UPPSALA
UNIVERSITET

Demography of Birch Populations across Scandinavia

Master Degree Project

Janek Sendrowski

Degree project in bioinformatics 30 hp, 2022
Biology Education Centre and Department of Ecology and Genetics,
Uppsala University Supervisors: Martin Lascoux and Jennifer James

Acknowledgements

I would like to express my thanks to my supervisor Martin Lascoux for helpful feedback, reading material, personal lectures and a conducive working atmosphere. I would also like to thank my co-supervisor Jennifer James for advising me on potential tools, helping create the project outline and kindly listening to my progress reports. At last, I also owe thanks to Luis Leal for a detailed introduction to the underlying data and useful recommendations.

Abstract

Boreal forests are particularly vulnerable to climate change, experiencing a much more drastic increase in temperatures and having a limited amount of more northern refugia. The trees making up these vast and important ecosystems already had to adapt previously to environmental pressures brought about by the repeated glaciations during past ice ages. Studying the patterns of adaption of these trees can thus provide valuable insights on how to mitigate future damage. This thesis presents and analyses population structure, demographic history and the distribution of fitness effects (DFE) of the diploid *Betula pendula* and tetraploid *B. pubescens* across Scandinavia. Birches–being widespread in boreal forests as well as having great economical importance–constitute superb model species. The analyses of this work confirm the expectations on postglacial population expansion and diploid-tetraploid introgression. They furthermore ascertain the presence of two genetic clusters and a remarkably similar DFE for the species. This work also contributes with a transparent, reproducible and reusable pipeline which facilitates running similar analyses for related species.

Contents

Acronyms

BCa Bias-Corrected and accelerated. 11, 13, 26, 31, 36, 74

BFGS Broyden-Fletcher-Goldfarb-Shanno. 25, 36

BWA Burrows-Wheeler Aligner. 7

DFE Distribution of Fitness Effects. 1, 2, 4, 7, 8, 11, 34–40, 72–74

FEEMS Fast Estimation of Effective Migration Surfaces. 8, 16, 20

FGM Fisher's Geometric Model. 35

GATK Genome Analysis Toolkit. 7

GIM Godambe Information Matrix. 25

HWE Hardy-Weinberg Equilibrium. 5, 7, 9, 10, 14, 48, 49

LD Linkage Disequilibrium. 18

LGM Last Glacial Maximum. 3, 4, 6, 9–11, 13, 23, 26–29, 31, 40, 63, 66, 67, 69, 70

LRT Likelihood-Ratio Test. 7, 13, 25, 26, 36, 39

MLE Maximum Likelihood Estimation. 8, 30, 34

PCA Principal Component Analysis. 5, 6, 8, 16–20, 22, 51, 53

SFS Site-Frequency Spectrum. 2, 6–12, 20, 23, 26, 29, 30, 33, 34, 36, 39, 61, 66, 73

SNP Single-Nucleotide Polymorphism. 5, 7, 11, 12, 18, 35, 48

UMAP Uniform Manifold Approximation and Projection. 5, 8, 16–18, 52

VCF Variant Call Format. 5, 7, 8, 12, 46

List of Figures

List of Tables

1 Introduction

1.1 Background

Observed patterns of genetic variability and structure in present day populations are shaped by a complex interplay of many factors, including individual dispersal ability, environmental conditions, and past demographic events. Climatic events such as glaciations are particularly noteworthy: they had a broad impact on many species, leading to reductions in population size and range contractions [25]. Many species then gradually recolonised areas as global temperatures increased [32]. The route that individuals took in recolonising, the distance from the refuge area, and the uneven inheritance of ancestral variation can all lead to patterns in population-level genetic data [25].

However, some of the genetic differences we observe between populations could also be the result of adaptation. After a glaciation, for example, species can recolonise previously inhospitable areas where they may encounter different climatic zones. This can result in species adapting to local environmental conditions [29]. Maintaining such adaptation can also lead to persistent population structure.

Gaining a better understanding of the complex factors that determine patterns of polymorphism within species is the broad aim of this project which will focus on birch species in Scandinavia. Trees are keystone species with limited dispersal ability; as sessile organisms, they are not able to migrate in the face of unfavourable climatic conditions, and thus might experience strong selective pressure to adapt to local environmental conditions. Birch trees have been the focus of a recent sequencing effort, where individual trees were sampled over a latitudinal gradient in Scandinavia that includes a transition between different climatic zones.

1.2 Project Importance

Birches are rather common in cool temperate and boreal forests which in turn constitute a large fraction of the global forest cover [33]. Boreal forests are particularly vulnerable to global warming as temperatures are projected to increase much more drastically in the north and as there may be no cooler refugia to retreat to northward [29]. Birches are moreover of economical importance, being used for timber [33]. This profile makes them excellent

model species in the context of climate change and studying past patterns of migration and adaptation of these large, slowly growing and dispersing species could aid in predicting of what is likely to happen to such species as climate conditions change in the future.

This thesis is also of importance to the understanding of population structure and demographic history of birch in particular and will, apart from this written report, contribute with a transparent analysis pipeline that will be made publicly available on GitHub [30]. Both of which will hopefully be useful to collaborators and others in the field.

1.3 Objectives & Report Outline

The goals of the project can be broken down into three parts. At first, we look at geographic differentiation, migration and isolation by distance, which we compare with established results from literature (cf. section 4). Subsequently, models of the demographic history are fit to the data. Relevant demographic models will have to be determined beforehand. The inferred demography is then checked for accordance with the expectations (cf. section 5). At last, the Distribution of Fitness Effects (DFE) will be estimated and compared for all samples as a whole as well as between different subpopulations. We will also compare DFE models with different assumptions regarding their shape (cf. section 6).

The three above-mentioned parts are preceded by this introductory part, presenting background, importance and expectations (cf. section 1). This is followed by an introduction to the underlying data, an outline of the workflow and pipeline as well as some basic summary statistics like nucleotide diversity and heterozygosity that could not be properly placed elsewhere (cf. section 2). Before proceeding with the actual analyses, the derivation of the site frequency spectra (SFS) is discussed which is essential for the inference of demography and DFE (cf. section 3). In this section we will also treat the identification of ancestral alleles as well as the different degeneracy and synonymy classes which are all necessary for the derivation of the SFS.

1.4 Species

We are working with two different birch species, *Betula pendula* and *Betula pubescens*, which are both monoecious, wind-pollinated boreal forest trees,

widespread across Eurasia in addition to being pioneer species [25]. *B. pendula*, the silver birch, is recognisable by its eponymous silverish bark and sometimes pendulous leaves. Further north, it is gradually replaced by its more cold-resistant cousin *B. pubescens*, also named downy birch because of its downy shoots. This species also makes up the treeline in northern Scandinavia [32]. We will henceforth refer to these two species by either using their Latin names or simply by "the two species" or "both species".

B. pendula is diploid whereas *B. pubescens* is tetraploid which constitutes a substantial reproductive barrier. There are two possible scenarios for diploid-tetraploid gene flow: backcrossed triploid hybrids or direct fertilisation of an unreduced diploid gamete from the diploid species with a conventional diploid gamete from the tetraploid species. The latter enables gene flow from diploid to tetraploid and the former favours gene flow in the same direction, noting that fertilised triploid hybrids primarily produce tetraploid offspring [12, 37]. A recent paper confirms that there is mainly unidirectional gene flow from *B. pendula* into *B. pubescens* [37]. The polyploid origin of *B. pubescens* is disputed, with potential parents being *B. pendula*, *B. platyphylla*, *B. nana*, *B. humilis* or *B. lenta* [34, 29]. A third birch species, *B. nana*, the dwarf birch, is endemic to Scandinavia where it occurs in northern and upland climes. It is easily recognisable by its stunted non-woody growth and has been found to hybridise readily with *B. pubescens* [2, 37]. Sequence data for this species were not available although including them would provide an interesting comparison between the amount and nature of introgression of *B. nana* and *B. pendula* into *B. pubescens*.

1.5 Expectations

We generally do not expect a high level of genetic differentiation among populations partly due to the large dispersal distances of the pollen [29, 25]. We do, however, expect to observe different population clusters owing to the recolonisation of different populations after the Last Glacial Maximum (LGM). In a previous study based on cpDNA, two principal clusters were found among European *B. pendula* samples, indicating an eastern and western wave of recolonisation [25]. Another recent study on Norway spruce (*Picea abies*) also found two principal clusters in Sweden–a northern and a southern one which originated from the Baltics and northern Russia, respectively [17]. This is unlike other less cold-resistant tree species like oak whose populations were restrained to a limited number of refugia in south-

ern Europe during the LGM [33]. In any case, the recolonisation should leave detectable traces of population expansion in the not too distant past. Besides that, we expect to observe some gene flow between the two species, possibly only unidirectionally from the *B. pendula* into *B. pubescens*. In addition, we do not expect the DFE to differ substantially between the two birch species or their subpopulations [5]. This expectation stems from the fact that both species have a similar geographical extend and that a species similar to *B. pendula* was likely one of *B. pubescens*'s ancestors.

2 Basics

2.1 Pipeline

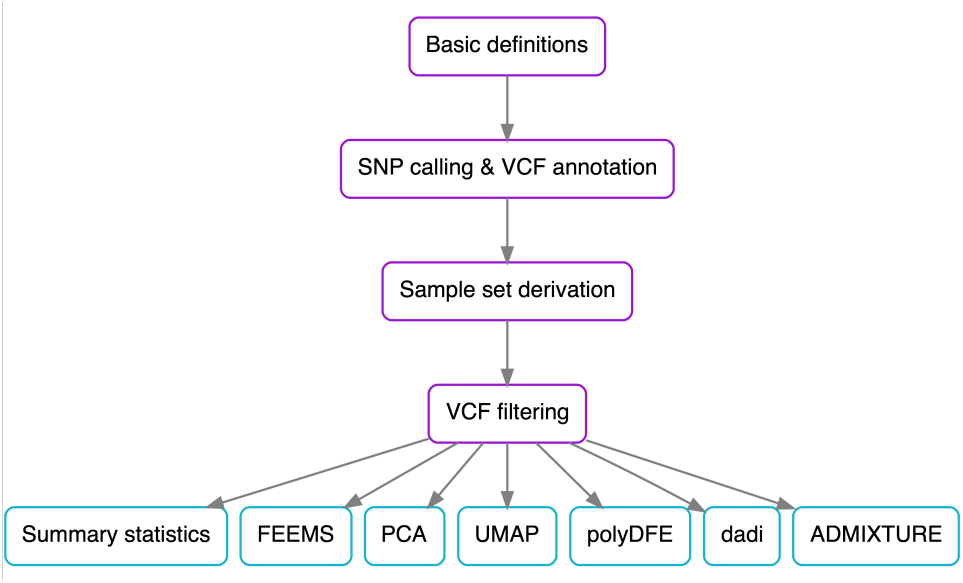
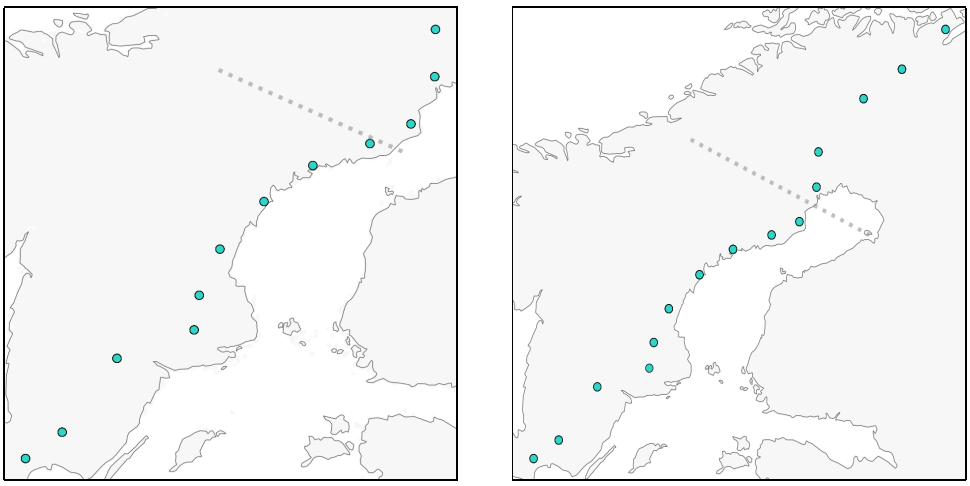


Figure 1: A schematic of the pipeline's subworkflows. Each subworkflow can be executed independently. The VCF filtering relies on the previous annotation of the VCF files as well as the derived sample sets (i.e. which individuals to include). The lowermost subworkflows only require a VCF file and possibly some sample metadata and subpopulation information.

The brunt of the working time was concerned with the implementation of a pipeline to perform the various analyses [30]. A scalable, parallelised and fully-automatic pipeline was implemented using Snakemake for workflow management which ensures reproducibility, integrability, portability and seamless integration with Python [20]. The package manager Conda is used for all non-trivial dependencies. The pipeline was successfully tested on both Linux and Darwin (macOS), with the exception of EST-SFS whose code did not compile well on Darwin and polyDFE which is only offered for Linux on Conda, despite executables for Darwin being available elsewhere. Most resource-intensive tasks were run on Uppmax (Uppsala University's computing center). The pipeline is divided into several subworkflows whose

structure is apparent from fig. 1. Each subworkflow can be executed independently and care was taken to provide an easy interface of input files.

2.2 Data



(a) *B. pendula* (b) *B. pubescens*

Figure 2: Sample locations for the two species. Each dot indicates a population from which several samples where collected. It is apparent that *B. pendula* has a more southern distribution than *B. pubescens*. The dotted lines indicate the subpopulation boundaries introduced later (cf. section 4).

The data consist of Illumina paired-end short read exome data for 155 and 218 samples of *B. pendula* and *B. pubescens*, respectively. The samples were collected as part of a project examining how birches might adapt to climate change and originate from different locations across Sweden and Norway. Fig. 2 shows the sample locations with several samples being taken from each location, representing a population. The samples are roughly distributed along a straight line which is pointing north-northeast. The *B. pendula* reference genome, annotation file and short reads for the two elder outgroup species (*A. incana* & *A. glutinosa*) where obtained from a Finnish genome sequencing project [29]. The reference for *B. pendula*'s rather small genome counts 435 Mbps divided into 5 642 contigs [28].

2.3 Workflow Outline

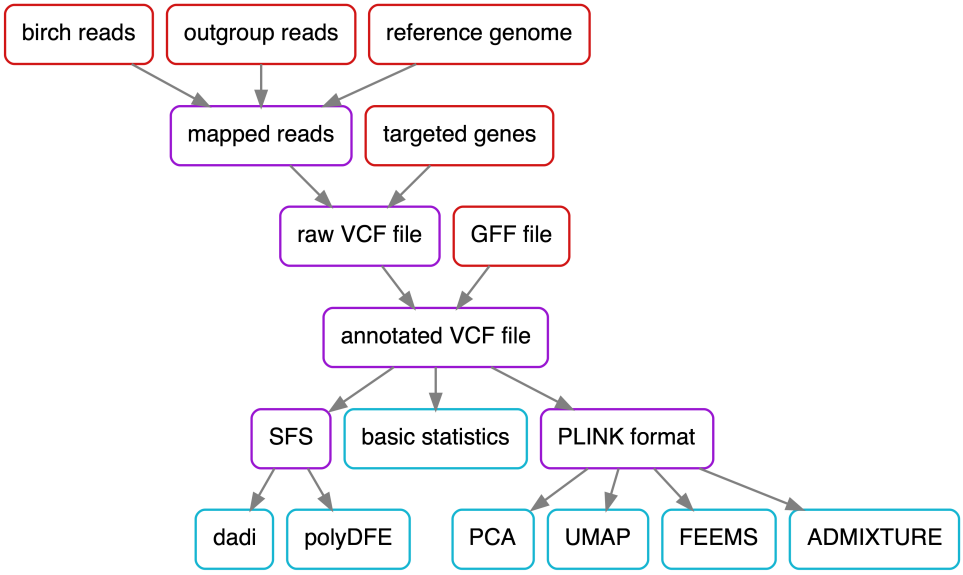


Figure 3: A schematic of the workflow. Resource data are marked red, intermediate steps are purple and final results are blue.

Fig. 3 shows a schematic of the pipeline's workflow. All analyses are based on bi-allelic SNP data whose derivation is as follows: The initial short reads were trimmed with Trimmomatic, mainly to remove the adapter sequences whence their quality was confirmed with FastQC [4, 3]. The mapping was done with BWA using the *B. pendula* reference genome [28]. In doing so, we also expect to capture most variation within *B. pubescens* whose polyploid origin likely involved *B. pendula* (cf. section 1.4). We then jointly called the variants using GATK's GVCF workflow to obtain a raw VCF file (including monomorphic sites) [23]. The variant calling was done twice: once treating all samples as diploid and once specifying their proper ploidy (*B. pubescens* is tetraploid). This was done as some analyses cannot handle polyploid individuals. The resulting variants were subsequently hard-filtered according to GATK's best practices. The sites were then annotated with regards to their degeneracy and synonymy using custom scripts and the ancestral alleles were inferred with EST-SFS [16]. All time-consuming tasks were parallelised to make the pipeline more efficient and scalable. The resulting annotated VCF file was then used for the derivation of various sample sets that we used for

further analyses. For each sample set configuration a separate VCF file was created in order to easily run analyses over different sample sets. Sites with more than 50% of missing values were removed from the sample sets if not specified otherwise. We then began by calculating some basic summary statistics like nucleotide diversity π, fixation index F_{ST} and Tajima's D. The population structure was visualised with PCA and UMAP (a non-linear dimension reduction technique) [22]. Migration surfaces were estimated with FEEMS [19]. Thereafter, we modelled the demographic history where δaδi was used to perform Maximum Likelihood Estimation (MLE) on various demographic models which we were required to specify explicitly [14]. Subsequently, polyDFE, which is also SFS-based, was used to estimate the Distribution of Fitness Effects (DFE) for different subpopulations which we then compared to each other [31]. For both analyses, we needed to differentiate between synonymous and non-synonymous sites to obtain a Site-Frequency Spectrum (SFS) for each class. To additionally infer the DFE for beneficial mutations and to increase the accuracy of the demographic modelling, we also needed to derive an unfolded SFS by distinguishing between ancestral and derived alleles.

2.4 Nucleotide Diversity

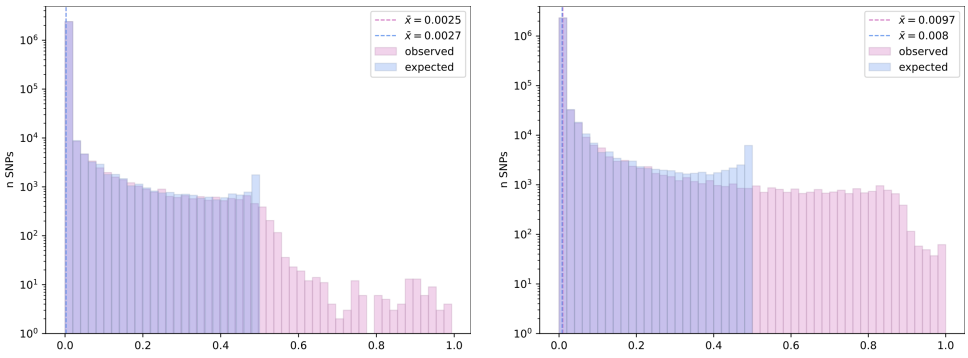
Table 1 shows the nucleotide diversity π and number of segregating sites S for both birch species. Both values are significantly higher for *B. pubescens* which is expected for polyploids, having more homologous sites where alleles can differ. This would be especially so for heteropolyploids with considerably diverged ancestors. The calculations were performed with δaδi and are SFS-based implying that each sample could be treated with respect to its proper ploidy [14]. We have $\pi = \frac{n}{n-1}\frac{1}{l}\Sigma_i 2p_i(1-p_i)$ where n is the number of haplotypes, l the total number of surveyed sites and p_i the frequency of any of the two alleles at the ith segregating site [11]. This formula can be interpreted as the probability that two randomly chosen sites are segregating and can thus directly be applied to polyploid samples, the only difference being the increased number of haplotypes n [9]. The ratio π_N/π_S roughly denotes the fraction of non-synonymous mutations that were not removed by purifying selection i.e. the fraction of mutations that are selectively neutral or more strongly advantageous. Selectively neutral means here that selection is sufficiently weak for drift to predominate. We note that π_N/π_S is slightly larger for *B. pubescens* which could be explained by relaxed purifying selection in polyploids due to higher gene redundancy. It is also noteworthy that, for

both species, π_N/π_S is rather low in comparison to other tree species [7, 8].

