Population genomics of a timberline conifer, subalpine larch (*Larix lyallii* Parl.)

by

Marie Vance
M.Sc., University of Neuchâtel, 2011
B.Sc., University of Victoria, 2009

A Dissertation Submitted in Partial Fulfillment
of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Biology

Supervisory Committee

Population genomics of a timberline conifer, subalpine larch (*Larix lyallii* Parl.)

by

Marie Vance
M.Sc., University of Neuchâtel, 2011
B.Sc., University of Victoria, 2009

Supervisory Committee

Dr. Patrick von Aderkas, Department of Biology
Co-supervisor

Dr. Barbara Hawkins, Department of Biology
Co-Supervisor

Dr. Gerry Allen, Department of Biology
Departmental Member

Dr. Brian Starzomski, School of Environmental Studies
Outside Member

Abstract

Supervisory Committee
Dr. Patrick von Aderkas, Department of Biology
Co-supervisor
Dr. Barbara Hawkins, Department of Biology
Co-Supervisor
Dr. Gerry Allen, Department of Biology
Departmental Member
Dr. Brian Starzomski, School of Environmental Studies
Outside Member

Subalpine larch (*Larix lyallii* Parl.) has a narrow ecological niche at timberline in the Cascade Range and the Rocky Mountains of western North America. Demographic factors, including a long generation time (average 500 years) and a late arrival at sexual maturity (100-200 years), make it unlikely that this species will be able to adapt to predicted climate change. A better understanding of genetic structure and genetic diversity is necessary in order to effectively manage this species for future generations. Foliage from 62 populations of subalpine larch was collected in order to elucidate the range-wide population genomics of the species. DNA was extracted and a next-generation sequencing method, restriction site associated DNA sequencing (RAD-seq), was used to generate genome-wide single nucleotide polymorphism (SNP) marker data. Three genetically differentiated clusters were identified via principal components analysis, a discriminant analysis of principal components and Bayesian STRUCTURE analysis: the Cascade Range, the southern Rocky Mountains and the northern Rocky Mountains. A monophyletic group in the central Rocky Mountains was also identified in a dendrogram of genetic distance but this group had weak bootstrap support (49%), meaning genetic differentiation depends on relatively few genetic variants. Genetically differentiated groups should be prioritized for future management and conservation efforts. Negative values of Tajima's D and preferred demographic scenarios generated by coalescent simulations indicated that 15 populations all have a recent history of expansion. Genetic diversity within these populations was found to be moderate (H_O = 0.15 – 0.20), inbreeding coefficients were found to be high (F_{IS} = 0.15 – 0.25) and genetic differentiation among populations was found to be high (average F_{ST} = 0.18).

These results indicated that fragmentation driven by Holocene warming may have resulted in reduced effective population sizes. Smaller populations experience stronger genetic drift and an increased likelihood of inbreeding, which may hinder an adaptive response to natural selection. Still, parameter estimates for preferred demographic scenarios suggested a minimum effective population size of around 20,000 individuals, which is not considered small by most conservationists. A final study of 18 populations found local adaptation to cold temperature in the northern portion of the species range. In all seasons, populations from the northern Rocky Mountains had significantly higher cold tolerance than populations from the central Canadian Rocky Mountains and the northern Cascades. Winter cold tolerance showed strong clines associated with the frost-free period and degree days below zero. These two climate variables explained 65% of the explainable variance in phenotype when redundancy analysis models were conditioned on geography. Seven SNPs were identified that explained a significant portion of the variance in winter cold tolerance. Range-wide, additional SNPs were identified as F_{ST} outliers and/or as significantly correlated with environmental gradients, even after correcting for neutral genetic structure. Together, the results of this work indicate that dispersal, neutral evolutionary processes and natural selection have all played important roles in shaping patterns of genetic variation across the natural range of subalpine larch. All of these factors should be considered during the development of management and conservation strategies for this high-elevation conifer species.

Table of Contents

List of Tables

List of Figures

Acknowledgments

Many people have provided me with invaluable assistance in seeing this project through to completion. First, I would like to acknowledge the support I received over the years from my PhD co-supervisors, Dr. Patrick von Aderkas and Dr. Barbara Hawkins, as well as my committee members, Dr. Gerry Allen and Dr. Brian Starzomski. Feedback from my external examiner, Dr. Sally Aitken, were also very much appreciated. Dr. Jean Richardson has been a rock and I would like to acknowledge that her support was essential for the successful completion of this thesis. Deep conversations about bioinformatics and genomic data analysis were very much appreciated. Dr. David Marques was extremely generous in sharing his expertise and code for genomics data analysis. The Aitken lab has also been extremely generous about sharing information, as well as hosting me for one semester at UBC. The Koop lab at UVic allowed me to perfect my DNA extraction protocol and standardize my extractions using their facilities. Thanks especially to Dr. Eric Rondeau and Mr. David Minkley for their support. I would like to acknowledge that my field season would not have been the success that it was without the enthusiasm and hard work of Ms. Genoa Alger. Mr. Barry Jaquish and Ms. Val Ashley very kindly collected tissue for my cold tolerance experiments from the breeding arboretum that they established Kalamalka Forestry Centre. In preparation for my field season, valuable intel was generously shared by Dr. Steve Arno, Mr. Barry Jaquish, Dr. Joe Antos and many others, including collaborators at provincial and national parks in both Canada and the United States. I would also like to recognize my current employer, the BC Ministry of Forests, Lands, Natural Resource Operations and Rural Development,

and especially my colleagues in the Forest Improvement and Research Management Branch, for their support while I finished my thesis. Finally, this acknowledgement section would not be complete without mentioning all of the people who were willing to discuss my project with me over beers: all of the members of the UVic evolutionary genetics journal club, as well as Dr. Evelyn Jensen and Dr. Cherie Mosher. I have very much appreciated the opportunity to join such an excellent community of scientists.

Dedication

I would like to dedicate this thesis to my grandmother, Iva Ruth Kvam, who was the first woman in our family to get a university degree, and to my parents, who have always emphasized the value of education.

CHAPTER 1: INTRODUCTION

Species evolve to occupy specific niches. Understanding why species occupy their current niches requires knowledge of evolutionary history, dispersal and environmental change. Conifers underwent a major radiation through the Mesozoic Era (250 – 65 Ma). Today, there are between 600 and 630 extant conifer species (Wang and Ran 2014). They account for 39% of the world's forests and are considered keystone species in many ecosystems (Armenise et al. 2012). Understanding why individual conifer species grow within their current distributions is essential for informed management and can provide critical context for predicting how tree species will cope with future environmental change, including anthropogenic climate change.

The most diverse group of conifers is the family Pinaceae, which includes between 220 and 250 species in 11 genera. Members of the Pinaceae dominate terrestrial ecosystems in the temperate, boreal and montane regions of the Northern Hemisphere. Fossil-calibrated molecular-clock estimates derived from transcriptome sequencing data suggest that the most recent common ancestor (MRCA) for extant Pinaceae lived 206 Ma during the Early Jurassic Period (Ran et al. 2018). Recent advances in molecular genetics have resolved the phylogeny of the Pinaceae within two clades, the abietoid clade and the pinoid clade (Lin et al. 2010; Ran et al. 2018). The abietoid clade includes the genera *Cedrus*, *Keteleeria*, *Abies*, *Pseudolarix*, *Tsuga* and *Nothotsuga*. The pinoid clade includes the genera *Pinus*, *Cathaya*, *Picea*, *Pseudotsuga* and *Larix*. Within the pinoids, the *Pseudotsuga-Larix* group is also estimated to have diverged during the Early Jurassic, approximately 185 Ma. The identification of *Pseudotsuga* and *Larix* as sister taxa

confirms previous phylogenies that grouped these genera based on striking similarities in pollen morphology, wood and bark anatomy, and the structure of the female gametophyte (Doyle 1918; Price et al. 1987).

