Patterns of Population Structure and Hybridization within and between *Populus trichocarpa* and *Populus balsamifera*

Muhammed F. Can

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Forestry

Jason Holliday, Chair
Amy Brunner
Eric Hallerman

December 15, 2021
Blacksburg, Virginia

Keywords: *Populus trichocarpa, Populus balsamifera*, population structure, hybridization.

Copyright 2021, Muhammed F. Can

Patterns of Population Structure and Hybridization within and between *Populus trichocarpa* and *Populus balsamifera*

Muhammed F. Can

(ABSTRACT)

The genus *Populus* consists of many ecologically and economically important forest tree species. Their rapid growth makes them one of the most productive hardwoods growing in temperate latitudes. *Populus* spp. frequently hybridize where their ranges overlap, and poplar hybrids are the most frequently planted genotypes for fiber production. To better understand the genomics of hybridization in *Populus*, we sampled and sequenced the genome of 574 poplar trees from six east-west transects across the hybrid zone between *Populus trichocarpa* and *Populus balsamifera* in western North America. I used these data to characterize population structure within and between transects, and hybridization between the species. There was a consistent transition from greater *P. balsamifera* ancestry in the north and east to greater *P. trichocarpa* ancestry in the south and west. Hybridization between the species was common across each of the six transects, though more common in colder climates. The results also showed that both latitude and longitude affect the genetic structure of this species complex, and that subtle introgression from *P. balsamifera* may facilitate adaptation of *P. trichocarpa* to colder climates.

Patterns of Population Structure and Hybridization within and between *Populus trichocarpa* and *Populus balsamifera*

Muhammed F. Can

(GENERAL AUDIENCE ABSTRACT)

The genus *Populus* has many ecologically and economically important forest tree species. Balsam poplar (*Populus balsamifera*) and black cottonwood (*Populus trichocarpa*) are two such species, both for fiber production and models for understanding tree biology and adaptation. Whereas black cottonwood is distributed close to the west coast of North America from California through Alaska, balsam poplar mostly occurs across the interior of Canada from Newfoundland through Alberta. Where their ranges overlap, the species often hybridize. In this study, we used genome sequencing of trees collected across six east-west transects from Washington state through British Columbia, Canada, and Alaska to understand genetic variation and the geography of hybridization. I found evidence of widespread hybridization across all transects. While the influence of *P. balsamifera* was extensive in northern populations, a large number of pure *P. trichocarpa* were found in southern populations. The transition from *P. trichocarpa* to *P. balsamifera* was also steeper in the south than the north, with a narrower hybrid zone in the south. Additionally, I found that gene flow among some populations was limited by temperature and geographical barriers. Taken together, my results suggest genetic structure and hybridization within and between these species is driven by climate variation, and that *P. balsamifera* ancestry may help northern *P. trichocarpa* populations adapt to their local environments.

Dedication

I dedicate this thesis to my family.

Acknowledgments

Firstly, I would like to thank the Ministry of National Education of the Republic of Turkey for giving me this opportunity through their funding.

I offer my sincere gratitude to my committee supervisor, Dr. Jason Holliday, for his support, encouragement, patience, and guidance throughout my graduate studies at Virginia Tech.

I would like to thank Dr. Eric Hallerman and Amy Brunner for serving on my Advisory Committee.

I would also like to thank the members of Holliday lab and my friends, Haktan, Ernie, and Sue for their help and friendship.

Finally, I would like to thank my family for always supporting me in everything I do, especially my wife, Berna; my son, Emir; my mother, Azize; my father, Arif; my sister, Melike Feyza; and my brother, Ali Fuat.

Contents

List of Figures

List of Tables

List of Abbreviations

ARC	Advanced Research Computing
BAM	Binary Alignment Map
BIC	Bayesian Information Criterion
BLUP	Best Linear Unbiased Prediction
BWA	Burrows-Wheeler Aligner
cpDNA	Chloroplast DNA
CTAB	Cetyltrimethylammonium Bromide
DA	Discriminant Analysis
DAPC	Discriminant Analysis of Principal Components
DNA	Deoxyribonucleic Acid
EST	Expressed Sequence Tag
FFP	Frost-Free Period
GATK	Genome Analysis Tool Kit
GBS	Genotyping by Sequencing
GCR	Genome Complexity Reduction
gDNA	Genomic DNA

GPS	Global Positioning System
IBD	Isolation by Distance
LD	Linkage Disequilibrium
MAT	Mean Annual Temperature
MCMT	Mean Coldest Month Temperature
PCA	Principal Component Analysis
RI	Reproductive Isolation
SAM	Sequence Alignment Map
SeqCap	Sequence Capture
SNP	Single Nucleotide Polymorphism
VCF	Variant Call Format
WGS	Whole-Genome Sequencing

Chapter 1

Introduction

Understanding how biological species form and persist is a central goal of evolutionary biology, and of applied importance to their management in a rapidly changing environment (Hoffmann et al., 2015). Improved DNA sequencing methods have resulted in a new focus on identifying genomic regions underlying adaptation and speciation in hybrid regions (Ellegren, 2008). Due to near geographical proximity and lack of effective reproductive isolation (RI), when porosity between species is high, selection should favor the evolution of tight physical association between adaptively interacting genes in hybrid zones to counteract the homogenizing power of recombination (Yeaman and Whitlock, 2011). Despite various potential paths to RI, many tree species hybridize with their relatives, potentially contributing to local adaptation in transition environments (Bawa, 2017).

The genus *Populus* comprises many ecologically and economically important species of forest trees. Their rapid growth makes them among the most productive hardwood trees grown in temperate latitudes in North America. Although not as economically important as conifers, this fast growth coupled with coppice regeneration has made them a focal species for lignocellulosic bioenergy production. *Populus* also has emerged as a model system for studying woody perennial biology due to ease of genetic transformation, compact genome, and short rotation time (Ellis et al., 2010). These features have led to an active, worldwide *Populus* research community. Early in the genomic era, more than 100,000 expressed sequence tags (ESTs) were developed, and later *Populus trichocarpa* was the third plant species to have its genome completely sequenced (Ellis et al., 2010).

In addition to being models for tree biology and genomics, as a source of both traditional forest products and lignocellulose for bioenergy, poplar (*Populus*) species are under intensive cultivation (Jansson and Douglas, 2007). Almost all cultivated poplar genotypes are hybrids, but our understanding of how interspecific genomic associations produce valued phenotypic effects, such as fast growth and abiotic stress resistance, is still limited. The most productive hybrid poplar crosses are between the Tacamahaca and Aigeiros sections because they give rise to abundant heterosis, leading to significant height and biomass gains in early-generation (F_1) hybrids (Stettler et al., 1980). Although increased growth rate is the primary target of poplar hybrid breeding programs, there is a tension between rapid growth and traits that depend on the ecological context regarding fitness. These features include the proper timing of seasonal developmental transitions, as well as structural and inducible defenses against pathogens. Intraspecific patterns of local adaptation in temperate and boreal tree species are well characterized (Howe et al., 2003; Savolainen et al., 2007), while much less is known about specific hybridization of adaptive characters and environmental variation. While inter-sectional hybridization is often used in a production setting, intra-sectional hybridization is more common in the wild, and may contribute to local adaptation. For example, introgression of *Populus balsamifera* haplotypes into *P. trichocarpa* has been reported in eastern and northern *P. trichocarpa* populations, both in contact zones and in areas outside of the areas mentioned above (Suarez-Gonzalez et al., 2018). Some studies showed that introgression from *P. balsamifera* contributes to shaping the geographically and climatically related genetic variation patterns of *P. trichocarpa* (Geraldes et al., 2014). Introgression from *P. balsamifera* to *P. trichocarpa* may allow the latter to occupy colder climates than typical of the species (Suarez-Gonzalez et al., 2018).

In this study, we sampled branch cuttings from 574 poplar trees from six east-west transects across the hybrid zone between *P. trichocarpa* and *P. balsamifera* in western North America. They were planted in a replicated common garden in Critz, VA, and each was subjected to whole-genome re-sequencing. On the basis of the resulting genome-wide dataset, I addressed the following questions: (1) What is the scale and spatial patterns of population structure within and between *P. trichocarpa*, *P. balsamifera*, and their hybrids? (2) What is the relationship between geography, climate, and hybridization between the species? (3) What is the relationship between species identity, hybridization, and spring bud phenology?

Chapter 2

Literature Review

2.1 Studied species

Populus balsamifera and *Populus trichocarpa* belong to the Tacamahaca section of the genus. *P. trichocarpa* (Black cottonwood) is native tree to the western United States and Canada from northern California to southern Alaska (Figure 2.1) (Gornall and Guy, 2007). It is likewise found in the interior of the coastal range of mountains. During late Pleistocene glacial periods, the range of *P. trichocarpa* was limited to southern and coastal areas, with subsequent postglacial recolonization that leads to its extant range, reaching northwestern coastal areas of North America (Levsen et al., 2012). In the southeast, *P. trichocarpa* can hybridize with *P. fremontii*, *P. deltoides*, and *P. angustifolia*, and with other species of *Populus* under managed circumstances (DeBell, 1990). Due to limited winter hardiness and possibly water requirements, the range of *P. trichocarpa* is more restricted than that of *P. balsamifera* (Eckenwalder, 1996).

P. balsamifera ranges from Eastern Canada to Alaska (Figure 2.1), having recolonized the Canadian boreal region after the ice sheets retreated (Najar, 2017). *P. balsamifera* and *P. trichocarpa* both exhibit rapid growth provided they are in the light, near a water supply, and have very good drainage. Reproductive maturity of each species is generally reached in 7 - 10 years (Braatne et al., 1996). They can hybridize freely with each other where their distributions overlap (Viereck and Little, 1972). As *P. balsamifera* is an incredibly frost-resistant boreal species, tolerating a very wide range of extreme temperatures (-62 to 44°C) (Richardson et al., 2014), these hybridizations may allow *P. trichocarpa* to colonize colder habitats in northern and inland areas than are otherwise typical of the species range (Suarez-Gonzalez et al., 2016).

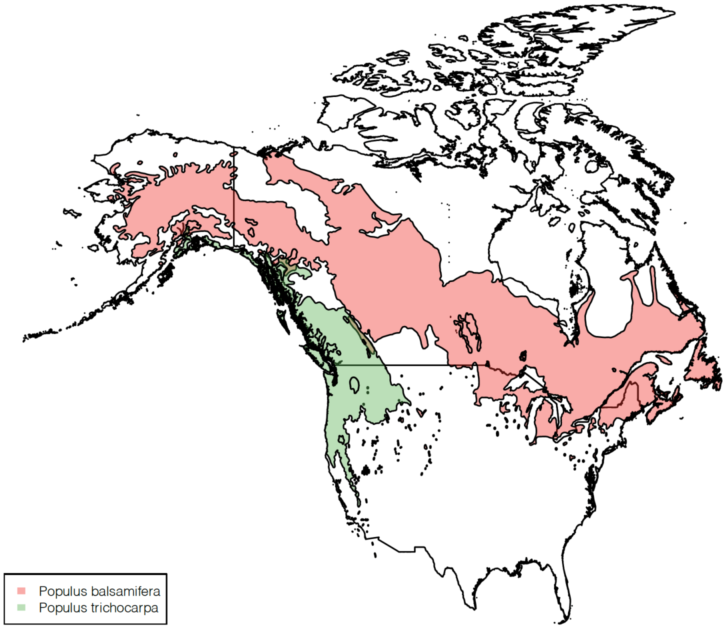


Figure 2.1: The ranges of *P. balsamifera and P. trichocarpa* (Little and Viereck, 1971)

2.2 Population structure

Population structure arises due to partial reproductive isolation between groups within a species and may impact the extent of physiological or behavioral adaptations of local populations to their environment (Schowalter, 2016). Population structure generally increases with decreasing gene flow and the duration of time over which populations have been isolated. In temperate and boreal forest trees, including most *Populus* species, gene flow via pollen or seed is usually highly efficient, which minimizes background genetic differentiation between populations (Rajora et al., 2005). Meirmans et al. (2017) analyzed intraspecific population structure in *P. trichocarpa* and *P. balsamifera*, which was assessed using principal component analysis (PCA), and found that the division between eastern and the central-western clusters was the most relevant subdivision of *P. balsamifera*, but more local spatial isolation still had an effect. They also found that admixture was directional from *P. balsamifera* to *P. trichocarpa* in the north, where the *P. trichocarpa* range is limited by minimum winter temperatures. Another study by Geraldes et al. (2014) sampled 498 putative *P. trichocarpa* trees originating mostly from British Columbia, Canada (in addition to two provenances from Washington/Oregon, USA), as well as 10 *P. balsamifera* reference samples, and genotyped these with a 34K SNP chip. There was a clear separation between *P. trichocarpa* and *P. balsamifera*, with a low level of admixture between the species. Within *P. trichocarpa*, they found that seven clusters best explained by patterns of variation, which were roughly partitioned into somewhat heterogeneous groups in Oregon, USA, southern British Columbia, Canada, and central/northwest British Columbia. A Mantel test suggested that gene flow across the entire range was limited due to isolation by distance (IBD). Finally, introgression with *P. balsamifera* was found only at the boundaries of the distribution of *P. trichocarpa*.

2.3 Hybridization

Genetic diversity and the evolutionary trajectory of populations and species can be considerably affected by hybridization (Meirmans et al., 2017). Many tree species hybridize with their congeners, which potentially contributes to local adaptation in transitional environments (Bawa, 2017). In temperate and boreal trees, the transition between coastal and continental climates represents an important environment where parent species may perform poorly relative to their hybrids (Zanewich et al., 2018). For example, De La Torre et al. (2014) studied the hybrid zone between *Picea glauca* (white spruce), which is adapted to cold and dry continental conditions, and *Picea engelmannii* (Englemann spruce), which is adapted to relatively warmer and wetter conditions, in western North America. They found high levels of admixture and introgression in the contact zone between these species, and almost all individuals had hybrid ancestry where the species ranges came into contact. They also found that these two species had a long history of hybridization and introgression, but strong environmental selection reduced contemporary interspecific gene flow. In addition to population structure, Meirmans et al. (2017) examined the effect of hybridization between *P. balsamifera* and *P. trichocarpa* to determine how it affects population structure and adaptation. They sampled 1517 individuals from these two species and genotyped them with a combination of 93 nuclear and 17 chloroplast DNA (cpDNA) SNPs. The results showed that introgression was typically restricted to the contact zone where the species' distributions overlap, although their sampling was distributed in such a way that it would have been difficult to resolve low levels of introgression in areas adjacent the contact zone. They also found extensive hybridization across the sampled transects, with a narrower zone of hybridization in the south, and a broader hybrid zone in the north. Furthermore, differentiation among eastern, central, and western *P. balsamifera* clusters suggested that historical influences affected the genetic structure of this species. The Central and Western clusters had a gradient of ancestry, but the border between the Eastern and Central clusters was sharp. The sharp border was also associated with climate, and they proposed that during the late Pleistocene ice ages, the detected clusters were isolated in three refuges. Finally, they concluded that the ancestry of *P. balsamifera* was shaped by historical variables, and the impact of interspecies hybridization was limited.

2.4 Relationships between adaptation, ancestry, and climate

Many *Populus* species have large geographic ranges and reveal strong adaptation to local environments (Soolanayakanahally et al., 2015). Differences in growth rates and morphological characteristics can be better determined using populations sampled over wide geographic ranges (Soolanayakanahally et al., 2015). The "common garden" method (Kawecki and Ebert, 2004) growing different groups under the same conditions is widely used to study local adaptation, and can also be used to model tree responses to climate change (Soolanayakanahally et al., 2015). Spring bud-flush is mainly driven by integration of warm temperatures over a period of weeks or months (Olson et al., 2013), the timing of which plays a critical role in the ecological trade-off between survival and growth (Frewen et al., 2000). Common garden studies reveal that bud-flush timing is related to latitude and elevation of origin (Frewen et al., 2000). For example, when planted in a warmer common garden site, trees from high latitude or elevation, which have a relatively low heat sum requirement, may flush earlier than in their native environment (Pauley and Perry, 1954).

The buds of *P. trichocarpa* generally open beginning in April through early May, and within a few weeks, the growth of the short shoot is complete (Critchfield, 1960). Then, a new terminal starts to develop (Critchfield, 1960). Low temperature-related injury is a serious problem affecting yields, quality, and survival of agricultural crops and forestry (Tsarouhas et al., 2003). Bud-flush timing can also be affected by low temperature-related injury because in spring, newly flushed shoots are at risk from low temperatures (Soolanayakanahally et al., 2015). For example, after a warm March, a cold April may cause frost damage. Finally, in order for plants to have frost resistance, they must be set early (Frewen et al., 2000). If the bud-flush of these plants occurs too early, the tissues can be damaged due to late frost

(Frewen et al., 2000).

2.5 Genome Re-sequencing strategies

Choosing an appropriate molecular marker is critical to accurately estimating patterns of population structure and hybridization, and the choice depends on several factors: genome size, reference genome quality, and cost. DNA sequence-based genotyping has become the standard, with the "genotyping by sequencing" (GBS) approach being the most cost-effective. GBS involves digesting genomic DNA (gDNA) with one or more restriction enzymes, followed by ligation of adapter oligonucleotides that contain individual barcode labels. This latter stop allows multiple samples to be pooled for sequencing and then for separation during computational processing (Elshire et al., 2011). Missing data, which is the disadvantage, is a problem due to polymorphic restriction sites and insufficient sequencing depth. The GBS method is most useful for genetic mapping studies in populations with high linkage disequilibrium (LD), where high coverage of the entire genome is not needed. Although GBS can yield many thousands of SNPs at a comparatively low cost, sequence capture (SeqCap) provides a variety of benefits that make it the genome complexity reduction (GCR) method of choice (Holliday et al., 2018). Missing data is less of an issue for SeqCap. It is useful for studies of adaptation due to its ability to target gene regions, which are the most likely targets of selection, but works better with a reference genome. However, it has a high cost compared to enzyme-based methods for the species.

The most comprehensive approach to sequence-based genotyping is to simply re-sequence the entire genome (whole-genome sequencing, or WGS). The WGS strategy makes it possible to score most variants, both rare and common, in both the coding and non-coding areas of the genome (Yin et al., 2019). As technology advances, the cost of whole-genome sequencing is decreasing, which makes it possible to produce genome-wide single nucleotide polymorphism (SNP) data at relatively low cost per sample. While WGS is best with a reference genome sequence for alignment, it is also possible without a reference genome. The cost of WGS has been declining, enabling genome-wide single nucleotide polymorphism data to be produced at relatively low cost per sample.