	B. pendula	*B. pubescens*
S	39101	225531
π	0.00148	0.00600
π_N	0.00095	0.00290
π_S	0.00378	0.00974
π_N/π_S	0.251	0.298

Table 1: The number of segregating sites S and the nucleotide diversity π over all sites. The synonymous and non-synonymous nucleotide diversity π_S and π_N were calculated over 4-fold and 0-fold degenerate sites, respectively.

2.5 Heterozygosity



(a) *B. pendula* (b) *B. pubescens*

Figure 4: Expected (under HWE) and observed site-wise heterozygosity over the set of bi- and mono-allelic sites. The observed excess heterozygosity for *B. pubescens* is caused by calling it as diploid. The expected site-wise heterozygosity is equivalent to the nucleotide diversity π. The dashed vertical lines indicate the average heterozygosities. Note that these plots are log-scaled. Linearly scaled plots can be found in the appendix (cf. fig. 30)

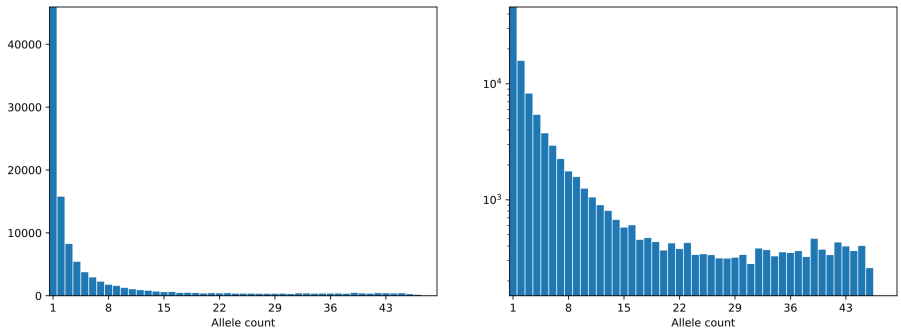
The heterozygosity at a given locus describes the fraction of individuals that are heterozygous there. An expected heterozygosity H can be calculated

from a site's allele frequency assuming it is in Hardy-Weinberg Equilibrium (HWE). We have $H_i = 2p_i(1 - p_i)$ where p_i is the frequency of any allele at the ith bi-allelic site [11]. This is exactly the same formula as for the nucleotide diversity. However, in contrast, when calculating the actual observed heterozygosity we confine our comparison to the two loci within each diploid individual to see how they associate. Fig. 4 shows the expected and observed heterozygosities for both species. By far most sites have values of 0 as they are monomorphic. Under HWE, at most half of all individuals are expected to be heterozygous provided that the sample size is greater than 1. H_i thus ranges from 0 to 0.5 where 0.5 is the theoretical maximum attained for $p = 0.5$. The observed heterozygosity exceeds 0.5 in some cases which is expected due to sampling variance but can, in principle, be indicative of outbreeding or heterozygote advantage if it does so consistently. An excess of homozygotes is more common, however, and can be caused by dynamics like inbreeding or population structure [11]. For *B. pendula*, the average observed heterozygosity is, in fact, slightly lower than its expectation (cf. fig. 4a). This can partly be caused by non-random mating due to isolation by distance which is expected for stationary species of large geographical extend. In any case, the two distributions are in rather good agreement (note the logarithmic scale). *B. pubescens*'s average observed heterozygosity, however, significantly exceeds its expectation which is, at least in part, due the bias introduced by calling it as a diploid (cf. fig. 4b). More precisely, it is caused by neglecting that polyploids are more likely to be heterozygous than diploids and by the overestimation of the frequency of rare alleles (cf. section 3.3 for a more detailed explanation). Fig. 31 shows site-wise p-values for the observed heterozygosity under Hardy-Weinberg Equilibrium. Both the heterozygosity and p-values were calculated using PLINK [27, 36].

3 Derivation of Site-Frequency Spectra

The Site-Frequency Spectrum (SFS) is a powerful summary statistic of bi-allelic SNP variation. It provides information on the number of sites in each frequency class, i.e. $\xi = (\xi_i)_{i=0}^{n}$ where ξ_i denotes the number of SNPs for which i out of n alleles are derived, i.e. not ancestral (cf. fig. 7). This is the definition of a so-called unfolded SFS as it distinguishes sites with i ancestral alleles from sites with i derived alleles (cf. folded SFS where frequency classes with i and $n - i$ derived variants are collapsed). Unfolded spectra are much more informative. For example, they provide information on the number of high-frequency derived alleles which–when numerous–can be indicative of positive selection or population expansion. Both δaδi and polyDFE obtain their information by means of an SFS and many summary statistics like S, π, F_{ST} and Tajima's D can be directly computed from it.

3.1 Ancestral State Identification



(a) linear (b) log-scaled

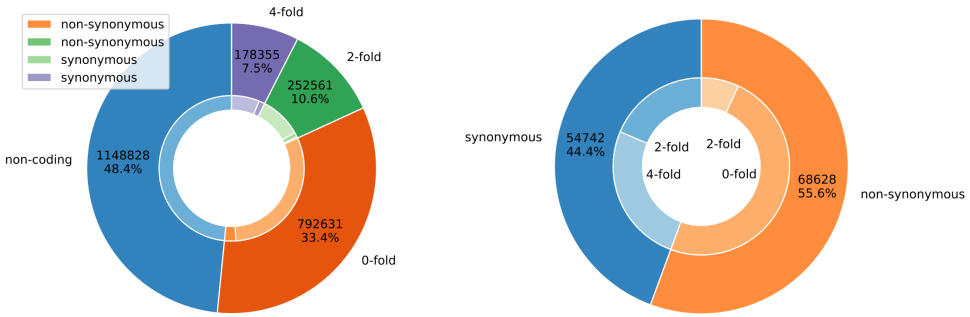
Figure 5: Linear and log-scaled site-frequency spectrum inferred with EST-SFS by using 50 ingroup and two outgroup samples. The spectrum is relatively smooth which was not the case for smaller sets of SNPs or a larger number of ingroup samples.

In order to obtain an unfolded SFS, we first need to infer the ancestral alleles which we did with the help of EST-SFS [16]. For the inference, 50 ingroup and two outgroup samples were used which resulted in an SFS that was

sufficiently smooth. The ingroup subsampling was done without replacement whenever possible and both birch species were used as we deemed them sufficiently close in comparison to the outgroup species. The Kimura 2-parameter nucleotide substitution model was used according to the article's recommendation. There were initial problems determining the ancestral variants. EST-SFS calculates separate prior probabilities for each frequency class based on the maximum likelihood of the tree topologies for the two different ancestral states. This is being done to correct for the fact that high-frequency derived alleles are much less common than high-frequency ancestral ones (e.g. due to having more slightly deleterious mutations at low frequencies that have not yet been removed by purifying selection). The large variance of the number of samples among the frequency classes caused the prior to favour high-frequency derived alleles in some cases. This was especially true for small chunks of VCF files that were initially used for the purpose of parallelisation. Choosing a larger number of ingroup samples also caused the variance to increase as we then have even more frequency classes containing fewer samples, each class receiving its own prior. On the other hand, reducing the ingroup sample size too much would introduce a lot of sampling variance. A moderate number of samples (i.e. 50) in addition to using all available SNPs at once, provided an SFS that was sufficiently smooth (cf. fig. 5).

3.2 Degeneracy & Synonymy

We may want to restrict our analyses to sites for which we have a more narrow notion of the underlying forces acting on them. For example, we might only be interested in sites not subject to direct selection or to compare different classes of sites. Coding sites can be classified in several ways: A polymorphic site is said to be synonymous if all alleles code for the same amino acid–it is said to be non-synonymous otherwise. Synonymous sites are often assumed to not be affected by direct selection so that their variational patterns are purely shaped by demography, drift and their genomic background. This is useful when demography is everything we are interested in or when investigating how selection affects variation by controlling for demography. For sites not necessarily polymorphic, we can instead look at their degeneracy, i.e. the potential of a point mutation to change the amino acid coded for. We distinguish between 0-fold, 2-fold, 4-fold degenerate sites. 4-fold degeneracy means that no mutation would result in a different amino acid, i.e. all mutations are silent. At 0-fold sites, any muta-



(a) Degeneracy (b) Synonymy

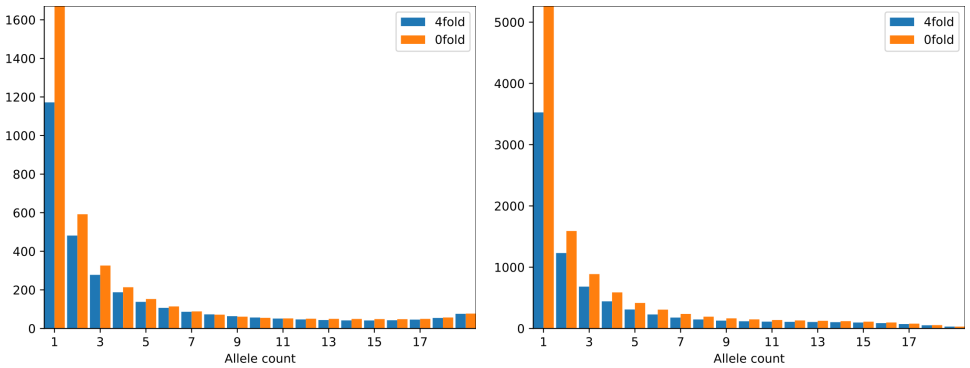
Figure 6: Classification of all targeted sites (left) as well as all coding bi-allelic sites (right). The whitish stretches in the inner disk of the left figure indicate the number of monomorphic sites.

tion would produce an amino acid change while at 2-fold sites, only two out of three possible mutations would cause such a change.

Fig. 6 shows how degeneracy and synonymy are allocated to all targeted sites and all coding bi-allelic sites, respectively. 0-fold degenerate sites seem to be much more common than 2-fold and 4-fold degenerate sites which can readily be confirmed by looking at a coding table. We nonetheless have a substantial amount of synonymous sites which can be explained by noting that they fix more easily (most non-synonymous mutations being deleterious). In the same vein, one might expect there to be more 2-fold non-synonymous than synonymous sites as the majority of changes (two out of three) are non-synonymous, after all. Again, we have to note that non-synonymous mutations are much more likely to get removed by purifying selection. The results obtained are based on custom scripts as no appropriate tool for determining the sites' degeneracy could be found. The synonymy of the mutations as well as the resulting codons were compared with results from VEP to confirm their validity [24].

3.3 Site Frequency Spectrum

Unfolded site-frequency spectra for 0-fold and 4-fold degenerate sites down-projected to a sample size of 20 are shown in fig. 7. Interestingly, *B. pendula* has many more high-frequency derived alleles for both classes. This u-shape



(a) *B. pendula* (b) *B. pubescens*

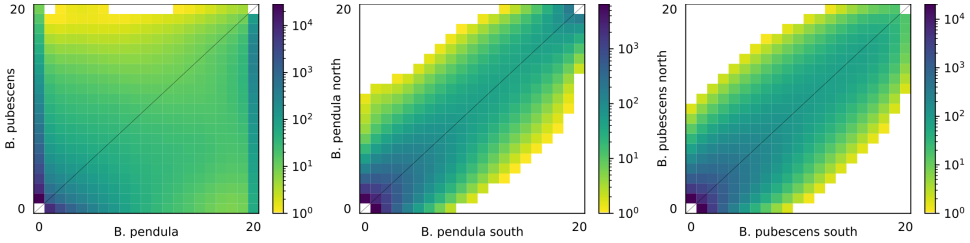
Figure 7: Unfolded site-frequency spectra for 0-fold and 4-fold degenerate sites down-projected to a sample size of 20. *B. pendula* has many more high-frequency derived alleles than *B. pubescens*.

can, in theory, be explained by population expansion, population structure, multiple-merger coalescents or gene flow from disparate populations [18]. In the case of gene flow, we would rather expect *B. pubescens* to have a u-shaped spectrum as we expect it to have a higher introgression rate. Population structure is unlikely to be the culprit either as not a lot of it can be found in birch. Another explanation could be ancestral misidentification but both species are based on the same ancestral states and *B. pubescens* rather exhibits a deficit of high-frequency derived alleles. *B. pubescens* furthermore has a relatively large amount of low-frequency derived alleles potentially owing to reduced purifying selection brought about by higher gene redundancy.

Treating *B. pubescens* as diploid when calling the variants biased the SFS toward having more intermediate- and fewer low-frequency alleles (cf. fig. 51). This is easiest explained by observing that low-frequency alleles will likely only have one copy in heterozygous individuals (assuming no strong deviation from HWE). Treating these individuals as diploid assumes that half of their haplotypes contain that allele, thus overestimating its frequency. Equivalently, we can say that for low-frequency alleles, the number of occurrences (i.e. roughly one per heterozygote) is counted to be the same but the total number of haplotypes is lower when assuming diploidy instead of a higher ploidy.

2-dimensional (2D) frequency spectra provide plenty of information on the

relatedness of two populations as they encode how their allele frequencies are associated site-wise. In general, the more recent their common ancestry or the more migration, the more sites we find in classes with similar frequencies. In fig. 8, we can see down-projected 2D frequency spectra for different sample sets. The spectrum in fig. 8a is rather spread-out, i.e. very disparate frequencies can be observed per site, which is indicative of little migration or relatedness. This makes perfect sense as we are looking at two separate species here and is in strong contrast with the spectra for *B. pendula* and *B. pubescens* whose subpopulations are highly admixed (cf. fig. 8b & 8c). These spectra also show much more symmetry about the line $x = y$, indicating similar allele frequencies in both populations.



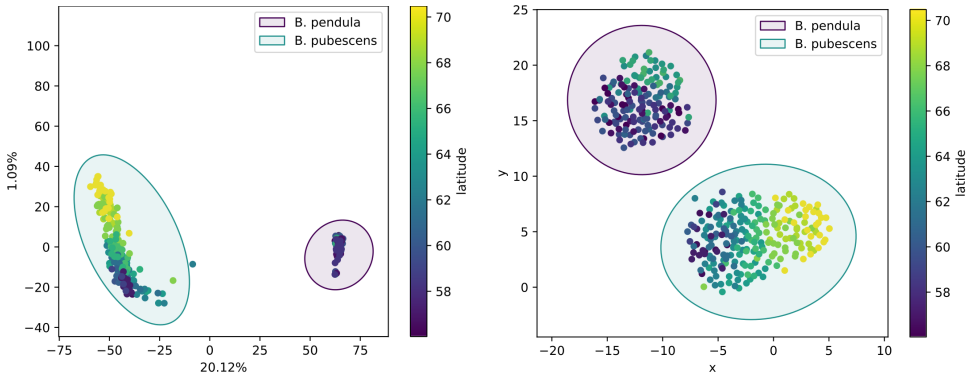
(a) both species (b) *B. pendula* (c) *B. pubescens*

Figure 8: Unfolded 2D site-frequency spectra whose subpopulations are shown on the axes. The ascertainment of these subpopulations is presented in the following section.

4 Population Structure

Population structure can be shaped by many forces, be it through local adaption, mating structure or demographic history. In this section we examine the structure and migration between and within populations of our two species. We do this by means of dimensionality reduction (PCA & UMAP), summary statistics (F_{ST}) and model-based approaches (ADMIXTURE, FEEMS). We mainly expect to find traces left behind by the postglacial recolonisation of different populations and adaption to different climatic environments as observed for Norway spruce [17].

4.1 PCA & UMAP

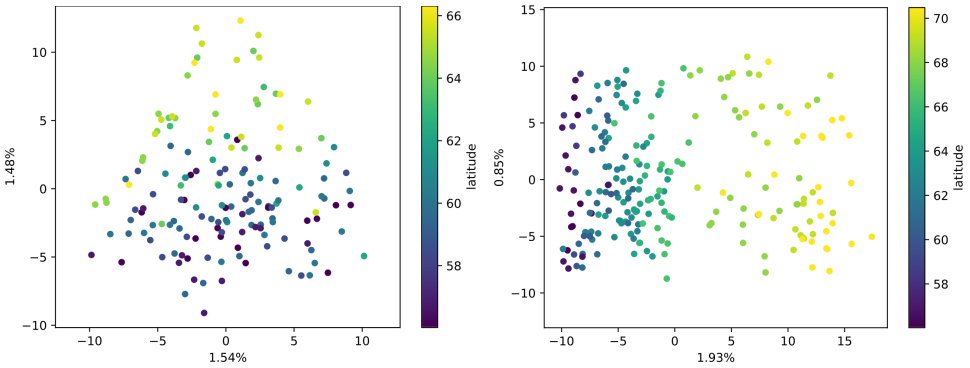


(a) PCA (b) UMAP

Figure 9: PCA and UMAP plots on the set of bi-allelic sites for *B. pendula* and *B. pubescens*. There is a clear separation between both species. For UMAP, the `spread` and `min_dist` parameters were set to 2 and 1.5, respectively.

Principal Component Analysis (PCA) employs linear dimensionality reduction thus retaining the proportions among distances and providing information on how much of the total variance is accounted for by each dimension. The Uniform Manifold Approximation and Projection (UMAP) analysis is more customisable at the cost of interpretability as the principal dimensions need not be linear. Fig. 9 shows PCA and UMAP plots for both species considered in tandem, and whose sites where restricted to be bi-allelic. There is a clear delineation separating the species indicating that hybrids are relatively

rare (with the exception of one or two intermediate samples). Considering the PCA plot, we see that roughly 20% of the total variance is explained by the type of species. The second principal component is much smaller with 1% and seems to account for latitudinal differentiation in *B. pubescens*. The UMAP plot looks similar but allows for a more detailed picture within each species brought about by specifying a minimum distance between samples (cf. fig. 9b). Here we can also observe weak latitudinal differentiation for *B. pendula*.



(a) *B. pendula* (b) *B. pubescens*

Figure 10: PCA plots on the set of synonymous sites for each of the two birch species. Two clusters are apparent for each species, especially for *B. pubescens*.

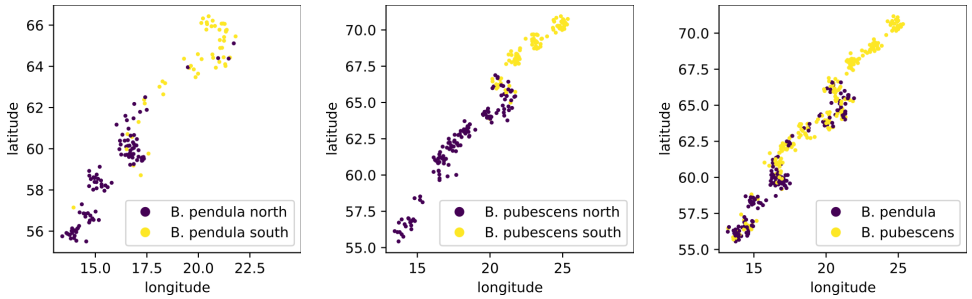
The two principal components of the intra-specific plots naturally account for much less of their total variance, being far more homogenous (cf. fig. 10 where sites were restricted to be synonymous for better visibility). For each species we can roughly see two clusters. This is especially apparent for *B. pubescens* where we observe a scarce region between the clusters and along the first principle component (cf. x-axis in fig. 10b). A smooth latitudinal gradient is also apparent along the same component indicating a northern and southern subpopulation. Restricting instead to bi-allelic or non-synonymous sites, the PCA and UMAP plots look very similar for *B. pubescens* (cf. figs. 35 - 37). For *B. pendula*, the second principal component varies latitudinally but much less smoothly so (cf. y-axis in fig. 10a). Using the set of non-synonymous sites instead, we do not observe any latitudinal gradient along the two principal components (cf. fig. 36a). This could indicate that the observed population structure in *B. pendula* is more due to past demography and not a result of local adaption . This could be in agreement with

the postglacial recolonisation of several distinct populations.

A subpopulation division seems appropriate, whose boundary was drawn along the 64th and 65th parallel for *B. pendula* and *B. pubescens*, respectively (cf. fig. 2). Interestingly, this boundary roughly corresponds to a transition from humid continental climate in the south to subarctic climate in the north according to the Köppen classification system. A boundary at a similar latitude was also determined for Norway spruce (cf. section 1.5) [17]. We also note that there is overall little population structure. Care has to be taken not to over-interpret the latitudinal gradient as the data itself mainly varies along that direction. Without prior knowledge or the scarce region separating the two clusters for *B. pubescens*, we could merely be led to conclude that there is isolation by distance.