Larches (*Larix*) are believed to have evolved in North America during the Late Cretaceous (99 – 65 Ma). Divergence between *Pseudotsuga* and *Larix* is estimated to have occurred 86 Ma (Ran et al. 2018). This is more recent than a previous estimate based on 49 chloroplast genes (187 Ma) but is more likely to be accurate given that 4,676 orthologous genic regions were included in the transcriptome-based analysis (Lin et al. 2010; Ran et al. 2018). Cretaceous-Era divergence is also broadly supported by paleontological and palynological evidence that points to extensive diversification within the Pinaceae as the continents began to move into their modern configuration (Florin 1963; Alvarez 1994). A North American origin is supported by fossil and molecular evidence. The oldest *Larix* fossils were discovered in the Buchanan Lake formation on Axel Heiberg Island, Canada, and date to the Middle Eocene (47 – 41 Ma; LePage and Basinger 1991; Jagels et al. 2001). The oldest *Pseudotsuga* fossils were discovered in Oregon, USA, and date to the Early Oligocene (approx. 32 Ma; Schorn 1994). In addition, several molecular phylogenies have identified monophyletic North American outgroups for both *Larix* and *Pseudotsuga* (Gernandt and Liston 1999; Semerikov et al. 2003; Wei and Wang 2004a,b; Gros-Louis et al. 2005; Wei et al. 2010). One study estimated that the MRCA for Asian species of *Pseudotsuga* lived during the Early Miocene, approximately 20 Ma (Wei et al. 2010). This relatively recent coalescence provides additional support for the hypothesis that *Pseudotsuga* originated in North America and later migrated to Asia via the Bering land bridge. It is possible that *Larix*

migrated after *Pseudotsuga*. Based on nucleotide substitution rates in eight nuclear genes, divergence between American and Eurasian lineages of *Larix* was estimated to have occurred 12 Ma during the Middle Miocene (Semerikov et al. 2013). Larches differ from other Pinaceae in striking ways. Although they resemble their close relatives in terms of needle morphology and canopy architecture, larches evolved a deciduous habit. Comparative studies show that deciduous larches produce a carbon-cheap, nitrogen-efficient and well-illuminated canopy in order to achieve net carbon gains similar to their evergreen competitors (Gower and Richards 1990). These and other adaptations have allowed larches to thrive.

Today, larches grow in a widely distributed circumboreal complex made up of ten widely accepted species and three hybrids (Ostenfeld and Larsen 1930; Farjon 1990). Seven larch species are native to Europe and Asia [*L. decidua* (European larch), *L. sibirica* (Siberian larch), *L. gmelinii* (Dahurian larch), *L. griffithiana* (Himalayan larch), *L. potaninii* (Potanin larch), *L. mastersiana* (Masters larch) and *L. kaempferi* (Japanese larch)] and three larch species are native to North America [*L. laricina* (tamarack), *L. occidentalis* (western larch) and *L. lyallii* (subalpine larch)]. Where their ranges overlap, natural hybridization occurs between Siberian and Dahurian larches in Russia (Semerikov and Lascoux 2003), Masters and Potanin larches in China and western and subalpine larches in western North America (Carlson 1965). Different species of larch have dramatically different distributions. Some high-latitude species are widely distributed, forming extensive lowland forests. Siberian and Dahurian larches dominate huge geographical areas in the western and eastern Russian boreal forests, respectively. In North America, tamarack is an equally important component of the Canadian boreal

forest, and western larch is widely distributed in the western montane forests. However other larch species have very restricted ranges, in part due to a reliance on high-elevation habitat at lower latitudes. European larch, for example, is restricted to timberline in the southern part of its range and Japanese larch only grows in the montane and subalpine forests of central Honshu. In the eastern Himalayas and adjacent high-elevation regions, Masters larch grows in the upper-montane forest while Himalayan larch and Potanin larch both grow in the subalpine. In western North America, subalpine larch is restricted to timberline over a relatively small geographic area. While boreal species are expanding their ranges in response to climate change, populations that depend on high-elevation habitat appear to be facing climate-associated retractions (Mamet et al. 2019).

Species with restricted distributions are more likely to suffer habitat loss under predicted climate change scenarios. In the mountains of western North America, subalpine larch is facing unprecedented rates of warming. Mean annual temperature is increasing, with the greatest increases occurring during winter months, leading to earlier snowmelt (IPCC 2013). Summer precipitation is decreasing (IPCC 2013). The combination of increased temperature, early snowmelt and decreased summer precipitation is expected to increase the frequency of late-summer drought events and wildfires (Westerling et al. 2006; Williamson et al. 2009). In addition to these abiotic challenges, subalpine larch will have increased competition from other conifer species as warmer temperatures and earlier snowmelt make high-elevation sites more suitable for colonization by lower-elevation species. Upward range shifts have already been documented at closely monitored sites around the world. In the Santa Rosa Mountains of California, vegetation surveys carried out in 1977 and 2006/2007 showed that nine of the

ten most widely distributed species in this region shifted their ranges upward by an average of 64.7 ± 33.8 m in elevation, with a corresponding median decrease in vegetation cover of 46% in the lower portions of their ranges (Kelly and Goulden 2008). Comparing species richness data from 66 European summits between 2001 and 2008, Pauli and colleagues (2012) found an average elevation gain of 2.7 m. While this is not a large shift, it nevertheless corresponds to significant decreases in species richness as a result of extirpations from Mediterranean mountain ranges. These trends do not bode well for subalpine larch, which is already growing at timberline with limited opportunities for upward migration. Furthermore, it is unlikely that subalpine larch will be able to migrate far or fast enough to track its shifting climate niche northward (Aitken et al. 2008). Bioclimate models comparing climate averages from 1997–2006 against a 1961–1990 reference period found that the climate niches of 15 tree species in British Columbia and Alberta may have already shifted northward by an average of 130 km and this could increase to 310 km by the 2020s (Gray and Hamann 2013). In western North America, conifers are estimated to migrate at a rate of 100 m per year (McLachlan and Clark 2004). Subalpine larch will therefore be required to adapt *in situ* in response to climate change. Unfortunately, this species may not be particularly adaptable.

Several factors limit subalpine larch's adaptive capacity. First, demographic factors such as a long generation time (average 500 years) and late arrival at sexual maturity (100-200 years) will slow any adaptive responses to selection (Arno and Habeck 1972). Second, low levels of genetic variation may limit the magnitude of any such response. A study of 19 populations in the northern part of the species range found that overall genetic diversity was low compared to western larch (*L. occidentalis*), a closely

related but more broadly distributed species (Khasa et al. 2006). The same study also found that populations of subalpine larch were more genetically differentiated than those of *L. occidentalis*: allele frequencies in seven populations deviated significantly from the predictions of an infinite-allele mutation model. This led the authors to conclude that subalpine larch is genetically depauperate as a result of recent bottleneck events and reproductive isolation between spatially discontinuous populations. Although most North American boreal and temperate conifer species are broadly distributed, with large population sizes and high levels of gene flow between populations, subalpine larch has a relatively small range in the interior North Cascade Range and Rocky Mountains (Arno 1990). In the North Cascades subalpine larch's range extends 193 km from the Wenatchee Mountains in Central Washington (47° 29' N) to just north of the Canadian border in British Columbia (49° 12' N). In the Rocky Mountains subalpine larch's range extends 700 km from the Salmon Mountains in Central Idaho (45° 28' N) to just north of Lake Louise in Alberta, Canada (51° 36' N). At their closest point these two mountain systems are separated by 200 km, creating an east-west disjunction. Within each major system, discontinuities are also present between mountain ranges high enough to provide suitable habitat. Subalpine larch's highly fragmented distribution may isolate populations, limiting adaptability in several ways. Isolation prevents the spread of beneficial alleles among populations and reduces the effective size of populations. Smaller effective size makes populations more vulnerable to environmental stochasticity, genetic drift and inbreeding depression. Genetic drift and inbreeding depression reduce the efficacy of selection in small populations. Together, these factors—demography, diversity, isolation and drift —could prevent subalpine larch from adapting to

environmental change. Failure to adapt in the face of sustained, directional change can result in increasingly maladapted populations that eventually decline into extinction (Lynch and Lande 1993).