Chapter 3

Materials and Methods

3.1 Plant material

Vegetative branch cuttings from 574 poplar trees were collected from six east-west transects (hereafter referred to as Alaska, Cassiar, Chilcotin, Jasper, Crowsnest, and Wyoming) across the hybrid zone between *Populus trichocarpa* and *Populus balsamifera* in January 2020 (Figure 3.1). The samples were obtained from across most of the species' latitudinal ranges, from 40° N to 65° N and -100° W to -150° W longitude. All plants were photographed, given a unique identifier, and the GPS (Global Positioning System) coordinates were recorded with the lower meter GPS registered. Approximately 20 - 30 cm vegetative branch cuttings were collected from each individual, which were subsequently rooted on a mist bench and planted in a 3x replicated common garden at the Reynolds Homestead Forest Resource Research Center located in Critz, VA (36° 37' N and 80° 09' W, elevation 360 m). The average of climate parameters for the nearest weather station of Patrick County showed that the average temperature in Patrick Springs is 57.3 °F (14.1 °C) for the year. The average temperature of the coldest and warmest month of the year is 39.6°F (4.2°C) and 75.3°F (24.1°C), respectively (weatherbase.com). It should be noted that this site is significantly warmer than the origins of most of our genotypes, which may have variable effects on expression of temperature-dependent phenotypes like bud-flush and growth, although relative rankings among genotypes should not be affected. Prior to planting, young leaves were collected for genomic DNA extraction. These tissue samples were stored in paper coin envelopes and placed immediately on dry ice. For long-term preservation, the samples were then transferred to a -80°C freezer.

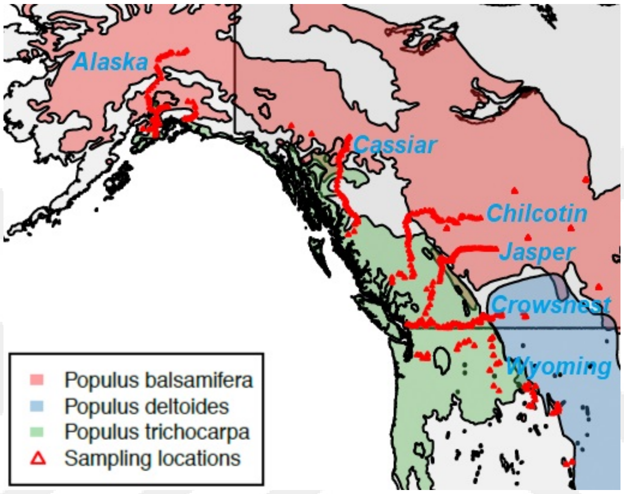


Figure 3.1: The origins of the collected samples. Green shading, red shading, and blue shading indicate western ranges of *P. trichocarpa*, *P. balsamifera*, and *P. deltoides* respectively. Red triangles indicate sampled trees.

3.2 Genomic Library Preparation and Bioinformatics

Approximately 100 mg leaf tissue (if very young, 80mg, if bigger, 90-110mg) was weighed depending on the leaf quality for each sample. The leaves were immediately ground into very fine powder when liquid nitrogen was almost gone by using a mortar and pestle and transferring the powder into microcentrifuge tube. Then, DNA was extracted from all samples using a modified Qiagen plant DNeasy kit extraction protocol, using phenol-chloroform extraction in place of a QIAshredder column. For samples with very low DNA concentration, a secondary extraction was performed using cetyltrimethylammonium bromide (CTAB), which is a cationic detergent. With a Nanodrop spectrophotometer, DNA concentrations and quality were quantified using the Qubit dsDNA HS Assay kit. DNA libraries were constructed at the Duke University Center for Genomic and Computational Biology, using an Illumina Nextera kit (Elshire et al., 2011). The libraries were sequenced on an S4 flow cell in 2x150bp format on an Illumina NovaSeq 6000 instrument with 64 samples per lane. De-indexing, QC, trimming adapter sequences, and sequence preprocessing were completed by the sequencing facility.

Subsequent bioinformatics tasks were performed on Virginia Tech's Advanced Research Computing System (ARC). Specifically, the Burrows-Wheeler Aligner (BWA) was used to map reads to the *P. trichocarpa* reference genome (v4.0), and the resulting SAM files were converted to BAM format with SAMtools (Li and Durbin, 2010). The Genome Analysis Tool kit (v3.7) Haplotype Caller algorithm was then used to generate individual gVCF files, which were merged in a single VCF file with the GATK Genotype GVCFs function. This raw VCF was quality-filtered for variants that had poor map quality, elevated strand bias, differential map quality between reference and alternative alleles, positional bias between reference and alternate alleles, or low coverage depth. Such variants were flagged and omitted as low-quality data. Finally, variants were removed that had >10% missing data across samples or that were present in only a single heterozygote (i.e., with minor allele frequency equal to 1/2N). This process yielded approximately 12 million SNPs. Because this quantity of SNPs is unnecessary for the analyses described below, we subsampled the total quality-filtered dataset randomly to include 1,000,000 SNPs.

3.3 Population Structure and Hybridization

Admixture proportion inference (Novembre, 2016) and principal component analysis are frequently used strategies to understand population structure (Cavalli-Sforza et al., 1994). In this study, patterns of population structure in the dataset were explored using ADMIXTURE

(Alexander et al., 2009), Discriminant Analysis of Principal Components (DAPC) (Jombart, 2008), and Principal Component Analysis (PCA). In addition, *Introgress* software (Gompert and Alex Buerkle, 2010) was used to estimate hybrid indexes for each sample, and F_{ST} (Willing et al., 2012) was calculated to estimate population differentiation depending on genetic structure. Finally, bud-flush phenotypes were recorded to assess which geographical, environmental, and genetic variables contribute to phenology changes.

Principal Component Analysis (PCA) reduces the dimensionality of large data sets and helps to define the relationship between geography and genetic variation. The data were coded as 0,1,2, where 0 implied homozygous reference; 1 was the heterozygote, and 2 was homozygous for the non-reference allele. The *adegraphics* R package (Dray et al., 2015) was used to summarize the most important axes of variation in this genotype matrix using PCA. Then, the PC results were compared with the geographical information of the sample's origin on a simple linear regression using the *ggpubr* package in R (Team, 2018). PCA was done within transect as well as across the entire sample. Lastly, the relationships between PC1 and PC2 as well as between PC2 and PC3 were observed to assess trends in PCs versus geography.

Admixture (Alexander et al., 2009) software was then used to estimate the ancestry of individuals. This approach considers that the samples originated from a hypothetical number *(K)* of ancestral populations. In this method, an unsupervised clustering algorithm is used to assign the genetic ancestry of each sample to these ancestral populations. For the admixture analysis, K-values were tested from K=2 to K=11, and the best K-value was selected with the help of a plot using cross validation values for the dataset. The results were plotted in R for the best K-value and arranged using the geographic information of their origin. To do this analysis, the *scatterpie* R package was used (Team, 2018).

To further assess population structure, the complementary approach DAPC was used to identify groups of genetically similar individuals in R. DAPC combines the data transformation of PCA with the division of variation between and within Discriminant Analysis (DA) groups under various population genetic models (Jombart et al., 2010). To do this, the *find.clusters* function in the *adegenet* package in R was used to select the best number of clusters (Jombart, 2008). The maximum number of clusters was set to 12 and the number of clusters was chosen based on the lowest Bayesian Information Criterion (BIC). In the next step, the *dapc* function (Jombart et al., 2010) was used to describe the clusters, and the leading 300 principal components were retained, which explained at about 80% of variance in the dataset.

Introgress analysis was used to estimate hybrid indices between 0 and 1 for each hybrid sample in R using *est.h* in the *Introgress* package (Gompert and Alex Buerkle, 2010). First, samples were identified as pure *P. balsamifera*, *P. trichocarpa*, and hybrid from admixture analysis. 44 samples were identified as *P. balsamifera* and were used as parental population 1, and 33 samples were identified as *P. trichocarpa*, which formed parental population 2 in the Introgress analysis. The other 497 samples were identified as hybrid poplar samples. According to the Introgress results, individuals were categorized as *P. balsamifera* (0.0 - 0.20), *P. trichocarpa* (0.80 - 1.0), or F_1 hybrids (0.4 - 0.6). In addition, hybrids backcrossed to *P. balsamifera* were identified according to a cutoff value between 0.2 and 0.4, and hybrids backcrossed to *P. trichocarpa* were identified according to the cutoff value between 0.6 and 0.8.

In addition, F_{ST} was calculated using VCFtools (Danecek et al., 2011). The *weir-fst-pop* function in vcftools was used to calculate Weir and Cockerham's (1984) F_{ST}. For this analysis, all hybrid individuals were removed from the dataset, and only pure *P. balsamifera* and *P. trichocarpa* were analyzed regardless of their location. First, the pure samples identified in the admixture analysis were added to the dataset. In addition, according to the results of *Introgress* analysis, the samples identified as *P. trichocarpa* (0.8 - 1.0) and *P. balsamifera* (0.0 - 0.2) were also added to the analysis. In the second step, the samples were separated according to their transect location, and the same thing was also analyzed between pure samples of *P. balsamifera* and *P. trichocarpa*. Furthermore, the relationship of *P. balsamifera* to *P. balsamifera* in other transects was examined separately, and the same was done for *P. trichocarpa*.

Relationships between genotype, geography, and climate were also tested. To infer which geographical, climate, and genetic variables contribute to the timing of phenology transitions, bud-flush phenotypes were collected from the *Populus* common garden grown under natural conditions in Critz, VA. Beginning in early April, the bud-flush stage was recorded three times a week until all trees completed the bud-flush. All plants completed bud-flush within 4 months. The timing of bud-flush was recorded as the number of days from January 1^{st} until the plants reached phenology stage three. The developmental stage of terminal buds of each tree was recorded on a categorical scale from 0 to 3 (Figure 3.2).

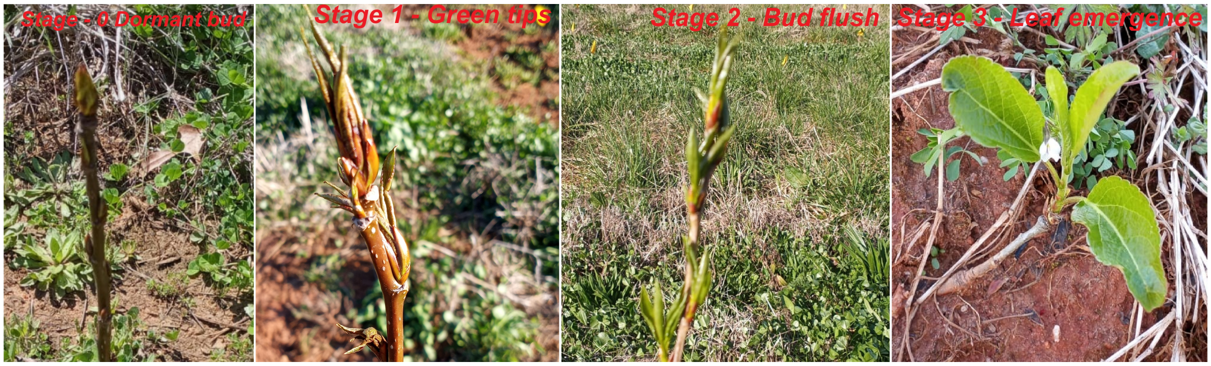


Figure 3.2: Spring phenology stages.

First, Best Linear Unbiased Prediction (BLUP) was used to estimate bud-flush timing per genotype, using the three clonal replicates as input data. BLUPs of each clone for each trait were estimated using the *lmerTest* package in R (Kuznetsova et al., 2017), in which the response variable was bud-flush date and the predictors were random effects for block and genotype. Negative BLUPs indicate early bud-flush, while positive BLUPs reflect late bud-flush. Statistical relationships between bud-flush timing BLUPs and the genotype climate of origin, as well as geography (latitude, longitude, and elevation), were analyzed. A simple linear model was used to understand the genotype-environmental association with the phenotypic BLUPs. First, the statistical relationship between bud-flush timing (BLUPs) and geographical variables was tested. The differences between the lowest and highest elevation, latitude, and longitude were 2000m, 25°, and 45°, respectively. To test the relationships between bud-flush timing and climate of origin, ClimateNA software was used (Wang et al., 2016). ClimateNA interpolates scale-free monthly, seasonal, and annual climate variables, which provides better resolution in mountainous areas such as those that characterize much of the *P. trichocarpa* range and hybrid zones with *P. balsamifera*. The 25 climate variables (DD<0: degree days below 0°C; DD>5: degree days above 5°C; MSP: May to September precipitation; TD: temperature difference between MWMT and MCMT, or continentality; MAP: mean annual precipitation; Eref: Hargreaves reference evaporation; CMD: Hargreaves climatic moisture deficit; MCMT: mean coldest month temperature; MAT: mean annual temperature; MWMT: mean warmest month temperature; DD<18: degree-days below 18°C, heating degree-days; DD>18: degree-days above 18°C, cooling degree-days; NFFD: the number of frost-free days; bFFP: the day of the year on which FFP begins; eFFP: the day of the year on which FFP ends; EXT: extreme maximum temperature over 30 years; PAS: precipitation as snow; EMT: extreme minimum temperature over 30 years; FFP: frost-free period; MAR: mean annual solar radiation; AHM: annual heat-moisture index; SHM: summer heat-moisture index; RH: mean annual relative humidity; CMI: Hogg's climate moisture index; DD1040: Degree-days above 10°C and below 40°C) were extracted from the elevation, latitude, and longitude of the sampled trees. The differences between the lowest and highest FFP, MAT, and MCMT were 200 days, 14°C, and 26°C, respectively. For relationships between bud-flush timing and their climate of origin, Mean Annual Temperature (MAT), Frost-Free Period (FFP), and Mean Coldest Month Temperature (MCMT) variables were used because they are expected to be the major climatic factors for the analysis for *P. balsamifera*. In the last step, Principal Component Analysis (PCA) for the 25 climate variables was done using R software. The first three principal components explained over 90% of the total variation. Therefore, the first three principal components were used for assessing the climate variables. The goal of this analysis was to summarize the many climate variables that are correlated in a smaller number of variables.

Lastly, the relationship between bud-flush timing and the proportions from admixture analysis was assessed. The admixture proportions were taken as the proportion of *P. balsamifera* ancestry at K=2, and the results were expressed as a simple linear regression within the transect. Specifically, the test was done to see whether proportionate ancestry from the two parental species explains bud-flush timing.

Chapter 4

Results

4.1 Population structure

In admixture analysis, I tested K-values from K=2 to K=11, and the best K-value was selected as K=4 with the help of a plot using cross-validation' values for the dataset. Admixture analysis at K=4, ordered by increasing distance from the Pacific coast, suggests genetic differentiation between individuals with increasing distance from the coast (Figure 4.1). This result was also supported at K=2 and K=3 (Figure 4.2 and Figure 4.3). The results at K=2 also showed a clear distinction between *Populus balsamifera* and *Populus trichocarpa.* In addition, the results at K=3 suggested an additional cluster (blue) that appeared in the southern half of the range but not in the north. The green cluster at K=4 was detected in the southern interior where some individuals are hybrids of *P. balsamifera* and *P. trichocarpa*. Lastly, individual cluster memberships varied depending on their distance from the coast rather than their location to the North or South because the number of pure individuals in the interior was found to be greater (Figure 4.2, Figure 4.3, and Figure 4.4). Hybrid individuals, on the other hand, were mostly found at intermediate locations from the coast, although this varied by transect.

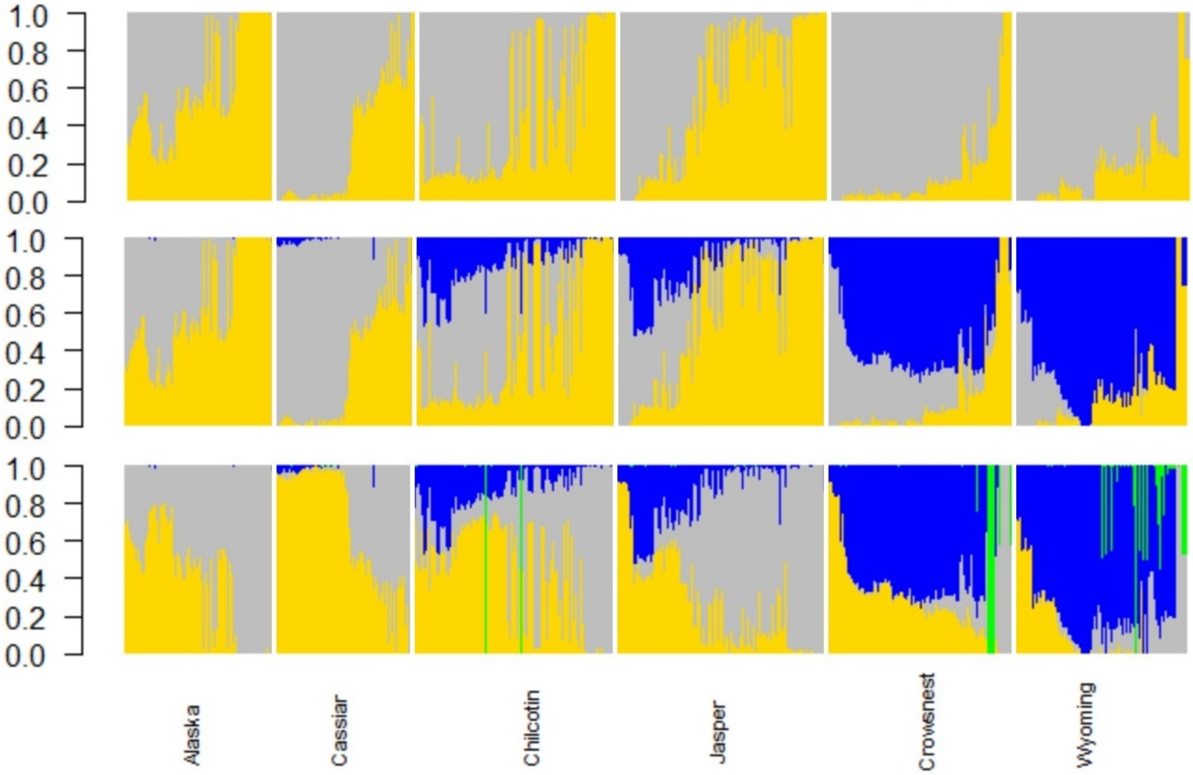


Figure 4.1: Admixture results for K=2 (top), K=3 (middle), and K=4 (bottom). Each vertical bar represents a single individual, with colors representing proportional cluster membership. Each graph was divided by transects and ordered each transect according to increasing distance from the Pacific coast.

