0,57 – 0,89 for migration ; 0,70 – 0,82 for SC), while only three sets supported IM (post. prob.: 0,63 – 0,75 for migration; 0,63 – 0,75 for IM). Strong support for models with migration was also observed based on the sets of CDS, albeit that five tests inferred isolation with a relatively low posterior probability (table S3).

The migration rate was inferred to be heterogeneous along the genome in nine out of the ten tests based on genomic windows assuming historical variation in population sizes, while under the assumption of constant population sizes it was only inferred to be heterogeneous for only three of the ten sets, suggesting that the historical variation of N_e during the divergence vs. heterogeneity along the genome are difficult to tease apart based on our simulations. In the tests on the coding sequences, the migration rate was detected heterogeneous along the genome in only four sets among the fifteen tests that inferred migration: once for scenarios assuming constant population sizes and three times for scenarios assuming variable population sizes.

b) Parameter estimations:

Under the best evolutionary scenario "Isolation with Migration", which was inferred across all tests under the assumption of constant effective population sizes, the estimated current effective population size of *P. davidiana* (N1) is larger than that of *P. rotundifolia* (N2), with N1 ranging from 28k – 59k, and N2 ranging from 10k – 29k (shown for tests on 10kb genomic windows in figure 4). Assuming variable population sizes, this pattern was not recovered, and posterior estimates for current effective population sizes had a large range (range of N1= 24k – 407k, range of N2= 12k – 738k) as compared to the estimates based on the scenarios assuming constant N_e. The opposite was observed for the ancestral effective population size (Na), that does not vary much under the assumption of variable population sizes, ranging from 28k – 60k across all 20 tests, but varied a lot assuming constant population sizes (range = 4k – 804k) (shown for tests on 10kb genomic windows in figure 5).

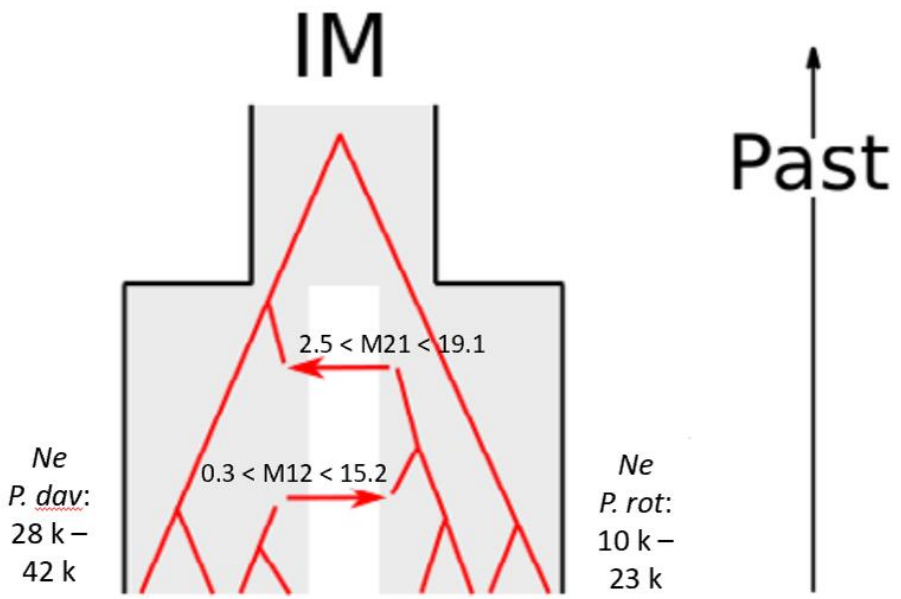


Figure 4: Best evolutionary scenario (IM= Isolation with Migration) inferred in the tests on 10 kb sliding genomic windows, and range of estimated median effective population sizes of P. davidiana (N1) and P. rotundifolia (N2) among the different sets, as well as the range of estimated median effective migration rates between them (M, in each direction, respectively) under the assumption of constant effective population sizes.

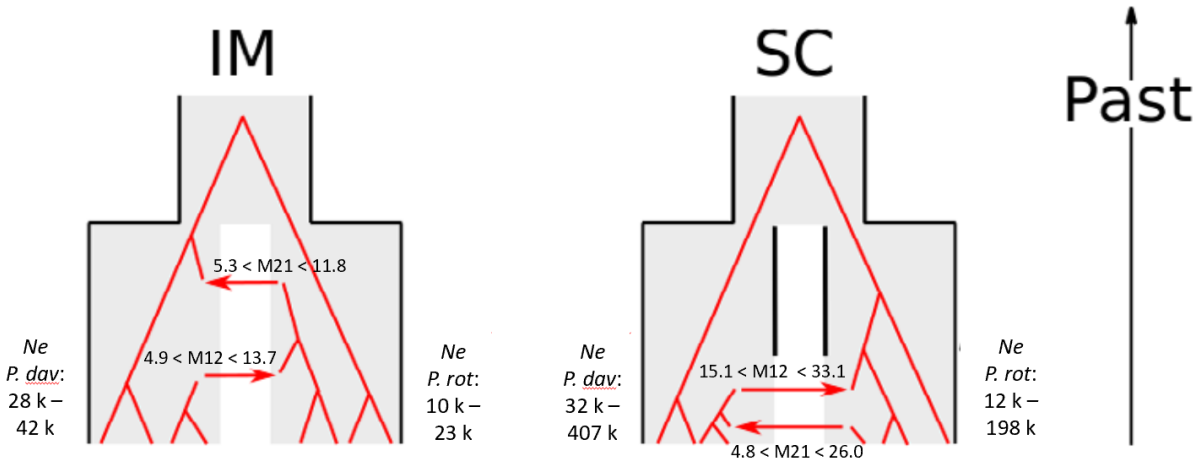


Figure 5: Best evolutionary scenarios (IM= Isolation with Migration or SC= Secondary Contact) inferred in the tests on 10 kb sliding genomic windows, and range of estimated median effective population sizes of P. davidiana (N1) and P. rotundifolia (N2) among the different sets, as well as the median range of estimated median effective migration rates between them (M, in each direction, respectively) under the assumption of variable effective population sizes.

For models assuming variable population sizes, recent population growth of *P. davidiana* has been inferred in nine out of ten tests on the coding sequences, as well as nine out of ten tests on the sliding windows. For the two last tests, population contraction was inferred. Similarly, population growth of *P. rotundifolia* was inferred in 7 out of 10 tests on the coding sequences and in 7 out of 10 tests on genomic windows, while population contraction was inferred for the 6 other tests.

c) Model checking:

Model checking was then performed to see if the best inferred scenarios fitted the data well: the expected and observed values differed significantly for only up to three of the 39 summary statistics used in each divergence test, respectively. Another possible validation is to look at the 2D-Site-Frequency-Spectra (also referred as a joint SFS: jSFS). Visual comparison between the observed jSFS and the jSFS for the best model (hereafter the expected jSFS) through their residuals also supported a good fit for all tests, except for the variants that are rare in one of the two species and absent in the other (figures 6 and 7). Two further examples are shown for the sets of coding sequences in the appendix in figures S1 & S2).