4.2 ADMIXTURE

To corroborate and refine the population clustering deduced from the PCA and UMAP plots, the model-based clustering tool ADMIXTURE was used [1]. As output we obtain the percental allocation to the specified clusters per individual with the number of clusters being specified by parameter K. The set of bi-allelic sites was filtered for sites in strong LD using PLINK where we used a window and step size of 500 and 50 SNPs, respectively as well as an r^2 threshold of 0.2. To detect the rather weak population structure of *B. pendula*, $r^2 = 0.99$ was used instead, for sufficiently many sites to remain. Fig. 11 shows the sample locations for two clusters ($K = 2$) where each sample has been labeled according to its most similar cluster (cf. fig. 39 to see how these clusters correlate with the PCA). We observe a subpopulation structure very similar to that in the previous section (cf. figs. 9 & 10). However, ADMIXTURE's cross validation procedure favours only a single cluster for the intra-specific cases (cf. fig. 38).



(a) *B. pendula* (b) *B. pubescens* (c) both species

Figure 11: Perturbed sample locations labeled with the ADMIXTURE clustering for $K = 2$. The population structure closely resembles the one deduced from the PCA plots.

Considering the bar plots, it is apparent that there is admixture from *B. pendula* into (mainly southern) *B. pubescens* populations but not vice versa (cf. fig. 12). This is entirely consistent with our expectations on diploid-tetraploid gene flow [12, 37]. The southern and northern subpopulations within each of the two species are furthermore visibly admixed although there is considerably more admixture near the contact zone for *B. pubescens* (cf. figs. 42 & 44). Admixture among *B. pendula*'s subpopulations, on the other hand, is apparent much farther away from the contact zone, notably from the northern population far into the south (cf. fig. 42). This is likely due to *B. pendula*'s very weak population structure. No sensible clustering was evident for larger values of K (cf. figs. 41, 43 & 45).

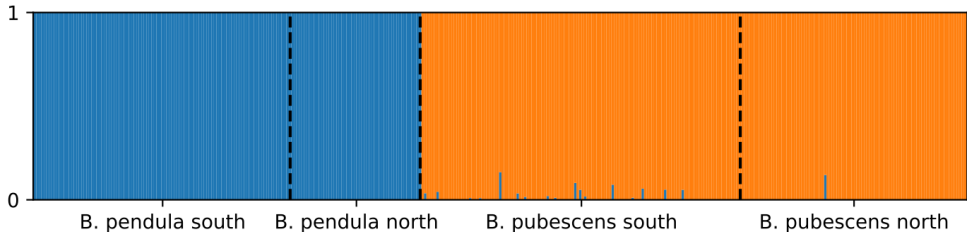


Figure 12: Bar plot for both species using $K = 2$. The samples are first sorted by the displayed (ADMIXTURE) subpopulations and second by latitude in ascending order. The clustering adheres to the species boundary. We also observe admixture from *B. pendula* into (mainly southern) *B. pubescens* individuals but not vice versa.

4.3 F_{ST}

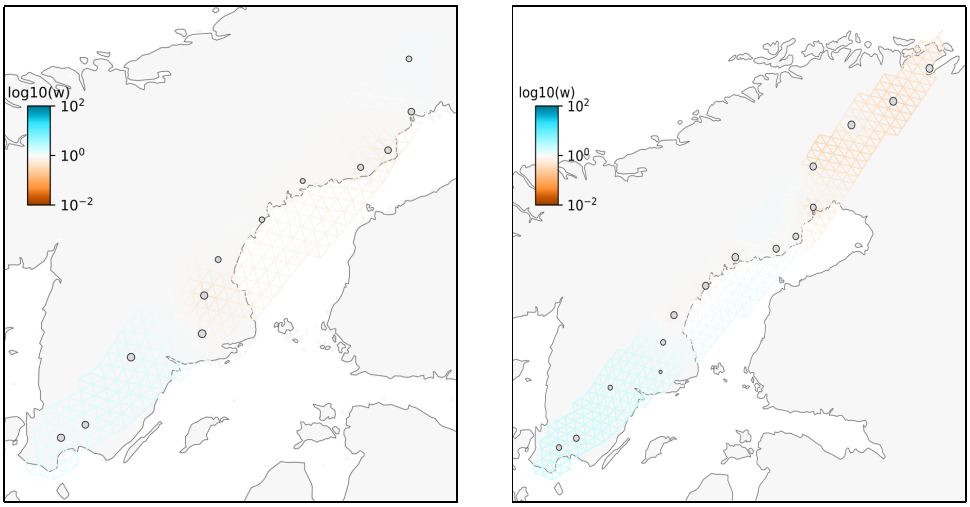
	mean F_{ST}	weighted F_{ST}
B. pendula / *B. pubescens*	0.0335	0.1951
B. pendula north / south	0.0026	0.0119
B. pubescens north / south	0.0025	0.0111

Table 2: Weighted as well as average F_{ST} per site.

Table 2 shows the population differentiation given by the fixation index F_{ST} which is a measure of subpopulation differentiation. The F_{ST} can be defined for a single site by $\sigma^2_{p_i}/\sigma^2_p$ where σ denotes the variance and p and p_i denote the frequency of a certain allele in the overall population and in the ith subpopulation, respectively. The denominator $\sigma^2_p = pq$ describes the variance of the site's allelic state in the overall population and the numerator $\sigma^2_{p_i} = \sum_j (p_j - p)^2/n$ the variance across the n different subpopulations (assuming they are of equal size)[11]. It is thus measuring how much of the total variance in allele frequencies is attributable to the variance across subpopulations. An F_{ST} of 0 indicates that there is no subpopulation differentiation of any kind while an F_{ST} of 1 would mean that all variance is due to subpopulation structure, i.e. there is no variance within subpopulations. *B. pubescens* was treated as a diploid for the site-wise calculations as VCFtools cannot handle polyploidy [26]. The weighted F_{ST} was inferred from the SFS using δaδi, and thereby properly reflects the different ploidies [14]. Both tools base their calculations on the method of Weir and Cockerham who advocate the weighted method [35]. We observe rather little differentiation between the northern and southern subpopulations within each of the two species as expected. A more marked subpopulation differentiation for *B. pubescens* as suggested by the PCA plots (cf. fig. 10) is not reflected in the values. Frequency distributions of the site-wise F_{ST} for *B. pendula* and *B. pubescens* can be found in the appendix (cf. fig. 28).

4.4 FEEMS

To determine the migration intensity between samples, FEEMS was used, which employs a model-based approach [19]. Performing the analysis, all samples are first assigned to the nearest node of the underlying spatial grid (cf. fig. 13). The smoothing parameter λ controls how much emphasis is put



(a) *B. pendula* (b) *B. pubescens*

Figure 13: Migration surface plots for both species with smoothing parameter $\lambda = 10$. The red and blue colours indicates below-average and above-average migration, respectively. Note that a buffer region has been added here for reasons of visibility. Unbuffered plots can be found in the appendix (cf. fig. 49).

on local structure and can be determined with leave-one-outcross-validation (cf. fig. 46). That is, one node is held out from the set of nodes for which the genotype is inferred from the remaining nodes under consideration of the estimated migration intensity between them. The final cross-validation error is then taken to be the average difference between expected and inferred genotype for each held-out node. Using warm starts for the cross-validation, where runs are initialised with the migration rates of the previous run, did not produce very different values of λ compared to cold starts where each run uses the same initialisation. Increasing the extend of the underlying grid by using a buffer of size 1 was furthermore necessary for the results to be better visible. Removing the buffer region, the results are optically less visible but we keep observing above and below-average migration in the south and north, respectively (cf. fig. 49). Preparing the data, sites with an allele frequency lower than 1% and a missingness greater than 50% were removed. The analysis was repeated several times for different thresholds yielding similar results although the cross-validation procedure often pro-

duced implausible values of λ (cf. fig. 47). We thus had to look for features in the migration surfaces that are visible for a large set of smoothing values and resort to common sense to gauge when a plot was overfitted. This approach is also recommended in the underlying article [19]. Choosing $\lambda = 10$ for both species provided rather smooth migration surfaces whose rates did not change unreasonably abrupt (cf. fig. 13). Choosing smaller values, we start to observe small isolated low-migration patches around some of the nodes and very abrupt changes in migration intensity, which is indicative of overfitting (cf. fig. 48).

Additional care has to be taken when interpreting these results as all the samples are distributed along a single latitudinal gradient which can potentially lead to artifacts. In fig. 13b and more clearly in fig. 48b, we observe a lower migration intensity along the axis comprising the samples compared to the buffer region for which there exist no additional samples. This does not seem entirely sensible and could be resolved by obtaining more samples that vary longitudinally. We can also argue, however, that migration between more distant populations is taken into account in the outer buffer regions where a lower migration rate would make sense.

In summary, we observe above-average migration in the south and below-average migration in the north. A partial barrier to gene flow, separating the northern and southern subpopulations apparent in the PCA plots, is not visible. Migration seems to be rather uniform, especially for *B. pendula* which can be interpreted to be in line with the less marked population structure in fig. 10a. The relatively homogenous migration rates also agree with there being overall little population structure. By choosing sufficiently large smoothing values, however, all data can be made to look homogenous. Nevertheless, it is reassuring that we obtain a similar picture for both species and different allele frequency cut-offs.

5 Demographic History

Ignoring selection and drift, what we are left with is a species' demography. Demography can be influenced by many factors like geographical boundaries, heterogeneous environmental pressures or mating system. In our case, we are mainly interested in population size changes and gene flow between different populations. We assume the mating structure to play only a minor role– birches being quite promiscuous with regards to where their pollen land. Here we attempt to find population size expansion after the Last Glacial Maximum and possibly gene flow between *B. pendula* and *B. pubescens* (cf. sections 1.4 & 1.5).

For our demographic analysis, we used an SFS-based tool called δaδi which explicitly models demographic scenarios using diffusion approximations [14]. The underlying equation essentially is a deterministic partial differential equation modelling the effects of drift, population size, selection and migration on the allele frequencies. Independent Poisson likelihoods are used for the number of sites in each frequency class. For our results not to be confounded by selection, we confine our analysis to synonymous sites which we assume to be neutrally-evolving. There being so many possible demographic scenarios, a particularly shaped SFS can result in many different ways. Choosing a particular model, e.g. a set of discrete population size changes in one population, we force the SFS to be modelled in only that manner.

5.1 Methods

Fig. 14 shows a schematic of the sample sets and population scenarios that were used for δaδi. We consider both one and two-population scenarios, with variable and fixed time and discrete as well as continuous population size changes. In the one-population scenarios we mostly consider population size changes whereas we can include migration in the two-population scenarios. For the fixed-time scenarios, the time has been fixed to roughly coincide with the end of the Last Glacial Maximum (about 16 000 BP). Variable-time models should provide better fits in most cases but fixing the time was nonetheless considered useful as it facilitated model comparison. That being said, the time since the LGM has to be interpreted with caution as our parameters are scaled by mutation rate and effective population size for which no accurate estimates are available.

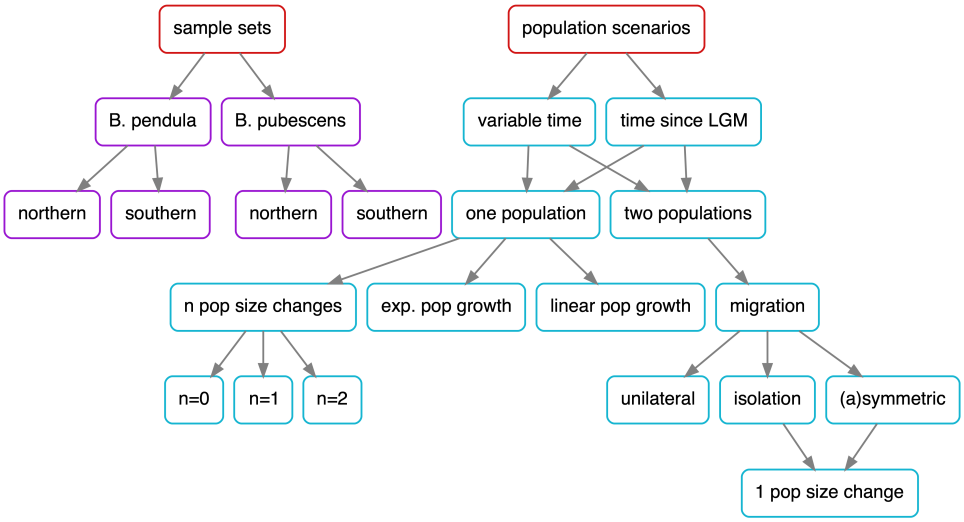


Figure 14: Schematic of the sample sets (left) and population scenarios (right) which were used for δaδi.

A mutation rate of $\mu = 10^{-9}$ per site per generation and generation time of $t = 20$ years was used [29]. The effective population size N_e was determined to be 945 883, 2 433 981 and 2 235 669 for *B. pendula* and *B. pubescens* and both species together, respectively. Here we calculated N_e based on the synonymous nucleotide diversity (i.e. $\pi_N = 4N_e\mu \Leftrightarrow N_e = \pi_N/(4\mu)$). This is a rather rough estimate and was solely used for the fixed-time scenarios. *B. pubescens*' effective population size is likely higher than *B. pendula*'s owing to a larger number of haplotypes per individual and introgression from *B. pendula* and *B. nana*. However, the above formula for π can be interpreted as μ multiplied by twice the expected coalescent time for two homologous sites. These coalescent times are likely higher between *B. pubescens*' two subgenomes which can lead to inflated estimates of N_e when assuming otherwise.

parameter	symbol	lower bound	upper bound	scaling
population size	ν	0.001	1000	N_e
migration rate	m	0	2	$2N_e$
time	t	0	1	$2N_e$

Table 3: Parameter bounds. Parameter ν is a fraction relative to the initial population size N_e. Migration is given by $m = 2N_e m_f$ where m_f is the fraction of individuals in the recipient population that come from the other population each generation. It is thus the effective number of individuals that migrate. Time is given by t in units of $2N_e$ generations.

Table 3 shows all commonly used parameters as well as a description and their upper and lower bounds. Most of the models are nested allowing for model comparison by means of Likelihood-Ratio Tests (LRTs). Apart from the likelihood, we also use the average Anscombe residual as an estimator of the models' goodness of fit. This quantity is not directly optimised for but should provide a good indication. The Anscombe residual quantifies the difference between the number of modelled and observed sites per frequency class [13]. It also normalises this difference with respect to its variance in addition to making it follow a normal distribution.

The parameter estimation for ∂a∂i consists of two steps (cf. fig. 52). First, to obtain solid estimates, a semi-global optimisation routine was performed by using 1000 parallelised BFGS local-optimisation runs that were randomly distributed in parameter space. For parameters with very wide bounds, a log-uniform sampling distribution was used. Basin hopping, a global optimisation technique, was also tested but did not yield satisfactory results. Bootstrap samples were generated using ∂a∂i's helper functions by sampling with replacement from several hundred contiguous chunks the genome was partitioned into. Using chunks instead of single sites is justified as it increases the samples' independence, linkage within chunks being greater than between [13]. A good chunk size was found to be 1 000 000 bases across, smaller sizes producing too many empty chunks. The bootstrap samples were then each initialised with the best result from the initial optimisation routine and optimised semi-locally using 5-10 separate BFGS runs of sequentially perturbed initial values. The first run produced the highest likelihood in most cases. ∂a∂i's standard deviation estimates by means of the Godambe Information Matrix deviated far from the manual bootstrap estimates.

5.2 One-Population Scenarios

scenario	$log(L)$	t	ν_0	ν_1	
constant size	-108 ± 9.5	0.33 ± 0.04	2.4 ± 0.13		
linear growth	-88 ± 6.0	0.21 ± 0.02	0.0012 ± 0.00059		$c = 488 \pm 111$
exp growth	-92 ± 7.1	0.049 ± 0.032	0.0034 ± 0.0035		$e = 25 \pm 4.6$
1-size change	-83 ± 5.4	0.067 ± 0.029	0.0044 ± 0.0017	39 ± 32	$s = 0.22 \pm 0.071$
constant size LGM	-115 ± 10	0.17	2.7 ± 0.21		
linear growth LGM	-93 ± 7.5	0.17	0.001 ± 2e-05		$c = 346 \pm 15$
exp growth LGM	-93 ± 7.2	0.17	0.016 ± 0.00047		$e = 18 \pm 2.1$
1-size change LGM	-82 ± 5.0	0.17	0.0013 ± 0.00025	354 ± 50	$s = 0.015 \pm 0.0026$

Table 4: 1D population scenarios for *B. pendula* after the LGM. The population size grows from ν_0 in the past to ν_1 in the present during time t. Parameter s denotes the fraction of t at which a discrete time change occurs and c and e parametrise the slope in the continuous growth scenarios. These parametrisations reduce to the constant size scenario and are thus appropriate for LRTs. The standard deviation has been calculated from 100 Bias-Corrected and accelerated (BCa) bootstraps and the point estimates are the average values thereof.

Table 4 shows the maximum likelihood estimates of some one-population scenarios for *B. pendula*. The standard deviation is reasonbly small indicating that a good local optimum could be found. The observed and modelled SFS as well as the population size trajectory of the fixed-time scenarios in table 4 are illustrated in figs. 15 - 18. The best fit was achieved by modelling a single population size change. Letting the time vary yielded similar time estimates whose likelihoods were not necessarily higher (cf. figs. 57 - 60). A comparison between fixed- and variable-time models by means of LRT's can be found in the appendix (cf. p-values in fig. 55). All tested scenarios for *B. pendula* indicate population expansion in the not too distant past.

Exponential population growth is the most natural growth model for unconstrained populations as it assumes that the future population size is a constant multiple of the current size. However, this type of growth can hardly be maintained over long periods of time for positive growth (e.g. since the end of the LGM) as there are limited amounts of resources and is thus not so realistic. Modelling two population size changes did not provide a significantly better fit than modelling only one change and yielded a slightly lower likelihood in some cases. This happened even though the two models

are nested and is likely due to a larger parameter space (2 more parameters) of the more complex model resulting in worse optimisation. Moreover, jointly modelling 1D scenarios for both birch species produced less stable albeit similar estimates.

The graphs for *B. pubescens* show somewhat different dynamics. The fixed-time models partly yielded very low likelihoods compared to the variable-time cases (cf. figs. 61 & 62). *B. pubescens*' large N_e estimate caused the estimated fixed time after the LGM to be rather short. To fit the model properly thus required a lot of mutations whence an unreasonably large relative effective population size ν was necessary so that the upper bound was reached ($\nu_{up} = 1000$). Such sizes seem rather large, even for drastic population changes, and suggest that *B. pubescens*' high polymorphism cannot be modelled sensibly in such a short amount of time. The fixed-time scenarios nevertheless indicate population expansion in all cases. *B. pubescens*' variable-time time models, in contrast, unanimously indicate population decline as well as providing a better fit. In light of *B. pubescens*' naive N_e estimate, these scenarios span several glaciations, however, so that this result is not necessarily contradictory. It also has to be noted that the signature left behind by the polyploid ancestry of *B. pubescens* is more complex which can complicate inference.

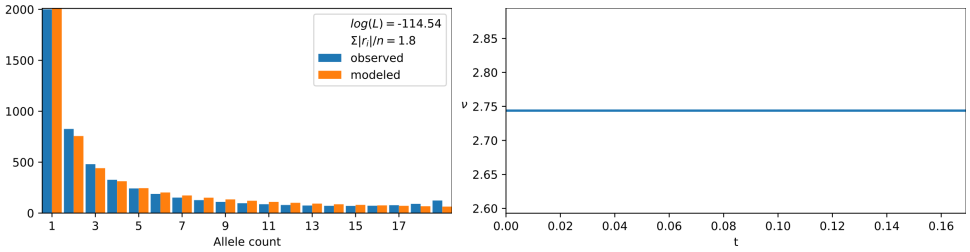


Figure 15: Constant population size scenario after the LGM for *B. pendula*. We can see the observed and the modelled SFS (left) as well as the population size trajectory (right). $\Sigma|r_i|/n$ denotes the average Anscombe residual and $log(L)$ the log-likelihood [13].

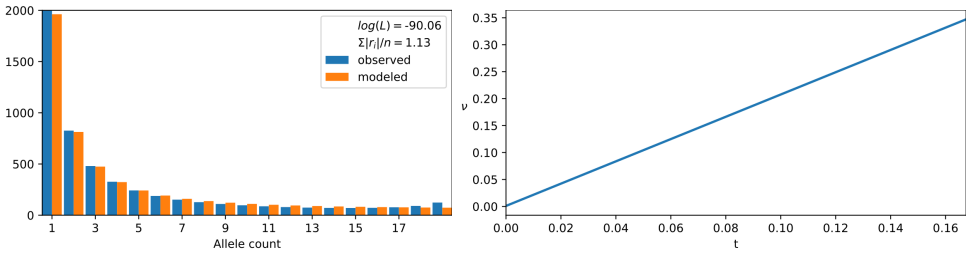


Figure 16: Linear population growth after the LGM for *B. pendula*. This model naturally performs better than the constant-size model but does not provide a very good fit overall.