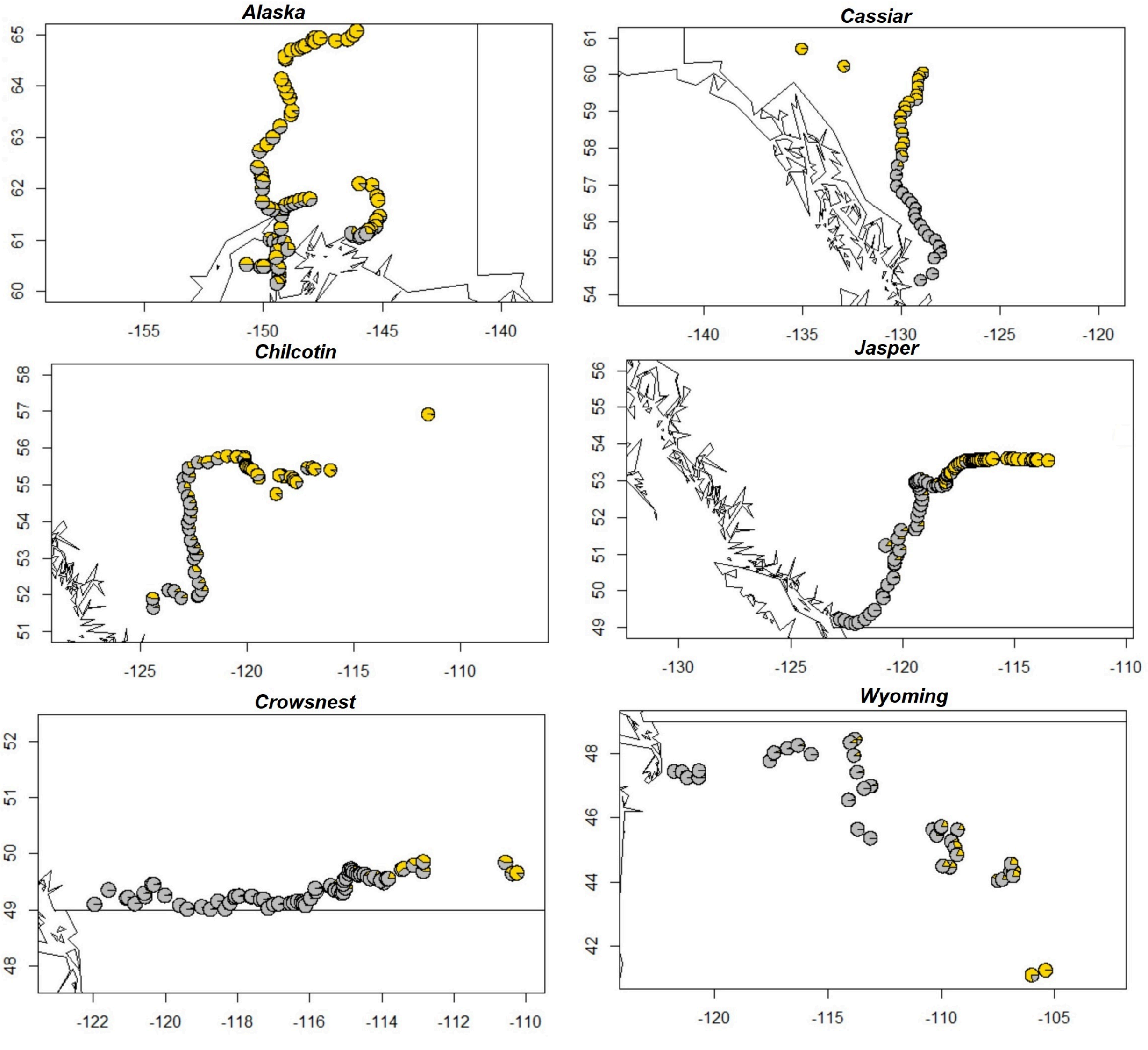


Figure 4.2: Admixture results for the 574 samples, organized by transect, at K=2, mapped according to the samples' GPS coordinates. The y-axes show latitude, and the x-axes indicate longitude.

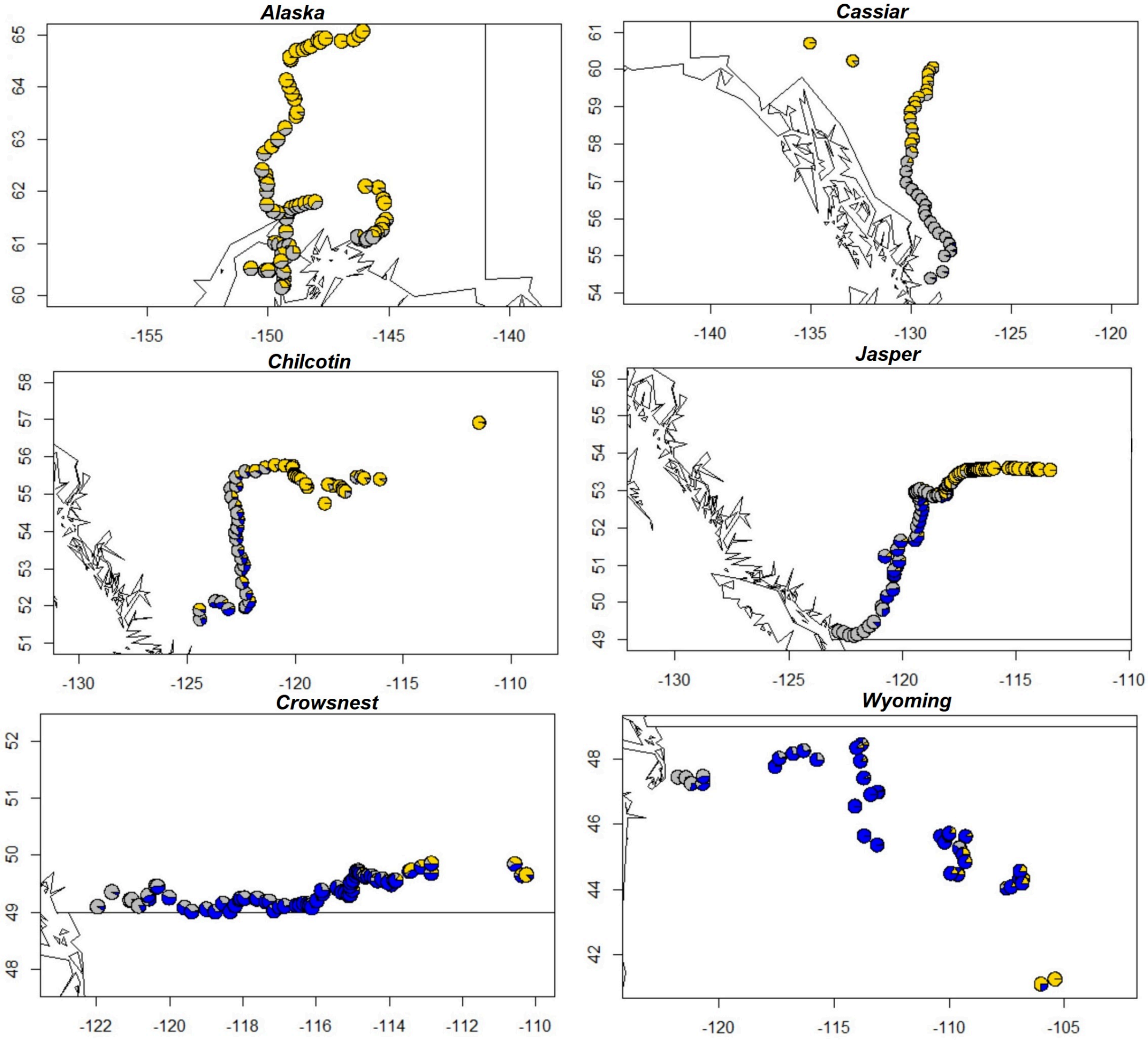


Figure 4.3: Admixture results for the 574 samples, organized by transect, at K=3, mapped according to the samples' GPS coordinates. The y-axes show latitude, and the x-axes indicate longitude.

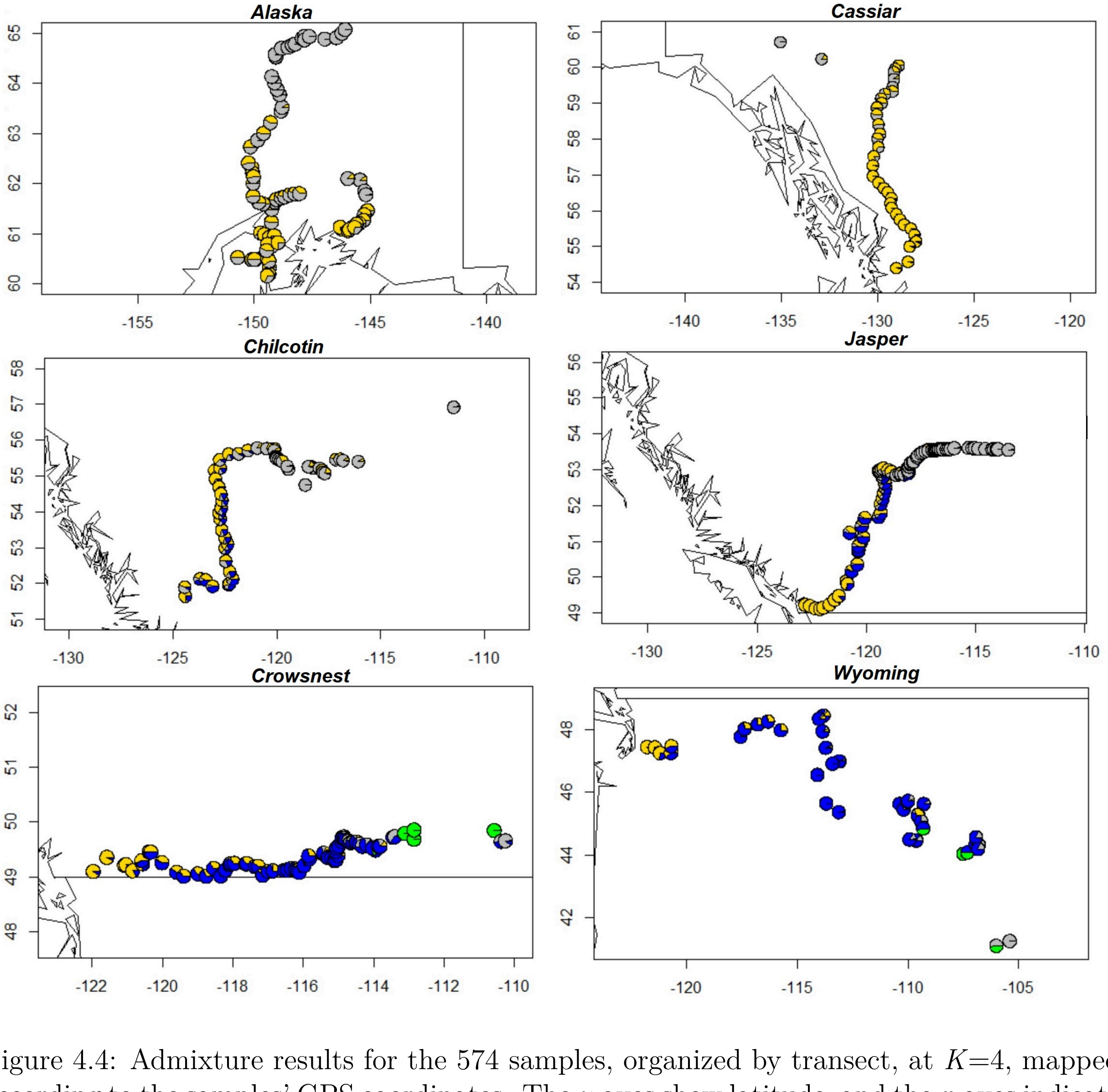


Figure 4.4: Admixture results for the 574 samples, organized by transect, at K=4, mapped according to the samples' GPS coordinates. The y-axes show latitude, and the x-axes indicate longitude.

Principal component analysis of genotypes was applied across all 574 samples to obtain a broad overview of population structure and was plotted in relation to geographical variables, including latitude, longitude, and elevation. The first three principal components (PC1, PC2, and PC3) were picked for the genotype variables because it explained approximately 70% of the variation among the 1,000,000 SNPs in the 574 poplar samples (Figure 4.5 and Figure 4.6). With the exception of PC1 and longitude ($P > 0.01$), there was a significant relationship between all geographical variables and all PC scores ($P < 0.01$) (Figure 4.7). These relationships had a positive correlation as indicated by the positive parameter values for the relationships between PCs (PC2 and PC3) and latitude. All the other relationships had a negative correlation with the positive parameter values. According to the graphs showing the relationship between PC1 and PC2 as well as between PC2 and PC3 colored by hybrid indices, more negative values on PC1 indicated greater *P. balsamifera* ancestry, and a more negative value on PC2 showed greater *P. trichocarpa* ancestry. PC3 also showed that when a value was close to 0, it resembled *P. balsamifera* ancestry. In the next analysis, we looked at the relationships within the six transects (Alaska, Cassiar, Chilcotin, Jasper, Crowsnest, and Wyoming) (Figure A.1, Figure A.2, and Figure A.3). While all PC scores had a strong relationship with longitude (all $P < 0.01$), the samples collected from the Alaska transect showed that there were not any statistical relationships between all PC scores and longitude (all $P > 0.01$). In addition, the samples collected from Crowsnest did not have any relationship with elevation and latitude (all $P > 0.01$). For latitude, only PC3 was found to have no relation, and for latitude, PC1 had no relation to the samples collected from Crowsnest.

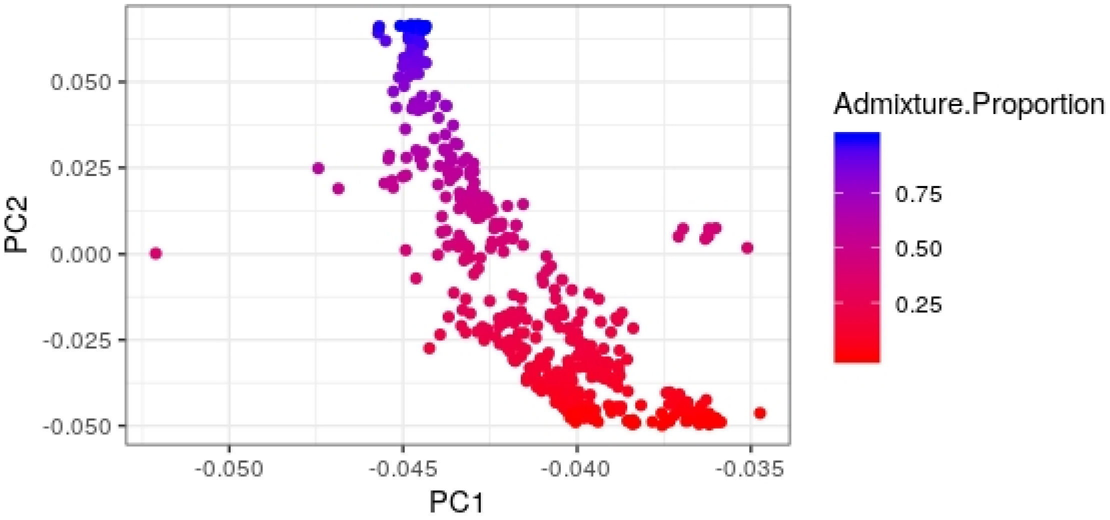


Figure 4.5: The relationship between PC1 and PC2 by admixture proportions. While the admixture proportion close to 0 resembles *P. trichocarpa* ancestry, close to 1 resembles *P. balsamifera* ancestry.

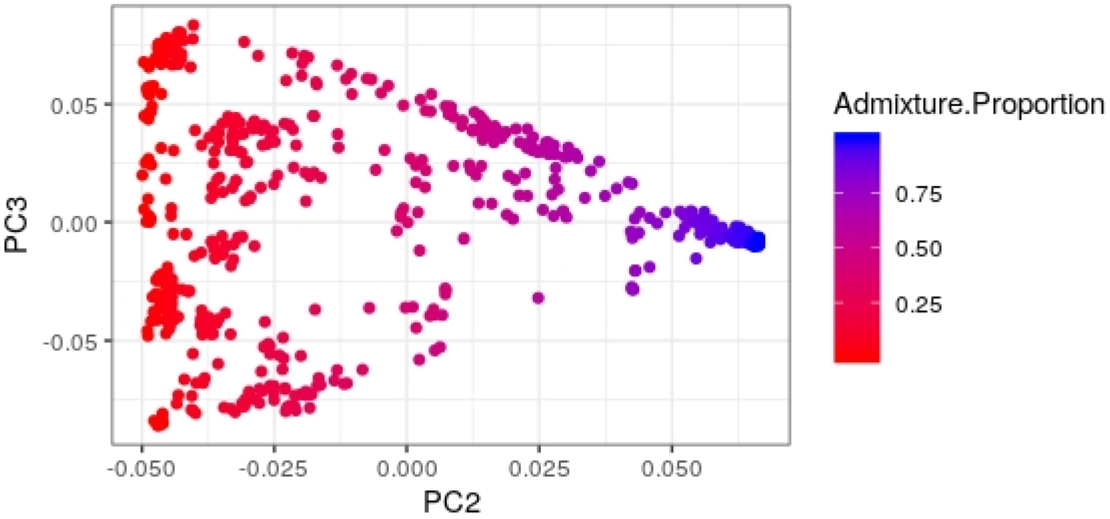


Figure 4.6: The relationship between PC2 and PC3 by admixture proportions. While the admixture proportion close to 0 resembles *P. trichocarpa* ancestry, close to 1 resembles *P. balsamifera* ancestry.

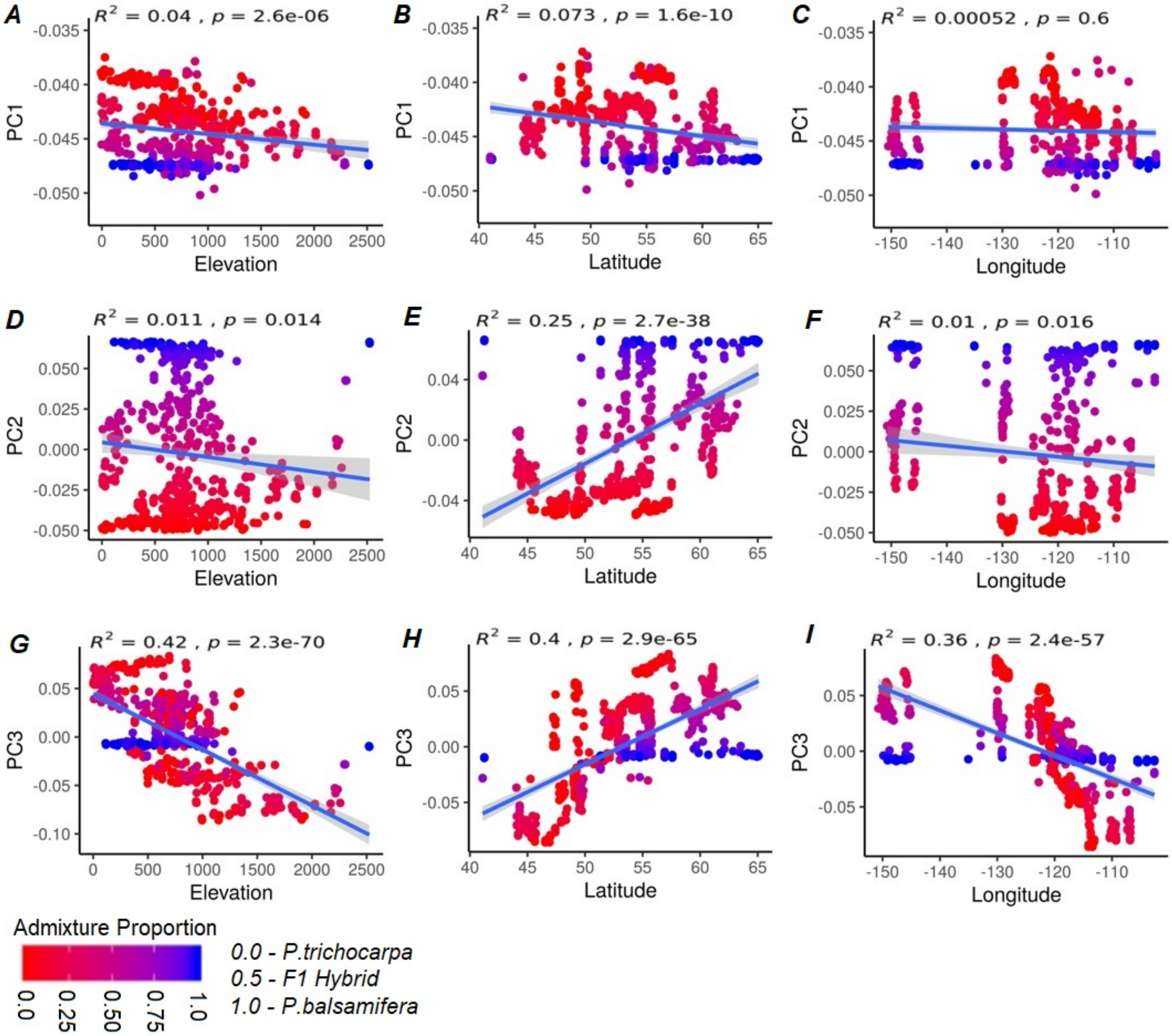


Figure 4.7: Relationship between PCA of genotype and geography. The results of PC1 are shown in graphs A, B, and C. PC2 results are shown in graphs D, E, and F, and the PC3 results are also represented in graphs G, H, and I. The y-axes show the PC scores, and x-axes represent the geographical variables. An admixture proportion close to 0 resembles *P. trichocarpa* ancestry, close to 1 resembles *P. balsamifera* ancestry.