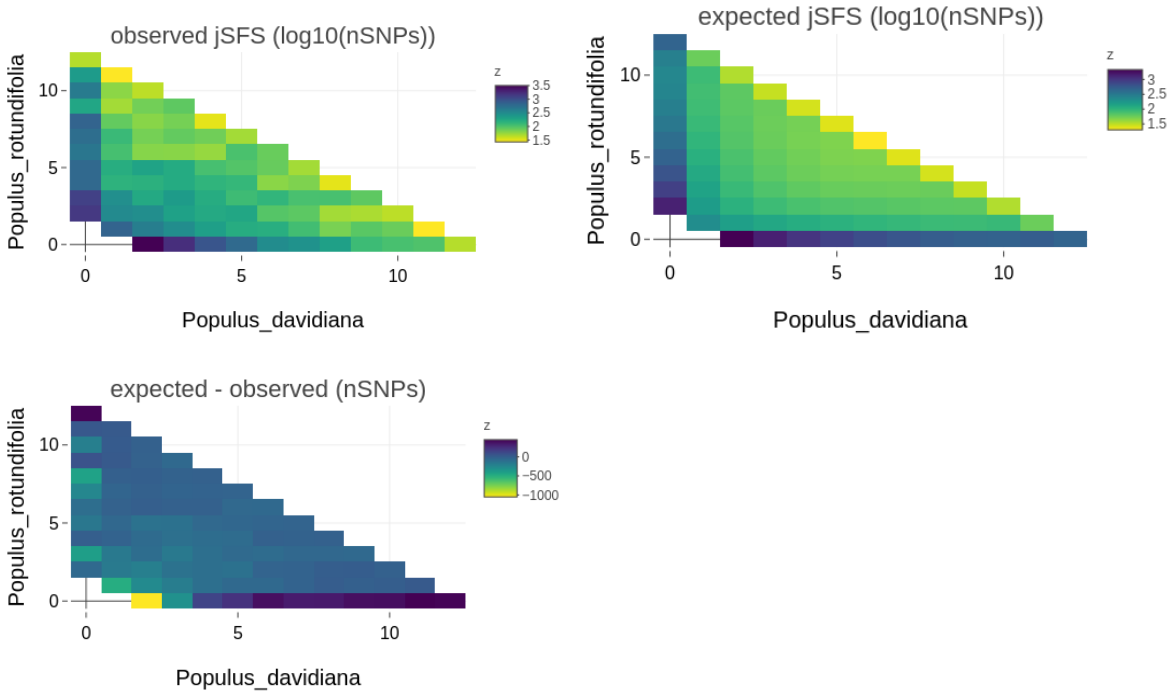


Figure 6: 2D-Site-Frequency-Spectra for the real data (top left), the best model (top right) and the residuals between them (model-data, bottom), respectively, shown for one of the divergence tests on 10 kb sliding windows (under the assumption of constant population sizes).

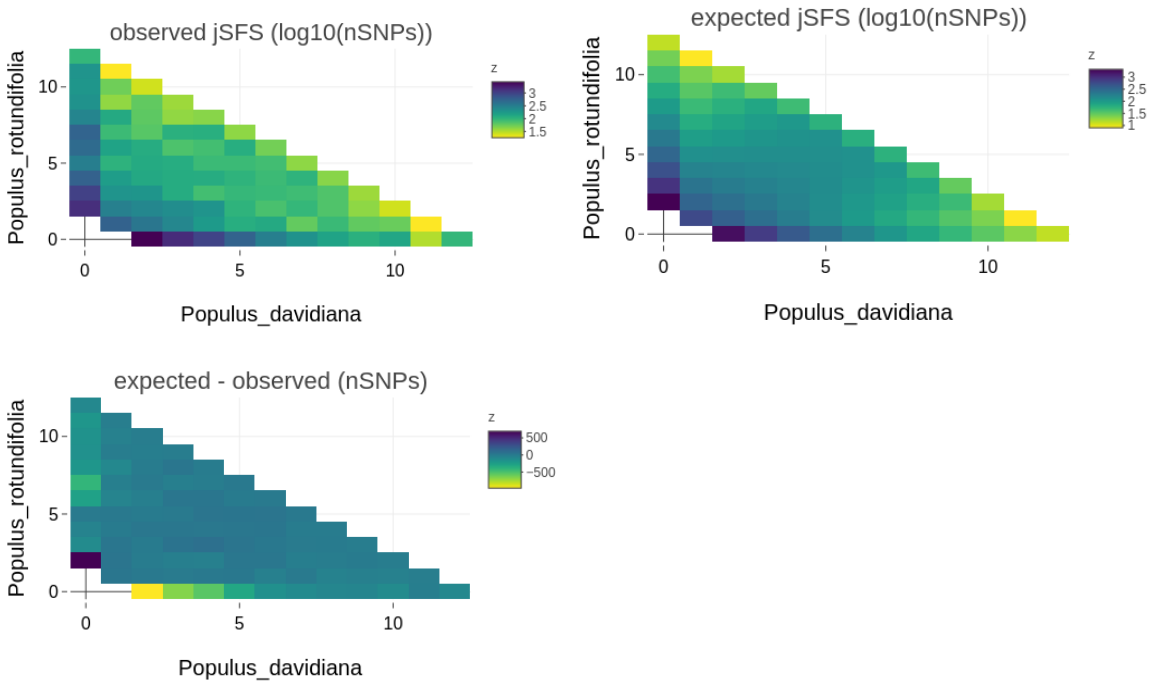
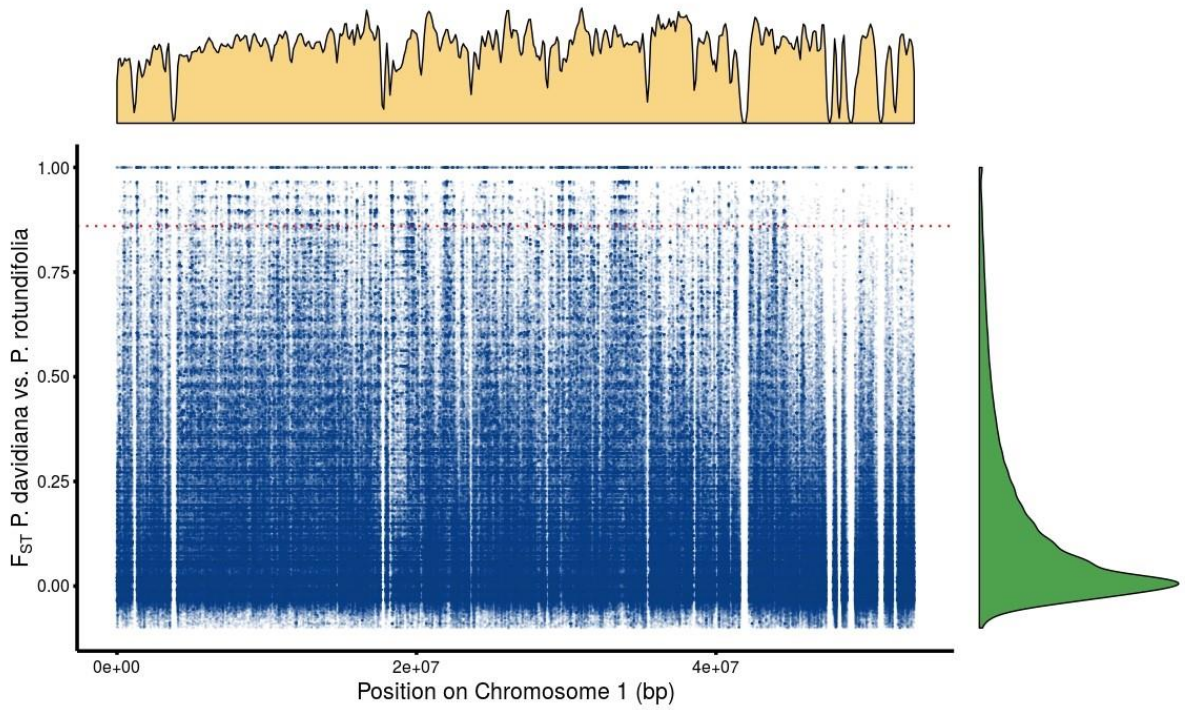


Figure 7: 2D-Site-Frequency-Spectra for the data (top left), the model (top right) and the residuals between them (model-data, bottom), respectively, shown for the best inferred scenario in one of the tests on 10 kb sliding windows (under the assumption of variable population sizes).