To effectively manage subalpine larch, it is necessary to develop an understanding of how it came to occupy its current distribution. However, subalpine larch is not well studied—even its position within the *Larix* phylogeny remains unclear. Molecular phylogenies based on allozyme data and ITS sequence data group subalpine larch and western larch, then tamarack (Gernandt and Liston 1999; Semerikov et al. 2003). Phylogenies based on AFLP data and cpDNA data group western larch and tamarack, then subalpine larch (Semerikov et al. 2003; Gros-Louis et al. 2005). Morphology and geographical proximity support grouping subalpine larch with western larch, but further research is needed to clarify the taxonomic relationship of the North American larches. The biogeographic history of subalpine larch is also largely unknown, although some general inferences can be made. Fossil evidence suggests that *Larix* was broadly distributed in western North America prior to the major glaciation events of the Pleistocene. In addition to the Middle Eocene fossils found on Axel Heiberg Island, *Larix* macrofossils have been identified at two other locations: a Miocene formation at Snake River Basin, Idaho (Axelrod 1965; Axelrod 1968), and a Pliocene formation at Birch Creek, Alaska (Miller and Ping 1994). Climate during the Eocene was much warmer than it is today but conifers were dominant at high latitudes and at high elevations farther south (Axelrod 1990). The uplift of the Rocky Mountains that began in the Late Cretaceous (70 Ma) and ended during the Late Eocene (37 Ma) provided high-elevation habitat at lower latitudes (Elias 2002). It seems likely that subalpine larch diverged as the

Rockies rose, adapting to high-elevation habitat as it came into existence. When global climate began to cool at the end of the Eocene, conifers began to expand their ranges southward. Further expansions occurred during the late Miocene and the Pliocene as mean annual temperature continued to drop and precipitation patterns became more seasonal (Graham 1998). Finally, the uplift of the Sierra Nevada and the Cascade Range, which ended in the Pliocene, would have provided habitat of significant relief in the southwest (McKee 1972). Thus at the beginning of the Quaternary Period (1.8 Ma – Present), subalpine larch was probably far more widely distributed than it is today.

North American conifer species, including subalpine larch, underwent dramatic range contractions during the Pleistocene (1.8 – 0.01 Ma). Cooling global temperatures led to the formation of major ice sheets on both sides of the Rocky Mountains—the Cordilleran Ice Sheet in the west and the Laurentide Ice Sheet in the east. The Cordilleran Ice Sheet extended from the mountains of southeastern Alaska into northern Washington and northwestern Montana (Booth et al. 2003). During the Pleistocene, the Cordilleran Ice Sheet underwent at least 16 large-scale advances, achieving its last glacial maximum approximately 11,000 years ago. Advancing ice would have extirpated subalpine larch from the northern part of its pre-Pleistocene range. More southerly sites in both the North Cascades and the Rocky Mountains remained unglaciated, providing refugia. When the ice retreated at the end of the Pleistocene, subalpine larch, like other conifer species, was once again able to expand its range northward. Cool climates at the end of the Pleistocene and the beginning of the Holocene would have provided continuous timberline habitat and therefore a good opportunity for northward movement. Unfortunately, it is not possible to reconstruct subalpine larch's migration routes using palynology techniques, as

has been done for other conifers, because *Larix* shares a nonsaccate pollen morphology with its closest relative, *Pseudotsuga*. The two pollen types are indistinguishable from one another in the fossil record (Simak 1966).

Most recently, subalpine larch is believed to have undergone altitudinal retreat. A general warming trend through the Holocene (10,000 years ago – Present) has pushed subalpine larch populations to high elevation, creating discontinuities between mountain ranges high enough to provide suitable habitat. Today, subalpine larch only grows at timberline in the North Cascade Range and Rocky Mountains of western North America. A poor competitor in mixed stands, it has carved out its niche in the transitional zone between the forest and alpine tundra biomes, above the altitudinal limits of other tree species (1,500 – 3,000 m). Local extirpation events during warmer interglacial periods such as the Hypsithermal (8,000 – 3,000 years ago) may explain why subalpine larch is absent from sites within its range where present-day climatic conditions could support it (Arno and Habeck 1972). Alternatively, environmental perturbations such as disease or fire may have extirpated isolated populations. The continued presence of subalpine larch on isolated, lower-elevation peaks in the Pioneer Range and the Savage Mountains of Montana is difficult to explain if the species is experiencing altitudinal retreat. However Arno and Habeck (1972) suggested that chance dispersal events may explain the colonization of these sites. Indeed, long-distance dispersal events may occur more frequently than previously suspected by biogeographers. A recent phylogenetic study of two acacias, one in the Hawaiian archipelago and one on Réunion Island, found that these two endemic species are populations of the same species that were transferred between islands by a single long-distance dispersal event of 18,000 km (Le Roux et al. 2014). In

light of such dramatic evidence, it's possible that altitudinal retreat, stochastic disturbance and long-distance dispersal have all played roles in shaping the current distribution of subalpine larch.

In the absence of fossil evidence, the history of subalpine larch needs to be elucidated by molecular population genetics. Modern patterns of genetic variation are themselves the result of dispersal and of evolutionary forces such as genetic drift and natural selection. Based on existing information, several hypotheses can be formulated regarding expected patterns of genetic variation in subalpine larch:

> *H_1: Isolation in disjunct Pleistocene refugia led to genetic divergence across the natural range of subalpine larch.*
>
> *H_2: Post-Pleistocene expansion generated gradients of genetic variation along migration routes.*
>
> *H_3: Natural selection has led to genetic differentiation between populations of subalpine larch in the northern part of the species range.*

First, I hypothesize that isolation in disjunct Pleistocene refugia led to genetic divergence between groups of subalpine larch. To explore this, patterns of genetic variation are examined across the entire species range. Like other western conifers, subalpine larch is likely to have survived the Pleistocene in unglaciated refugia in the southern part of its range. There is already evidence that at least two refugia existed—one in the North Cascades and one in the Rocky Mountains (Khasa et al. 2006). The chance retention of different polymorphisms in different refugia can produce divergence between populations. Without migration, genetic drift in isolated populations will lead to the random fixation of alleles. Thus, populations expanding out of a single refugium should be genetically more similar to one another than populations expanding out of separate

refugia. If eastern and western genetic clusters are identified, it will support the existence of Pleistocene refugia in both the Rocky Mountains and the Cascade Range. If additional genetic sub-clusters are identified, it may provide evidence that multiple refugia existed within mountain systems.