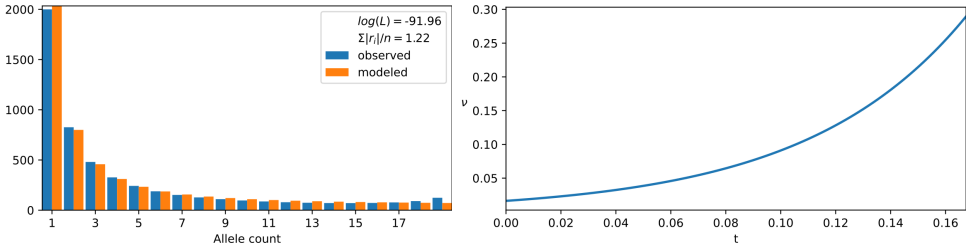


Figure 17: Exponential population growth after the LGM for *B. pendula*. Exponential growth is unlikely to be sustainable over long periods of time. This model performs worse than a single discrete population size change (cf. fig. 18).

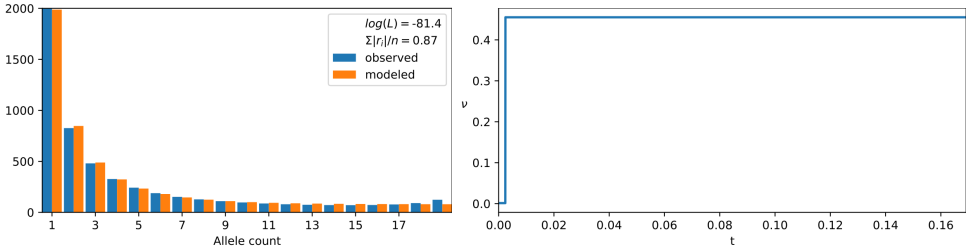
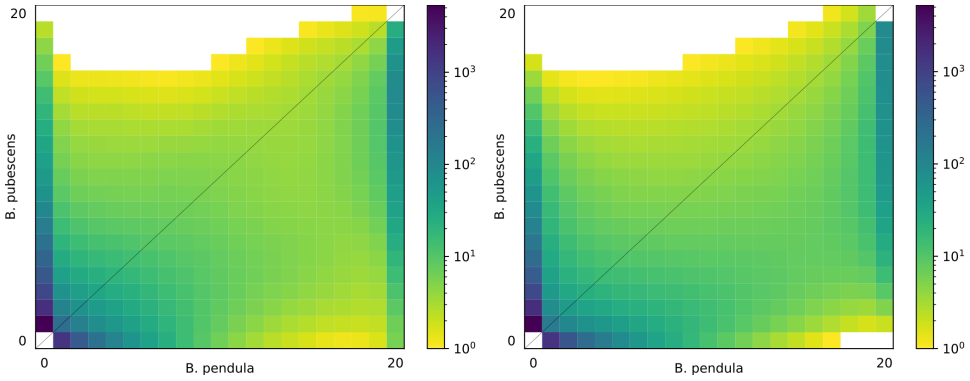


Figure 18: One population size change after the LGM for *B. pendula*. This scenario provides the best fit among all tested 1D population scenarios.

The one-population models presented here are rather simple but there is not much more that can be included in the absence of selection. None of the

models manages to completely capture the excess of high-frequency derived alleles we observe for *B. pendula* (cf. figs. 15 - 17). This excess could be due to a more fine-scale population structure, ancestral misidentification errors, selection acting directly on synonymous sites or multiple-merger coalescents where some individuals produce disproportionally many offspring. We nevertheless obtain lower average residuals than for *B. pubescens* for which we have a deficit of high-frequency derived alleles (cf. figs. 61 - 68). In summary, the results indicate population expansion in the not too distant past which is in line with postglacial recolonisation (given that we approximated the time after the LGM reasonably well). The models were also tested for different fixed time values which suggested similar results. *B. pubescens*' estimates may very well be confounded by the signatures of its polyploid ancestors but the results also indicate recent population expansion.

To confirm our hypothesis about population expansion, Tajima's D was calculated whose values were determined to be -0.95, -1.49 and -1.73 for *B. pendula*, *B. pubescens* and both species together, respectively. VCFtools was used for the calculations, specifying a window size of 1000 bps from which a weighted average was calculated (cf. fig. 29 for frequency distributions). Inferring Tajima's D from the SFS using δaδi provided values of -1.36, -1.79 and -1.92 which are consistently lower and should be less biased as *B. pubescens*'s proper ploidy could be used. Negative values of Tajima's D are common in practice, but could, in principle, be indicative of population expansion. They are, however, well within the margins of the basic coalescent's expectations [15, 11].



(a) observed (b) modelled

Figure 19: Observed and modelled SFS for asymmetric migration between the two species together with population growth modelled by a single population size change. This is the model with the highest likelihood and lowest average residual among all 2D scenarios.

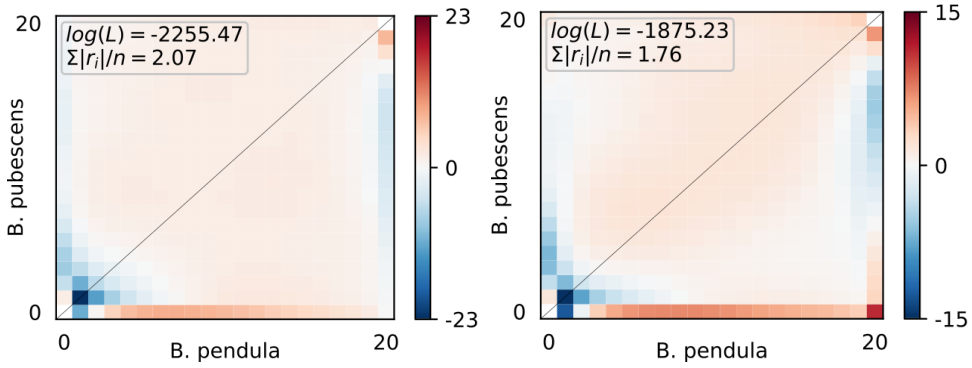
5.3 Two-Population Scenarios

Modelling scenarios involving two populations, we can include migration which causes the site-wise allele frequencies of the populations to be correlated. Migration cannot so easily be confounded with other dynamics as the only other explanation for correlated allele frequencies is a common origin or an extremely similar demography. It is, however, more difficult to know when migration occurred and at which intensity. Intuitively, extensive recent admixture can have a similar signature to more moderate admixture further in the past. Table 5 displays parameter estimates for several 2D scenarios involving migration between *B. pendula* and *B. pubescens*. Here we consider different types of migration and possibly include population growth. Modelling migration between the southern and northern subpopulations of each species continuously hit the upper migration rate boundaries (cf. table 3). High rates being computationally very expensive, no attempts were made to estimate them further. Some MLE estimates for t are close to their upper bound ($\nu_{up} = 1$) but we found estimates for larger values of ν_{up} to be quite similar.

scenario	$log(L)$	t	ν_{pub}	ν_{pen}	$m_{pen\to pub}$	$m_{pub\to pen}$
migr_{none}	-2260 ± 89	0.24 ± 0.0083	4.9 ± 0.21	0.43 ± 0.023	0	0
$\text{migr}_{pen\to pub}$	-1371 ± 51	0.95 ± 0.047	3.2 ± 0.14	0.78 ± 0.038	1.3 ± 0.06	0
$\text{migr}_{pub\to pen}$	-1995 ± 80	0.4 ± 0.024	4.5 ± 0.15	0.35 ± 0.017	0	1.1 ± 0.069
migr_{sym}	-1435 ± 58	0.9 ± 0.05	4.0 ± 0.19	0.47 ± 0.028	0.69 ± 0.035	0.69 ± 0.035
migr_{asym}	-1280 ± 50	1.0 ± 0.0044	3.3 ± 0.13	0.65 ± 0.02	1.1 ± 0.045	0.22 ± 0.019
migr_{sym} + growth	-1227 ± 46	0.82 ± 0.058	3.2 → 1.4	0.19 → 3.5	0.73 ± 0.033	0.73 ± 0.033
migr_{asym} + growth	-1091 ± 33	0.99 ± 0.031	0.59 → 6.6	0.37 → 2.5	1.1 ± 0.037	0.31 ± 0.016

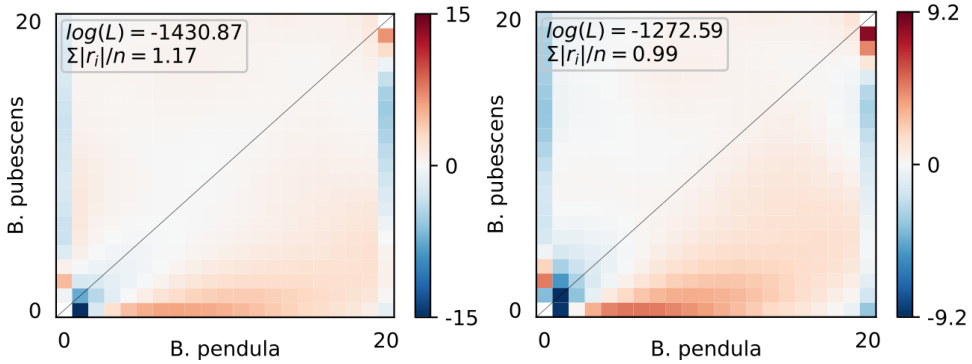
Table 5: 2D migration scenarios. Parameters ν_{pub} and ν_{pen} denote the relative population size of *B. pubescens* and *B. pendula*, respectively. Migration from *B. pendula* to *B. pubescens* is denoted by $m_{pen\to pub}$ and vice versa. The subscripts *sym*, *asym* and *none* denote symmetric, asymmetric and no migration, respectively. Population growth is modelled by a single discrete population size change. Everything takes place during time t and $log(L)$ denotes the log-likelihood. For the scenarios including growth we write $\nu_0 \to \nu_t$ for the values of ν at time 0 and t. The standard deviation has again been calculated from 100 Bias-Corrected and accelerated (BCa) bootstraps and the point estimates are the average values thereof.

Migration from *B. pendula* into *B. pubescens* is consistently estimated to be higher than vice versa which agrees with the expectations about polyploid introgression (cf. section 1.4). Figs. 20 - 22 visualise the residuals for some of the models. The most complex model–asymmetric migration with population growth–provides the highest likelihood (cf. fig. 19). P-value comparisons for the nested models in table 5 as well as for fixed vs variable-time models can be found in the appendix (cf. fig. 56). Similar to the one-population case for *B. pubescens*, we obtain significantly lower likelihoods when trying to fix the time to the end of the LGM (cf. figs. 69 - 71 and p-values in fig. 56a). This is again due the very short time span integrated over.



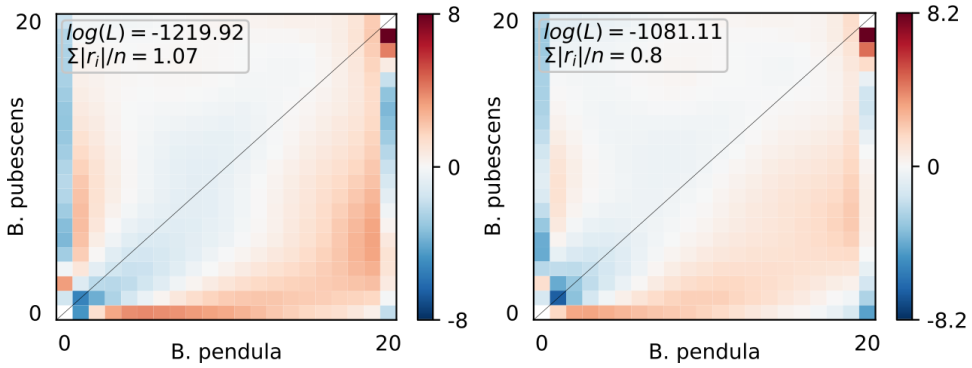
(a) no migration + constant size (b) no migration + 1-size change

Figure 20: Anscombe residuals for 2D scenarios without migration over variable time.



(a) symmetric migration (b) asymmetric migration

Figure 21: Anscombe residuals for 2D scenarios with (a)-symmetric migration and a constant population size over variable time.



(a) sym. migration + 1-size change (b) asym. migration + 1-size change

Figure 22: Anscombe residuals for 2D scenarios with (a)-symmetric migration and population growth over variable time. These scenarios are the most complex ones and achieve significantly higher likelihoods. The observed and modelled SFS of fig. 22b are shown in fig. 19.

5.4 Discussion

The demographic analyses mostly confirmed our expectations on recent population expansion and asymmetric gene flow between the two species. The lack of precise parameter estimate for N_e and μ make more precise inferences difficult within this framework. Coalescent simulations could shed light on more fine-scale population structure or more detailed population size trajectories but the problem of N_e remains. Having more precise trajectories, one might be able to discern several glaciations. The models also support gene flow from *B. pubescens* to *B. pendula*, although at lower rates which is not implausible. In fact, there has also been found to be bi-directional gene flow between *B. pubescens* and *B. nana* in Iceland [2].

6 The Distribution of Fitness Effects

In section 3.2, we already mentioned that most non-synonymous mutations are deleterious. This makes sense intuitively for coding sites considering that a protein with its extremely fine-tuned structure will most likely perform worse when randomly changing one of its amino acids. If we are interested in how likely a novel mutation modifies the carrier's fitness by a certain amount, we are led to the concept of the Distribution of Fitness Effects (DFE) which is simply a frequency distribution of selection coefficients for derived genetic variants. The DFE can be either determined empirically by conducting mutagenesis experiments or theoretically by inferring it from polymorphism data. Knowing the DFE of a certain organism can help resolve an array of important questions regarding the accumulation of deleterious mutations or the nature of quantitative traits [10].

6.1 Methods

We used polyDFE for the inference [31]. As input, an unfolded SFS for both synonymous and non-synonymous sites is required. The synonymous sites are assumed to be neutrally-evolving, in addition to assuming that demography affects synonymous and non-synonymous sites in the same manner. The difference in the non-synonymous SFS compared to the synonymous one is thus caused by selection so that we can infer the strength of selection by controlling for all other factors, i.e. demography. A similar principle is put to use in the well-known McDonald-Kreitman test where the ratio of non-synonymous to synonymous polymorphism within a species is compared to the ratio of non-synonymous to synonymous substitutions across species [21]. In polyDFE, however, passing divergence data is optional and was avoided in this analysis. Besides that, we are required to specify the total number of surveyed sites. Maximum Likelihood Estimation (MLE) is again used to find the best fitting parameters with the expected allele frequencies being inferred from Poisson Random Field theory which is diffusion-based. This is similar to how ∂a∂i works. Poisson likelihoods are used again, but this time for each site independently (assuming that we do not choose to model variable mutation rates). This is opposed to independent Poisson likelihoods per frequency class in ∂a∂i. Nuisance parameters are used to correct for demography, and ancestral misidentification can optionally be included by introducing a single parameter modelling the probability of any site being assigned the wrong ancestral state (cf. fig. 23).

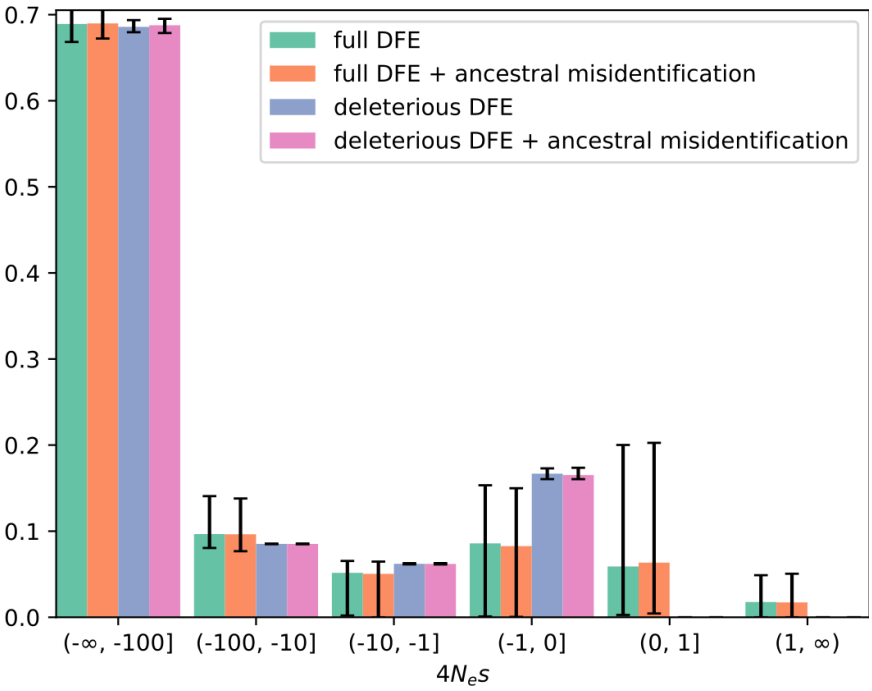


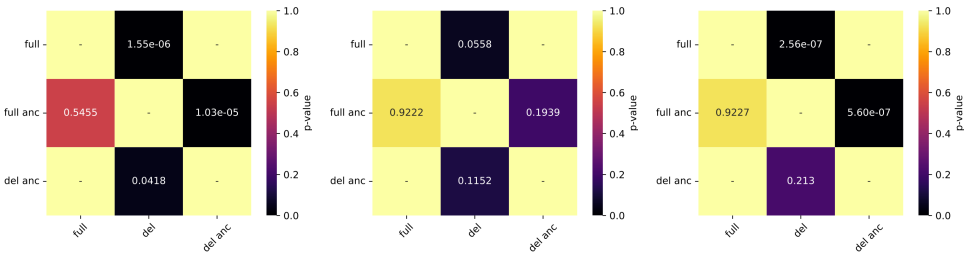
Figure 23: Different types of DFE for *B. pendula*. Correcting for ancestral misidentification does not change the distribution considerably.

The DFE's default shape is assumed to follow a reflected gamma distribution for non-positive selection coefficients and an exponential distribution for positive ones. This model is used for all displayed DFEs if not otherwise specified and is polyDFE's default model. Different models and a brief comparison between them can be found in the appendix (cf. figs. 73 - 75). The exponential shape for positive selection coefficients, i.e. a much higher probability of slightly advantageous mutations compared to strongly advantageous ones, seems plausible when considering Fisher's Geometric Model (FGM). That is, if we assume the fitness of an organism to be close to a local maximum as well as being determined by many interdependent parameters (e.g. SNPs), then a large fitness increase by changing only one of the parameter seems unlikely. The distribution for negative selection coefficients is more complex, possibly being multimodal (i.e. having several maxima). This could result from there being several component distributions that overlap [10]. A gamma distribution, being more flexible and able to assume both the shape of an exponential distribution and a skewed normal-like distribution, is thus

more appropriate. In polyDFE, we can also distinguish between estimating only a deleterious DFE, i.e. only considering negative selection coefficients, and a full DFE where we also consider beneficial mutations (cf. fig. 23). These models are nested and are thus eligible for model comparison through Likelihood-Ratio Tests (cf. fig. 24).

Confidence intervals were determined using 100 Bias-Corrected and accelerated (BCa) bootstraps where the bootstrap samples' allele frequencies were drawn from a Poisson distribution (cf. fig. 77 for a comparison with percentile bootstraps). The BFGS algorithm was used which performs local optimisation and is polyDFE's default algorithm. The species' proper ploidies are also reflected properly, the input being SFS-based.

6.2 Results



(a) both species (b) *B. pendula* (c) *B. pubescens*

Figure 24: Visualisation of p-values of LRTs for various DFE models. The null hypothesis assumes that the more complex model (positioned on the vertical axis) does not provide a better fit to the data. The axis labels *full anc* and *del* denote the *full DFE + ancestral misidentification* and *deleterious DFE*, respectively.

Fig. 23 displays *B. pendula*'s DFE for various nested models. Including ancestral misidentification does not seem to affect the distribution significantly which is reflected in the rather large non-significant p-values in fig. 24a. Including advantageous mutations, however, does provide a much better fit which is again apparent in the minute p-values, allowing us to reject the null hypothesis. The significantly larger variance of the effect size and proportion of beneficial mutations visible in fig. 23 can be explained by the scarcity of such mutations, having a much smaller impact on the SFS in comparison

to deleterious mutations [6]. Strongly advantageous mutations moreover quickly reach fixation, further reducing the number of possible observations [31].