3.3 Genome scans for regions of high differentiation

Regarding the pattern of relative divergence, I observed a lot of fixed differences (F_{ST} =1) at the SNP level throughout the genome (figure 8). As a consequence, it was impossible to use a classic genome scan procedure to identify the SNPs exhibiting peaks in F_{ST}. Therefore, a different strategy was used by focusing only on the local enrichment in SNPs exhibiting F_{ST} values that are in the top 1% of all F_{ST} values, corresponding to an F_{ST} of at least 0.860 (red dotted line, figure 8), by computing the proportion of these highly differentiated SNPs across the genome using 10kb genomic windows. As an example, the variation in Hudson's F_{ST}, the proportion of highly differentiated SNPs and $\pi_{Between}$ are shown for chromosome 17 in figure 9. We can clearly identify a region exhibiting a majority of SNPs exhibiting a very high F_{ST}, one of the highest F_{ST} values among all genomic windows and a very high $\pi_{Between}$ (red square, Figure 9). I generalized this approach by looking at all windows exhibiting these patterns: see below for a scatterplot between the proportion of highly differentiated SNPs and $\pi_{Between}$ (figure 10). (The scans for all chromosomes are shown in figure S3.)



<u>Figure 8</u>: F_{ST} values calculated for each SNP along chromosome 1 (as an example) (SNPs shown in blue). A lot of highly differentiated SNPs between P. davidiana and P. rotundifola can be identified throughout the genome, includingSNPs with F_{ST} values equal to one. Density curves of F_{ST} values and for the density of SNPs are shown on the the left and on top, respectively. The threshold value (0.860) corresponding to the top 1% of F_{ST} values among all SNPs (<u>i.e.</u> not only on chromosome 1) is indicated with the red dotted line.

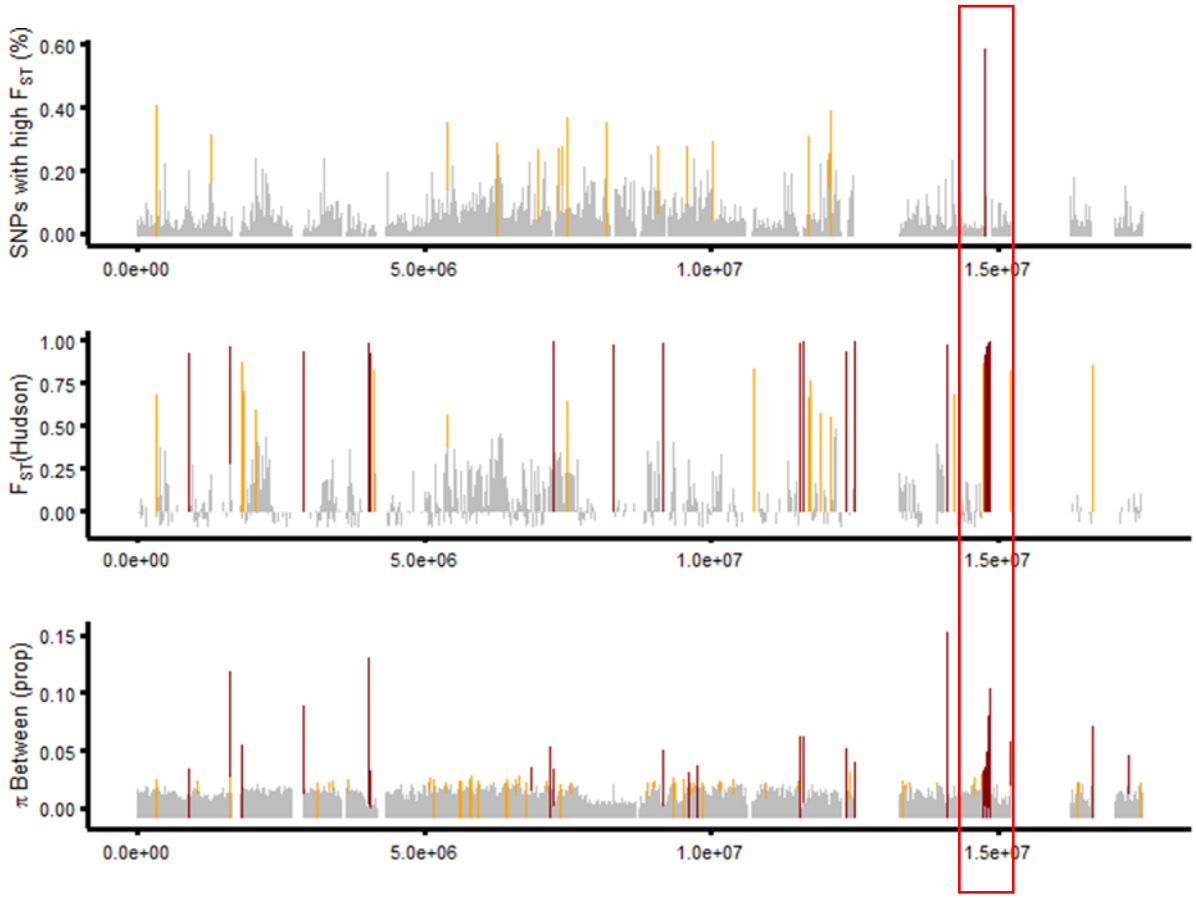


Figure 9: Genome scans for the proportion of highly differentiated SNPs (top panel), Hudson's F_{ST} (middle panel) and $\pi_{Between}$ (bottom panel), computed over 10kb non-overlapping genomic sliding windows for chromosome 17 (here shown as an example). Colored bars indicate windows within the top 5% (orange) or 1% (red) of $\pi_{Between}$ or top 5% (orange) and 1% (red) Hudson's F_{ST} values, or with a high proportion of highly differentiated SNPs: 0.25 (orange) or 0.5 (red). The red rectangle shows a region of interest, which was searched for reproductive barrier loci to gene flow.

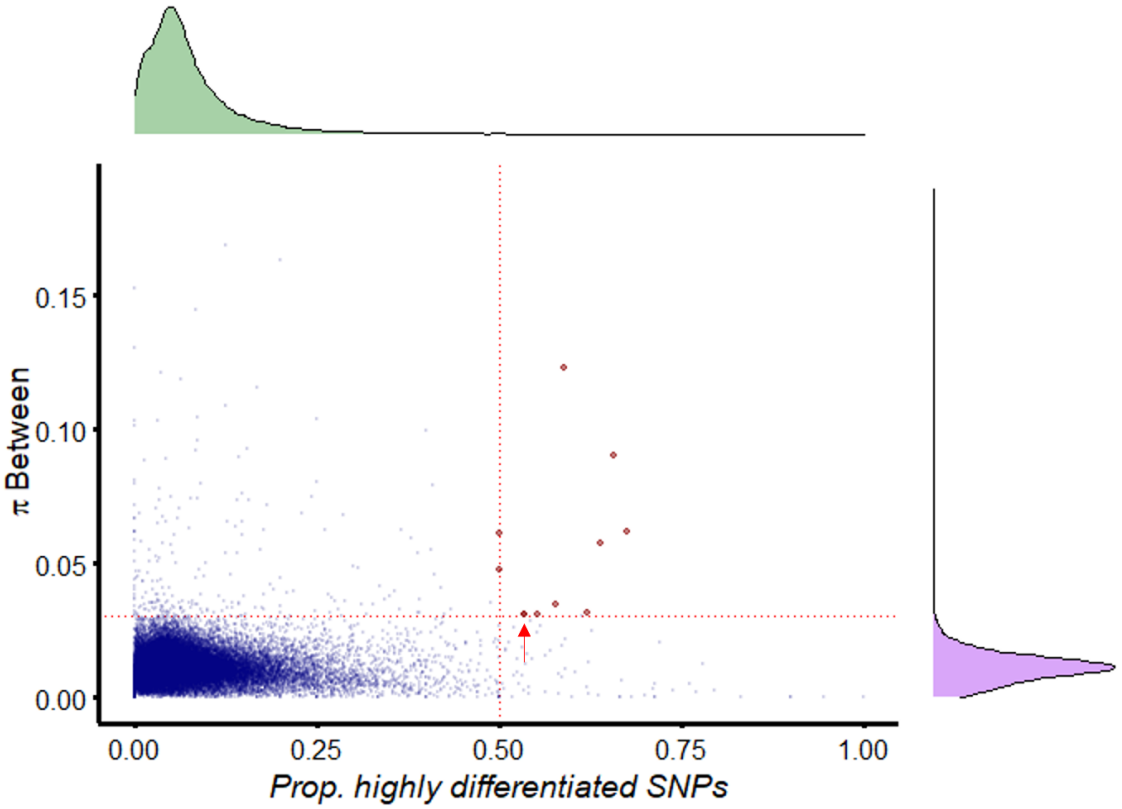


Figure 10: Scatterplot between the proportion of highly differentiated SNPs and $\pi_{Between}$. The red dotted lines indicate the thresholds of the top 1 percentile of $\pi_{Between}$ (0.0306) and a proportion of highly differentiated SNPs of 0.5. The red dots indicate the eleven windows above these two thresholds (there is one dot which represents two windows with very similar values, indicated by the red arrow). Density curves are shown for the proportion of highly differentiated SNPs (green) and $\pi_{Between}$(purple), respectively.