Second, I hypothesize that clines of genetic diversity were generated along post-Pleistocene expansion routes. When the Cordilleran Ice Sheet retreated approximately 11,000 years ago, subalpine larch was able to expand its range into previously glaciated habitat. Expansion can generate clines in diversity via successive founder events. Theory predicts that rapid population growth and enhanced genetic drift at expanding range margins will reduce genetic variation, creating clines of high to low diversity between core and range-front populations (Peischl et al. 2013). If subalpine larch populations survived the Pleistocene in southern refuges and then expanded northward, a north-south gradient of genetic variation, with lower levels of genetic diversity in the north, should have been established. This pattern has been observed in many conifers found in western North America including Engelmann spruce (Ledig et al. 2006), Sitka spruce (Mimura and Aitken 2007), western redcedar (O'Connell et al. 2008) and western white pine (Nadeau et al. 2015).

Finally, I hypothesize that natural selection has contributed to genetic differentiation among populations of subalpine larch. In Chapter 4, variation in a phenotypic trait, cold tolerance, is assessed for 18 populations growing together in a common garden. Cold tolerance is known to be a crucial trait for determining the range boundaries of many conifer species. At timberline, subalpine larch relies on physiological adaptation to cold to survive within its environmental niche. Although it grows slowly,

with growth rates of one radial centimeter every 8 – 10 years considered impressive (Arno and Habeck 1972), subalpine larch has managed to form extensive stands on cool, moderate exposures. It has also succeeded in colonizing avalanche chutes, snowdrift sites, bedrock, coarse talus and bogs. Clearly this is a species that can cope with environmental extremes. What is not clear is how well it will cope with predicted environmental change. A better understanding of local adaptation would inform the conservation and management of the species.

The overall goal of this thesis is to assess patterns of genetic variation among and within populations of subalpine larch throughout the entire species range. Genetic factors that may threaten subalpine larch's long-term persistence (e.g., low genetic variation, reproductive isolation, small effective population sizes and inbreeding) will be assessed. Genetically unique populations that may harbor different adaptive features will be identified. Demographic and biogeographic histories will be elucidated. This body of work will represent a substantial step forward in our understanding of subalpine larch's history and its odds of long-term persistence under climate change.

CHAPTER 2: GENETIC STRUCTURE

Introduction

Conservation priorities must be set with an understanding of spatial context and connection in order to capture intraspecific genetic variation and the processes that generate it. Molecular genetics can be used to elucidate patterns of genetic variation on different scales across the landscape. Ultimately, molecular approaches seek to identify meaningful conservation units that preserve genetic diversity. Such diversity represents evolutionary history, underpins local adaptation and provides the raw material for future responses to natural selection. The International Union for Conservation of Nature recommends conserving genetic diversity in order to maintain biological interactions and ecological processes (IUCN 2017). In this study, range-wide patterns of genetic variation are elucidated for a high-elevation conifer species, subalpine larch (*Larix lyallii* Parl).

Subalpine larch is a long-lived North American conifer with several characteristics that make it vulnerable to climate change. First, this species has a restricted range in western North America (Arno and Habeck 1972). In the Rocky Mountains, the species' range extends approximately 700 km from the Salmon Mountains in central Idaho (45° 28' N) to just north of Lake Louise in Banff National Park, Alberta (51° 36' N). In the Cascade Range, it extends from the Wenatchee Mountains in central Washington (47° 29' N) to just north of the Canadian border (49° 12' N). This relatively restricted geographic range makes subalpine larch vulnerable to local disturbances, which are predicted to increase with climate change (IPCC 2013). For example, climate-associated outbreaks of pests and pathogens have already been

observed in the forests of western North America (Bentz et al. 2010) and wildfires will continue to increase in frequency and severity (Westerling et al. 2006; Williamson et al. 2009). Second, subalpine larch is adapted to a very narrow ecological niche. This species is a poor competitor in mixed stands and only grows at timberline above the altitudinal limits of other tree species (1,520 – 3,010 m). Unfortunately, alpine habitat is shrinking in western North America (Hamann and Wang 2006). General circulation models predict warmer mean annual temperature, earlier snowmelt and decreased summer precipitation in the mountains of western North America (IPCC 2013). Subalpine larch will likely face more frequent late-summer drought events as well as increased competition from lower elevation tree species as they shift their ranges upwards (Pauli et al. 2012). Third, subalpine larch is demographically challenged. Although it is long-lived—trees live an average of 500 years—sexual maturity is not reached until 100-200 years of age (Arno and Habeck 1972). It is therefore unlikely that subalpine larch will be able to adapt quickly enough to avoid maladaptation (Lynch and Lande 1993). Finally, subalpine larch may lack the standing genetic variation necessary to respond to natural selection. Because this species is restricted to timberline, its distribution is highly fragmented. Generally, population subdivision leads to lower effective population sizes and stronger genetic drift, which can result in the random fixation of deleterious alleles. Maladaptation further reduces population size. Mating between close relatives is more likely in small populations and can lead to inbreeding depression. The accumulation of deleterious or loss-of-function alleles reduces fecundity as well as the survival of inbred offspring. Together, genetic drift and inbreeding depression in small populations contribute to an "extinction vortex", whereby small populations have reduced fitness, which leads to

further reductions in population size, and so on until extinction (Gilpin and Soulé 1986). How this may apply to larch species has received limited study.

In one previous study, Khasa and colleagues (2006) examined spatial patterns of genetic variation in 19 populations of subalpine larch from the Canadian portion of the species' range. Individuals were genotyped using seven microsatellite markers. Two genetically divergent clades were identified based on a dendrogram of chord distances: one comprising populations from the Cascades and south-central BC and the other composed of populations from the Rocky Mountains. A genetically distinct sub-cluster was also identified in the southeastern Rockies. Estimates of genetic diversity per locus within populations were lower in subalpine larch than western larch (*Larix occidentalis*), a closely related sister species with a broader distribution at lower elevation, while estimates of genetic differentiation between populations of subalpine larch were higher. In addition, seven of nineteen subalpine larch populations deviated significantly from mutation-drift equilibrium (under the assumptions of the infinite allele mutation model) whereas no western larch populations deviated. This led the authors to suggest that patterns of variation in subalpine larch are the result of founder effects during post-glacial expansion and/or ongoing reproductive isolation and genetic drift. However, this study was limited by two factors: the relatively small geographic scope of the study area and the relatively small number of markers used to make inferences.

Marker choice affects the quality and quantity of information obtained. Microsatellite markers are popular for phylogeographic analysis because they are neutral and highly polymorphic but they are also costly and difficult to develop, which limits their availability. Furthermore, they are difficult to score, which limits their utility.

Interpretations of electropherogram peaks are somewhat subjective and neither homoplasy resulting from insertions or deletions in flanking regions nor point mutations can be detected. Next-generation sequencing (NGS) technologies appear to provide an exciting alternative for assessing genomic variation in non-model organisms. High-throughput sequencing allows large numbers of loci and individuals to be genotyped. Large genomes can be subsampled using reduced representation methods that rely on restriction enzymes, such as restriction enzyme associated DNA sequencing (RAD-seq; Miller et al. 2007; Baird et al. 2008; Davey et al. 2011). In a one-to-one comparison, single nucleotide polymorphism (SNP) markers are less informative than microsatellites but they do provide high statistical power in large numbers. A study of parentage in the black-throated blue warbler found that 40 – 97 SNPs were as powerful as six microsatellites for assigning paternity (Kaiser et al. 2017). A phylogeographic analysis of European carp found that SNP data for 18 % of the samples included in the full microsatellite dataset produced robust phylogeographic inferences that emphasized fine-scale structure among populations (Jeffries et al. 2016).