In DAPC analysis, the best-supported number of clusters was chosen to be 4 based on the lowest Bayesian Information Criterion (BIC) result. Then, I analyzed the DAPC results among transects and the species (*P. balsamifera* and *P. trichocarpa*). The results showed that there was little distinction in clusters along transects (Figure 4.8). Similar to the admixture results, the differentiation of samples was mostly attributable to coastal versus interior habitats. Interestingly, the same result was found among individuals collected from the interior parts of the Alaska, Chilcotin, and the Jasper transects. The same figure colored by *P. balsamifera*, *P. trichocarpa*, and hybrids indicated that the samples identified as *P. balsamifera* was in only cluster 1, *P. trichocarpa* was in clusters 2 and 4, and hybrids were in clusters 1,2,3, and 4 (Figure 4.9).

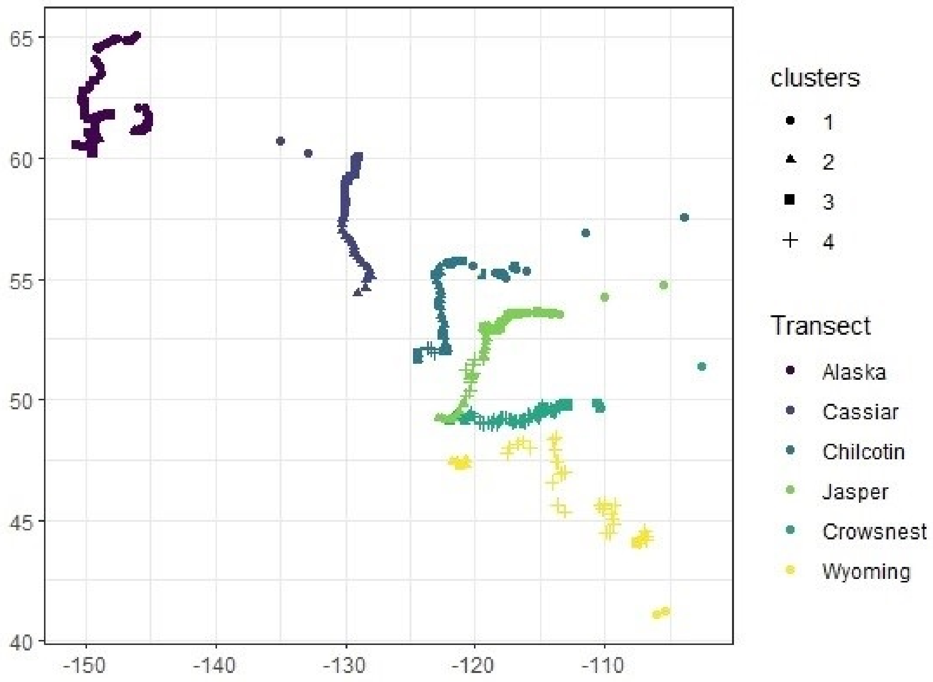


Figure 4.8: The result of DAPC analysis at K=4 clusters. Clusters are indicated by different shapes. Transects are labeled with different colors.

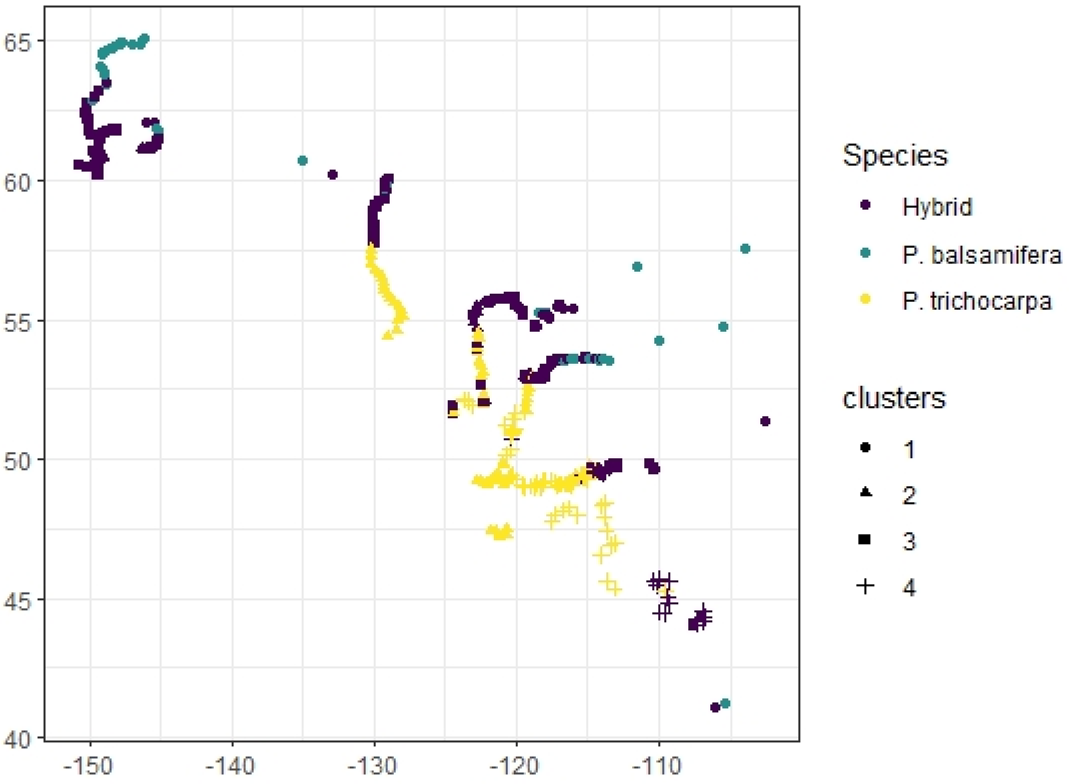
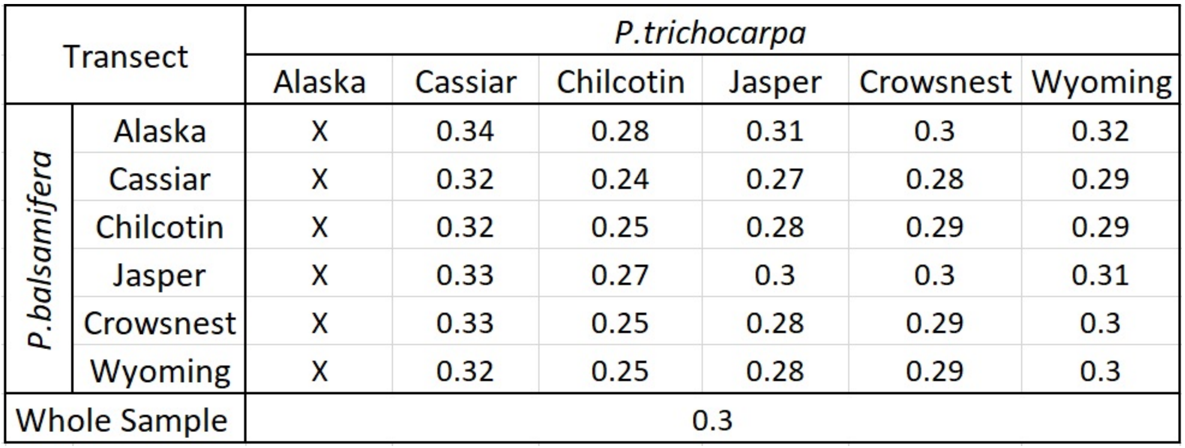


Figure 4.9: The result of DAPC analysis at K=4 clusters. Clusters are indicated by different shapes. Hybrids, *P. trichocarpa*, and *P. balsamifera* are labeled with different colors.

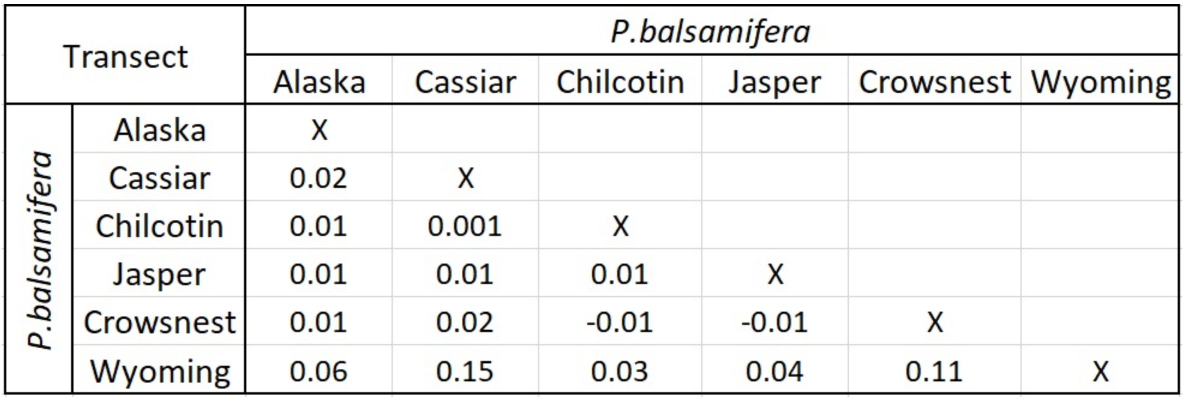
The average F_{ST} estimate between *P. balsamifera* and *P. trichocarpa* individuals across the entire sample was 0.30. F_{ST} estimates between *P. trichocarpa* and *P. balsamifera* individuals within transects varied between 0.24 and 0.34 (Table 4.1). While the highest F_{ST} value was between the individuals of *P. balsamifera* collected from the Alaska transect and individuals of *P. trichocarpa* collected from the Cassiar transect, the lowest F_{ST} value was between individuals of *P. balsamifera* collected from the Cassiar transect and individuals of *P. trichocarpa* collected from the Chilcotin transect. Furthermore, the average F_{ST} estimate among *P. balsamifera* individuals collected among transects showed that the value was between 0 and 0.15 (Table 4.2). The highest value (0.15) was between the individuals collected from the Wyoming and Cassiar transects. The lowest values (0) were among individuals collected from Crowsnest and Chilcotin, as well as Crowsnest and Jasper. In addition, the average

F_{ST} estimate among *P. trichocarpa* individuals collected among transects indicated that the range was between 0.01 and 0.06 (Table 4.3). The most similar individuals were those collected between Crowsnest and Wyoming, Crowsnest and Jasper, and Jasper and Chilcotin, while the most different were those collected between Wyoming and Cassiar transect.



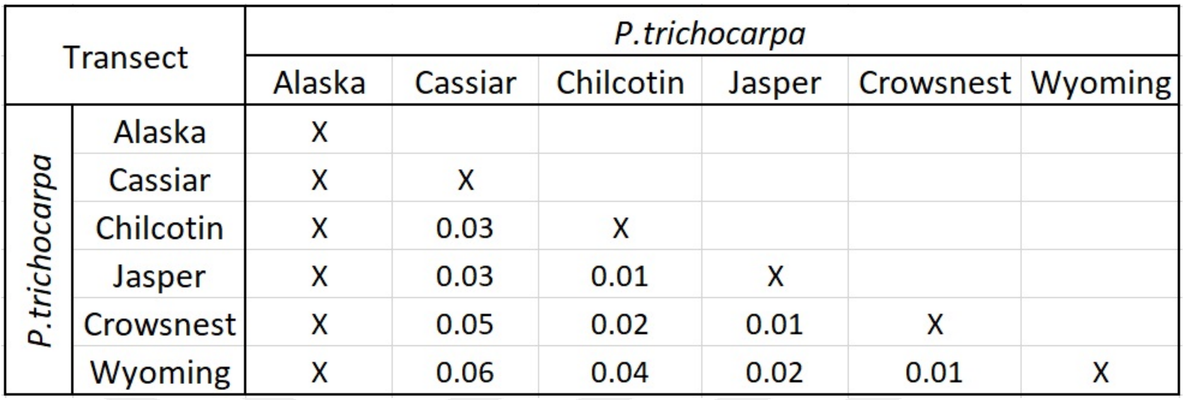
Transect		*P.trichocarpa*					
		Alaska	Cassiar	Chilcotin	Jasper	Crowsnest	Wyoming
P.balsamifera	Alaska	X	0.34	0.28	0.31	0.3	0.32
	Cassiar	X	0.32	0.24	0.27	0.28	0.29
	Chilcotin	X	0.32	0.25	0.28	0.29	0.29
	Jasper	X	0.33	0.27	0.3	0.3	0.31
	Crowsnest	X	0.33	0.25	0.28	0.29	0.3
	Wyoming	X	0.32	0.25	0.28	0.29	0.3
Whole Sample		0.3					

Table 4.1: F_{ST} metrics between individual *P. trichocarpa* and *P. balsamifera* among transects. The last row represents the overall F_{ST} result between all individuals of *P. trichocarpa* and *P. balsamifera*.



Transect		*P.balsamifera*					
		Alaska	Cassiar	Chilcotin	Jasper	Crowsnest	Wyoming
P.balsamifera	Alaska	X					
	Cassiar	0.02	X				
	Chilcotin	0.01	0.001	X			
	Jasper	0.01	0.01	0.01	X		
	Crowsnest	0.01	0.02	-0.01	-0.01	X	
	Wyoming	0.06	0.15	0.03	0.04	0.11	X

Table 4.2: F_{ST} metrics between individual *P. balsamifera* among transects.



Transcect		*P.trichocarpa*					
		Alaska	Cassiar	Chilcotin	Jasper	Crowsnest	Wyoming
P.trichocarpa	Alaska	X					
	Cassiar	X	X				
	Chilcotin	X	0.03	X			
	Jasper	X	0.03	0.01	X		
	Crowsnest	X	0.05	0.02	0.01	X	
	Wyoming	X	0.06	0.04	0.02	0.01	X

Table 4.3: F_{ST} metrics between individual *P. trichocarpa* among transects.

4.2 Hybridization

Results of *Introgress* analysis indicated a high frequency of interspecific hybridization (53%) among the sampled individuals. Among 574 samples, 266 samples were genetically identified as hybrids (Figure 4.10, Figure 4.11, and Table 4.4). Hybrid proportions and species composition showed differentiation by location. While the lowest average hybrid indices (suggesting greater balsam poplar influence) were in the samples collected from northern regions, Alaska, Cassiar, Chilcotin, and Jasper (54%, 43%, 55% and 41%, respectively), the highest average hybrid indices (suggesting greater black cottonwood influence) were in southern regions, Crowsnest and Wyoming' samples (59% and 67%, respectively). Of the 574 samples, 308 were identified as pure *P. balsamifera* or *P. trichocarpa.* The most numerous pure *P. balsamifera* samples (29 samples) were found in the Jasper transect and *P. trichocarpa* in the Crowsnest transect (73 samples). These results were mapped according to the admixture results, and geographically widespread hybridization was obvious within and among the transects (Figure 4.2, Figure 4.3, and Figure 4.4). The results of admixture analysis also showed a clear separation between the pure samples of *P. balsamifera* and *P. trichocarpa* at K=2. In addition, the level of hybridization/introgression increased with distance from the Pacific Ocean. Lastly, DAPC analysis confirmed these results, showing a genetic discontinuity between samples collected from coastal areas and the samples collected from inland areas (Figure 4.8 and Figure 4.9).

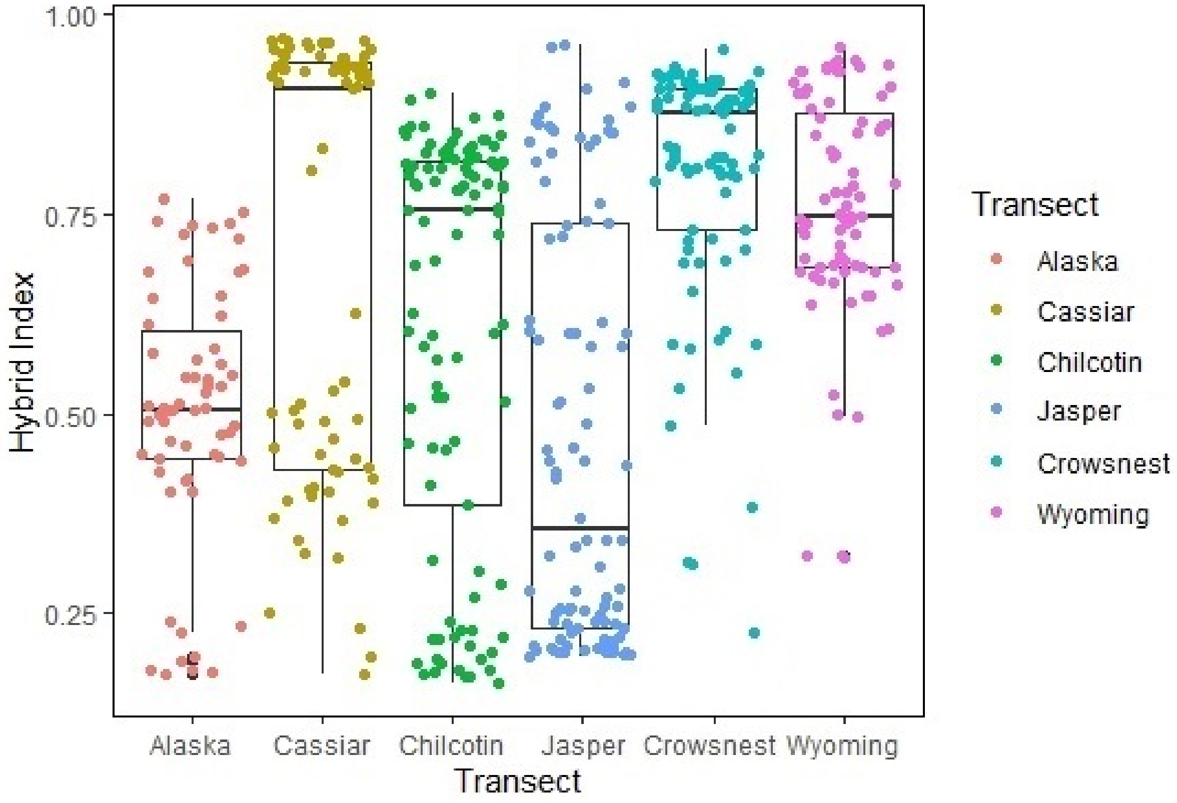


Figure 4.10: Boxplot of hybrid indices obtained from *Introgress* analysis, organized by transect. The graph does not include the 77 individuals selected as parental population from admixture analysis.