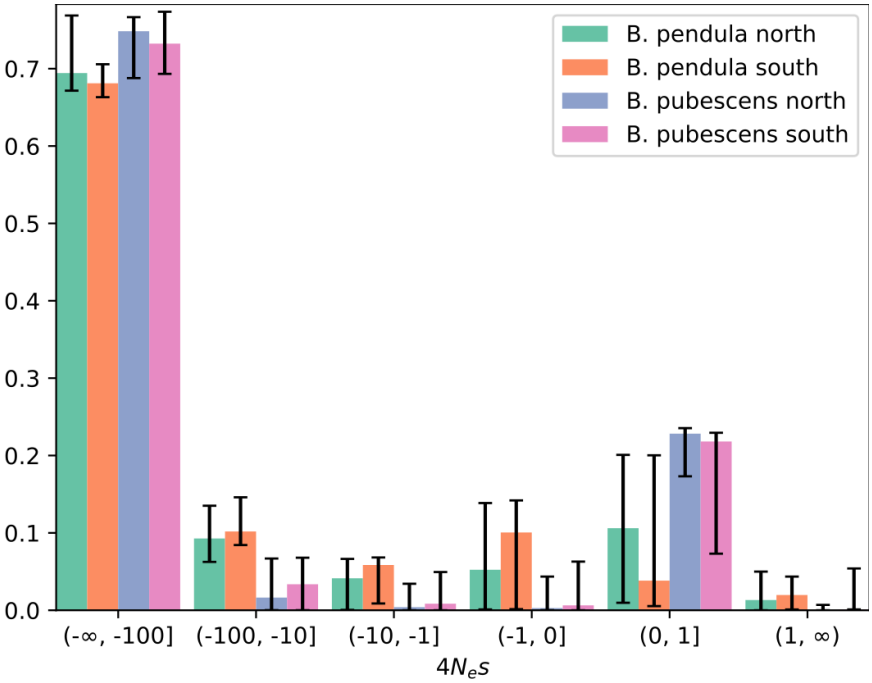


Figure 25: The full DFE for the four different subpopulations. The vertical bars indicate 95% confidence intervals. There seem to be more slightly advantageous mutations for *B. pubescens*.

A comparison of the DFE for the two southern and northern subpopulations is shown in fig. 25. There seem to be rather large differences in the number of beneficial mutation between the two species. *B. pubescens* is estimated to have many more slightly advantageous mutations but all confidence intervals, in fact, overlap. This is in contrast with the deleterious DFEs which have very narrow confidence intervals and are strikingly similar among all subpopulations (cf. fig. 26). For all analyses, the optimisation runs completed with a gradient reasonably small and larger gradients do not seem to introduce a bias to the estimated parameters in general (cf. fig. 78). Furthermore, the DFE seems to be rather robust to the bias introduced by calling *B. pubescens* as diploid (cf. fig. 76).

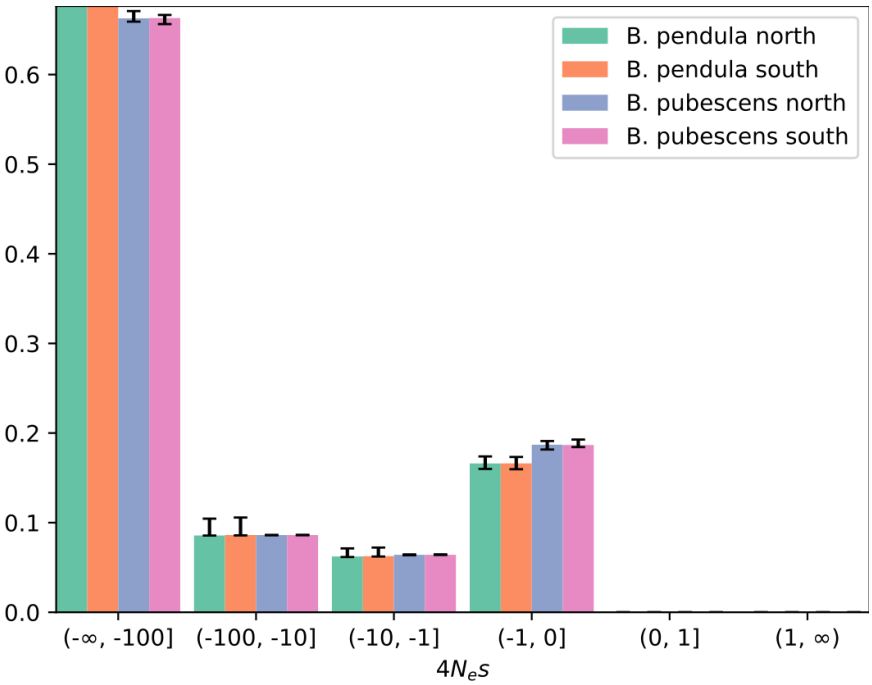


Figure 26: The deleterious DFE for the four different subpopulations. The shape is remarkably similar between the two species.

6.3 Discussion

The shape of the DFE is strikingly similar between the two species (cf. fig. 26). This is perhaps not that surprising when considering the species' related ancestry, in particular, if a species similar to *B. pendula* was the autopolyploid ancestor of *B. pubescens* (cf. section 1.5). In a recent paper, the shape of the deleterious DFE has been found to be very similar across several great ape species [5]. This similarity between closely related species could be confirmed for the two birch species. Shared ancestry thus seems to be an important determinant in the similarity of the DFE, the bulk of allele frequencies taking a long time to change after speciation, after all. It can also be argued, however, that closely related species have a rather similar genomic architecture and hence the DFE is unlikely to change considerably between them. That being said, different species could be subject to environmental pressures of a very different nature and intensity after speciation, which could change the DFE significanly.

The full DFE differs substantially in the amount of beneficial mutations among the two species although the variance in the amount of such mutations is rather high. The amount of beneficial mutations inferred from the SFS is very sensitive to ancestral misidentification as such mutations are rare and expected to segregate at high frequencies [31]. polyDFE's ancestral misidentification parameter can properly correct for that error provided that the probability of misidentification is the same over all frequency classes which might not be the case, especially when using more elaborate methods for inferring the ancestral states. In the great apes paper, Likelihood-Ratio Tests did not favour a DFE including beneficial mutations which, in our case, is true for *B. pendula* but not for *B. pubescens* (cf. fig. 24). Moreover, the fraction of selectively neutral ($|4N_e s| < 1$) or more strongly advantageous ($4N_e s \geq 1$) mutations is consistently estimated to be larger in *B. pubescens* (compared to *B. pendula*) which is in line with a higher value of π_N/π_S (cf. section 2.4). In all cases, the fraction of such mutations is nevertheless lower than suggested by π_N/π_S which could be due to the fixation of more strongly deleterious mutations ($4N_e s \leq 1$).

7 Closing Words

I hope to have provided useful analyses as well as a clear overview of the demography and population structure of *B. pendula* and *B. pubescens*, two important boreal species, not only to Scandinavia but to large parts of Eurasia as a whole. The polyploid nature of *B. pubescens* presented significant obstacles to correct inference but offered, at the same time, many interesting and thought-provoking dynamics. Working with two species in tandem also provided useful comparisons and sanity checks although it was not always immediately apparent where differences stem from. There were also hurdles in the inference of demography, partly because little is known about important parameters like mutation rate, effective population size and generation time. The developed pipeline supplies some useful abstractions to the tools used in this work and will hopefully find repeated use, if only by the author.

To summarise, this work confirms the existence of a northern and a southern cluster in each species, unidirectional introgression from *B. pendula* into *B. pubescens*, and population expansion after the LGM. The DFE of new mutations between the two species is also remarkably similar and the amount of deleterious mutations seems acceptable. There is thus hope that these species will be able to adapt to future environmental challenges and that *B. pendula* continues being a good and sustainable resource for timber. Much remains to be done to unravel the complex interplay and history of these rather flexible species whose prosperity is crucial to a functioning boreal ecosystem.

8 References

[1] D.H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19:1655–1664, 2009.

[2] Kesara Anamthawat-Jonsson and Aegir Thorsson. Natural hybridisation in birch: Triploid hybrids between betula nana and b. pubescens. *Plant Cell Tissue and Organ Culture*, 75:99–107, 11 2003.

[3] S Andrews. Fastqc: A quality control tool for high throughput sequence data. 2010.

[4] Anthony Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics (Oxford, England)*, 30, 04 2014.

[5] David Castellano, Moisès Coll Macià, Paula Tataru, Thomas Bataillon, and Kasper Munch. Comparison of the full distribution of fitness effects of new amino acid mutations across great apes. *Genetics*, 213:genetics.302494.2019, 09 2019.

[6] Jun Chen, Thomas Bataillon, Sylvain Glémin, and Martin Lascoux. What does the distribution of fitness effects of new mutations (dfe) reflect? insights from plants. *New Phytologist*, 10 2021.

[7] Jun Chen, Sylvain Glémin, and Martin Lascoux. Genetic Diversity and the Efficacy of Purifying Selection across Plant and Animal Species. *Molecular Biology and Evolution*, 34(6):1417–1428, 02 2017.

[8] Jun Chen, Lili Li, Pascal Milesi, Gunnar Jansson, Mats Berlin, B. Karlsson, Jelena Aleksić, Vendramin Giovanni Giuseppe, and Martin Lascoux. Genomic data provide new insights on the demographic history and the extent of recent material transfers in norway spruce. *Evolutionary Applications*, 12, 09 2019.

[9] Richard Durrett. *Probability Models for DNA Sequence Evolution*. 01 2008.

[10] Adam Eyre-Walker and Peter Keightley. The distribution of fitness effects of new mutations. *Nature reviews. Genetics*, 8:610–8, 09 2007.

[11] John H. Gillespie. *Population Genetics: A Concise Guide*. The Johns Hopkins University Press, Baltimore, Maryland, 2004.

[12] Stebbins GL. *Chromosomal Evolution in Higher Plants.* Edward Arnold, London, 1971.

[13] Ryan Gutenkunst. dadi - read the docs. `https://dadi.readthedocs.io/en/latest/`, 2021.

[14] Ryan Gutenkunst, Ryan Hernandez, Scott Williamson, and Carlos Bustamante. Gutenkunst rn, hernandez rd, williamson sh, bustamante cd. inferring the joint demographic history of multiple populations from multidimensional snp data. plos genet 5: e1000695. *PLoS genetics*, 5:e1000695, 10 2009.

[15] Carsten Wiuf Jotun Hein, Mikkel H. Schierup. *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory.* Oxford University Press Inc., New York, 2005.

[16] Peter Keightley and Benjamin Jackson. Inferring the probability of the derived versus the ancestral allelic state at a polymorphic site. *Genetics*, 209:genetics.301120.2018, 05 2018.

[17] Lili Li, Pascal Milesi, Mathieu Tiret, Jun Chen, Janek Sendrowski, John Baison, Zhiqiang Chen, Linghua Zhou, B. Karlsson, Mats Berlin, Johan Westin, Maria García-Gil, Harry Wu, and Martin Lascoux. Teasing apart the joint effect of demography and natural selection in the birth of a contact zone, 01 2022.

[18] Nina Marchi and Laurent Excoffier. Gene flow as a simple cause for an excess of high-frequency-derived alleles. *Evolutionary Applications*, 13, 10 2020.

[19] Joseph Marcus, Wooseok Ha, Rina Barber, and John Novembre. Fast and flexible estimation of effective migration surfaces. *eLife*, 10, 07 2021.

[20] Felix Mölder, Kim Jablonski, Brice Letcher, Michael Hall, Christopher Tomkins-Tinch, Vanessa Sochat, Jan Forster, Soohyun Lee, Sven Twardziok, Alexander Kanitz, Andreas Wilm, Manuel Holtgrewe, Sven Rahmann, Sven Nahnsen, and Johannes Köster. Sustainable data analysis with snakemake. *F1000Research*, 10:33, 04 2021.

[21] John H. McDonald and Martin Kreitman. Adaptive protein evolution at the adh locus in drosophila. *Nature*, 351:652–654, 06 1991.

[22] Leland McInnes and John Healy. Umap: Uniform manifold approximation and projection for dimension reduction. 02 2018.

[23] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark DePristo. The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20:1297–303, 09 2010.

[24] Gil L. Hunt S.E. et al. McLaren, W. The ensembl variant effect predictor. *Genome Bio 17, 122*, page genetics.301120.2018, 06 2016.

[25] Anna Palme, Qiao Su, Anja Rautenberg, F Manni, and Martin Lascoux. Postglacial recolonization and cpdna variation of silver birch, betula pendula. *Molecular ecology*, 12:201–12, 02 2003.

[26] Goncalo Abecasis Cornelis A. Albers Eric Banks Mark A. DePristo Robert Handsaker Gerton Lunter Gabor Marth Stephen T. Sherry Gilean McVean Richard Durbin Petr Danecek, Adam Auton and 1000 Genomes Project Analysis Group. The variant call format and vcftools. *Bioinformatics*, 2011.

[27] Shaun Purcell, Benjamin Neale, Katherine Todd-Brown, Lori Thomas, Manuel Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul Bakker, Mark Daly, and Pak Sham. Plink: A tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, 81:559–75, 10 2007.

[28] Jarkko Salojarvi. Betula pendula scaffold assembly. `https://genomevolution.org/CoGe/GenomeInfo.pl?gid=35079`, 2017.

[29] Jarkko Salojarvi, Olli-Pekka Smolander, Kaisa Nieminen, Sitaram Rajaraman, Omid Safronov, Pezhman Safdari, Airi Lamminmäki, Juha Immanen, Tianying Lan, Jaakko Tanskanen, Pasi Rastas, Ali Amiryousefi, Balamuralikrishna Jayaprakash, Juhana Kammonen, Risto Hagqvist, Gugan Eswaran, Viivi Hassinen, Juan Alonso-Serra, Fred Asiegbu, and Jaakko Kangasjärvi. Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch. *Nature Genetics*, 49, 05 2017.

[30] Janek Sendrowski. Code and supplementary material. `https://github.com/Sendrowski/BirchesScandinavia`, 2022.

[31] Paula Tataru and Thomas Bataillon. *polyDFE: Inferring the Distribution of Fitness Effects and Properties of Beneficial Mutations from Polymorphism Data*, volume 2090, pages 125–146. 01 2020.

[32] Camille Truong, Anna Palme, and François Felber. Recent invasion of the mountain birch betula pubescens ssp. tortuosa above the treeline due to climate change: Genetic and ecological study in northern sweden. *Journal of evolutionary biology*, 20:369–80, 02 2007.

[33] Yoshiaki Tsuda, Vladimir Semerikov, Federico Sebastiani, Vendramin Giovanni Giuseppe, and Martin Lascoux. Multispecies genetic structure and hybridization in the betula genus across eurasia. *Molecular Ecology*, 26, 10 2016.

[34] Nian Wang, Laura Kelly, Hugh Mcallister, Jasmin Zohren, and Richard Buggs. Resolving phylogeny and polyploid parentage using genus-wide genome-wide sequence data from birch trees. *Molecular Phylogenetics and Evolution*, 160:107126, 02 2021.

[35] Bruce Weir and C. Cockerham. Weir bs, cockerham cc. estimating f-statistics for the analysis of population-structure. evolution 38: 1358-1370. *Evolution*, 38:1358–1370, 11 1984.

[36] Janis Wigginton and Goncalo Abecasis. A note on exact tests of hardy-weinberg equilibrium. *American journal of human genetics*, 76:887–93, 06 2005.

[37] Jasmin Zohren, Nian Wang, Igor Kardailsky, James Borrell, Anika Joecker, Richard Nichols, and Richard Buggs. Unidirectional diploid-tetraploid introgression among british birch trees with shifting ranges shown by rad markers. *Molecular ecology*, 25, 04 2016.

9 Appendix

9.1 Pipeline

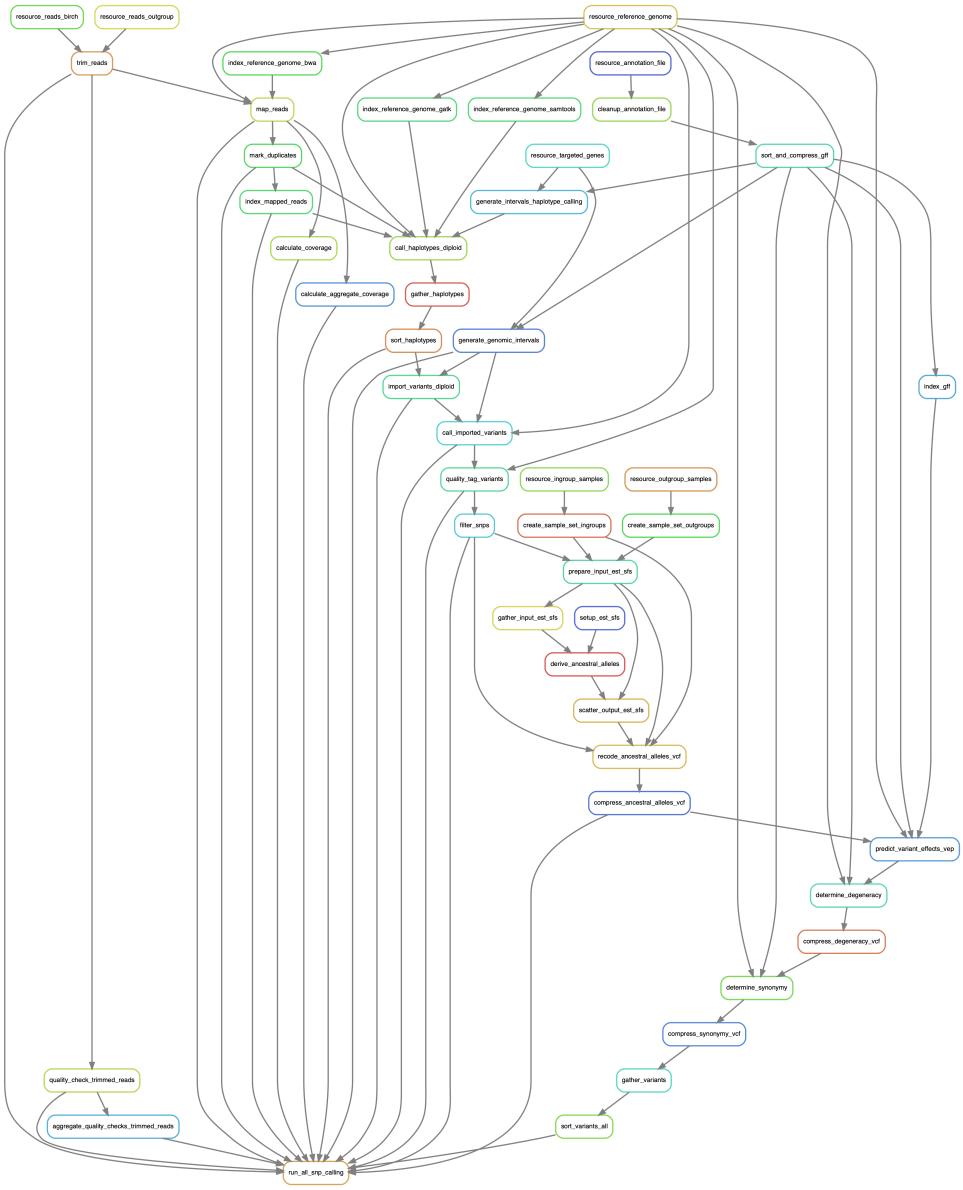
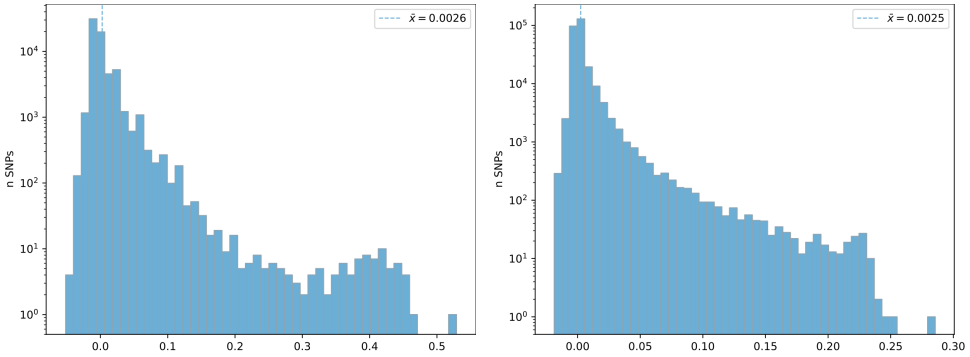


Figure 27: Workflow to obtain the initial VCF file from the raw reads.