Among the 40,090 genomic windows, we found eleven genomic windows that have a proportion of highly differentiated SNPs of at least 0.5, as well as a $\pi_{Between}$ value within the top 1% of values (figure 10, table S4). These eleven windows exhibited a median proportion of highly differentiated SNPs of 0.5776, a median Hudson's F_{ST} of 0.9660, and a median $\pi_{Between}$ of 0.0479. Over the whole genome, the median proportion of highly differentiated SNPs is 0.0586, and the median Hudson's F_{ST} and $\pi_{Between}$ are 0.0893 and 0.0112.

3.4 Candidate genes for reproductive isolation

An early attempt to identify some candidate genes was then performed. First, among these eleven windows/regions with elevated divergence levels between the two species or in their immediate vicinity (5kb flanking regions), I identified twelve coding sequences, with up to 4 genes per region. These genes are observed on six of the 11 regions (table 2). Using blastn on NR (NCBI), I observed e-values=0 for 10 genes (one CDS sequence has no blast results on NR and another is only described as an uncharacterized locus). Extensive literature searches were performed on these genes as reported in the discussion.

Table 2: Coding sequences found in the eleven genomic windows listed in table S4 or in their immediate vicinity (5kb flanking regions). For each coding sequence, their genomic position (chromosome, start and end position), their annotated gene name retrieved through a blastn against the NR database from NCBI, the species in which the gene was found, as well as the E-value (similar to a P-value) of the blastn search are indicated.

chr	start position	end position	Gene name	Species (blastn)	E-value (blastn)
chr1	18345174	18346251	cucumber peeling cupredoxin	*Populus trichocarpa*	0
chr1	18351047	18354593	At2g27500 glucan endo-1,3-beta-glucosidase 14-like	*Populus alba*	0
chr12	11746551	11753714	**RHD3=Root Hair Defective 3 homolog 2**	*Populus trichocarpa*	0
chr13	8852252	8853097	At5g39130 Germin-like protein subfamily 1 member 16	*Populus alba*	0
chr13	13435590	13436624	S-adenosylmethionine synthase 2	*Populus alba*	0
chr13	13440666	13444128	uncharacterized LOC118060889	*Populus alba*	0
chr13	13444873	13445700	no blast results on NR		
chr13	13448071	13454060	probable receptor-like protein kinase At2g42960 or At5g18500	*Populus alba/tremula*	0
chr16	12108071	12111855	ASNAP alpha-soluble NSF attachment protein-like	*Populus trichocarpa*	0
chr16	12111737	12120591	PIGA phosphatidylinositol N-acetylglucosaminyltransferase subunit A	*Populus alba*	0
chr4	10495995	10500830	**WRKY3, probable WRKY transcription factor 3**	*Populus trichocarpa*	0
chr4	10509745	10513880	uncharacterized in Populus ,referred as a WRKY11 (Dimocarpus longan, evalue=0)	*Populus alba*	0

4. Discussion

4.1 Divergence with gene flow between *P. davidiana* and *P. rotundifolia*

The first objective of this work was to better understand the evolutionary history of *P. davidiana* and *P. rotundifolia,* in particular regarding the levels of interspecific gene flow. The results obtained from the demographic analyses strongly support ongoing gene flow between *P. davidiana* and *P. rotundifolia*: This result was supported by all tests on our ten sets of genomic windows. However, some of the tests on the sets of coding sequences supported current isolation between the two species (2 tests assuming constant population sizes, and three tests assuming variable population sizes). DILS has been shown to

distinguish well between scenarios with ongoing migration vs. isolation (Fraïsse *et al.*, 2021), so this result is surprising. It suggests that this is caused by specific properties of the sets of coding sequences used for these two tests, for which isolation has been inferred. They could, for example, contain several loci involved in reproductive isolation for which levels of local gene flow are strongly reduced. It should be noted that the sets supporting isolation exhibit a relatively low posterior probability, suggesting an ambiguous situation between isolation and migration, which could be expected for inferences based on a combination of neutral and barrier loci. That fact that one of these tests even inferred the model "Strict Isolation" (SI), should not be overinterpreted, since DILS is not able to perfectly distinguish between the two sub-models of the main model isolation (i.e. Ancient Migration and Strict Isolation or Isolation with Migration and Secondary Contact, Fraïsse *et al.*, 2021). With the notable exception of these two simulations, all our results are consistent with ongoing gene flow and this evidence for ongoing gene flow between *P. davidiana* and *P. rotundifolia* is also consistent with the literature (Shang, 2021; Zheng *et al.*, 2017).

4.2 Historical and genome-wide variations in effective population sizes

All ABC inferences I performed support a larger effective population size of *P. davidiana* than of *P. rotundifolia*. This result is consistent with the more elevated nucleotide diversity observed in *P. davidiana* than in *P. rotundifolia* (figure 2). Indeed, the levels of genetic diversity depend on both the mutation rates and N_e. Given their genetic proximity (divergence ~ 1.3 mya), the two species likely share roughly the same mutation rates, and as a consequence, this difference is expected to be associated with a difference in effective population size. This difference in effective population size is also consistent with the wider species distribution of *P. davidiana* in the lowlands of northeastern and central China, while *P. rotundifolia* is limited to mountainous regions in the Himalaya and adjacent areas (Zheng *et al.*, 2017).

For both species, positive median genome-wide Tajima's D values were estimated, indicating a recent population contraction for each of the species (figure 3). The higher median Tajima's D of *P. rotundifolia* indicates that there is a greater deficit of rare alleles in *P. rotundifolia*, which could be due to a more sudden population contraction than in *P. davidiana*. This is consistent with the results of Huiying Shang (2021), who inferred using SMC++ a more abrupt change in effective population size for *P. rotundifolia* than for *P. davidiana* soon after their divergence. Surprisingly, when variable population sizes along divergence were modeled, DILS was unable to recover this population contraction and found rather the opposite. In 18 and 14 of the 20 tests performed, DILS inferred recent population growth for *P. davidiana* and for *P. rotundifolia*, respectively. There is therefore more work to be done to infer the demographic trajectories. But it should be noted that these inferences could be erroneous and caused by very recent changes in the effective population sizes of the two species, since the accuracy of DILS inferences suffers under these circumstances (Fraïsse *et al.*, 2021).

Effective population sizes were inferred to be heterogeneous along the genome in all DILS tests on coding sequences, which is not surprising, as linked selection is expected to

have a more important effect on these genes on average than on the rest of the genome, in such a way that some genes will exhibit variation in N_e depending on selection coefficients (i.e. strength of selection). For example, in the case of purifying selection (and therefore background selection), genes that are more important (or even essential) for the fitness of a species, should be under stronger purifying selection. In the case of positive selection, genes that are more beneficial to the fitness of a species will contributed to a more contrasted N_e locally in the genome. Overall, these marked effects of linked selection will increase the heterogeneity of the landscape of genetic diversity. It should however be noted that only twelve of the 20 tests on 10kb genomic windows inferred a heterogeneous N_e along the genome.