In this study, 61 populations of subalpine larch were sampled to assess range-wide genetic structure. NGS methods were used to generate hundreds of SNPs. Spatial patterns of genetic variation were analyzed and three genetically unique geographic regions were identified. These regions should be prioritized for future management and conservation efforts.

Methods

Sampling

Foliage was collected from two populations of western larch and 44 populations of subalpine larch distributed across the species range (Figure 1; Table 1). Populations of western larch were sampled to provide an outgroup for phylogeographic analysis. An additional 18 populations of subalpine larch were sampled from clones grafted *ex situ* at the BC Ministry of Forests, Lands, Natural Resource Operations and Rural Development (FLNRORD) Kalamalka Forestry Centre in Vernon, BC (Figure 1; Table 2).

For each population sampled *in situ*, foliage was collected from approximately 30 individuals at a minimum inter-tree sampling distance of 50 m. Distances were measured using the "find waypoint" function on a handheld Garmin eTrex20 GPS unit (Olathe, KS, USA). This sampling strategy was used to obtain representative genetic samples and to avoid sampling spatially aggregated relatives. Although mean minimum inter-tree distance measured in the field was 71 m, it was not always possible to sample 30 trees at this distance (Table 1). Either the population was too small or the trees were too difficult to access due to the steepness of the terrain. For example, mean inter-tree sampling distance for Holland Pass, MT, was only 41 m. Because this population was so small, the minimum inter-tree sampling distance was reduced to 20 m in order to collect more individuals. Three additional populations (Skylark Lake, MT; Preston Park, MT; Indian Head, MT) had more than two individuals sampled at an inter-tree distance of less than 40 m.

Orchard populations had fewer individuals (median = 10) than populations sampled in the field (median = 30). Orchard populations had diminished since the

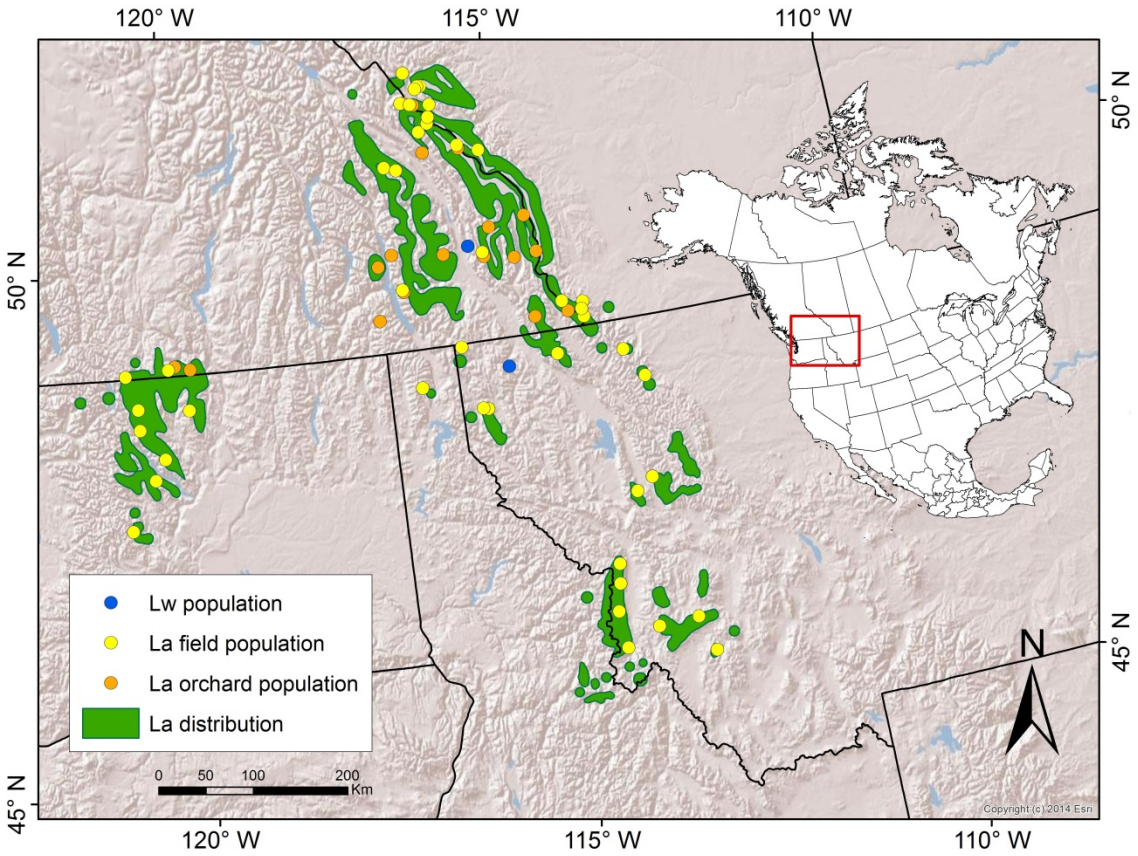

Figure 1. Two populations of western larch (Lw) and 44 populations of subalpine larch (La) were sampled in the field. An additional 18 populations of subalpine larch were sampled from clones grafted at the Kalamalka Forestry Centre in Vernon, BC.

Table 1. Foliage samples were collected from 44 populations of subalpine larch and two populations of western larch in the field; a subset of individuals from each population were selected for sequencing (N).

Pop.	Location	Latitude (ddº)	Longitude (ddº)	Elevation (m)	No. Trees	Av. Min. Measured Dist. (m)	Av. Min. Calculated Dist. (m)	N
01	Frosty Mountain, E.C. Manning Provincial Park, BC	49.0265	-120.8275	2038	33	54	40	5
02	Quiniscoe Lake, Cathedral Lakes Provincial Park, BC	49.0632	-120.2020	2098	30	66	61	5
03	Tiffany Mountain, WA	48.6632	-119.9421	2247	30	69	52	5
04	Hart's Pass, WA	48.7037	-120.6762	2000	30	67	51	5
05	Blue Lake, WA	48.5082	-120.6705	1912	30	75	53	5
06	Upper Eagle Lake, WA	48.2131	-120.3413	2161	30	77	78	5
07	Big Hill, WA	48.0183	-120.5016	2016	30	66	63	5
08	Windy Pass, WA	47.5494	-120.8689	2061	33	114	58	5
09	Carlton Ridge Research Natural Area, MT	46.6921	-114.2067	2483	30	64	65	5
10	McCalla Lake, MT	46.5041	-114.2433	2477	30	61	44	5
11	Canyon Lake, MT	46.2442	-114.3264	2229	30	72	43	5
12	Mosquito Meadows, MT	46.0410	-113.8127	2438	30	64	67	5
13	Trapper Peak, MT	45.8859	-114.2824	2804	30	70	54	5
14	Stine Mountain (Grouse Lakes), MT	45.7213	-113.1007	2734	30	69	47	5
15	Storm Lake, MT	46.0670	-113.2661	2536	30	80	43	5
16	Skylark Lake, MT	47.3522	-113.7968	2046	15	70	43	5
17	Holland Pass, MT	47.4655	-113.5552	2287	28	41	30	5
18	Northwest Peak, MT	48.9708	-115.9457	1929	33	94	49	5
19	Roman Nose, ID	48.6345	-116.5804	1925	30	73	52	5
20	Gray Creek Pass, BC	49.5849	-116.6859	2020	30	71	61	5
21	Sparkle Lake, Top of the World Provincial Park, BC	49.8386	-115.4481	2027	30	84	40	5