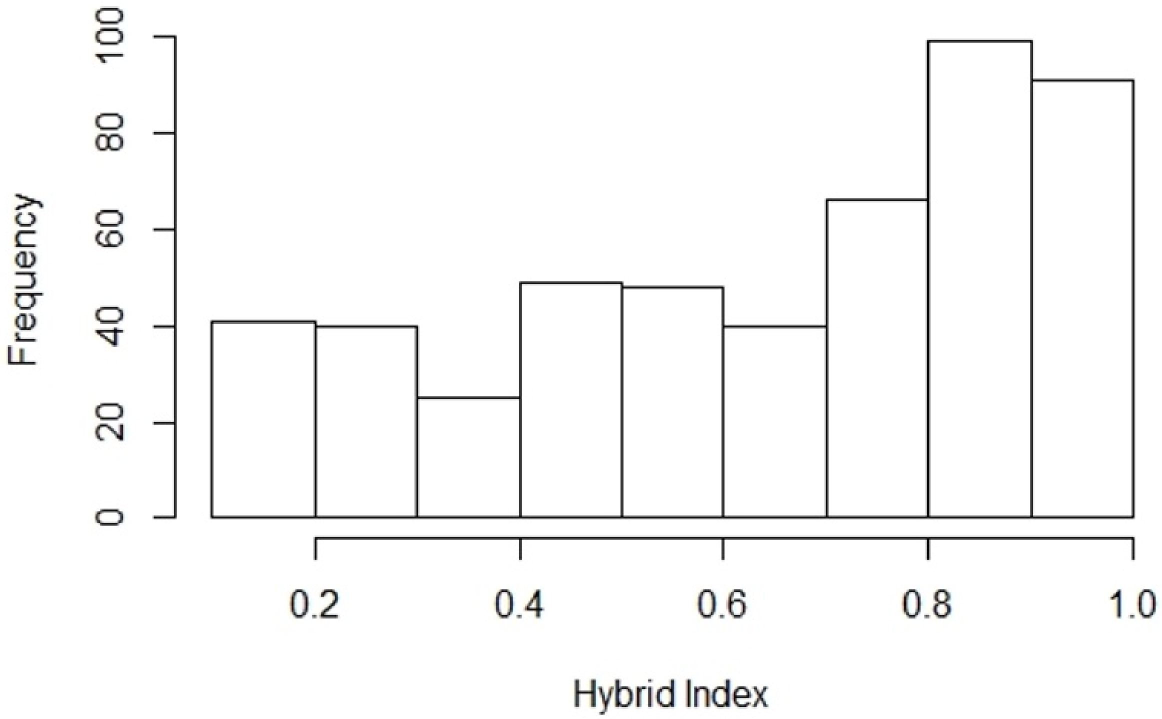
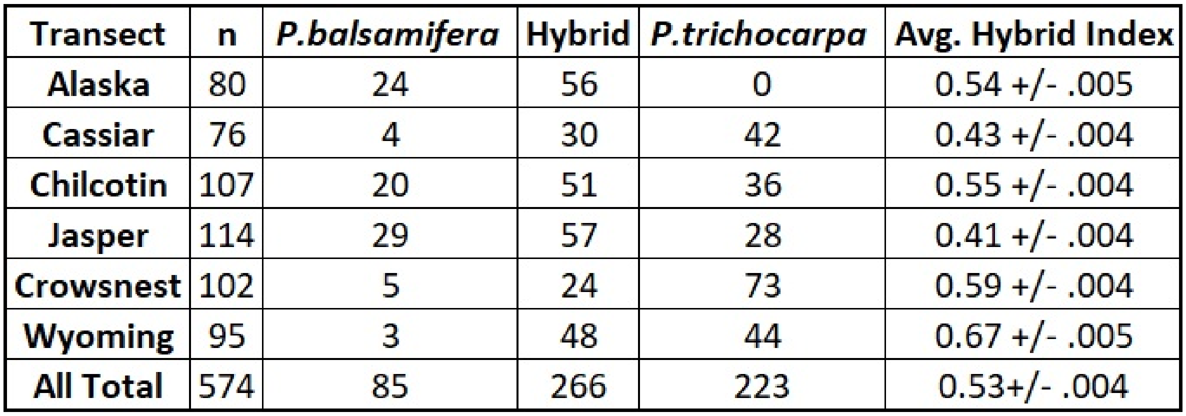


Figure 4.11: Histogram of hybrid indices among all individuals obtained from *Introgress* analysis.



Transect	n	*P.balsamifera*	Hybrid	*P.trichocarpa*	Avg. Hybrid Index
Alaska	80	24	56	0	0.54 +/- .005
Cassiar	76	4	30	42	0.43 +/- .004
Chilcotin	107	20	51	36	0.55 +/- .004
Jasper	114	29	57	28	0.41 +/- .004
Crowsnest	102	5	24	73	0.59 +/- .004
Wyoming	95	3	48	44	0.67 +/- .005
All Total	574	85	266	223	0.53+/- .004

Table 4.4: Categorization of samples genetically identified as *P. balsamifera*, *P. trichocarpa*, and hybrid. The results were taken with the help of admixture and *Introgress* analysis. Average hybrid indices were calculated from the results of *Introgress* analysis, excluding individuals identified as *P. balsamifera*, and *P. trichocarpa*, which were used as parental populations in the analysis.

4.3 Relationships between adaptation, geography, and climate

Statistical relationships between bud-flush timing using BLUP, geography, and climate of origin were tested. The relationship between geography (latitude, longitude, and elevation) and bud-flush timing based on a simple linear regression was statistically significant in all cases (all $P < 0.01$) (Figure 4.12). There was a positive correlation between bud-flush timing and longitude and elevation, and a negative correlation with latitude. In addition, the relationship between bud-flush timing and climate of origin (Mean Annual Temperature (MAT), Frost Free Period (FFP), and Mean Coldest Month Temperature (MCMT)) showed that there was a significant effect of all these climate variables on the date of bud-flush timing (all $P < 0.01$) (Figure 4.12), and that relationship had a positive correlation for MAT, FFP, and MCMT.

I also tested these relationships within each of the six transects (Alaska, Cassiar, Chilcotin, Crowsnest, Jasper, and Wyoming). For the Alaska transect, there was a relationship between bud-flush timing and geography (elevation, latitude, and longitude) ($P < 0.01$). There was a similar relationship between bud-flush timing and climate of origin (MAT and MCMT) (all $P < 0.01$), but not with FFP ($P > 0.01$). The relationship between bud-flush timing and all geographical variables showed a negative correlation, indicating earlier bud-flush in areas with higher elevation, latitude, and longitude. In addition, there was a negative correlation between bud-flush timing and all climate variables, indicating earlier bud-flush in areas with higher FFP, MAT, MCMT. For the Cassiar transect, there was a significant relationship between bud-flush timing and all geographic variables (all $P < 0.01$), as well as between bud-flush timing and all climate variables of origin (all $P < 0.01$). While there was a negative correlation between bud-flush timing and geographical variables (elevation and latitude), the other variables (longitude, FFP, MAT, and MCMT) had a positive correlation in this relationship. The third transect, Chilcotin, showed no statistically significant relationships between bud-flush timing and all geographic variables (all $P > 0.01$). Further, there were no relationships between bud-flush timing and all climate variables of origin (all $P > 0.01$). In the samples collected from the Crowsnest transect, a strong statistical relationship between bud-flush timing and longitude ($P < 0.01$) was observed, but there was no relationship with latitude, elevation, or climate variables (all $P > 0.01$) except for MCMT ($P < 0.01$). These relationships had a negative correlation with MCMT, and a positive correlation with longitude. Finally, the samples collected from the Jasper and Wyoming transects showed no statistically significant relationships between bud-flush timing and all geographic variables (all $P > 0.01$), and there were no relationships between bud-flush timing and all climate variables of origin (all $P > 0.01$). All these results are shown in Figure B.1 and Figure B.2.

In addition, I performed PCA for the 25 climate variables to see their relationships with bud-flush timing and geography (elevation, latitude, and longitude). First, a simple linear regression was used to test the relationship between the leading PCs (e.g., PC1, PC2, PC3) and bud-flush timing in the overall sample (Figure 4.13) as well as within transects (Figure B.3). There was a significant relationship between these leading PCs and bud-flush timing in overall sample (all $P < 0.01$). Within-transects, PC1 had a significant negative relationship ($P < 0.01$) with bud-flush timing in the Cassiar transect. There was also a relationship between PC2 and bud-flush timing in the Alaska, Cassiar, and Crowsnest transects (all $P < 0.01$). While PC2 was positively correlated with bud-flush timing in Alaska and Cassiar, the relationship was negatively correlated with bud-flush timing in the Crowsnest transect. For PC3, the result showed a significant relationship with bud-flush timing in Alaska, and the correlation was positive.

In the last step, the relationship between bud-flush timing and proportions from admixture analysis showed that there was a strong relationship in three transects (Figure 4.14). The samples collected from the Alaska, Cassiar, and Crowsnest transects had a strong relationship with bud-flush timing. There was no such relationship in other transects. The samples of *P. balsamifera* within transects tended to have earlier bud-flush timing than samples of *P. trichocarpa* in Alaska and Cassiar (P < 0.01 for both). This result was seen only in Alaska and Cassiar, the two northernmost regions. In addition, the samples collected in the Crowsnest transect have showed the opposite trend.

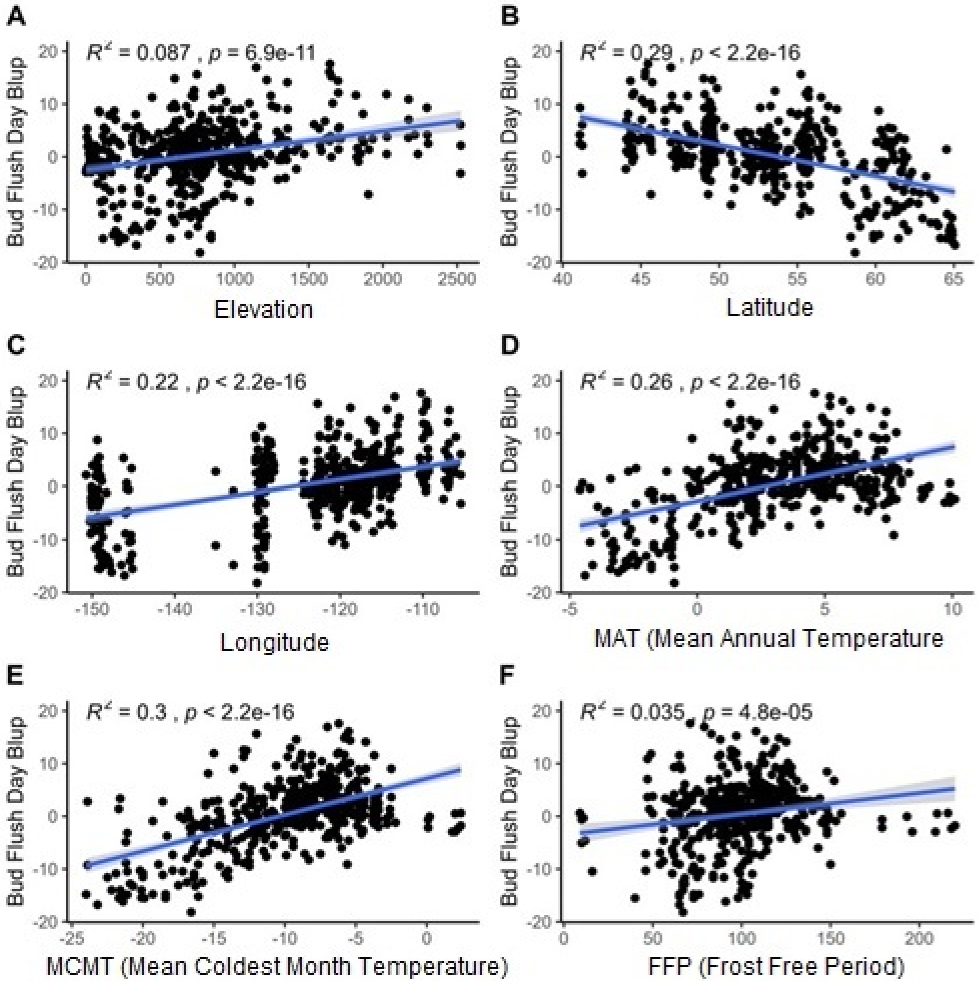


Figure 4.12: The relationship of bud-flush timing (BLUP) with climate and geography in the whole sample. The results of geography are shown in graphs A, B, and C, and climate variables are represented in graphs D, E, and F. The y-axes are BLUPs for date of bud-flush. Negative values are early bud-flush, positive are late bud-flush.

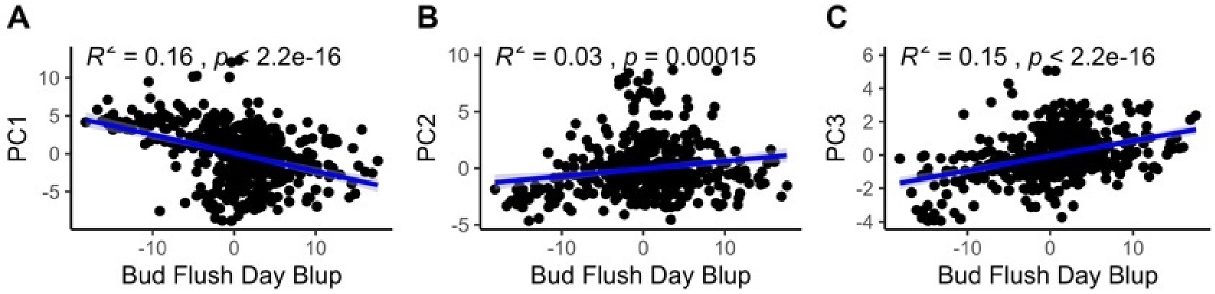


Figure 4.13: Relationship between PCA of 25 climate variables and bud-flush timing (BLUP). The y-axes indicate the PC scores and x-axes are BLUPs for date of bud-flush timing. Negative values are early bud-flush, positive are late.

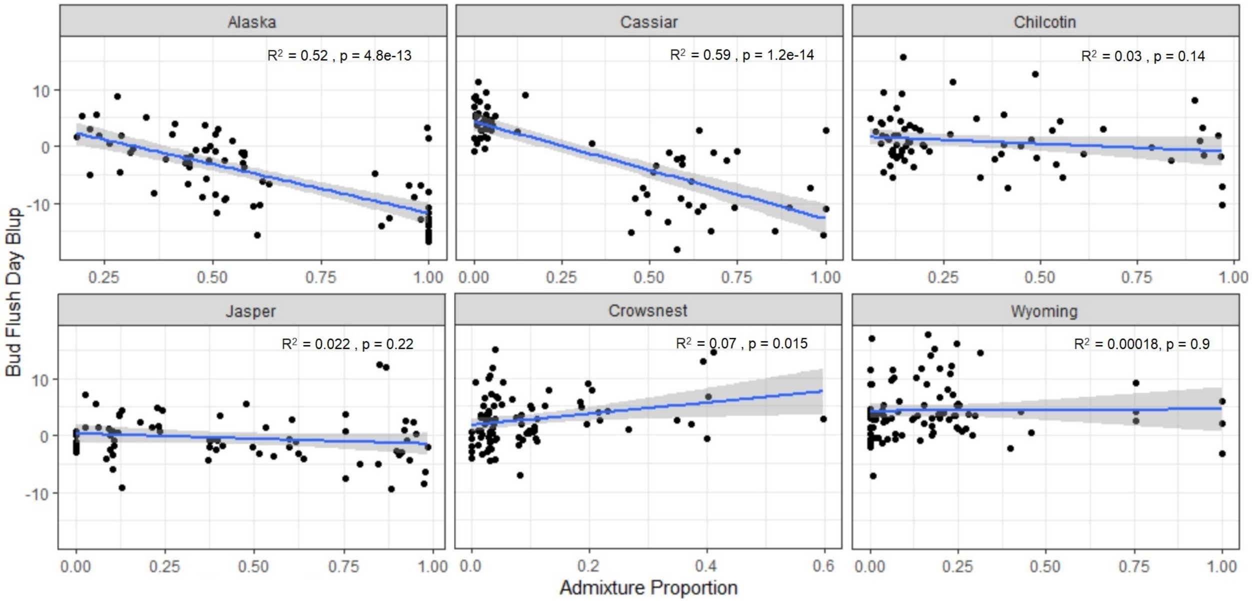


Figure 4.14: The relationship between bud-flush timing and hybrid proportion from result of admixture analysis. The graphs are divided according to transects and ordered in their latitudinal position. The y-axes are BLUPs for date of bud-flush. Negative values are early bud-flush, positive are late bud-flush.

Chapter 5

Discussion

5.1 Population structure

My results revealed complex patters of genetic structure across the sampled transects that shed light on the geographic and climatic factors affecting gene flow and reproductive isolation in *Populus trichocarpa* and *Populus balsamifera*. Principal component analysis showed that both latitude and longitude impacted genetic structure. Elevation also had an impact, but the relationships were more idiosyncratic as there was more variation in elevation in some transects than others. In addition, latitude, longitude, and climatic variables were confounded with elevation. Strong evidence of isolation was found in the northern populations, particularly in Alaska, which may reflect the unique climate in this area. While the influence of *P. balsamifera* was extensive in northern populations, a large number of pure *P. trichocarpa* were found along the Cassiar transect in northwest BC, which is a colder environment than is otherwise typical of this species. A study by Geraldes et al. (2014) found that introgression occurred mostly at the boundaries of the distributions of *P. trichocarpa* and *P. balsamifera*, and there was limited gene flow in their sampled area. Across our six transects, we observed a transition from *P. trichocarpa* near the Pacific coast, through a transition zone of hybridization that was steep in the south and more shallow in the north, and ending in pure *P. balsamifera* stands east of the Rocky Mountains. Our results also suggest that there was subtle introgression from *P. balsamifera* throughout central and northern British

Columbia. Interestingly, the admixture runs that supported four genetic cluster suggest that from central British Columbia through the intermountain west of the USA, two clusters exist for *P. trichocarpa*, which may reflect limited gene flow between the coast and interior, or possibly subtle effects of *P. balsamifera* ancestry in the interior (although this second interior cluster did not appear from the admixture analysis to reflect *P. balsamifera* ancestry).