4.3 Evidence for variable effective migration rates along the genome

Effective migration rates were inferred to be heterogenous along the genome in the majority of the sets based on the genomic windows (12/20). This could be an indication that these sets of coding sequences or genomic windows could contain loci contributing to reproductive isolation which represent barriers to gene flow. Surprisingly, among the 15 scenarios supporting ongoing migration based on the sets of CDS, only four scenarios supported a heterogeneous level of gene flow. Two mutually non-exclusive reasons could explain this result: First, the populations we used are not from the contact zone and the levels of gene flow are expected to be reduced as compared to the levels of effective migration rates in the zones of contact. As a consequence, the overall low migration rates can contribute to lower differences in the migration rates between barriers and neutral loci. Second, despite often presented as a pair at the early stage of speciation (Shang *et al.*, 2020; Shang, 2021), this pair seems to be sufficiently distant to be inferred as isolated for some simulations. Although Shang *et al.* (2020) clearly found that this pair of species is the closest among all *Populus section Populus* species, the fact that this pair represents an early stage of speciation could probably be revised.

In practice, when the ABC software DILS infers a heterogeneous effective migration landscape along the genome, it is then theoretically possible to perform a more advanced analysis to identify the genomic markers present in the real data that support isolation rather than gene flow, and then use this strategy to search regions, and virtually directly genes, contributing to reproductive isolation. However, given that this heterogeneity was not inferred for all the tests I performed, we decided to use a different strategy thanks to genome scanning for elevated divergence levels in order to identify these regions potentially involved in reproductive isolation.

4.4 Genome scanning for elevated differentiation

A lot of fixed differences (F_{ST}=1) were found between *P. davidiana* and *P. rotundifolia* (figure 8). This suggests that the species pair is already at a more advanced stage of speciation. However, the fact that the sampled populations are far from the contact zone (especially in the case of *P. davidiana*) is not ideal. Using a different strategy by focusing only

on regions enriched in highly differentiated SNPs (SNPs with F_{ST} values in the top 1% of all values along the genome), many regions could be identified, where this proportion is above 0.5 (figures 9 & S3), which further supports a high degree of differentiation between the two species.

4.5 Candidate genes for reproductive isolation

Within the eleven regions exhibiting an elevated $\pi_{Between}$ and proportion of highly differentiated SNPs (and their 5kb flanking regions), we found twelve genes including several genes already described. Below, we discussed three genes that might contribute to differential adaptation of *P. davidiana* and *P. rotundifolia* to their respective environments. We focused on ecological barriers.

One of these two genes is the Root Hair Defective 3 Gene (RHD3). Wang *et al.* (1997) showed that normal cell expansion in many tissues of the model plant *Arabidopsis thaliana* depends on this gene. Plants that are homozygous for a mutation in this gene show a reduction in cell size, and therefore exhibit a smaller phenotype. This is due the rhd3 mutant being slightly insensitive to the plant growth hormones auxin and ethylene (Wang *et al.*, 1997). It has been shown that the growth of root hairs also strongly depends on the RHD3 gene (Parker *et al.*, 2000). Root hairs considerably increase the root surface of a plant and are thus essential to the uptake of water and nutrients. Water and nutrient availability could differ between the contrasting environments at different elevations inhabited by *P. davidiana* and *P. rotundifolia*. Therefore, the RHD3 gene could possibly contribute to differential adaptation of this species pair to different edaphic conditions.

Another identified candidate gene was the WRKY (probable WRKY transcription factor 3) gene. This gene belongs to a family of transcription factors characterized in rice (*Oryza sativa*) and *Arabidopsis thaliana* that have been shown to regulate plant responses to herbivory, pathogen attack, as well as abiotic stresses in several plant species (e.g. Skibbe *et al.*, 2008; Chujo *et al.*, 2007; Park *et al.*, 2006). The WRKY3 gene, specifically, is involved in the regulation of Jasmonic Acid (JA)-dependent signaling in response to herbivore attacks (Skibbe *et al.*, 2008). Plant species growing at lower elevations have been shown to be subjected to increased herbivore pressure and invest more into defense mechanisms as compared to plants growing at higher altitudes (e.g. Pellissier *et al.*, 2016). Since *P. rotundifolia* grows at a higher elevation that *P. davidiana* (Zheng *et al.*, 2017; Sun, Gong & Liu, 2020), the WRKY3 gene could thus contribute to differential adaptation of the two species to the different altitudes they occupy.

Another gene found in one of the regions of high differentiation is the gene encoding the S-adenosylmethionine synthase 2, which has been linked to drought, as well as flooding responses (Wang, Oh & Komatsu, 2016). *P. davidiana* has been shown to exhibit a high resistance to drought (Hou *et al.*, 2018), and there might be different hydrological conditions at the different elevations at which *P. davidiana* and *P. rotundifolia* occur. Therefore, this gene could be another good candidate for an ecological barrier between the two species.

This functional annotation work represents a first step toward the characterization of the reproductive barriers at play, but much remains to be done. Of course, there are likely additional ecological and non-ecological barrier loci to gene flow, that remain undiscovered after this thesis. The fact that no candidate gene was found in the region on chromosome 17, which contains many 10kb genomic windows exhibiting a peak in the proportion of highly differentiate SNPs, $\pi_{Between}$, as well as Hudson's F_{ST} (figure 9), suggests thatthe thresholds used are probably too restrictive or that additional candidate genes might be close to the searched regions, but outside the considered 5kb flanking regions. Therefore, it could be beneficial to increase the length of the flanking regions, in order to identify additional candidate genes possibly contributing to reproductive isolation. Furthermore, in this thesis, regions exhibiting peaks in Hudson's F_{ST} were not considered for the detection of the regions, this criterion could be combined in future analyses with regions showing peaks in $\pi_{Between}$ in order to identify additional regions, and therefore candidate genes that could contribute to reproductive isolation.

5. Conclusion

This thesis provides a first detailed description of the demographic history of these two Asian poplar species. My analyses strongly support present-day gene flow between the species *P. davidiana* and *P. rotundifolia*, which is consistent with the literature, as well as migration along divergence. Regarding the effective population size, all my results are consistent with a higher N_e in *P. davidiana* as compared to *P. rotundifolia*, even if the very recent trajectories of both species remain unclear after this thesis. The migration rate was inferred to be heterogeneous along the genome, which supports the presence of reproductive barriers between these species. As a consequence, genome scanning for regions associated with both high relative and absolute differentiation represents a good strategy that allowed me to identify several regions of primary interest, and several genes within these regions, including *rhd3*, *wrky3* and the gene for the S-adenosylmethionine synthase 2, which possibly contribute to differential adaptation of the two species to their contrasted ecological environments.

6. References

- Akey, J M, Zhang, G, Zhang, K, Jin, L, & Shriver, M D. (2002). Interrogating a high-density SNP map for signatures of natural selection. *Genome Research, 12*(12), 1805-1814.
- Beaumont, M A, & Nichols, R A. (1996). Evaluating Loci for Use in the Genetic Analysis of Population Structure. *Proceedings of the Royal Society. B, Biological Sciences, 263*(1377), 1619-1626.
- Beaumont, M A, & Balding, D J. (2004). Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology, 13*(4), 969-980.