Pop.	Location	Latitude (ddº)	Longitude (ddº)	Elevation (m)	No. Trees	Av. Min. Measured Dist. (m)	Av. Min. Calculated Dist. (m)	N
22	Walter Lake, Bugaboo Provincial Park, BC	50.7638	-116.7241	2108	30	63	42	5
23	Tiger Pass, BC	50.7258	-116.5514	2187	30	62	51	5
24	Floe Lake, Kootenay National Park, BC	51.0569	-116.1376	2076	32	59	53	5
25	Tower Lake, Banff National Park, AB	51.3030	-115.9164	2115	30	69	61	5
26	Molar Pass, Banff National Park, AB	51.6327	-116.2538	2312	23	183	74	5
27	Lake O'Hara, Yoho National Park, B.C.	51.3491	-116.3481	2165	30	59	63	5
28	Larch Valley, Banff National Park, AB	51.3277	-116.2105	2294	30	66	75	5
29	Baker Lake, Banff National Park, AB	51.4897	-116.0193	2168	30	70	45	5
30	Boulder Pass, Banff National Park, AB	51.4900	-116.0661	2334	30	71	44	5
31	Halfway Hut, Banff National Park, AB	51.4662	-116.1020	2150	30	85	41	5
32	Ball Pass, Banff National Park, AB	51.1277	-115.9890	2132	31	58	46	5
33	Gibbon Pass, Banff National Park, AB	51.1866	-115.9636	2286	30	57	46	5
34	Wonder Pass, Banff National Park, AB	50.8833	-115.5891	2273	31	75	59	5
35	Chester Lake, Peter Lougheed Provincial Park, AB	50.8088	-115.2844	2187	30	81	65	3
36	Middle Kootenay Pass, AB	49.2648	-114.4089	1939	30	78	54	5
37	Bovin Lake, AB	49.2300	-114.1120	1978	30	62	75	5
38	Lone Lake, Waterton Lakes National Park, AB	49.0859	-114.1261	2139	30	58	43	5

Pop.	Location	Latitude (ddº)	Longitude (ddº)	Elevation (m)	No. Trees	Av. Min. Measured Dist. (m)	Av. Min. Calculated Dist. (m)	N
39	Avion Ridge Trail, Waterton Lakes National Park, AB	49.1602	-114.1425	2136	30	58	45	5
40	Preston Park, Glacier National Park, MT	48.7128	-113.6510	2161	19	57	32	3
41	Paradise Park, Glacier National Park, MT	48.4302	-113.4061	2024	31	65	46	5
42	Lake Mountain, MT	48.7757	-114.5888	2271	30	71	46	5
43	Indianhead Mountain, Cabinet Range, MT	48.3527	-115.6888	2119	32	86	49	5
44	Dome Mountain, Cabinet Range, MT	48.3662	-115.7561	2182	30	64	53	5
45*	Premier Lake Provincial Campground, BC	49.91540	-115.6466	927	10	127	84	5
46*	Peck Gulch Campground, MT	48.72572	-115.3053	797	10	77	50	5
				Tot:	**1321**	**Av: 73**	**Av: 53**	**226**

*Western larch populations sampled as an outgroup for phylogeographic analysis

Table 2. Foliage was collected from 18 populations of subalpine larch grafted *ex situ* at the BC Ministry of FLNRORD Kalamalka Forestry Centre, Vernon, BC.

Pop.	Location	Latitude (decimal degrees)	Longitude (decimal degrees)	Elevation (m)	No. Trees	No. Sequenced
AL01	Baldy Mountain, BC	49.32	-117.07	1981	16	5
AL02	Burdett Peak-Gray Pass, BC	49.57	-116.67	2134	13	4
AL03	Mount Kaslo, BC	49.93	-116.78	2149	8	3
AL04	Fletcher Creek, BC	49.83	-116.99	2012	12	5
AL05	Inverted Ridge, BC	49.16	-114.83	2164	11	3
AL06	Sunkist Mountain, BC	49.16	-114.34	2210	5	2
AL07	Racehorse Pass, BC	49.77	-114.66	2210	11	2
AL08	Mount Kuleski, BC	49.75	-115.00	2179	10	2
AL09	Mount Dingley, BC	49.80	-115.43	2195	7	3
AL10	Mount Gass, BC	50.13	-114.76	2256	8	2
AL11	Mount Mike-Quinn Range, BC	50.07	-115.30	2377	12	4
AL12	Luxor Pass-Mount Crook, BC	50.86	-116.12	2164	21	5
AL13	Mount Assiniboine, BC	50.85	-115.58	2210	11	5
AL14	Mount Bradford, BC	49.87	-116.02	2454	14	3
AL15	Twin Buttes, BC	49.09	-120.11	2270	8	0
AL16	Lake O'Hara, BC	51.35	-116.32	2200	8	4
AL17	Moraine Lake, AB	51.32	-116.17	2200	4	4
AL20	Cathedral Lake Provincial Park, BC	49.05	-119.88	2200	9	3
				Total:	**188**	**59**

original cuttings collected by Mr. Barry Jaquish were grafted to western larch rootstock in 1996 (described in Khasa et al. 2006). Mortality also continued to occur in subsequent years due to unknown causes. Orchard populations therefore represent a subset of survivors from an original random sample.

Green foliage was collected into paper coin envelopes and stored in plastic bags containing silica gel. Silica gel was replaced regularly until samples were completely dry, at which point it was removed. Dry samples were sealed in plastic and stored at ambient temperature until the end of the field season. Samples were then moved into -20 ºC freezers at the University of Victoria. Note that liquid nitrogen is the preferred method for DNA preservation but was not feasible to use over a relatively long field season (> 2 months) that involved traveling between remote wilderness locations in the Cascade Range and Rocky Mountains.

Dried tissue totalling several hundred milligrams was collected from each of 1,489 subalpine larch trees representing 62 populations, as well as 20 western larch trees representing two populations, for a total of 1,509 trees. Latitude, longitude, elevation, aspect, diameter at breast height (1.3 m) and growth-form data were recorded for 1,321 trees sampled in the field. These data were not available for the clones grafted in the orchard.

Coordinate data for individual trees were loaded into the R statistical environment (R Core Team 2017) and distance matrices were calculated between individuals within populations. Flat distances were calculated using the *geodDist* function in the R oce package (Kelley and Richards 2018), which accounts for the ellipsoidal curvature of the earth. Differences in elevation were incorporated using the Pythagorean theorem. Based

on coordinate data, the average minimum inter-tree sampling distance was 53 m (Table 1). Two populations had mean inter-tree sampling distances of less than 40 m (Holland Pass, MT; Preston Park, MT). Several factors may explain why calculated distances were less than those observed in the field. First, calculated distances are straight lines and do not account for landscape topography. Second, calculated distances exacerbate error. Tests showed that distances read from the "find waypoint" function were accurate within -0.65 ± 5 m but distances calculated from coordinate data underestimated actual distance by -7 ± 22 m. Therefore calculated distance matrices generally represent conservative estimates of inter-tree distance but provide important positional information for all trees within a particular population (i.e. not just those sampled in sequence).