5.2 Hybridization

In general, my sample was comprised of more pure *P. trichocarpa* than pure *P. balsamifera* genotypes, which is consistent with greater sampling coverage of the putative geographic *P. trichocarpa* range. However, hybridization between these species was common across each of the six transects. I observed a transition in the sample from greater *P. balsamifera* ancestry in the north, to more *P. trichocarpa* in the south, consistent with the former inhabiting colder climates than the latter. Indeed, no pure *P. trichocarpa* was detected in Alaska, although many hybrids were identified in the coastal portion of this transect, which suggests an adaptive advantage of *P. trichocarpa* ancestry in this relatively warm area. While interior Alaska experiences are extremely cold in winter, like much of the *P. balsamifera* range, the coast is much more temperate. For example, Valdez, Alaska, on the coast, has an average low temperature of -7°C in January, which would be similar to many of the mountainous areas we sampled much further south. By contrast, Fairbanks, Alaska, in the interior, has an average low temperature in January of -27°C. The next northernmost transect, Cassiar, follows river valleys that are part of the Stikine and Nass watersheds from the coast, with balsam hybridization being mostly absent south of Iskut, British Columbia (57.5°N, 129.5°W). There is a transition from the river valley protected from gene flow by mountains on both the east and north, to an area north of Iskut where gene flow from balsam could come down from the north via a broad valley. This connectivity may explain why we see the transition to balsam hybridization in that area. While most of the Cassiar transect (excepting the area near the coast at Prince Rupert, British Columbia) is an area with severe winters, *P. balsamifera* hybridization does not appear to be required for *P. trichocarpa* to persist. The next transect south, Chilcotin, had much greater *P. balsamifera* introgression in western and southern areas, which may reflect an unobstructed path for gene flow through a break in the Rocky Mountains in northeast British Columbia, which leads to the broad Chilcotin and Cariboo plateaus in central BC, reaching to the Coast Mountains. It is also in the Chilcotin transect that we observed an additional *P. trichocarpa* cluster when admixture analysis allows three clusters. This may be a consequence of isolation by adaptation of *P. trichocarpa* population in the warmer western sampling locations compared with colder central areas. In the Jasper transect, the interior has a substantial number of pure samples of *P. balsamifera* even though we cannot see any pure samples of *P. balsamifera* in comparable areas of the northern transects, Cassiar and Chilcotin. Lastly, in the Crowsnest and Wyoming transects, we have an additional cluster, unique to these transects, when K=4 in the admixture analysis. This cluster was likely indicative of additional hybridization with either *P. deltoides* or *P. angustifolia*, the ranges of which extend into the eastern extent of both of these transects. However, we could not finely characterize the influence of these additional species as they were not included in the baseline sequencing.

5.3 Relationships between adaptation, geography, and climate

Our common garden experiment revealed significant differences in bud-flush timing in relation to geography, climate, and ancestry. For instance, the relationship of bud-flush timing

(BLUPs) with climate and geography across the whole sample had significant relationship. When this relationship is considered within regions, the effect of climate and geography on bud-flush timing was determined only in the two northernmost regions (Alaska, and Cassiar). While northern clones from Alaska and Cassiar formed bud-flush in April, southern clones from Chilcotin, Jasper, Crowsnest, and Wyoming exhibited bud-flush in Critz common garden later (May - June). It could be a result of cold weather in the north. The results also showed the late bud-flush for high elevation and longitude. For elevation, there is correlation with latitude and longitude on bud-flush timing. For longitude, while high latitude showed earlier bud-flush timing, high longitude which is generally colder did not show the same result. It could be the result of other environmental variables (soil conditions, and precipitation). The relationship between bud-flush timing and hybridization proportion from the admixture analysis also showed the same result in the regions. While the pure species of *P. balsamifera* and *P. trichocarpa* in the two northernmost regions (Alaska and Cassiar) had a significant relationship, there is no relationship in the other transects. In this study, cold weather is probably the most important predictor of bud-flush timing.

Chapter 6

Conclusions

Hybridization affects the genetic diversity of species, and there may also be genetic clusters within species where gene flow is limited, which also may affect the efficiency of natural selection (Meirmans et al., 2017). In this study, widespread hybridization was found between *Populus balsamifera* and *Populus trichocarpa* throughout our six transects. Hybridization was more common in northern and inland areas, suggesting a benefit for *P. trichocarpa* to colonize these colder habitats. The geographic idiosyncrasies of our sampled transects appeared to modulate the degree of *P. balsamifera* ancestry, with physical barriers to gene flow partly determining the extent and distribution of such introgression into *P. trichocarpa* populations. In the future, the data we generated will be used to test how particular genomic regions contribute to hybrid fitness through genotype-environment association analysis. Genomic regions uncovered by this approach can be directly tested for their contribution to hybrid fitness through controlled crosses, which many enable genome-informed hybrid breeding of *Populus* cultivars in the future.

Bibliography

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9):1655–1664.

Bawa, R. K. (2017). *Evolutionary Genomics of Populus trichocarpa (Western Poplar)*. PhD thesis, Virginia Tech.

Braatne, J. H., Rood, S. B., and Heilman, P. E. (1996). Life history, ecology, and conservation of riparian cottonwoods in north america. *Biology of Populus and its Implications for Management and Conservation*, (Part I):57–85.

Cavalli-Sforza, L. L., Cavalli-Sforza, L., Menozzi, P., and Piazza, A. (1994). *The History and Geography of Human Genes*. Princeton University Press.

Critchfield, W. B. (1960). Leaf dimorphism in *Populus trichocarpa*. *American Journal of Botany*, 47(8):699–711.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., et al. (2011). The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158.

De La Torre, A. R., Roberts, D. R., and Aitken, S. N. (2014). Genome-wide admixture and ecological niche modelling reveal the maintenance of species boundaries despite long history of interspecific gene flow. *Molecular Ecology*, 23(8):2046–2059.

DeBell, D. S. (1990). *Populus trichocarpa* Torr. & Gray, black cottonwood. *Silvics of North America*, 2:570–576.

Dray, S., Siberchicot, A., Thioulouse, J., and Julien-Laferrière, A. (2015). adegraphics: An s4 lattice-based package for the representation of multivariate data. *URL https://github.com/sdray/adegraphics, r package version*, pages 1–0.

Eckenwalder, J. E. (1996). Systematics and evolution of *Populus*. *Biology of Populus and its implications for management and conservation*, 7:32.

Ellegren, H. (2008). Sequencing goes 454 and takes large-scale genomics into the wild.

Ellis, B., Jansson, S., Strauss, S. H., and Tuskan, G. A. (2010). Why and how *Populus* became a "model tree". In *Genetics and Genomics of Populus*, pages 3–14. Springer.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., and Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, 6(5):e19379.

Frewen, B. E., Chen, T. H., Howe, G. T., Davis, J., Rohde, A., Boerjan, W., and Bradshaw Jr, H. (2000). Quantitative trait loci and candidate gene mapping of bud set and bud flush in populus. *Genetics*, 154(2):837–845.

Geraldes, A., Farzaneh, N., Grassa, C. J., McKown, A. D., Guy, R. D., Mansfield, S. D., Douglas, C. J., and Cronk, Q. C. (2014). Landscape genomics of *Populus trichocarpa*: The role of hybridization, limited gene flow, and natural selection in shaping patterns of population structure. *Evolution*, 68(11):3260–3280.

Gompert, Z. and Alex Buerkle, C. (2010). Introgress: A software package for mapping components of isolation in hybrids. *Molecular Ecology Resources*, 10(2):378–384.

Gornall, J. L. and Guy, R. D. (2007). Geographic variation in ecophysiological traits of black cottonwood *(Populus trichocarpa)*. *Botany*, 85(12):1202–1213.

Hoffmann, A., Griffin, P., Dillon, S., Catullo, R., Rane, R., Byrne, M., Jordan, R., Oakeshott, J., Weeks, A., Joseph, L., et al. (2015). A framework for incorporating evolutionary genomics into biodiversity conservation and management. *Climate Change Responses*, 2(1):1–24.

Holliday, J., Hallerman, E., and Haak, D. (2018). Genotyping and Sequencing Technologies in Population Genetics and Genomics. In *Population Genomics*, pages 83–125. Springer.

Howe, G. T., Aitken, S. N., Neale, D. B., Jermstad, K. D., Wheeler, N. C., and Chen, T. H. (2003). From genotype to phenotype: unraveling the complexities of cold adaptation in forest trees. *Canadian Journal of Botany*, 81(12):1247–1266.

Jansson, S. and Douglas, C. J. (2007). *Populus*: A model system for plant biology. *Annu. Rev. Plant Biol.*, 58:435–458.

Jombart, T. (2008). adegenet: A r package for the multivariate analysis of genetic markers. *Bioinformatics*, 24(11):1403–1405.

Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC genetics*, 11(1):1–15.

Kawecki, T. J. and Ebert, D. (2004). Conceptual issues in local adaptation. *Ecology Letters*, 7(12):1225–1241.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. (2017). lmertest package: tests in linear mixed effects models. *Journal of Statistical Software*, 82(1):1–26.

Levsen, N. D., Tiffin, P., and Olson, M. S. (2012). Pleistocene speciation in the genus *Populus (Salicaceae)*. *Systematic Biology*, 61(3):401.

Li, H. and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, 26(5):589–595.

Little, E. L. and Viereck, L. A. (1971). *Atlas of United States Trees*. Number 1146. US Department of Agriculture, Forest Service.

Meirmans, P., Godbout, J., Lamothe, M., Thompson, S., and Isabel, N. (2017). History rather than hybridization determines population structure and adaptation in *Populus balsamifera*. *Journal of Evolutionary Biology*, 30(11):2044–2058.

Najar, A. (2017). Breaking the wall of silence of tree Mining metabolomics to describe hybridization and predict performance in the *Populus–Sphaerulina musiva* pathosystem.

Novembre, J. (2016). Pritchard, stephens, and donnelly on population structure. *Genetics*, 204(2):391–393.

Olson, M. S., Levsen, N., Soolanayakanahally, R. Y., Guy, R. D., Schroeder, W. R., Keller, S. R., and Tiffin, P. (2013). The adaptive potential of *Populus balsamifera* to phenology requirements in a warmer global climate. *Molecular Ecology*, 22(5):1214–1230.

Pauley, S. S. and Perry, T. O. (1954). Ecotypic variation of the photoperiodic response in *Populus*. *Journal of the Arnold Arboretum*, 35(2):167–188.

Rajora, O. P., Mann, I. K., and Shi, Y.-Z. (2005). Genetic diversity and population structure of boreal white spruce *(Picea glauca)* in pristine conifer-dominated and mixedwood forest stands. *Botany*, 83(9):1096–1105.

Richardson, J., Isebrands, J., Ball, J., et al. (2014). Ecology and physiology of poplars and willows. *Poplars and willows: Trees for Society and the Environment*, pages 92–123.

Savolainen, O., Pyhäjärvi, T., and Knürr, T. (2007). Gene flow and local adaptation in trees. *Annu. Rev. Ecol. Evol. Syst.*, 38:595–619.

Schowalter, T. D. (2016). *Insect ecology: An Ecosystem Approach.* Academic Press, NY.

Soolanayakanahally, R. Y., Guy, R. D., Street, N. R., Robinson, K. M., Silim, S. N., Albrectsen, B. R., and Jansson, S. (2015). Comparative physiology of allopatric *Populus* species: Geographic clines in photosynthesis, height growth, and carbon isotope discrimination in common gardens. *Frontiers in Plant Science*, 6:528.

Stettler, R. F., Koster, R., and Steenackers, V. (1980). Interspecific crossability studies in poplars. *Theoretical and Applied Genetics*, 58(6):273–282.

Suarez-Gonzalez, A., Hefer, C. A., Christe, C., Corea, O., Lexer, C., Cronk, Q. C., and Douglas, C. J. (2016). Genomic and functional approaches reveal a case of adaptive introgression from *Populus balsamifera* (balsam poplar) into *P. trichocarpa* (black cottonwood). *Molecular Ecology*, 25(11):2427–2442.

Suarez-Gonzalez, A., Hefer, C. A., Lexer, C., Douglas, C. J., and Cronk, Q. C. (2018). Introgression from *Populus balsamifera* underlies adaptively significant variation and range boundaries in *P. trichocarpa*. *New Phytologist*, 217(1):416–427.

Team, R. C. (2018). R: a language and environment for statistical computing. r foundation for statistical computing, vienna. http s. *www. R-proje ct. org.*

Tsarouhas, V., Gullberg, U., and Lagercrantz, U. (2003). Mapping of quantitative trait loci controlling timing of bud flush in salix. *Hereditas*, 138(3):172–178.

Viereck, L. A. and Little, E. L. (1972). *Alaska Trees and Shrubs.* Number 410. US Forest Service.

Wang, T., Hamann, A., Spittlehouse, D., and Carroll, C. (2016). Locally downscaled and spatially customizable climate data for historical and future periods for north america. *PLoS One*, 11(6):e0156720.

Willing, E.-M., Dreyer, C., and Van Oosterhout, C. (2012). Estimates of genetic differentiation measured by FST do not necessarily require large sample sizes when using many SNP markers.

Yeaman, S. and Whitlock, M. C. (2011). The genetic architecture of adaptation under migration–selection balance. *Evolution: International Journal of Organic Evolution*, 65(7):1897–1911.

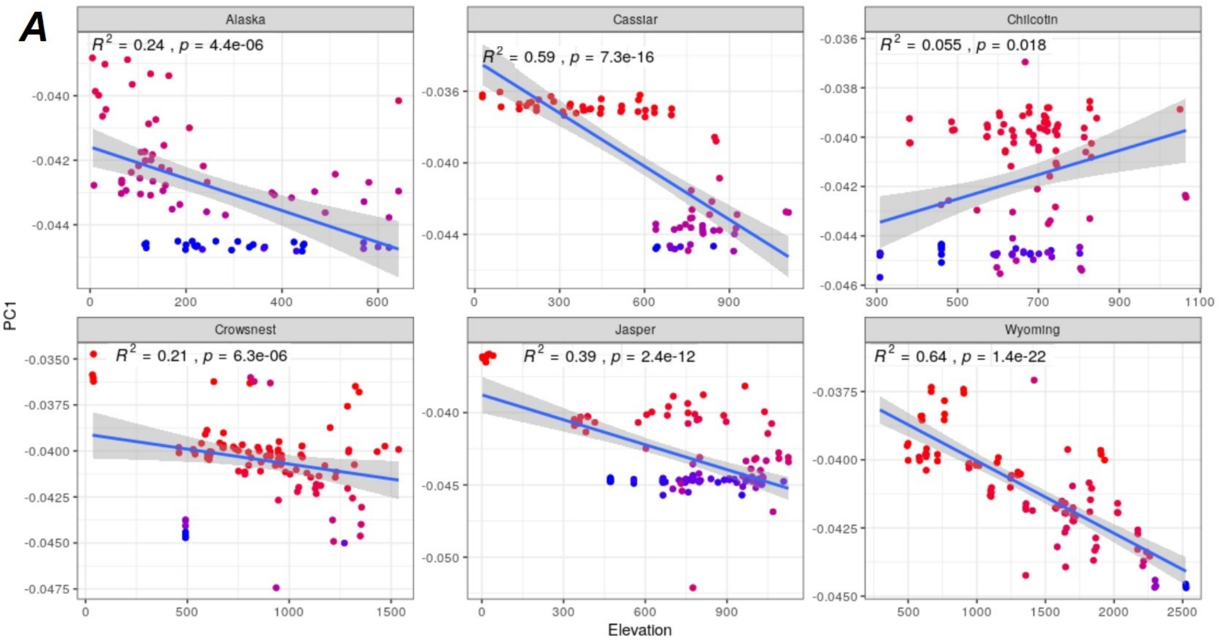
Yin, R., Kwoh, C. K., and Zheng, J. (2019). Whole genome sequencing analysis.

Zanewich, K. P., Pearce, D. W., and Rood, S. B. (2018). Heterosis in poplar involves phenotypic stability: Cottonwood hybrids outperform their parental species at suboptimal temperatures. *Tree Physiology*, 38(6):789–800.