- Charlesworth B. (1998) Measures of divergence between populations and the effect of forces that reduce variability. *Mol Biol Evol* 15, 538-43.
- Chujo, T, Takai, R, Akimoto-Tomiyama, C, Ando, S, Minami, E, Nagamura, Y, . . . Yamane, H. (2007). Involvement of the elicitor-induced gene OsWRKY53 in the expression of defense-related genes in rice. *Biochimica Et Biophysica Acta. Gene Structure and Expression, 1769*(7), 497-505.
- Darwin, C. (1859). On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. *London: John Murray*.
- Feder, J L, Egan, S P, & Nosil, P. (2012). The genomics of speciation-with-gene-flow. *Trends in Genetics, 28*(7), 342-350.
- Foll, M, & Gaggiotti, O. (2008). A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. *Genetics (Austin), 180*(2), 977-993.
- Fraïsse, C, Popovic, I, Mazoyer, C, Spataro, B, Delmotte, S, Romiguier, J, . . . Roux, C. (2021). DILS: Demographic inferences with linked selection by using ABC. *Molecular Ecology Resources,* Molecular ecology resources, 2021-01-15.
- Hewitt, G M. (1988). Hybrid zones-natural laboratories for evolutionary studies. *Trends in Ecology & Evolution, 3(7)*, 158-167.
- Hou, Z, Wang, Z, Ye, Z, Du, S, Liu, S, & Zhang, J. (2018). Phylogeographic analyses of a widely distributed *Populus davidiana*: Further evidence for the existence of glacial refugia of cool-temperate deciduous trees in northern East Asia. *Ecology and Evolution, 8*(24), 13014-13026.
- Hudson, R R, Slatkin, M, & Maddison, W P. (1992). Estimation of Levels of Gene Flow from DNA Sequence Data. *Genetics (Austin), 132*(2), 583-589.
- Irwin, D E, Mila, B, Toews, D P L, Brelsford, A, Kenyon, H L, Porter, A N, Grossen, C, Delmore, KE, Alcaide, M & Irwin, J H. (2018). A comparison of genomic islands of differentiation across three young avian species pairs. *Molecular Ecology 1:17*. DOI: 10.1111/mec.14858
- 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. *Molecular Ecology, 26*(12), 3093-3103.
- Leroy, T, Anselmetti, A, Tilak, M K, Bérard, S, Csukonyi, L, Gabrielli, M, Scornavacca, C, Milá, B, Thébaud, C & Nabholz, B. (2019). A bird's white-eye view on neo-sex chromosome evolution. *bioRxiv* 505610, ver. 4 peer-reviewed and recommended by PCI Evolutionary Biology. DOI: 10.1101/505610
- Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv, 1303.3997.
- Mayr, E. (1942). Systematics and the origin of species. *New York: Columbia University Press.*
- Nei, M. & Li, W-H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *PNAS*, 76: 5269-5273.
- Nosil, P. (2008). Speciation with gene flow could be common. *Molecular Ecology*, 17(9), 2103-2106.

- Park, C, Shin, Y, Lee, B, Kim, K, Kim, J, & Paek, K. (2006). A hot pepper gene encoding WRKY transcription factor is induced during hypersensitive response to Tobacco mosaic virus and *Xanthomonas campestris*. *Planta, 223*(2), 168-179.
- Park, S, Kim, J Y, Han, J Y, Ahn, C, Park, E, & Choi, Y E. (2017). Exploring genes involved in benzoic acid biosynthesis in the *Populus davidiana* transcriptome and their transcriptional activity upon methyl jasmonate treatment. *Journal of Chemical Ecology, 43*(11-12), 1097-1108.
- Parker, J S, Cavell, A C, Dolan, L, Roberts, K, & Grierson, C S. (2000). Genetic Interactions during Root Hair Morphogenesis in *Arabidopsis*. *The Plant Cell, 12*(10), 1961-1974.
- Pellissier, L, Moreira, X, Danner, H, Serrano, M, Salamin, N, Dam, N.M. van, & Rasmann, S. (2016). The simultaneous inducibility of phytochemicals related to plant direct and indirect defences against herbivores is stronger at low elevation. *The Journal of Ecology, 104*(4), 1116-1125.
- Quinlan, A R, & Hall, I M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England), 26*(6), 841-842.
- 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. *Journal of Evolutionary Biology, 30*(8), 1450-1477.
- 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.
- 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. *Philosophical Transactions. Biological Sciences, 375*(1806), 20190544.
- Shang, H, 2021. *Genomics of the speciation continuum in Eurasian Populus species.* PhD dissertation, University of Vienna.
- Shang, H, Rendón-Anaya , M, Paun, O, Field, D, Hess, J, Vogl, C, Liu, J, Ingvarsson*, P K., Leroy*, T, Lexer, C 1,8 †* *Conserved genomic landscapes of differentiation across Populus speciation continuum* (in preparation).
- Skibbe, M, Qu, N, Galis, I, & Baldwin, I T. (2008). Induced Plant Defenses in the Natural Environment: *Nicotiana attenuata* WRKY3 and WRKY6 Coordinate Responses to Herbivory. *The Plant Cell, 20*(7), 1984-2000.
- Stankowski, S, & Ravinet, M. (2021). Quantifying the use of species concepts. *Current Biology, 31*(9), R428-R429.
- Sun, C, Gong, X, & Liu, X. (2020). The complete chloroplast genome sequence of *Populus rotundifolia*, and a comparative analysis with other *Populus* species. *Mitochondrial DNA. Part B. Resources, 5*(3), 2463-2465.
- Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105(2): 437-460.
- Tajima, F. (1989). The effect of change in population size on DNA polymorphism. *Genetics*, 123(3): 597-601.

- Wang, H, Lockwood, S K, Hoeltzel, M F, & Schiefelbein, J W. (1997). The ROOT HAIR DEFECTIVE3 gene encodes an evolutionarily conserved protein with GTP-binding motifs and is required for regulated cell enlargement in *Arabidopsis*. *Genes & Development, 11*(6), 799-811.
- Wang, X, Oh, M W, & Komatsu, S. (2016). Characterization of S-adenosylmethionine synthetases in soybean under flooding and drought stresses. *Biologia Plantarum, 60*(2), 269-278.
- 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. *Molecular Biology and Evolution, 33*(7), 1754-1767. DOI:10.1093/molbev/msw051.
- Watterson, G A. (1975). On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.*, 7(2): 256-276.
- Wright, S. (1943). Isolation by distance. *Genetics*, 28: 114-138.
- Zhang, X, Wu, N, & Li, C. (2005). Physiological and growth responses of Populus davidiana ecotypes to different soil water contents. *Journal of Arid Environments, 60*(4), 567-579.
- Zheng, H, Fan, L, Milne, R I, Zhang, L, Wang, Y and 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
- Zheng, C, Tan, L, Sang, M, Ye, M, & Wu, R. (2020). Genetic adaptation of Tibetan poplar (*Populus szechuanica var. tibetica*) to high altitudes on the Qinghai–Tibetan Plateau. *Ecology and Evolution, 10*(20), 10974-10985.

7. Appendix

Table S1: Locations of sampling sites for all individuals used in this thesis (see also Huiying, 2021)

Code number	Species	Latitude (N)	Longitude (E)	Coverage
pdav73	*Populus davidiana*	45.32	127.35	25.89
pdav74	*Populus davidiana*	45.10	128.00	31.94
pdav75	*Populus davidiana*	45.02	128.18	29.83
pdav76	*Populus davidiana*	44.95	128.79	21.57
pdav77	*Populus davidiana*	44.93	128.97	19.95
pdav78	*Populus davidiana*	44.92	128.99	24.74
pdav79	*Populus davidiana*	44.53	129.79	27.09
pdav80	*Populus davidiana*	44.77	129.15	25.18
pdav81	*Populus davidiana*	44.93	128.73	13.71
pdav82	*Populus davidiana*	45.25	127.73	32.63
pdav83	*Populus davidiana*	45.25	127.71	28.47
pdav84	*Populus davidiana*	45.32	127.32	17.93
pdav85	*Populus davidiana*	45.33	127.27	33.84
prot261A01	*Populus rotundifolia*	27.14	99.39	19.06
prot261A02	*Populus rotundifolia*	27.14	99.39	35.85
prot261A03	*Populus rotundifolia*	27.14	99.39	36.47
prot261A04	*Populus rotundifolia*	27.14	99.39	34.46
prot261A05	*Populus rotundifolia*	27.14	99.39	25.95
prot261A06	*Populus rotundifolia*	27.14	99.39	33.79
prot261A07	*Populus*	27.14	99.39	28.73

	rotundifolia			
prot261A08	*Populus rotundifolia*	27.14	99.39	26.00
prot261A09	*Populus rotundifolia*	27.14	99.39	13.86
prot261A10	*Populus rotundifolia*	27.14	99.39	29.59
prot261A11	*Populus rotundifolia*	27.14	99.39	25.72
prot261A12	*Populus rotundifolia*	27.14	99.39	29.44
prot261A13	*Populus rotundifolia*	27.14	99.39	25.29
prot261A14	*Populus rotundifolia*	27.14	99.39	41.12
prot261A15	*Populus rotundifolia*	27.14	99.39	42.80
prot261A16	*Populus rotundifolia*	27.14	99.39	26.64
prot261A17	*Populus rotundifolia*	27.14	99.39	39.09

Table S2: Inferred models for all DILS tests performed on all sets of 10 kb sliding genomic windows, under the assumption of constant and variable population sizes, respectively. Posterior probabilities are indicated in brackets after each inferred model. (IM=Isolation with migration; Nhetero/Nhomo= a heterogeneous/homogeneous effective population size among loci; Mhetero/Mhomo= a heterogeneous/homogeneous migration rate among loci).