Molecular Techniques

High molecular weight DNA was extracted from silica-dried larch foliage. Protocol optimization was required to ensure that DNA extractions were pure and free of contaminant nucleic acids such as RNA and plastid/bacterial DNA. Extractions were carried out using the PL2 Extraction Protocol in the NucleoSpin 96 Plant II Core Kit (Machery-Nagel, ON, Canada) with the following modifications:

1) Dry tissue instead of fresh
2) Tissue weight increased to 30 ± 2 mg per sample
3) 1-2 minutes of bead beating at 30 beats per second on Retsch MM 400
4) 2 % PVPP added to the lysis buffer
5) 5 µL of 2-mercaptoethanol added to the lysis buffer immediately prior to use
6) Lysis incubation extended to 45 min.
7) RNase incubation for 30 min. at 37 ºC after SDS precipitation
8) Extra PW1 wash

9) Centrifuge extended to 2 min. following the second PW2 wash
10) Final wash using 400 µL of 100 % ethanol followed by a 10 min. centrifuge
11) Two elution steps with 50 µL of PE buffer were used in order to increase the concentration of the final product

This modified protocol was used to extract DNA from all individuals. Extractions were carried out in Dr. Sally Aitken's laboratory at the University of British Columbia using the EpMotion automated pipetting system, which allowed for two 96-well plates to be extracted simultaneously.

Extraction quality was assessed using a Nanodrop 8000 Spectrophotometer (Thermo Scientific, Wilmington, USA) and a Qubit 2.0 Fluorometer (Life Technologies, Burlington, Canada). DNA quantification was carried out using fluorometry because spectrophotometry is highly sensitive to sample purity and will often overestimate or underestimate DNA concentration. DNA was initially considered good quality if it had a Nanodrop 260/280 reading between 1.7 and 2.0, a Nanodrop 260/230 reading between 2.0 and 3.0, and a Qubit DNA concentration of at least 20 ng/µL. Based on these quality criteria (QC), 64% of samples were considered successfully extracted.

Range-wide data were obtained by sequencing 285 individuals representing 61 populations of subalpine larch and two populations of western larch. Note that a single orchard population (AL15) was excluded from sequencing because no samples passed initial QC. As a general rule, only samples that met initial QC were considered. However three orchard samples with slightly lower Nanodrop 260/230 values (1.76 – 1.92) were included in order to test this QC criterion. A median of five individuals from each of the remaining 61 populations of subalpine larch and the two populations of western larch were included (Table 1; Table 2). For orchard clones, up to five individuals were

randomly selected from each population. For field-collected samples, a custom R script was used to identify the five most spatially separated individuals within each population based on calculated distance matrices (Figure 2). Selected trees had an average inter-tree distance of 308 m between nearest neighbors and a minimum inter-tree distance of 75 m. DNA from selected individuals was normalized to a final concentration of 20 ng/µL and sent to Floragenex for library preparation (Floragenex, Inc., Eugene, USA).

Sequence data were generated using restriction enzyme associated DNA sequencing (RAD-seq), a reduced representation method that generates short reads (100 bp) adjacent to restriction endonuclease (RE) recognition sites distributed throughout the genome. Three multiplexed RAD libraries (C446, C447 and C448) were prepared using standard protocols (Figure 3). Briefly, genomic DNA was digested using the eight-base Sbf1 restriction enzyme (target sequence 5’ CCTGCAGG 3’). P1 adaptors were ligated onto the resultant sticky overhang sequences, also known as RAD tags. Note that P1 adaptors included ten-nucleotide barcode sequences that differed by at least four nucleotides from all other barcodes. Barcoded individuals were pooled into three multiplexed libraries containing 95 individuals each. A yeast control (*Saccharomyces bayanus* Saccardo, teleomorph 504.3) was added to each library for quality control purposes. DNA was sheared by sonication. Fragments between 300 and 500 base pairs were size-selected to ensure that DNA was a suitable length for PCR amplification and sequencing. Following DNA end repair, Illumina P2 sequencing adaptors were ligated onto fragments and DNA was amplified via selective PCR. Only fragments with both

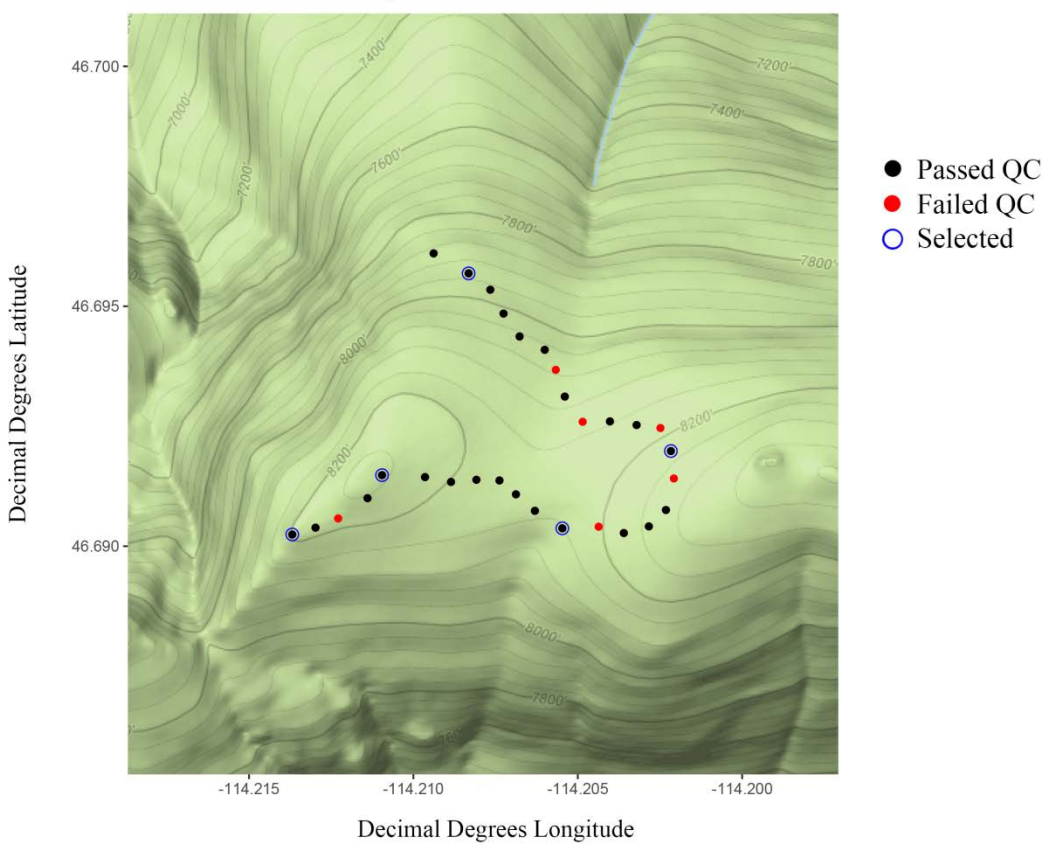

Figure 2. Example of selecting the five most spatially separated individuals among successful DNA extractions for subalpine larch samples collected from Carlton Ridge Research Natural Area, MT.