9.2 Summary Statistics

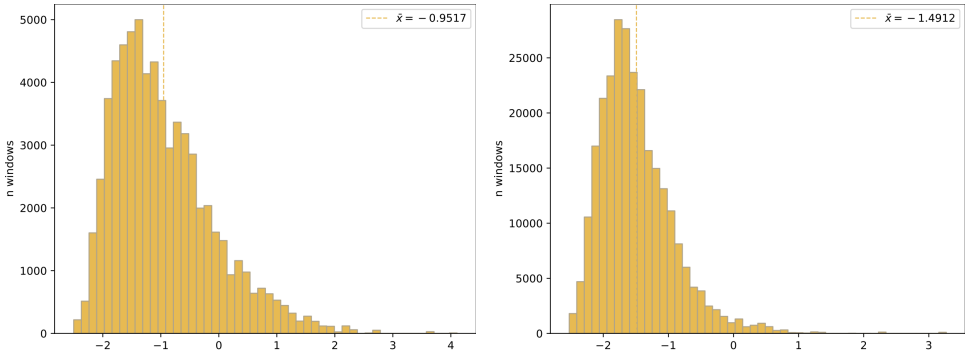
9.2.1 F_{ST}



(a) *B. pendula* (b) *B. pubescens*

Figure 28: Log-scaled frequency distributions for site-wise values of the F_{ST}.

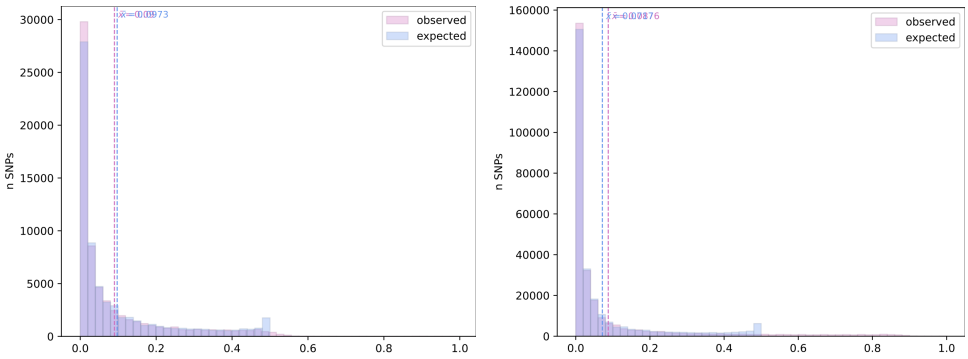
9.2.2 Tajima's D



(a) *B. pendula* (b) *B. pubescens*

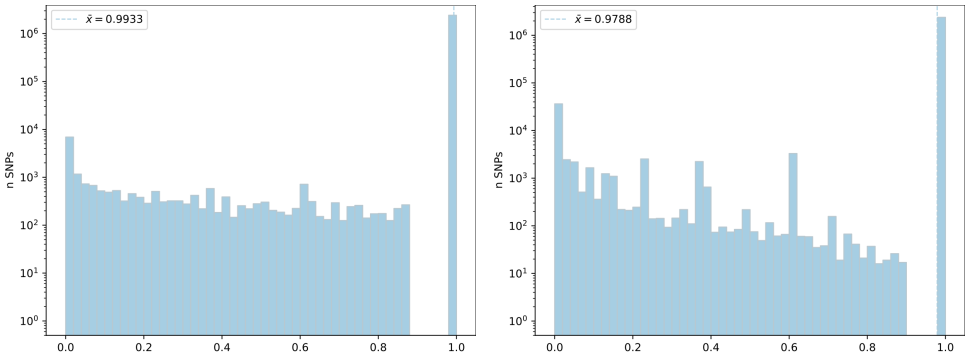
Figure 29: Frequency distributions for values of Tajima's D with a window size of 1000 bps.

9.2.3 Heterozygosity



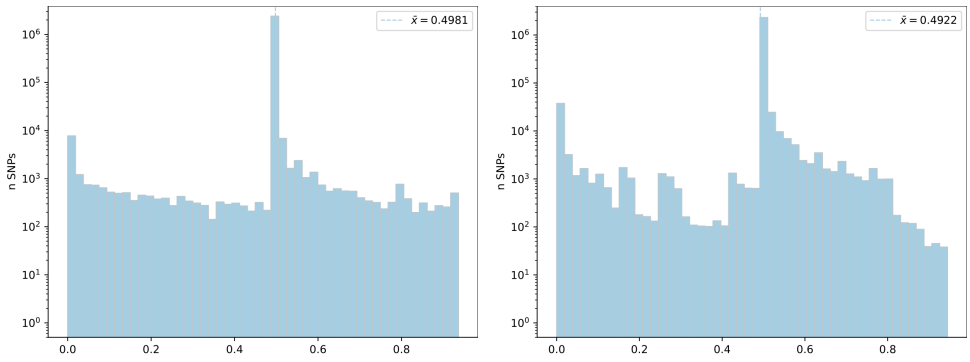
(a) *B. pendula* (b) *B. pubescens*

Figure 30: Linearly scaled plots of the expected and observed heterozygosity on the set of bi-allelic sites.



(a) *B. pendula* (b) *B. pubescens*

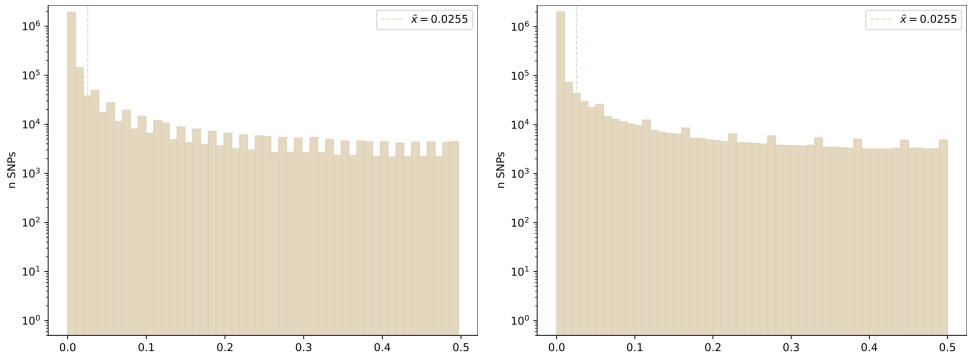
Figure 31: Site-wise p-values for observed heterozygosity under HWE. The two plots and their average p-value are only partly comparable, however, as they depend on the total number of SNPs and their frequencies. Almost all sites are either monomorphic for which we obtain a p-value of 1, or have alleles that segregate far from non-intermediate frequencies for which p-values of 1 are very likely since there is little statistical power to detect deviations from HWE.



(a) *B. pendula* (b) *B. pubescens*

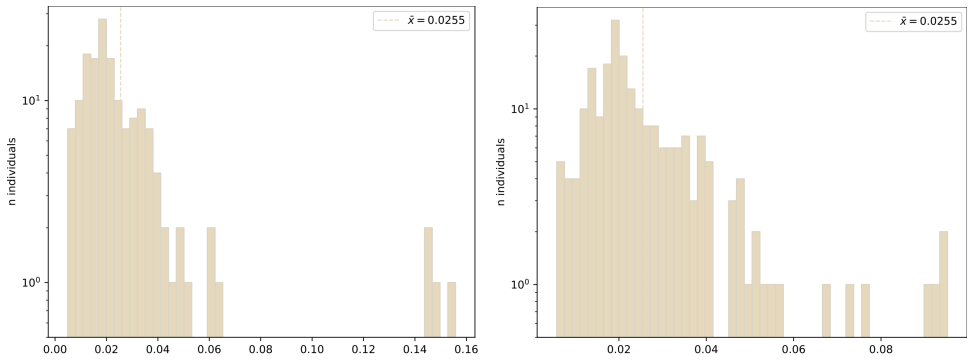
Figure 32: Mid-p-values for heterozygosity under HWE.

9.2.4 Missingness



(a) *B. pendula* (b) *B. pubescens*

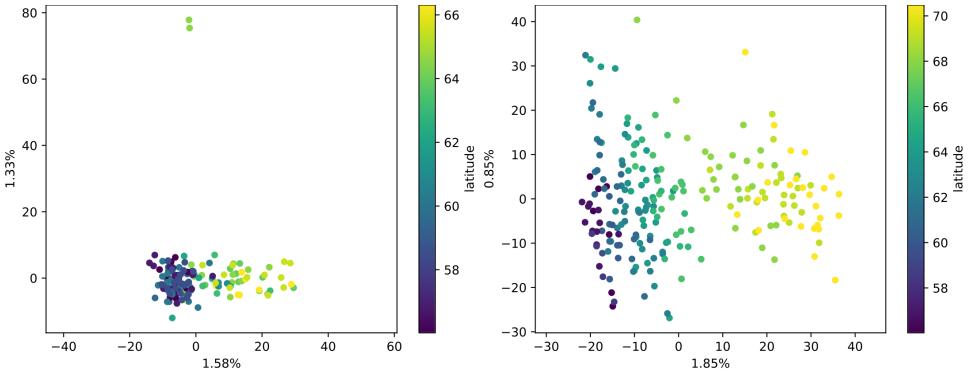
Figure 33: Missingness per site.



(a) *B. pendula* (b) *B. pubescens*

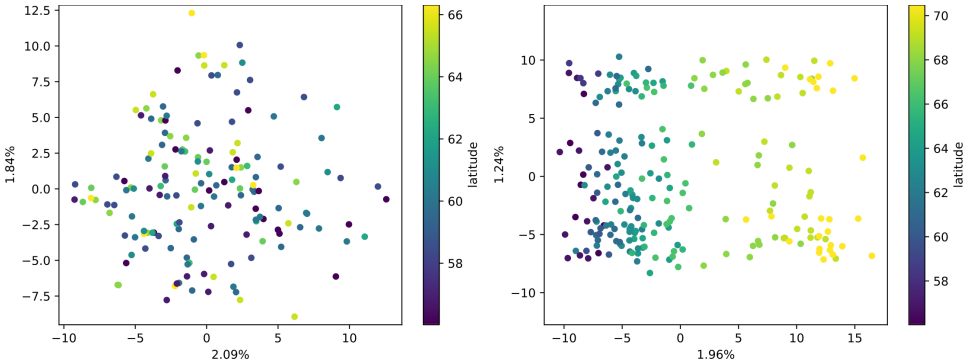
Figure 34: Missingness per individual.

9.3 PCA



(a) *B. pendula* (b) *B. pubescens*

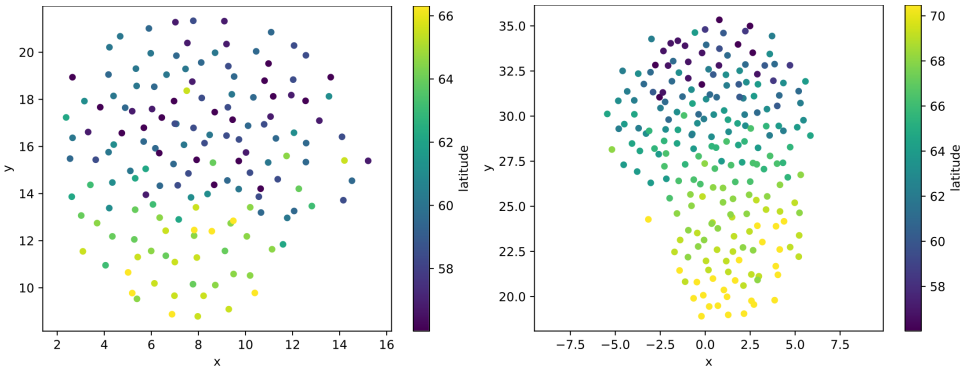
Figure 35: PCA plots on the set of bi-allelic sites. There are two outliers in the PCA plot for *B. pendula* which are not apparent in the plot for synonymous sites (cf. fig. 10a).



(a) *B. pendula* (b) *B. pubescens*

Figure 36: PCA plots on the set of non-synonymous sites. The rather weak population structure of *B. pendula* is not apparent at all in this case (cf. fig. 35a).

9.4 UMAP

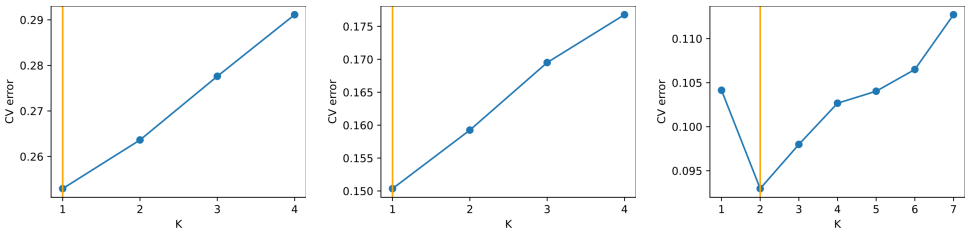


(a) *B. pendula* (b) *B. pubescens*

Figure 37: UMAP plots on the set of bi-allelic sites for each of the two birch species. The `spread` and `min_dist` parameters were set to 2 and 0, respectively.

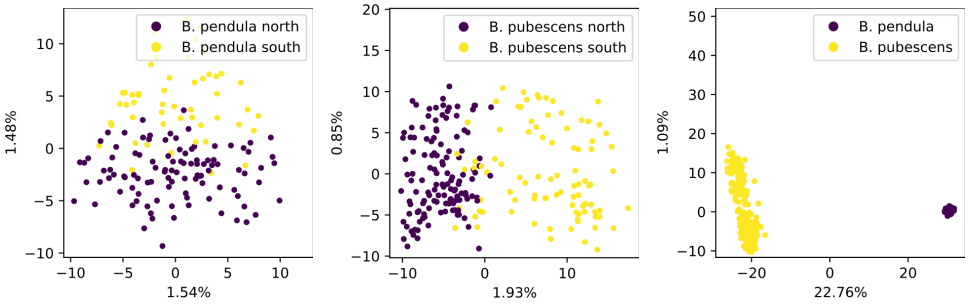
9.5 ADMIXTURE

All bar plots are first sorted by the displayed (ADMIXTURE) subpopulations and second by latitude in ascending order.



(a) *B. pendula* (b) *B. pubescens* (c) both species

Figure 38: 5-fold cross validation error over the number of clusters K. A single population ($K = 1$) is erroneously favoured within each of the two species. The set of bi-allelic sites was used for all analyses.



(a) *B. pendula* (b) *B. pubescens* (c) both species

Figure 39: PCA scatterplots labeled with the ADMIXTURE clustering for $K = 2$. A clustering very similar to the latitudinal gradients in figs. 9 & 10 is apparent.

9.5.1 *B. pendula* & *B. pubescens*

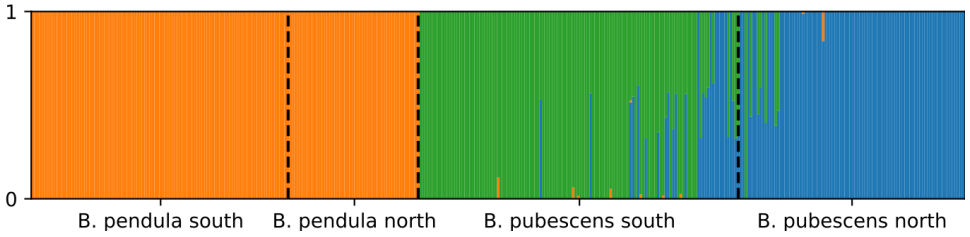


Figure 40: Bar plot for both species using $K = 3$. Compared to the two-cluster case we now additionally distinguish between a northern and a southern subpopulation for *B. pubescens*. *B. pendula*'s subpopulation structure is considerably weaker so detecting *B. pubescens*'s subpopulation structure first seems reasonable.

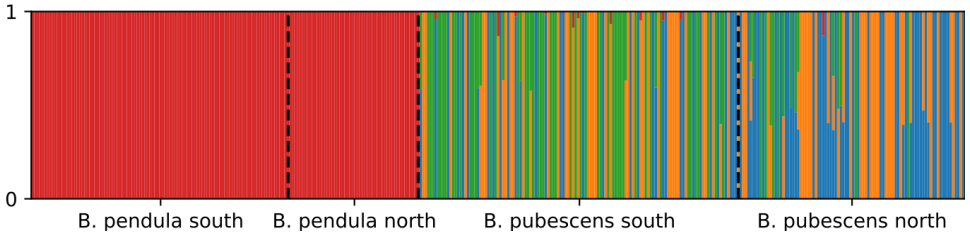


Figure 41: Bar plot for both species using $K = 4$. ADMIXTURE further partitions *B. pubescens* individuals but compared to $K = 3$, no latitudinal structure is apparent. This trend persists for higher values of K. The subpopulation structure of *B. pendula* is too weak to be detected considering both species together.

9.5.2 *B. pendula*

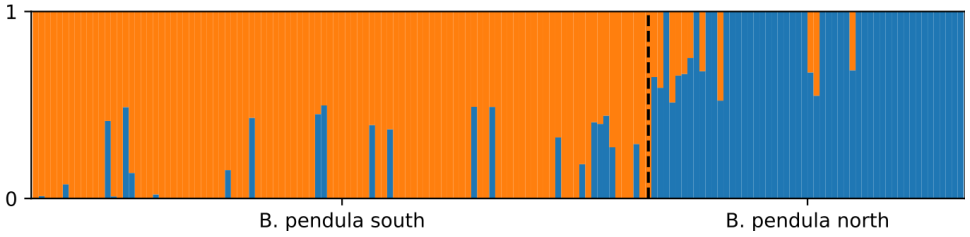


Figure 42: Bar plot for *B. pendula* using $K = 2$. The northern and southern subpopulation are clearly identified. There also seems to be considerable admixture between the two subpopulations, particularly from the northern population into the southern one.

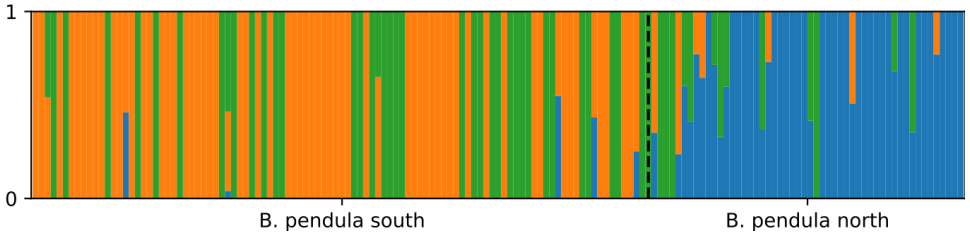


Figure 43: Bar plot for *B. pendula* using $K = 3$. A third subpopulation is roughly superimposed on the case for $K = 2$. This subpopulation does not seem to correlate with latitude, however. For higher values of K, we obtain even more scrambled images.

9.5.3 *B. pubescens*

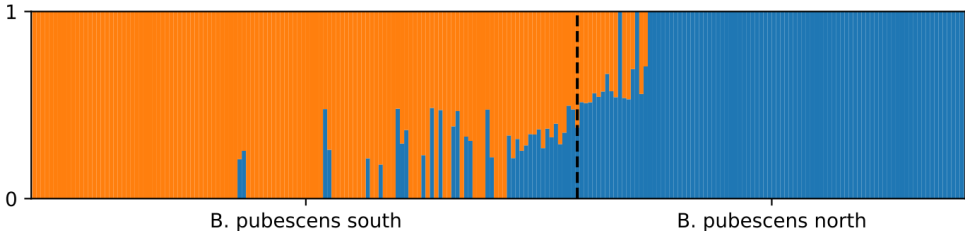


Figure 44: Bar plot for *B. pubescens* using $K = 2$. The northern and southern subpopulations are clearly identified with considerable admixture being apparent, especially near their contact zone.

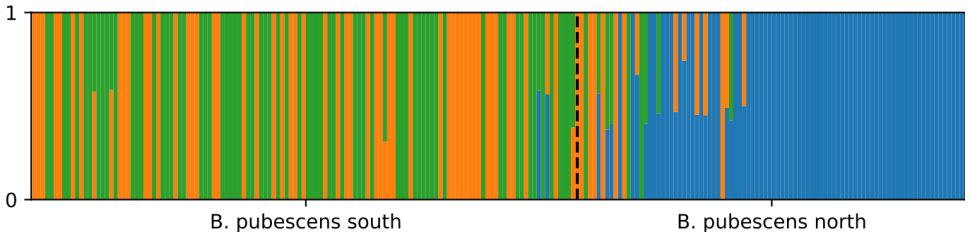
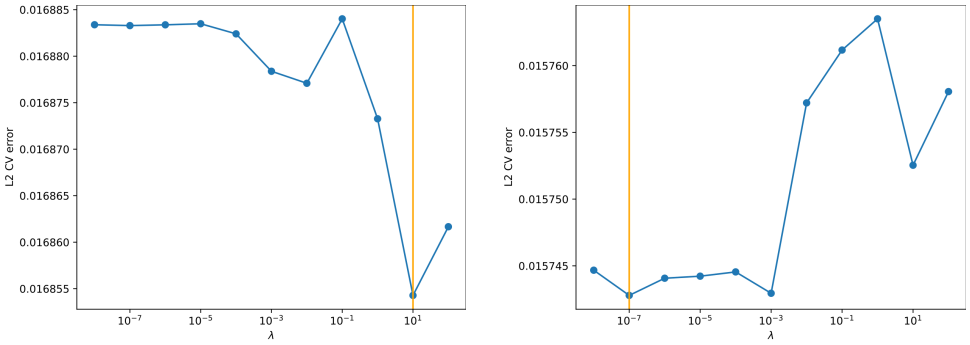


Figure 45: Bar plot for *B. pubescens* using $K = 3$. The southern population is now divided into two clusters which do not seem to correlate with latitude, however. For higher values of K, we obtain even more disordered images.