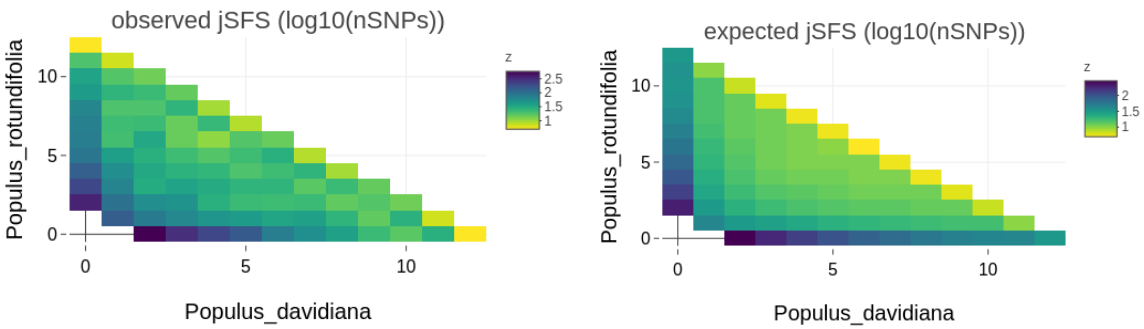
Dataset	Assumption about population sizes	Inferred Models (& posterior probabilities)			
1	constant	migration (0.77135)	IM (0.81377)	Nhetero (0.76598)	Mhomo (0.90398)
2	constant	migration (0.8301)	IM (0.75702)	Nhetero (0.74622)	Mhomo (0.802)
3	constant	migration (0.7213)	IM (0.82283)	Nhetero (0.81027)	Mhomo (0.82112)
4	constant	migration (0.81897)	IM (0.8798)	Nhetero (0.80497)	Mhetero (0.8816)
5	constant	migration (0.84353)	IM (0.7793)	Nhetero (0.74962)	Mhomo (0.85068)
6	constant	migration (0.7797)	IM (0.8023)	Nhetero (0.74267)	Mhetero (0.89313)
7	constant	migration (0.86062)	IM (0.83462)	Nhomo (0.70047)	Mhetero (0.81458)
8	constant	migration (0.87607)	IM	Nhomo	Mhomo

			(0.78062)	(0.71955)	(0.85413)
9	constant	migration (0.8751)	IM (0.75362)	Nhetero (0.77815)	Mhomo (0.86597)
10	constant	migration (0.92138)	IM (0.82117)	Nhomo (0.75688)	Mhomo (0.87343)
1	variable	migration (0.75125)	IM (0.67157)	Nhomo (0.84802)	Mhetero (0.84193)
2	variable	migration (0.72567)	SC (0.77227)	Nhetero (0.82512)	Mhetero (0.71627)
3	variable	migration (0.69378)	IM (0.6309)	Nhomo (0.86358)	Mhomo (0.85287)
4	variable	migration (0.57237)	SC (0.74257)	Nhetero (0.82168)	Mhetero (0.80158)
5	variable	migration (0.62527)	SC (0.82243)	Nhetero (0.79873)	Mhetero (0.70187)
6	variable	migration (0.71787)	SC (0.70177)	Nhetero (0.81047)	Mhetero (0.86568)
7	variable	migration (0.73212)	SC (0.78432)	Nhomo (0.8862)	Mhetero (0.83248)
8	variable	migration (0.73625)	SC (0.73123)	Nhetero (0.82855)	Mhetero (0.80382)
9	variable	migration (0.8916)	SC (0.72025)	Nhomo (0.78677)	Mhetero (0.7805)
10	variable	migration (0.62667)	IM (0.74555)	Nhomo (0.70828)	Mhetero (0.81545)

Table S3: Inferred models for all DILS tests performed on all sets of coding sequences, under the assumption of constant and variable population sizes, respectively. Posterior probabilities are indicated in brackets after each inferred model. (IM=Isolation with migration; SC=Secondary Contact; AM=Ancient Migration; SI=Strict Isolation; Nhetero/Nhomo= a heterogeneous/homogeneous effective population size among loci; Mhetero/Mhomo= a heterogeneous/homogeneous migration rate among loci).

Dataset	Assumption about population sizes	Inferred Models (& posterior probabilities)			
1	constant	migration (0.74588)	IM (0.59185)	Nhetero (0.6839)	Mhetero (0.86983)
2	constant	isolation (0.73508)	AM (0.54425)	Nhetero (0.80173)	
3	constant	migration (0.72567)	IM (0.76475)	Nhetero (0.74647)	Mhomo (0.82622)
4	constant	migration (0.7038)	IM (0.76687)	Nhetero (0.79067)	Mhomo (0.73727)
5	constant	isolation (0.7535)	AM (0.79097)	Nhetero (0.81352)	

6	constant	migration (0.81955)	IM (0.6793)	Nhetero (0.76818)	Mhomo (0.73497)
7	constant	migration (0.79347)	IM (0.804)	Nhetero (0.66147)	Mhomo (0.90385)
8	constant	migration (0.83533)	IM (0.72987)	Nhetero (0.71703)	Mhomo (0.8872)
9	constant	migration (0.70223);	IM (0.7582)	Nhetero (0.72788)	Mhomo (0.86945)
10	constant	migration (0.77007);	IM (0.67128)	Nhetero (0.68283)	Mhomo (0.84677)
1	variable	isolation (0.53093)	AM (0.73033)	Nhetero (0.91148)	
2	variable	isolation (0.50548)	AM (0.5462)	Nhetero (0.9279)	
3	variable	migration (0.56982)	SC (0.79395)	Nhetero (0.77973)	Mhetero (0.77208)
4	variable	migration (0.80817)	SC (0.7145)	Nhetero (0.82688)	Mhomo (0.72015)
5	variable	migration (0.74882)	SC (0.58903)	Nhetero (0.83423)	Mhomo (0.89087)
6	variable	isolation (0.51957)	SI (0.61475)	Nhetero (0.84188)	
7	variable	migration (0.73842)	SC (0.72072)	Nhetero (0.79387)	Mhetero (0.79503)
8	variable	migration (0.55972)	SC (0.78515)	Nhetero (0.80067)	Mhomo (0.69992)
9	variable	migration (0.4146);;	SC (0.69203)	Nhetero (0.73398)	Mhomo (0.64272)
10	variable	migration (0.6311)	SC (0.701)	Nhetero (0.7625)	Mhetero (0.84333)



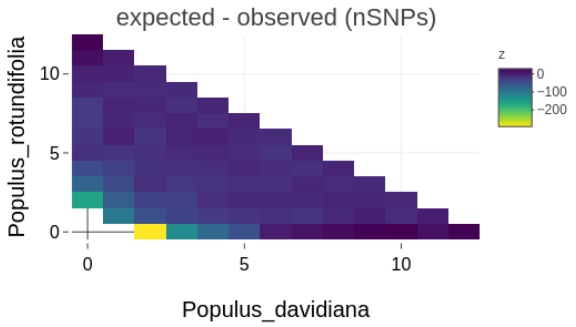


Figure S1: 2D-Site-Frequency-Spectra for the real data (top left), the best model (top right) and the residuals between them (model-data, bottom), respectively, shown for one of the divergence tests on coding sequences (under the assumption of constant population sizes).