Appendices

Appendix A

PCA of genotype vs geography among transects



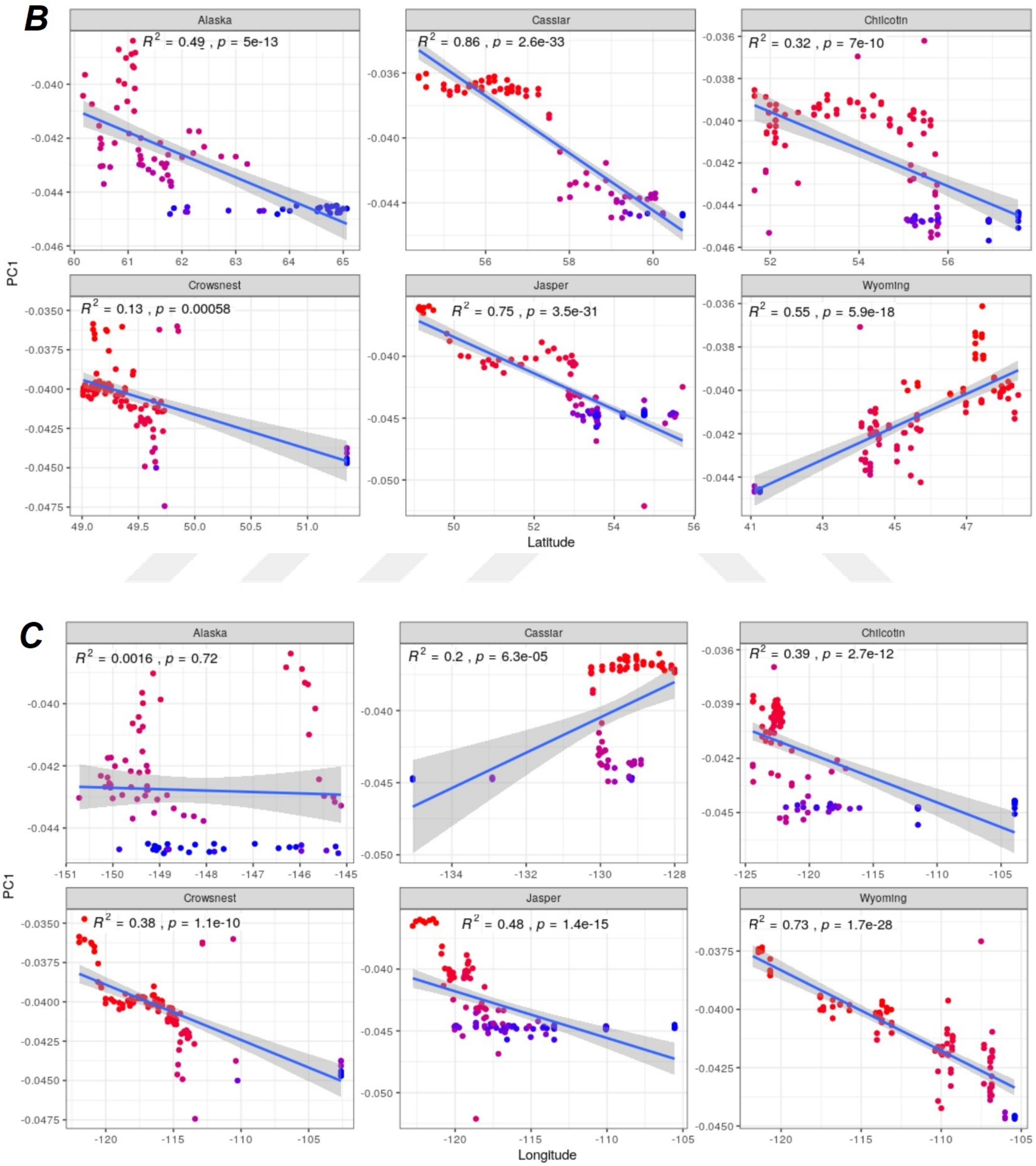
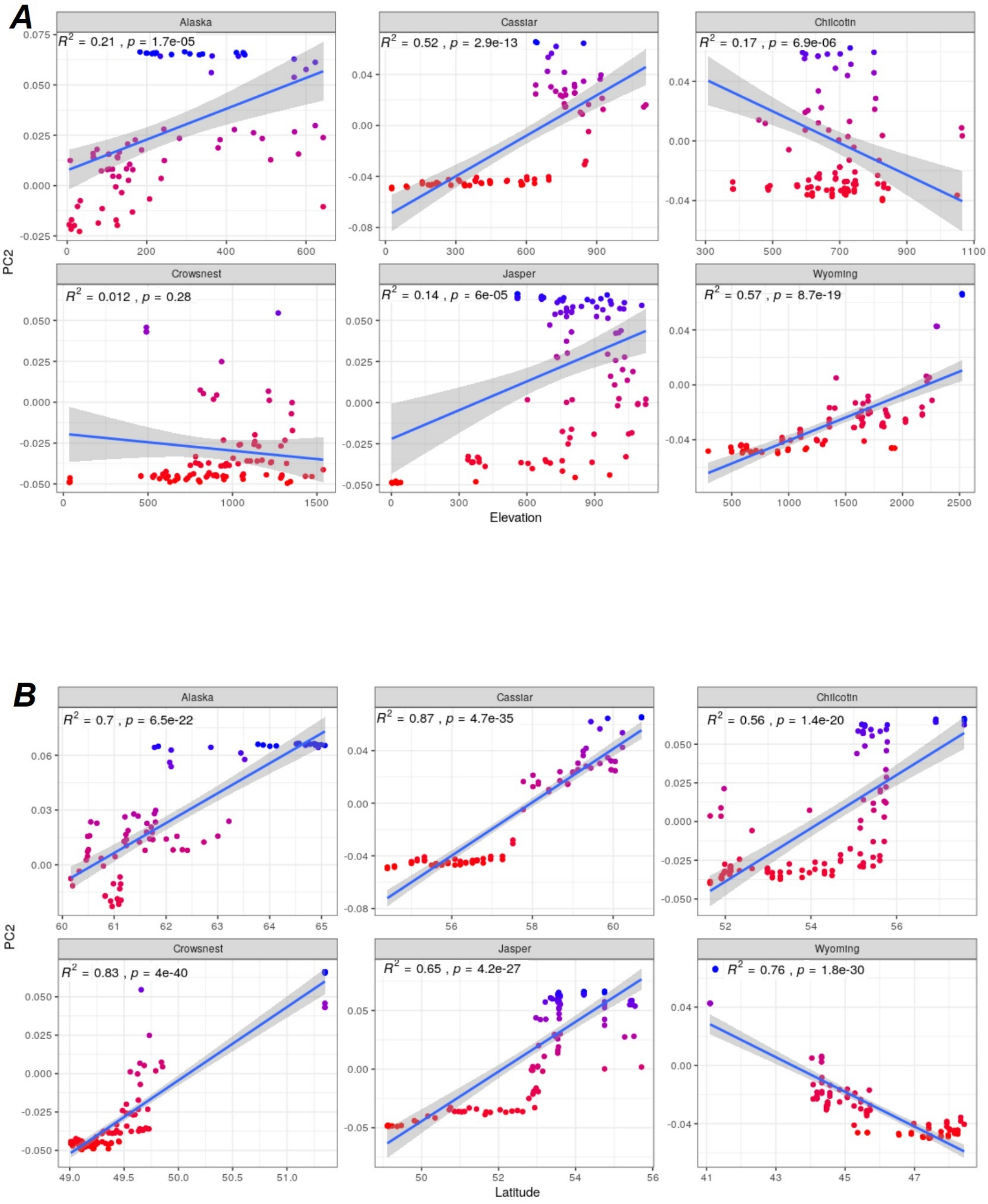


Figure A.1: The relationship between PCA of genotype and geography. The results of elevation, latitude, and longitude are shown in graphs A, B, and C, respectively. The graph is divided by transects. A value close to 0 means that the sample (tree) has both alleles that are different from the reference genome, while positive values far from 0 mean that it is close to resembling the reference genome.


A
Alaska
R^2 = 0.21 , p = 1.7e-05
Cassiar
R^2 = 0.52 , p = 2.9e-13
Chilcotin
R^2 = 0.17 , p = 6.9e-06
Crowsnest
R^2 = 0.012 , p = 0.28
Jasper
R^2 = 0.14 , p = 6e-05
Wyoming
R^2 = 0.57 , p = 8.7e-19
PC2
Elevation
0.075
0.050
0.025
0.000
-0.025
0
200
400
600
0.04
0.00
-0.04
-0.08
0
300
600
900
300
500
700
900
1100
-0.050
0
500
1000
1500
2000
2500
B
Alaska
R^2 = 0.7 , p = 6.5e-22
Cassiar
R^2 = 0.87 , p = 4.7e-35
Chilcotin
R^2 = 0.56 , p = 1.4e-20
Crowsnest
R^2 = 0.83 , p = 4e-40
Jasper
R^2 = 0.65 , p = 4.2e-27
Wyoming
R^2 = 0.76 , p = 1.8e-30
PC2
Latitude
0.06
0.03
0.00
60
61
62
63
64
65
56
58
60
52
54
56
49.0
49.5
50.0
50.5
51.0
0.05
0.00
-0.05
50
52
54
56
41
43
45
47

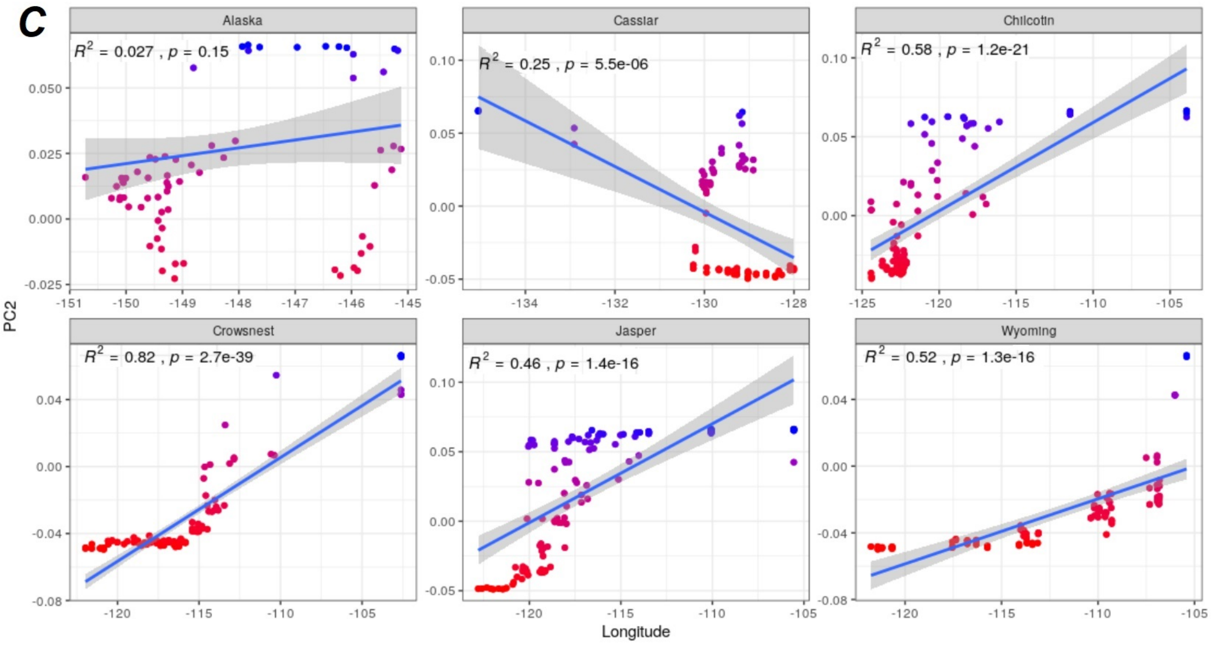
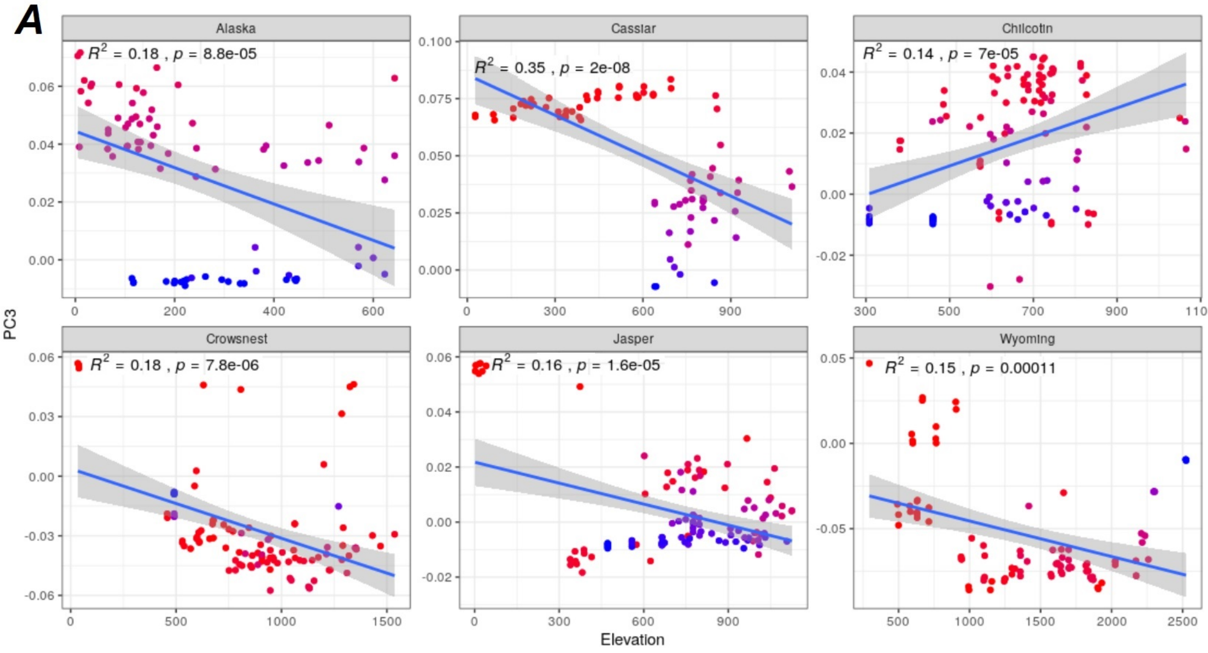


Figure A.2: The relationship between PCA of genotype and geography. The results of elevation, latitude, and longitude are shown in graphs A, B, and C, respectively. The graph is divided by transects. A value close to 0 means that the sample (tree) has both alleles that are different from the reference genome, while positive values far from 0 mean that it is close to resembling the reference genome.



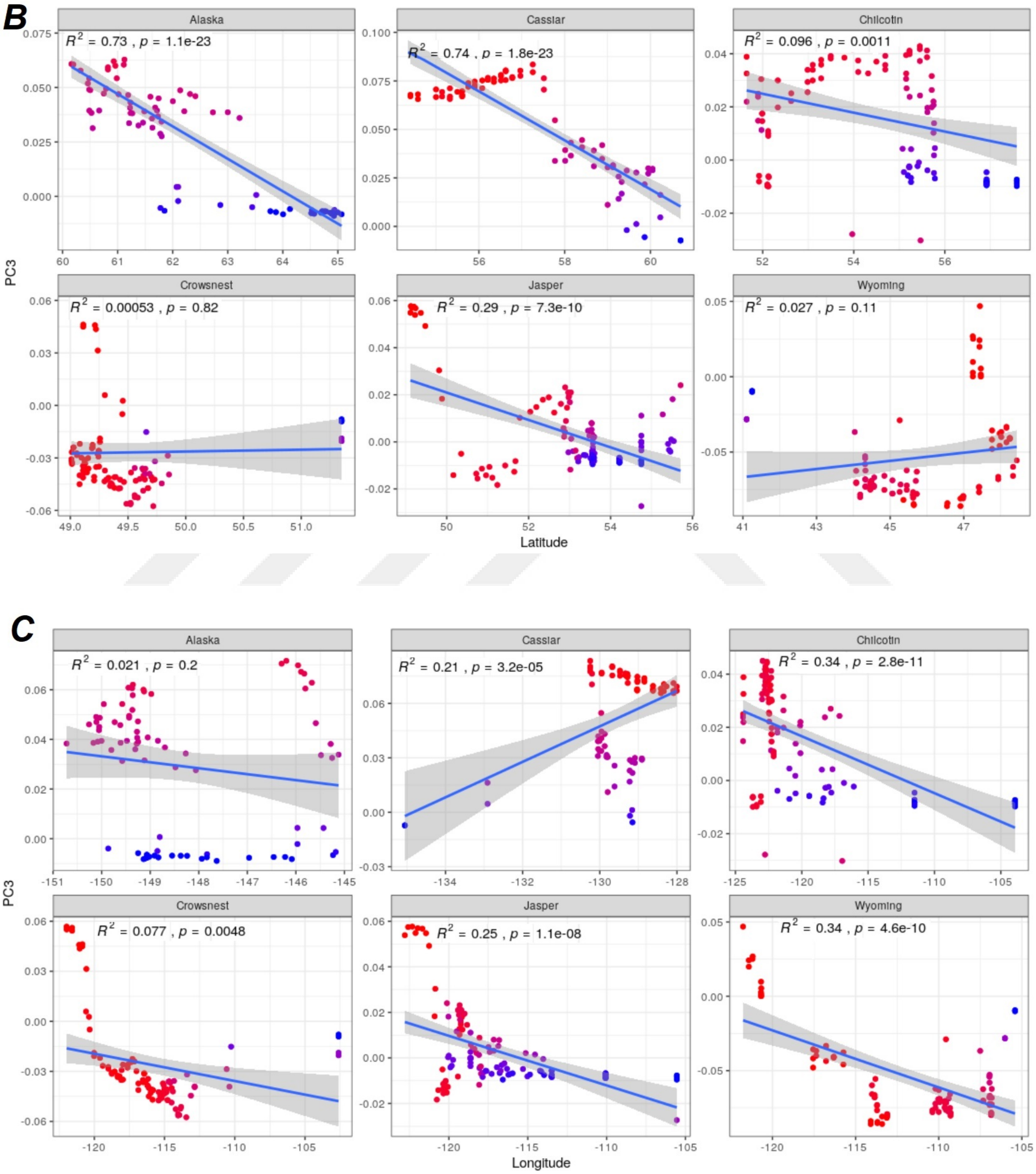
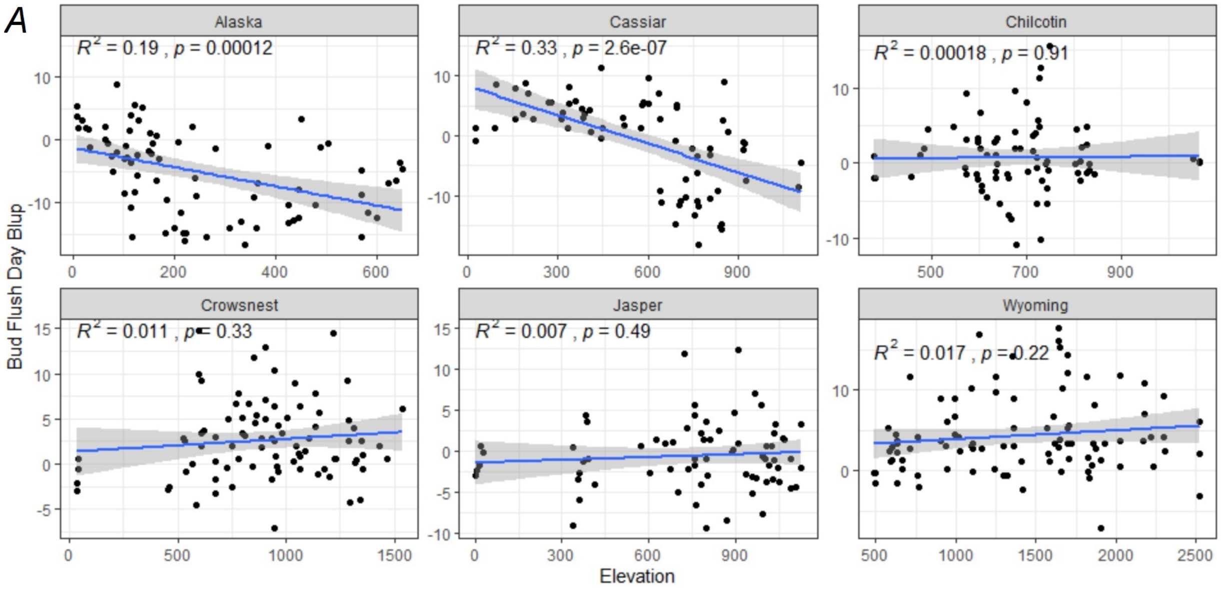


Figure A.3: The relationship between PCA of genotype and geography. The results of elevation, latitude, and longitude are shown in graphs A, B, and C, respectively. The graph is divided by transects. A value close to 0 means that the sample (tree) has both alleles that are different from the reference genome, while positive values far from 0 mean that it is close to resembling the reference genome.