9.6 FEEMS

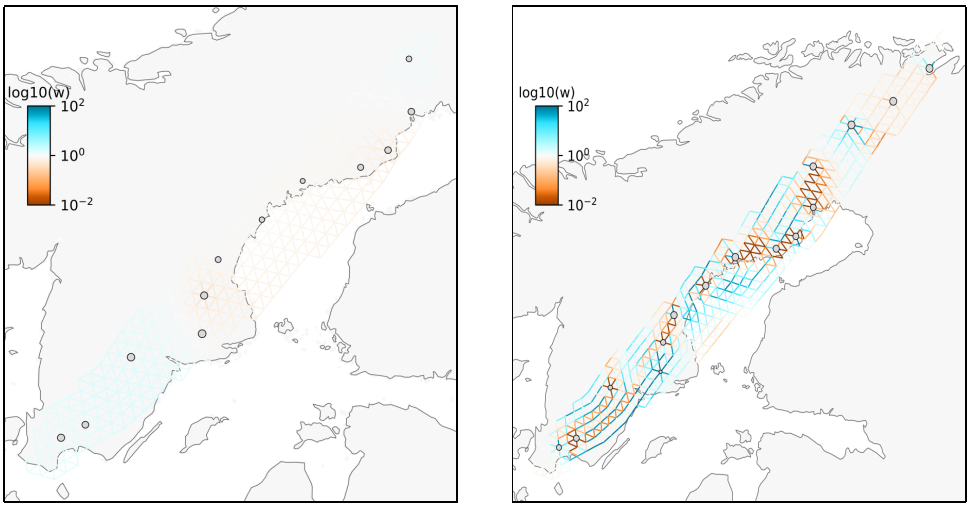
9.6.1 1% cut-off

In this subsection, alleles with a frequency lower than 1% have been removed. A buffer size of 1 was used if not otherwise specified.



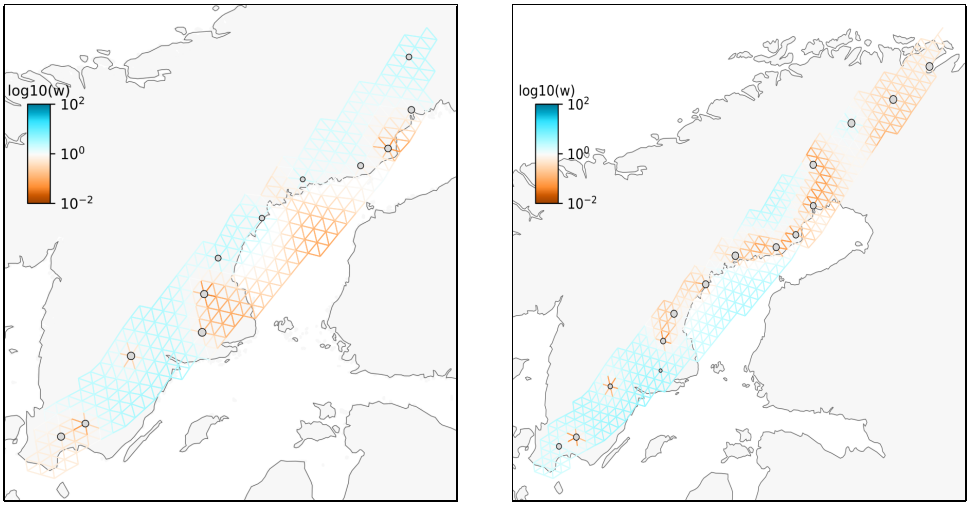
(a) *B. pendula* (b) *B. pubescens*

Figure 46: Cross validation error over the smoothing parameter λ using warm starts. We obtain similar values of λ using cold starts, although with increased variance.



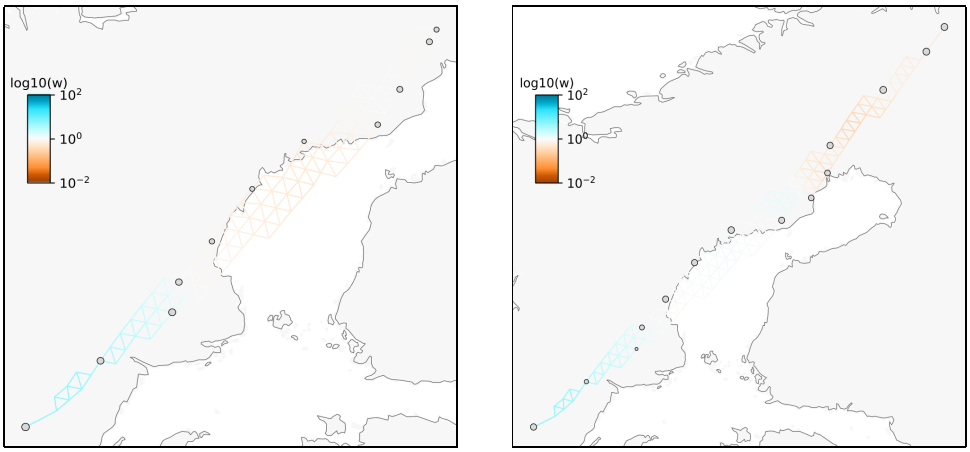
(a) *B. pendula* (b) *B. pubescens*

Figure 47: Migration surface plots for λ values of 10 and e^{-7} for *B. pendula* and *B. pubescens*, respectively which provided the lowest cross validation error (cf. fig. 46). The right plot seems highly overfitted.



(a) *B. pendula* (b) *B. pubescens*

Figure 48: Overfitted migration surface plots using $\lambda = 0.1$.

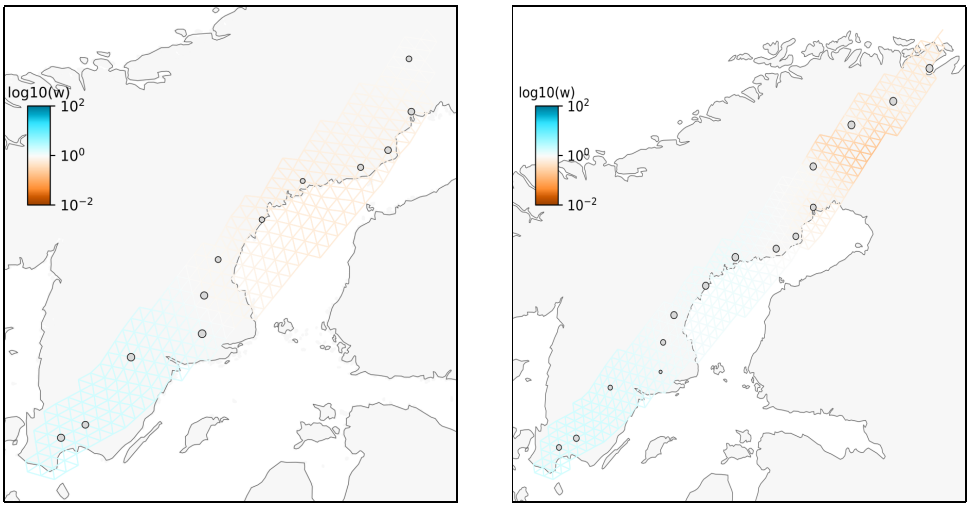


(a) *B. pendula* (b) *B. pubescens*

Figure 49: Migration surface plots with smoothing parameter $\lambda = 10$ and no buffer around the sampled locations. The results are much less visible but a similar trend of above-average migration in the south and below-average migration in the north can be observed (cf. fig. 13).

9.6.2 5% cut-off

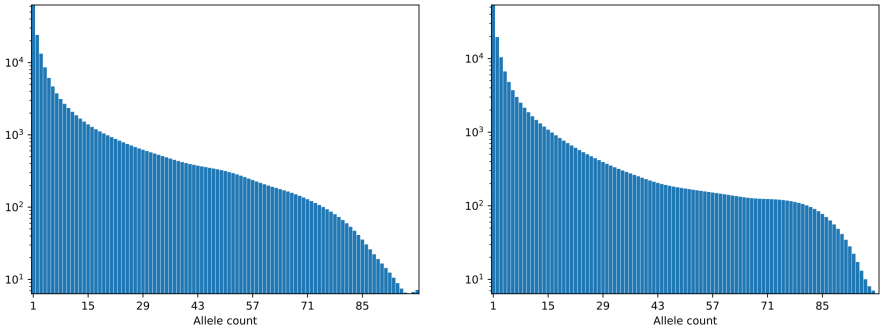
In this section, alleles with a frequency lower than 5% have been removed. The results are similar to those for a 1% cut-off, showing above-average migration in the south and below-average migration in the north. The cross validation favours highly overfitted plots for both species.



(a) *B. pendula* (b) *B. pubescens*

Figure 50: Migration surface plots for $\lambda = 10$.

9.7 SFS



(a) diploid calling

(b) tetraploid calling

Figure 51: Calling *B. pubescens* as diploid biases the SFS towards having more intermediate and fewer low-frequency alleles. The displayed SFS only comprises *B. pubescens* individuals.

9.8 δaδi

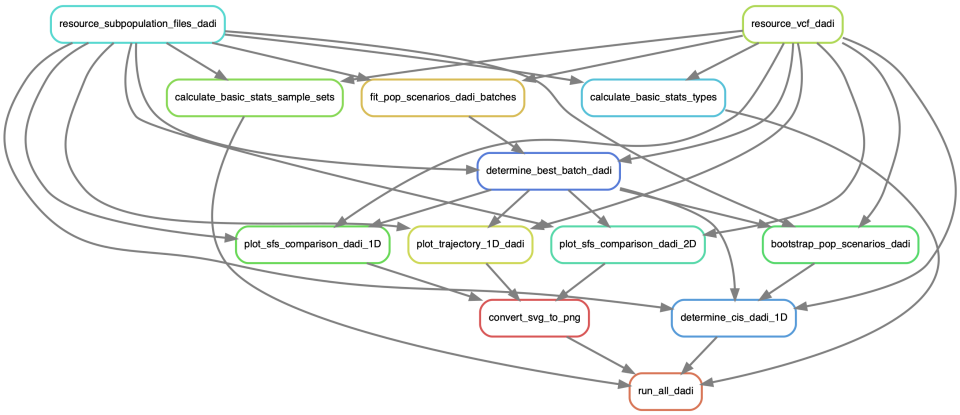
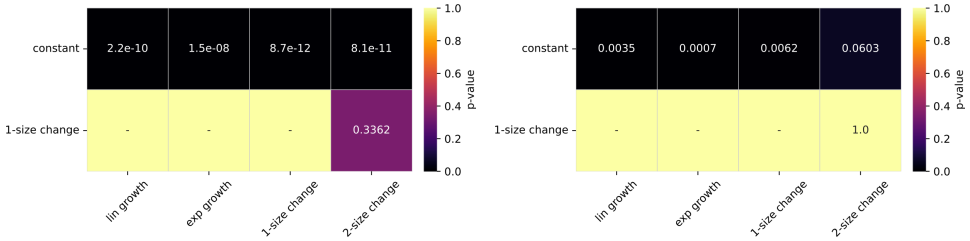
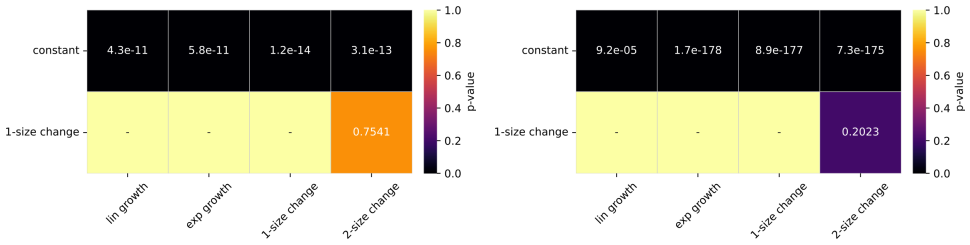


Figure 52: δaδi subworkflow.



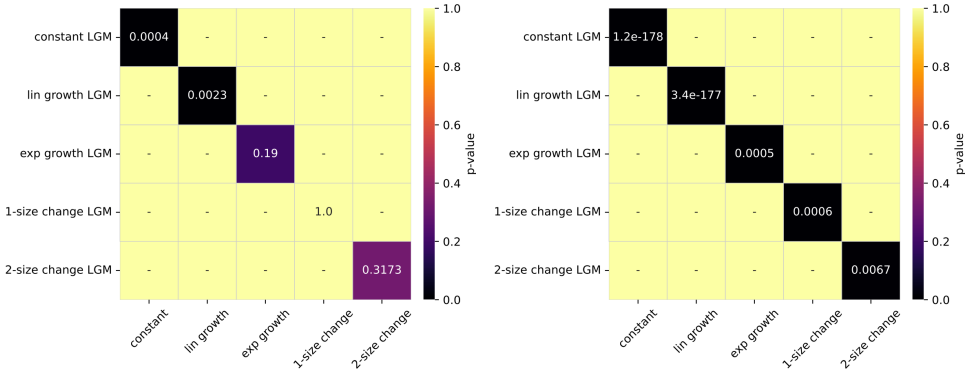
(a) *B. pendula* (b) *B. pubescens*

Figure 53: Nested one-population models. The more complex models are positioned on the horizontal axis. Not surprisingly, the non-constant growth scenarios provide a significantly better fit. Modelling two population size changes is not significantly better than modelling only one change. The average likelihood over all bootstrap samples was taken for the calculation of all p-values in this section.



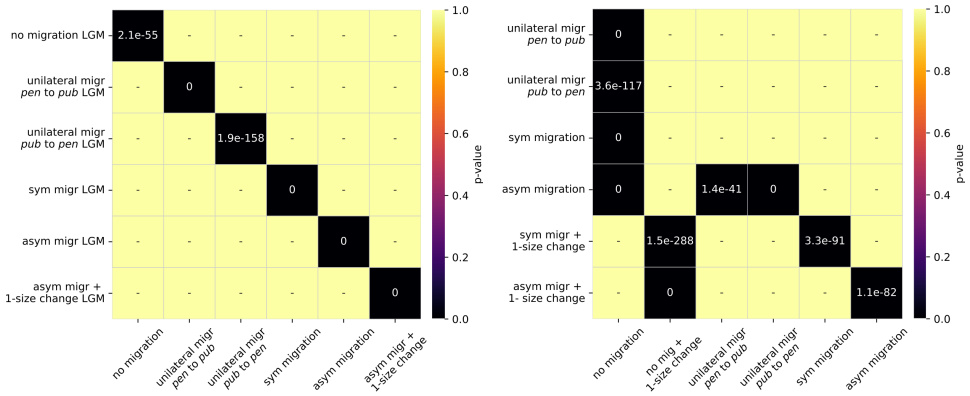
(a) *B. pendula* (b) *B. pubescens*

Figure 54: Nested one-population models after the LGM. The non-constant growth scenarios provide a significantly better fit in most cases. We observe qualitatively similar p-values to the variable-time cases above.



(a) *B. pendula* (b) *B. pubescens*

Figure 55: Variable-time vs. fixed-time one-population models where the time has been fixed to roughly coincide with the end of the LGM. The more complex variable-time models provide consistently better fits only for *B. pubescens*.



(a) variable vs. fixed-time

(b) nested fixed-time models

Figure 56: Nested two-population models comprising *B. pendula* & *B. pubescens*. The more complex models provide significantly better fits in all cases.

9.8.1 *B. pendula*

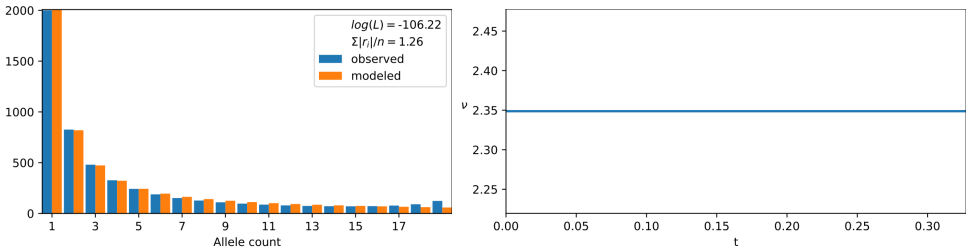


Figure 57: Constant population size scenario over variable time.

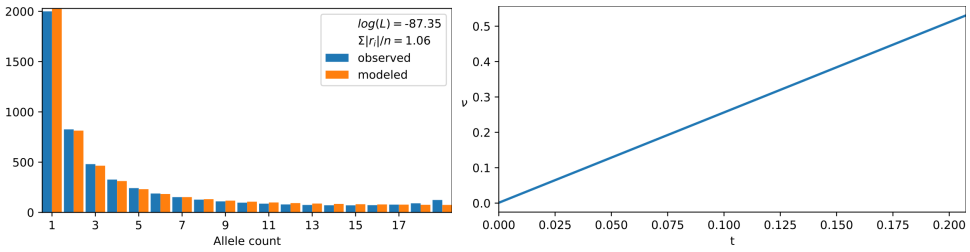


Figure 58: Linear population growth scenario over variable time.

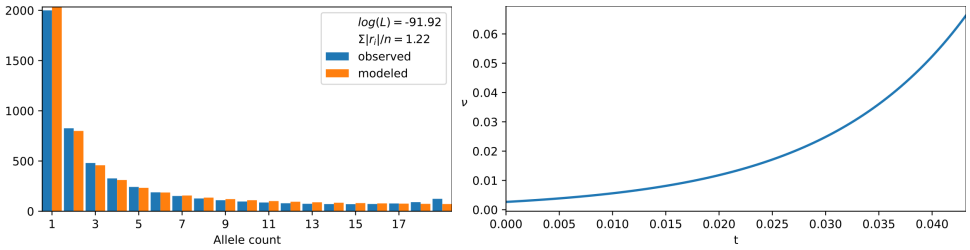


Figure 59: Exponential population growth scenario over variable time.

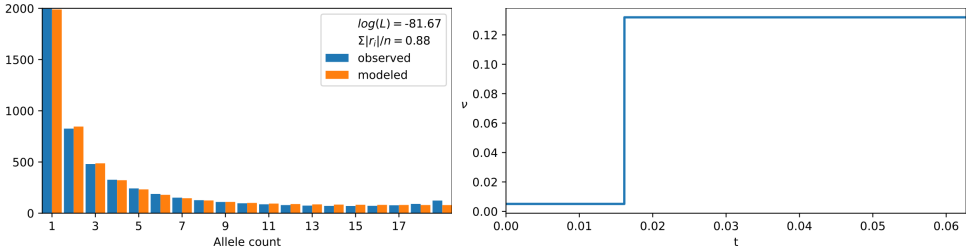


Figure 60: One population size change over variable time. This scenario provides the best fit among all variable-time models.

9.8.2 *B. pubescens*

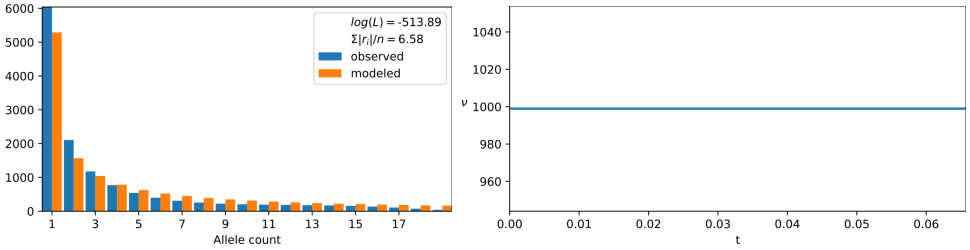


Figure 61: Constant population size after the LGM. The population size hits the upper bound for ν. The SFS cannot be properly fit for realistic values of ν over a time span that short. This would be possible for even larger values of ν, i.e. a higher mutation rate.