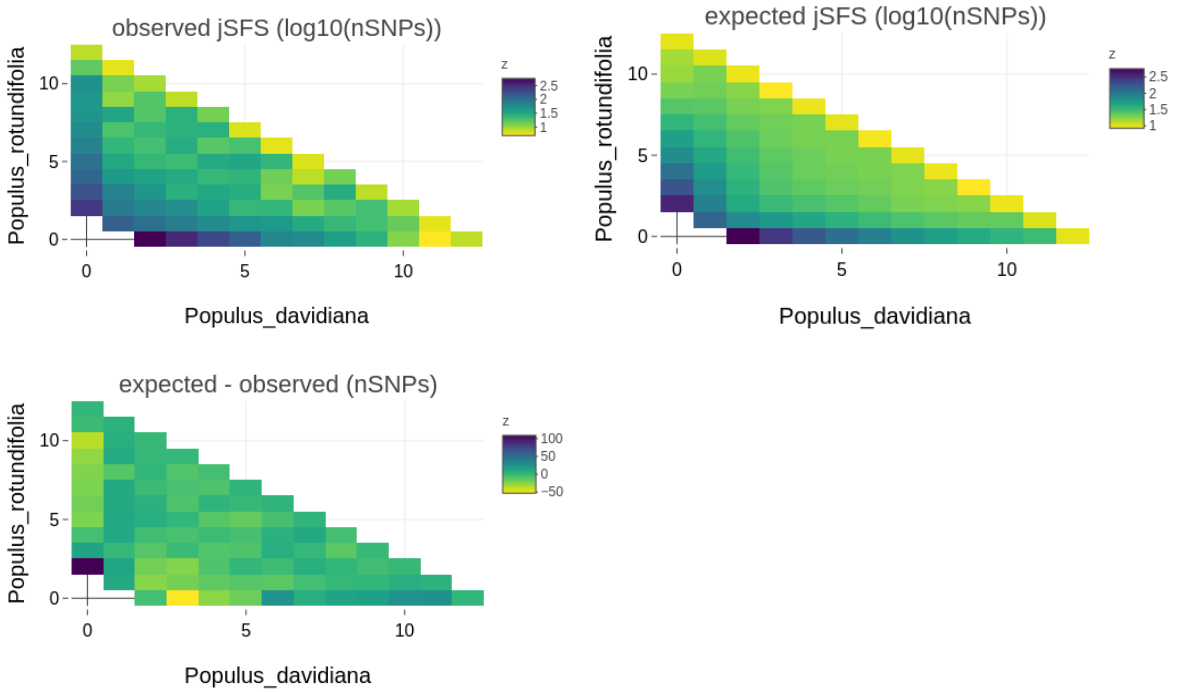
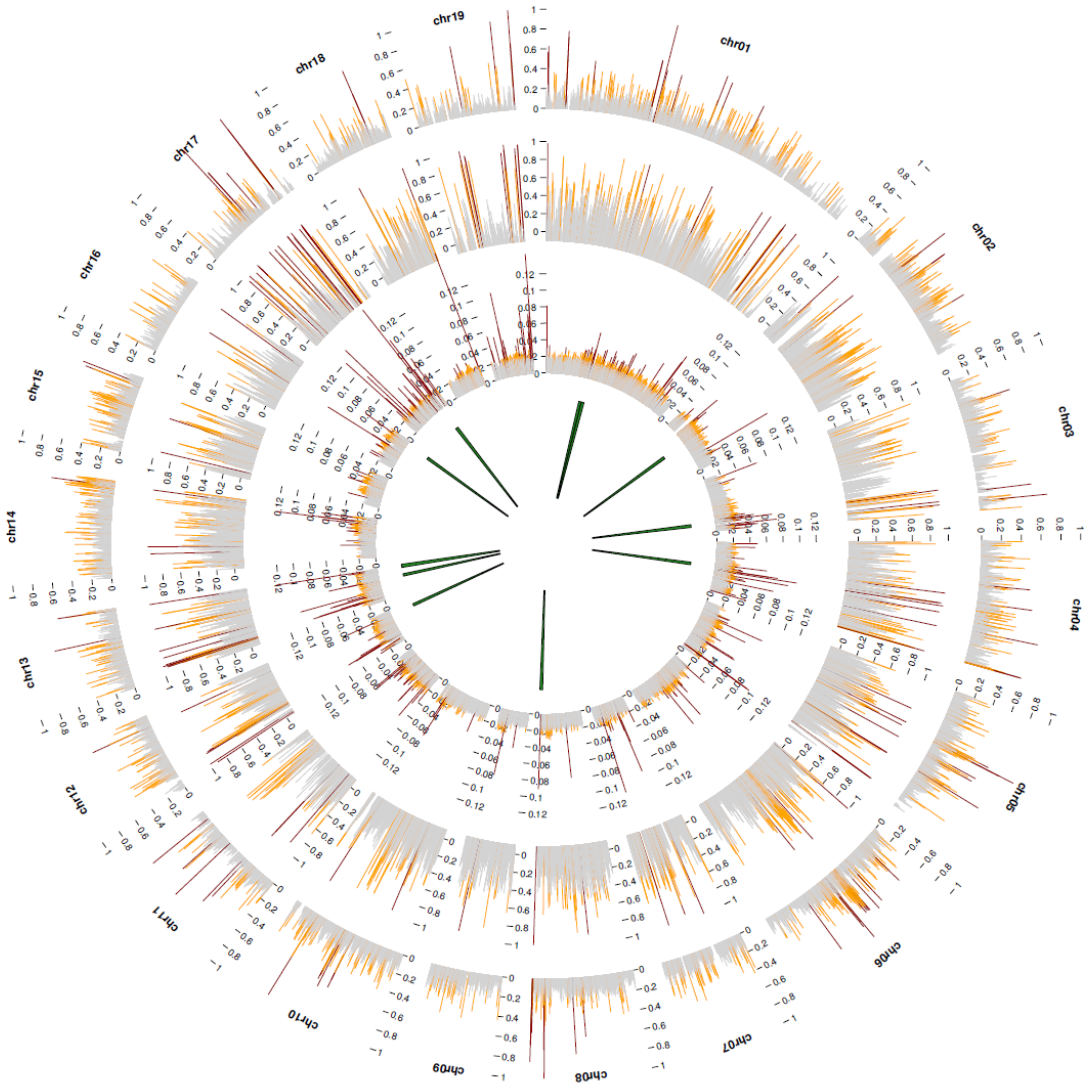


Figure S2: 2D-Site-Frequency-Spectra for the data (top left), the model (top right) and the residuals between them (model-data, bottom), respectively, shown for the best inferred scenario in one of the tests on coding sequences (under the assumption of variable population sizes).



<u>Figure S3</u>: Manhattan plots for the proportion of highly differentiated SNPs (outermost circle), Hudson's F_{ST} (circle in the middle), as well as $\pi_{Between}$ (innermost circle) shown for all 19 chromosomes of the species pair P. davidiana and P. rotundifolia. Dark red bars indicate windows within the top 1% of $\pi_{Between}$ or Hudson's F_{ST} values, or in which the proportion of highly differentiated SNPs is at least 0.5. Orange bars: windows within the top 5% of $\pi_{Between}$ or Hudson's F_{ST} values, or in which the proportion of highly differentiated SNPs is at least 0.25. The green rods in the middle point to the eleven regions of high differentiation (which are in the top 1% of Hudson's F_{ST}, as well as $\pi_{Between}$ values, and in which the proportion of highly differentiated SNPs is at least 0.5). It should be noted that two windows on chromosome 1 are very closed together and could therefore be mistaken for one window.

Table S4: Eleven 10kb genomic windows within the top 1% of $\pi_{Between}$ values, and in which the proportion of highly differentiated SNPs is at least 0.5. These windows were searched for potential reproductive barrier loci to gene flow.

chr	startwin	endwin
chr1	17310001	17320000

chr1	18340001	18350000
chr12	11750001	11760000
chr13	8840001	8850000
chr13	13440001	13450000
chr16	12110001	12120000
chr17	14810001	14820000
chr2	10410001	10420000
chr3	18840001	18850000
chr4	10500001	10510000
chr8	16880001	16890000