Appendix B

The relationships between bud-flush timing, climate of origin, and geography among transects



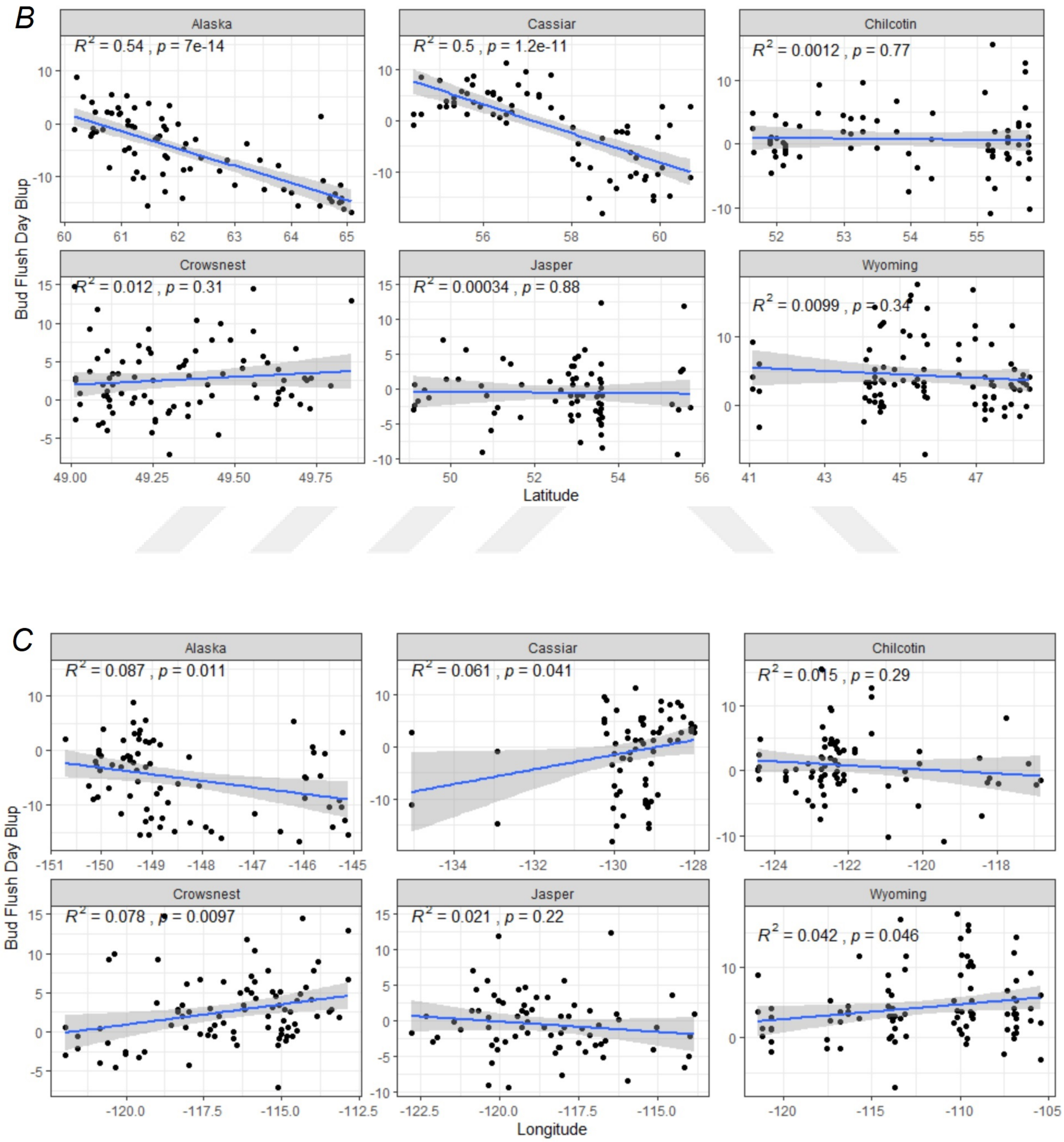
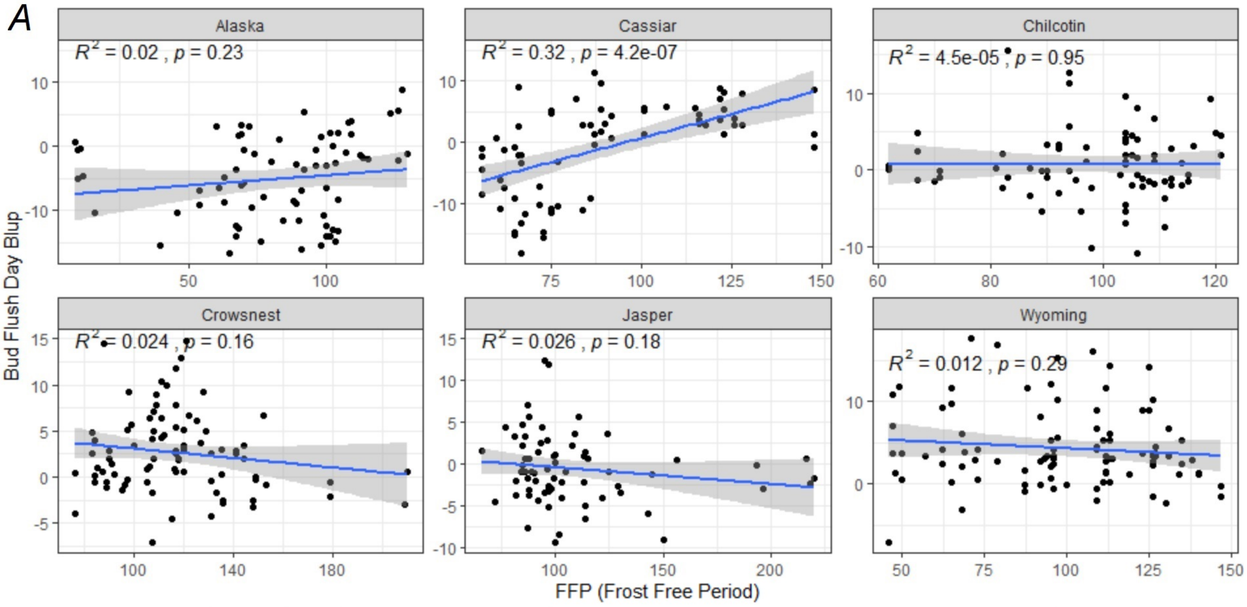
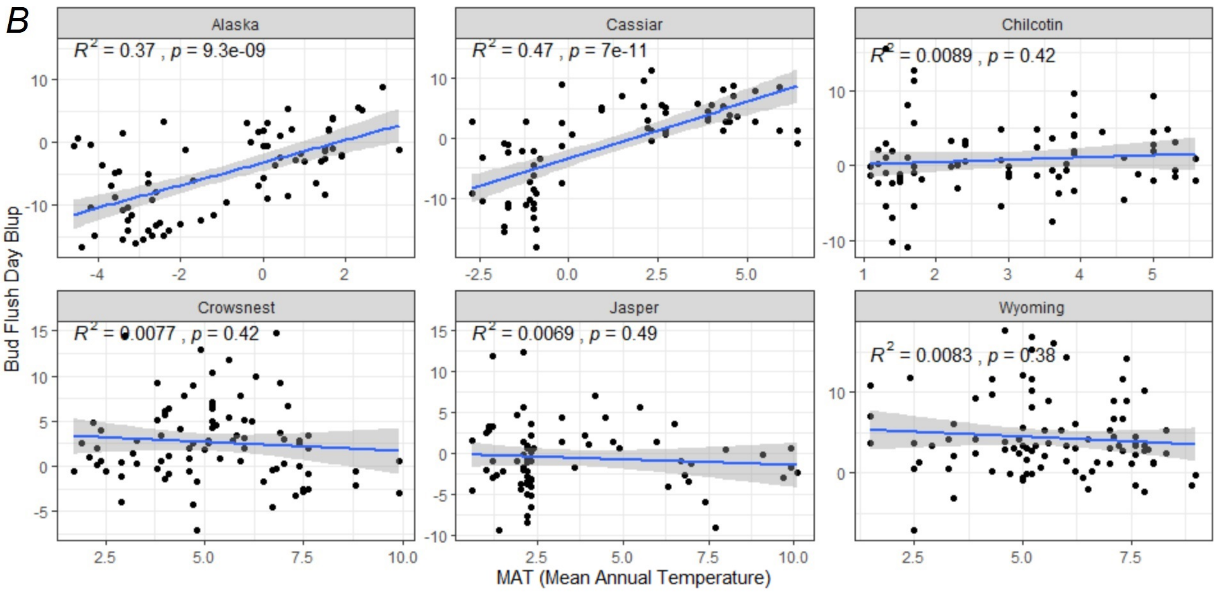


Figure B.1: The relationship between bud-flush timing and geography. The results of elevation, latitude, and longitude are shown in graphs A, B, and C, respectively. The graph is divided by transects. The *y*-axes are BLUPs for date of bud-flush. Negative values are early bud-flush, positive are late bud-flush.


A
Alaska
$R^2 = 0.02$, $p = 0.23$
Cassiar
$R^2 = 0.32$, $p = 4.2e-07$
Chilcotin
$R^2 = 4.5e-05$, $p = 0.95$
Crowsnest
$R^2 = 0.024$, $p = 0.16$
Jasper
$R^2 = 0.026$, $p = 0.18$
Wyoming
$R^2 = 0.012$, $p = 0.29$
Bud Flush Day Blup
FFP (Frost Free Period)
10
0
-10
50
100
75
125
150
60
80
120
15
5
-5
140
180
200


B
Alaska
$R^2 = 0.37$, $p = 9.3e-09$
Cassiar
$R^2 = 0.47$, $p = 7e-11$
Chilcotin
$R^2 = 0.0089$, $p = 0.42$
Crowsnest
$R^2 = 0.0077$, $p = 0.42$
Jasper
$R^2 = 0.0069$, $p = 0.49$
Wyoming
$R^2 = 0.0083$, $p = 0.38$
Bud Flush Day Blup
MAT (Mean Annual Temperature)
10
0
-10
-4
-2
2
-2.5
0.0
2.5
5.0
1
3
4
5
15
5
-5
7.5
10.0

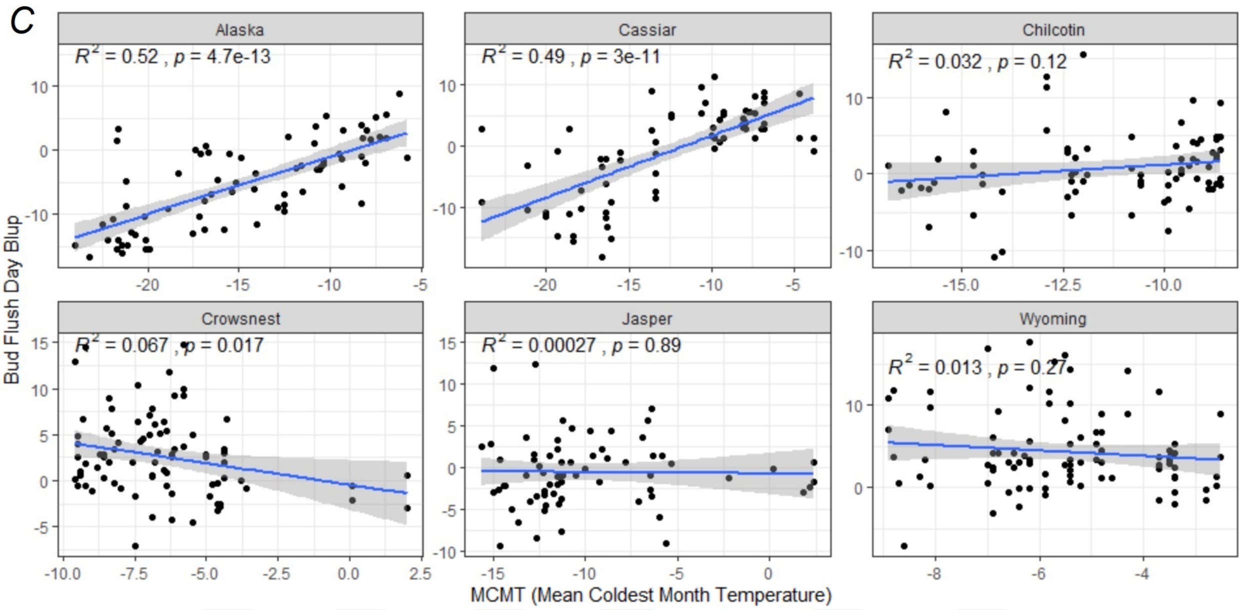
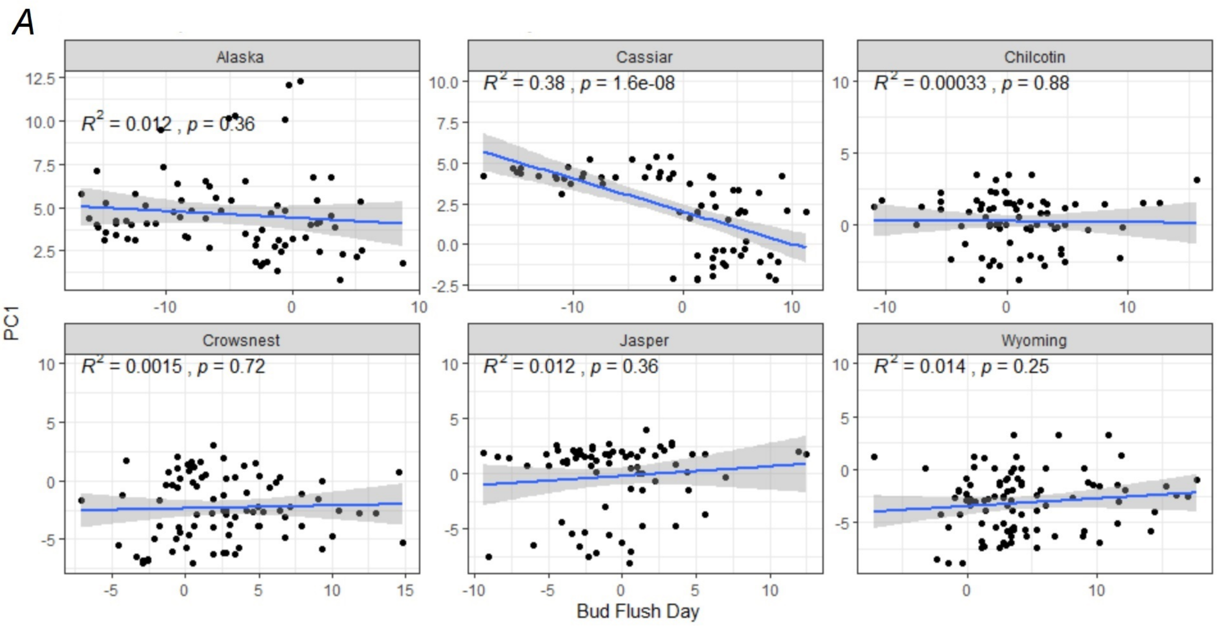


Figure B.2: The relationship between bud-flush timing and climate variables. The results of FFP, MAT, and MCMT are shown in graphs A, B, and C, respectively. The graph is divided by transects. The y-axes are BLUPs for date of bud-flush. Negative values are early bud-flush, positive are late bud-flush.



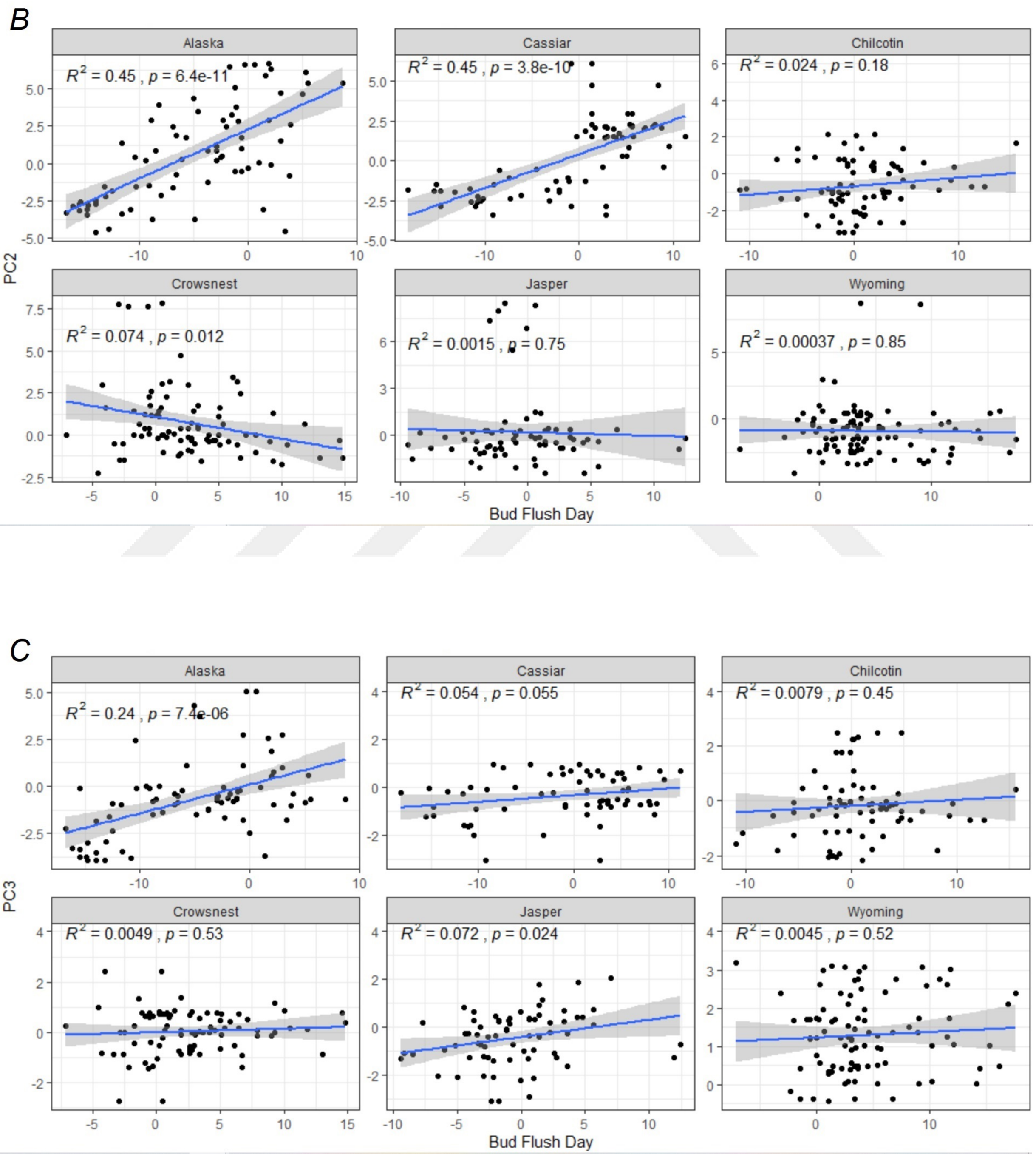


Figure B.3: The relationship between PCA of climate variables and bud-flush timing. The results of PC1, PC2, and PC3 are shown in graphs A, B, and C, respectively. The graph is divided by transects. The x-axes are BLUPs for date of bud-flush. Negative values are early bud-flush, positive are late bud-flush.