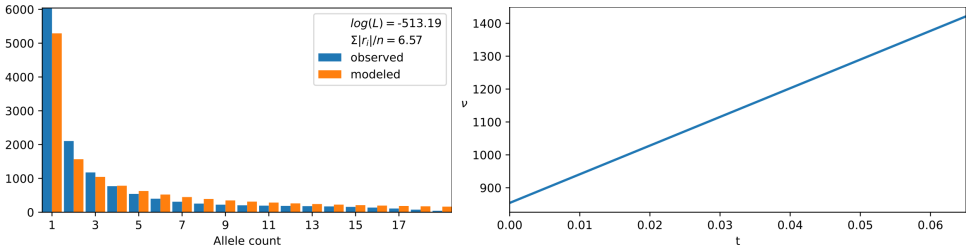


Figure 62: Linear population growth after the LGM. The initial population size ν_0 is close to the upper bound for ν. The SFS cannot be properly fit for realistic values of ν. Note that ν is allowed to exceed the upper bound for positive values of t as these values are modelled as a multiple of ν_0. This parametrisation allows for a comparison with the constant size model.

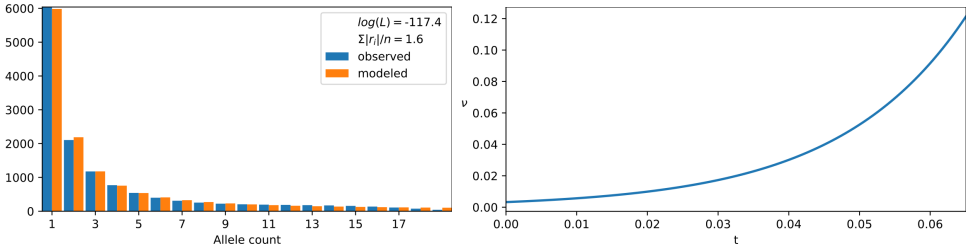


Figure 63: Exponential population growth after the LGM. This scenario provides a relatively good fit and indicates positive population growth.

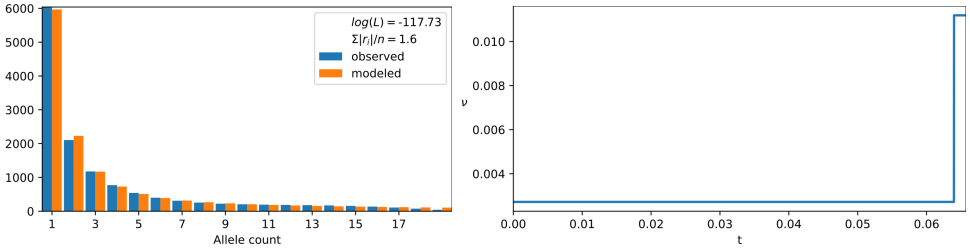


Figure 64: One population size change after the LGM. This scenario provides a relatively good fit and indicates positive population growth.

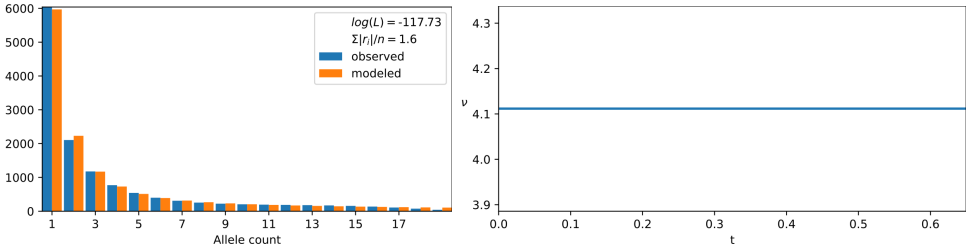


Figure 65: Constant population size over variable time.

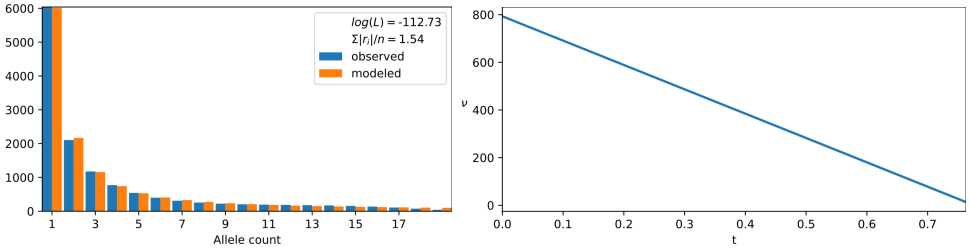


Figure 66: Linear population growth over variable time. Unlike the fixed-time counterpart, this model could be properly fit and indicates negative population growth (cf. fig. 62). Observe that the time parameter is much larger, likely spanning many glaciations.

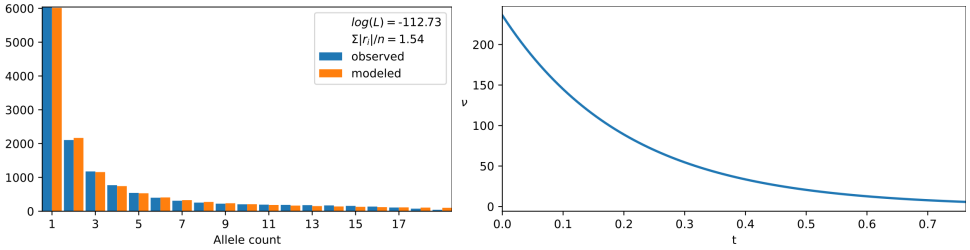


Figure 67: Exponential population decline over variable time.

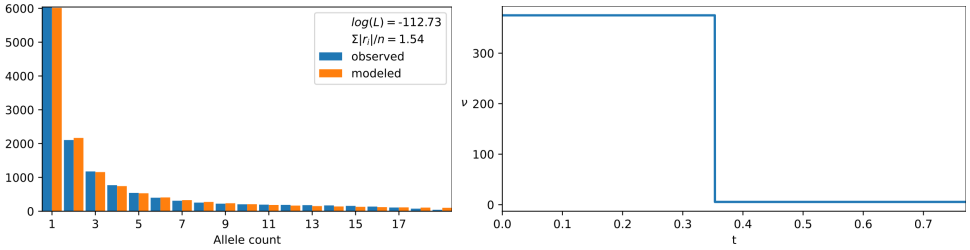
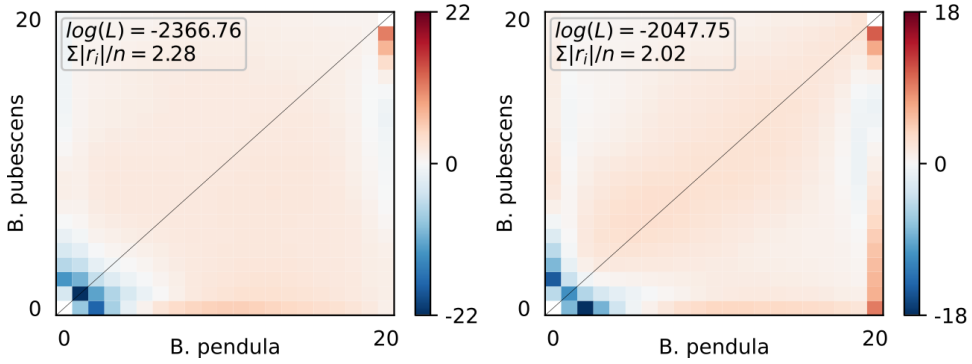


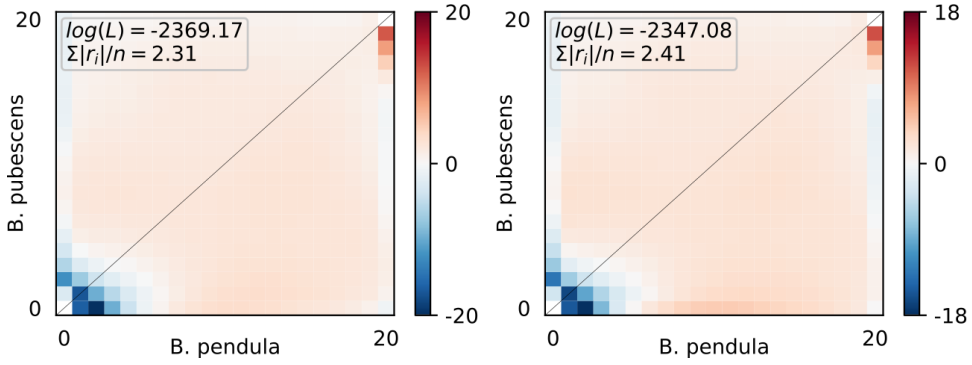
Figure 68: One population size change over variable time. We also observe population decline in this case. All non-constant variable-time scenarios for *B. pubescens* attain very similar likelihoods.

9.8.3 *B. pendula & B. pubescens*



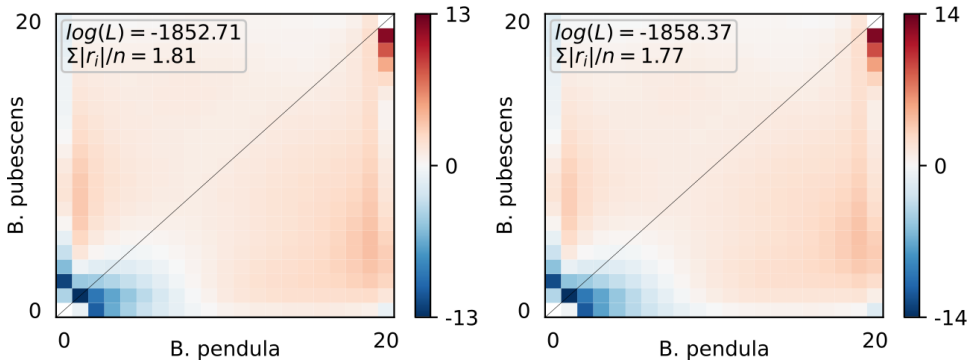
(a) no migration + constant size (b) no migration + 1-size change

Figure 69: Anscombe residuals for 2D scenarios without migration and time fixed to the end of the LGM. $\Sigma|r_i|/n$ denotes the average residual. The fixed-time scenarios have all much lower likelihoods than their variable-time counterparts. The relative effective population size ν is close to its upper bound. This is again due the very short time span integrated over.



(a) symmetric migration (b) asymmetric migration

Figure 70: Anscombe residuals for (a)-symmetric 2D scenarios with constant population size after the LGM. $\Sigma|r_i|/n$ denotes the average residual.

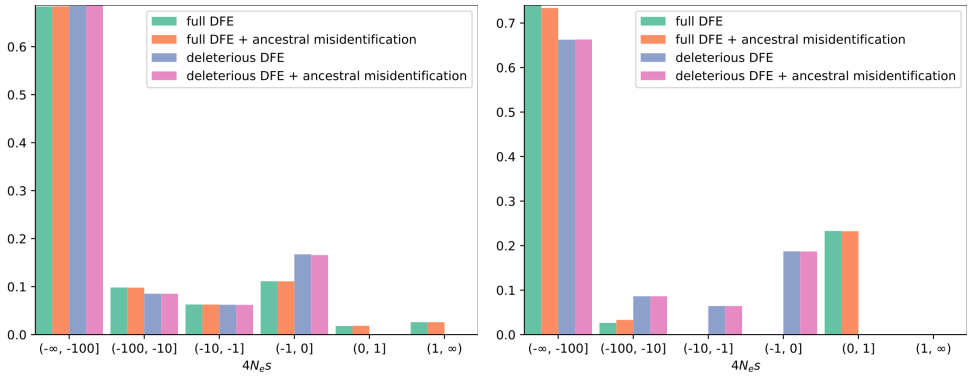


(a) sym. migration + 1-size change (b) asym. migration + 1-size change

Figure 71: Anscombe residuals for (a)-symmetric 2D scenarios with population growth after the LGM.

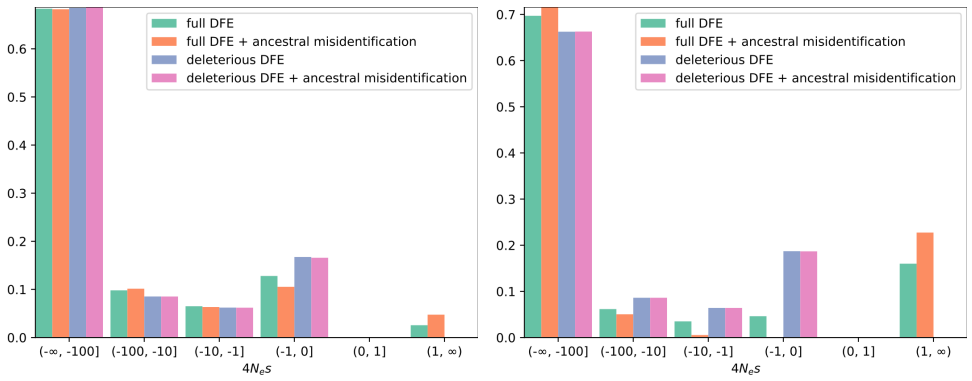
9.9 polyDFE

Figure 72: polyDFE subworkflow.



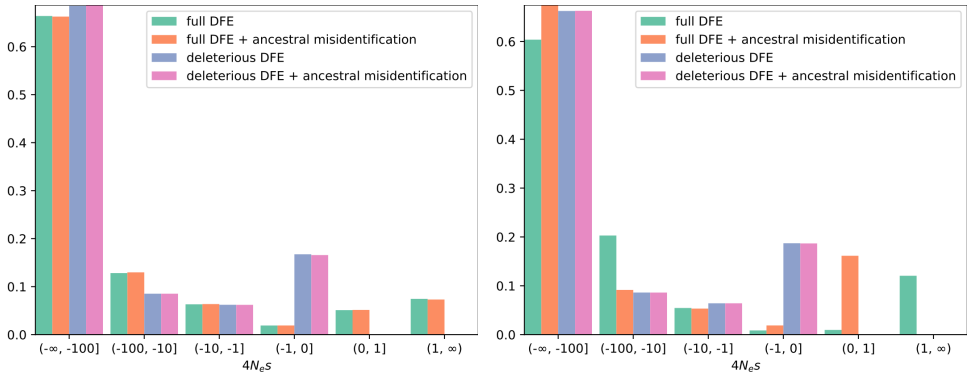
(a) *B. pendula* (b) *B. pubescens*

Figure 73: Species comparison of different DFE types for the default model, i.e. a reflected gamma and exponential distribution for non-positive and positive selection coefficients, respectively. The deleterious DFEs are rather similar but the full DFEs differ substantially in the amount of beneficial mutations.



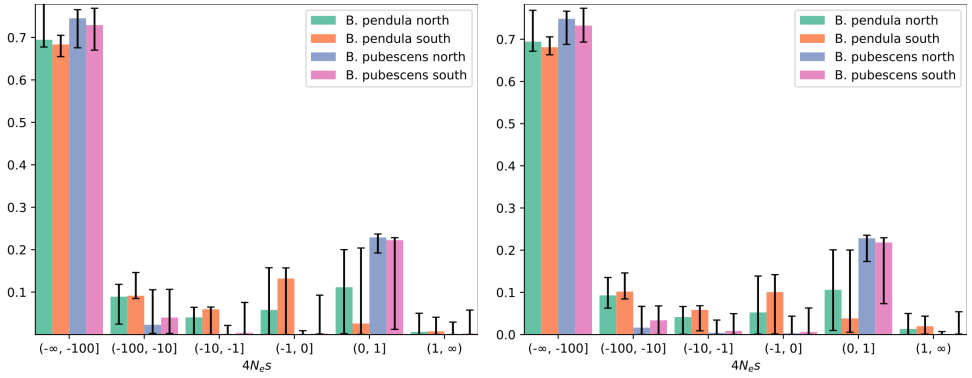
(a) *B. pendula* (b) *B. pubescens*

Figure 74: Species comparison of different DFE types where we assume a reflected gamma distribution and discrete distribution for non-positive and positive selection coefficients, respectively. Here the beneficial mutations have much larger selection coefficients compared to the default model (cf. fig. 73).



(a) *B. pendula* (b) *B. pubescens*

Figure 75: Species comparison of DFE types whose shape we assume to be a reflected displaced gamma distribution. Here, we obtain very different results when including ancestral misidentification to the full DFE for *B. pubescens*. Confidence intervals not being available, it is not apparent, however, whether this is caused by a large variance.



(a) diploid calling (b) tetraploid calling

Figure 76: DFE for various subpopulations where *B. pubescens* has been called as a diploid (left) and tetraploid (right). There is barely any difference despite the disparity in the SFS for *B. pubescens* (cf. fig. 51).

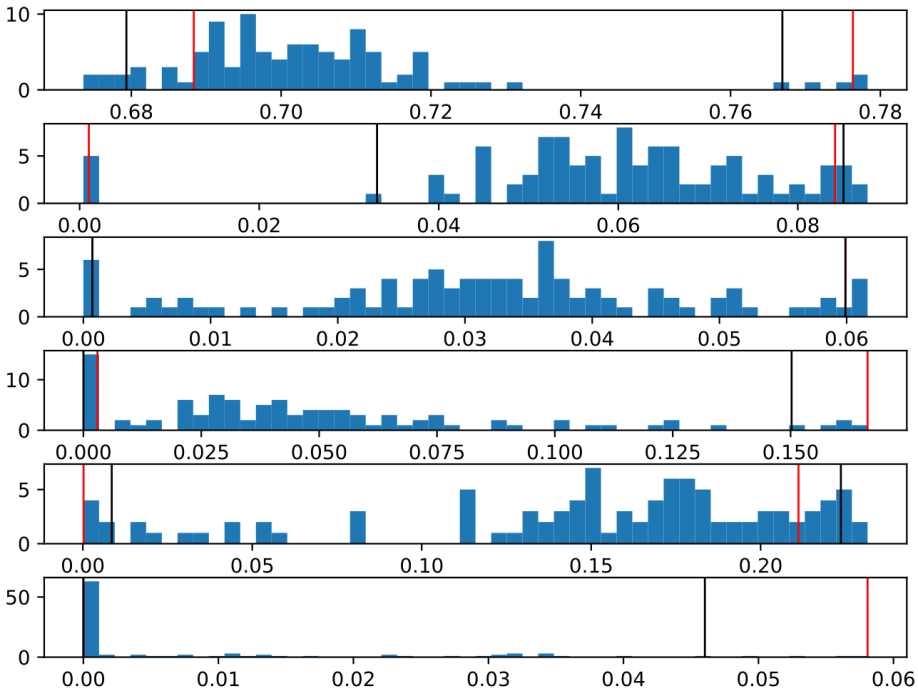


Figure 77: Distribution of bootstrap values for the selection coefficient intervals that were used for the DFE plots. A full DFE was jointly estimated for both species using the default model. The red and black vertical lines denote 95% confidence intervals using BCa and percentile bootstraps, respectively. We note that BCa bootstraps are more sensitive to outliers.

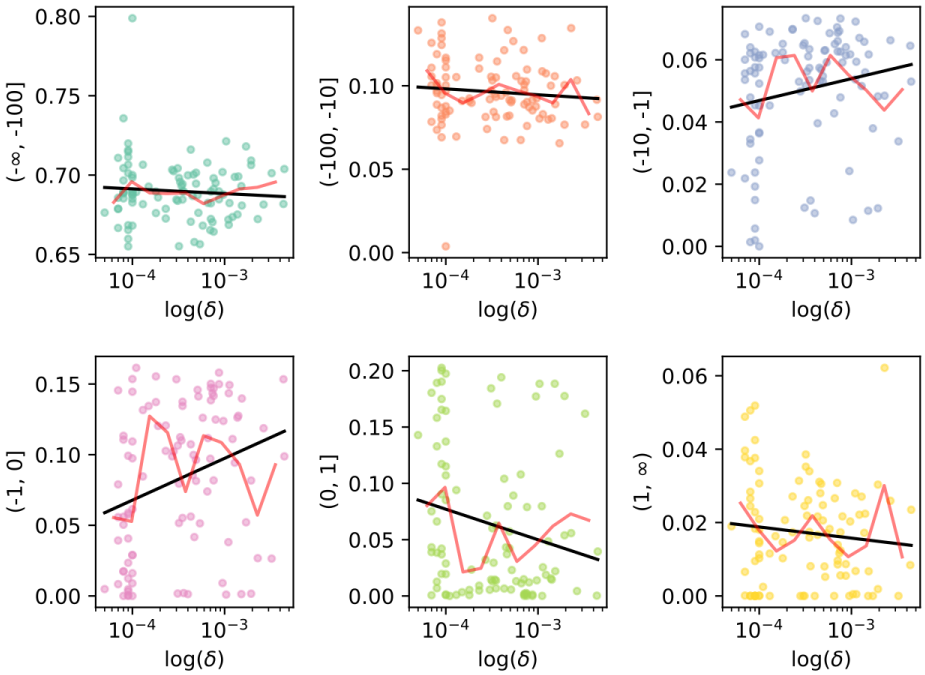


Figure 78: Values for selection coefficient intervals in *B. pendula* plotted against lowest gradient values from the optimisation routine. A larger gradient does not seem to introduce a significant bias. The black line denotes the linear regression solution and the red line the average per window. The graph looks similar for *B. pubescens*.