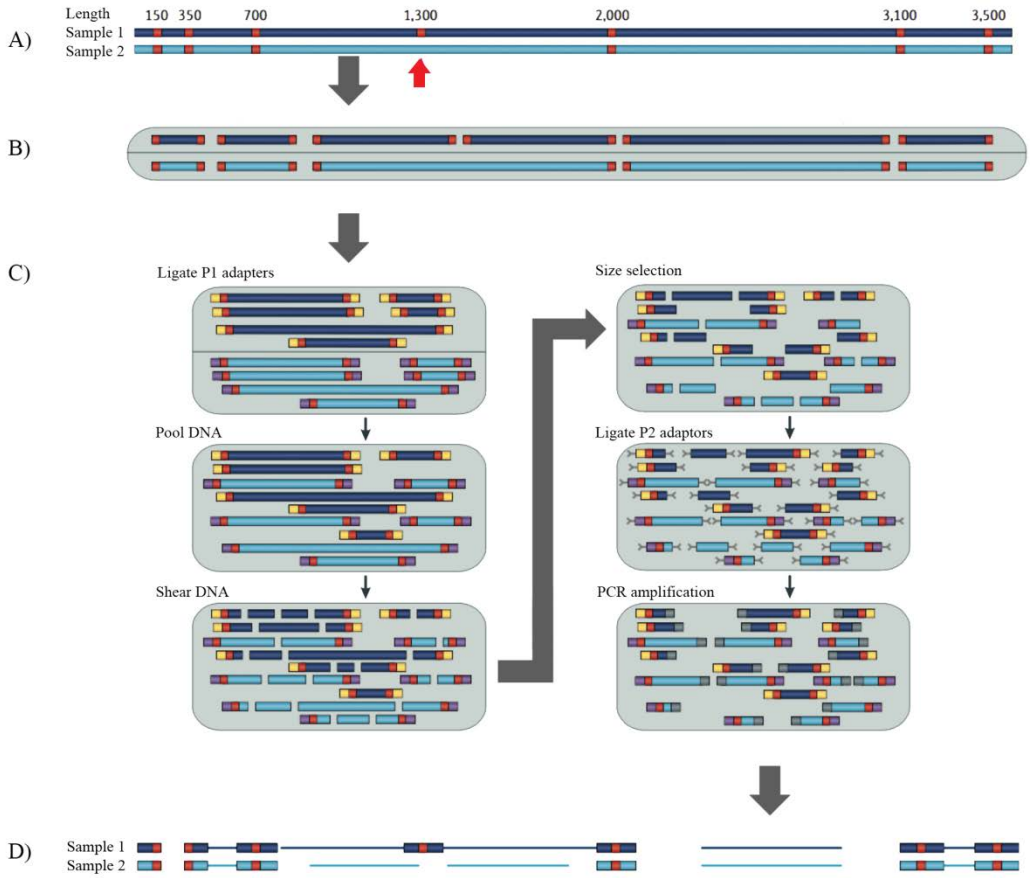

Figure 3. (A) DNA to be sequenced from two individuals (dark blue and light blue). Restriction endonuclease (RE) recognition sites in this genomic region are illustrated in red. Sample 2 has a variation in the cut site at 1,300 bases (red arrow) and so this site will not be cut. (B) **RE digestion of DNA. (**C**) Barcoded P1 adapters (yellow and purple) are ligated to the sticky overhangs left behind by digestion. Barcoded fragments are pooled, randomly sheared and size selected. P2 adaptors with divergent ends are ligated to the fragments with and without P1 adaptors. Fragments are amplified using P1- and P2-specific primers. The P2 adaptor is completed when fragments containing P1 adaptors are copied. The P2 primer only binds to completed P2 adaptors. Only fragments with P1 and P2 adaptors (i.e. fragments containing restriction sites) are amplified via PCR. (**D**) Sequenced markers are aligned to a reference genome. Thin lines indicate the region that would be covered by paired-end sequencing. Modified from Davey et al. 2011.**

adaptors—i.e. fragments with RAD tags—were amplified during selective PCR. After purification, Agilent Bioanalyzer traces were used to validate library integrity and quantitative PCR was used to estimate optimal Illumina sequencing plating concentrations.

Libraries were sequenced on an Illumina platform at the University of Oregon's Genomics and Cell Characterization Core Facility. RAD libraries typically have low complexity because sequences contain a common RAD tag at the same position. To prevent this from negatively affecting the accuracy of the Illumina base-calling software, 10% bacteriophage PhiX DNA was added to each sequencing lane to provide nucleotide diversity and to allow the Illumina software to call bases with greater confidence. In total, 285 individuals representing 61 populations of subalpine larch and two populations of western larch were sequenced on three lanes of Illumina HiSeq 2000.

Bioinformatics

Bioinformatics processing was carried out to clean and de-multiplex reads (Table 3). Overall library quality was assessed using FastQC (Andrews 2010). An average of 152 million reads was obtained per lane of sequencing for a total of over 450 million DNA sequences. Short-read sequences were 101 nucleotides in length. Per-base sequence quality plots confirmed that sequence data from all three libraries were generally high quality. Phred quality score is a measure of confidence that is logarithmically linked to the error probability of a given base call: Phred 10 means a base was called with 90% confidence and Phred 20 means a base was called with 99% confidence. Quality dropped below a mean Phred quality score of 20 at nucleotides 76, 98 and 76 in the three libraries,

Table 3. Reads lost and kept over successive stages of bioinformatics processing in three libraries generated using restriction site associated DNA sequencing (RAD-seq) with the Sbf1 enzyme.

Software	Process	Library C446	Library C447	Library C448	Average No. Reads	% Reads Lost	Av. Reads Retained	Av. % Reads Retained
FastQC	Count raw reads	149,804,952	151,303,463	155,435,946	152,181,454	0	152,181,454	100
Cutadapt	Trim adaptors	17,059,980	14,282,417	21,723,601	17,688,666	0	152,181,454	100
	Remove if < 50 bp	2,408,561	1,675,928	3,601,890	2,562,126	1.7	149,619,327	98.3
STACKS process_radtags	Remove low quality reads	19,360,834	16,213,820	22,515,467	19,363,374	12.7	130,255,954	85.6
	Remove reads with ambiguous barcode	2,889,372	2,864,835	2,779,999	2,844,735	1.9	127,411,218	83.7
	Remove reads with ambiguous RAD tag	73,818,847	80,431,249	70,164,899	74,804,998	49.2	52,606,220	34.6
	Remove yeast control reads	16,856,090	6,517,514	6,765,259	10,046,288	6.6	42,559,932	28.0
NextGenMap	Remove reads that did not align to reference genome	738,926	805,935	1,210,817	918,559	0.6	41,641,373	27.4
UnifiedGenotyper	Remove reads with mapping quality score = 0	9,627,033	12,443,808	14,074,683	12,048,508	7.9	29,592,865	19.4
UnifiedGenotyper	Remove reads with mapping quality score < 20	4,094,444	5,373,395	6,888,632	5,452,157	3.6	24,140,708	15.9

respectively. Decreases in quality were expected at the end of reads due to the accumulation of sequencing errors, which increased the ambiguity of the fluorescence signal produced by clusters on the flowcell. This drop in quality reflected a subset of sequences. However, mean quality remained above Phred 30 until the end of the reads in all three libraries.

Reads were trimmed to remove P1 and P2 adapter sequences using Cutadapt 1.10 with an error tolerance of 0.20 (Martin 2011). P1 adaptors can be sequenced on the 3' end of short reads if an adaptor has ligated to a nearby restriction enzyme cut site and no shearing occurred between sites. P2 adapters can be sequenced on the 3' end of short reads if sequences are less than 101 nucleotides in length (due to degradation or shearing). P2 adaptors should be rare because DNA sequences were size-selected for a minimum of 300 bases during library preparation. Adaptor sequences were identified and trimmed off the 3' end of 11.6% of reads. Cutadapt was also used to remove reads shorter than 50 nucleotides (1.7% of reads).

After adapter sequences were removed, sequences were trimmed, de-multiplexed and filtered using the *process_radtags* function in STACKS v 1.44 (Catchen et al. 2011; Catchen et al. 2013). After experimenting with different trim lengths (–t), reads were truncated to 91 bases. This trim length was chosen because the retention of additional bases led to 100% read loss during filtering. To remove low-quality reads, the –q option used a sliding-window approach to discard reads if average quality within a window dropped below Phred 10. On average, 13% of reads were discarded due to low quality, 2% of reads were discarded due to the presence of an ambiguous barcode sequence and 49% of reads were discarded due to the presence of an ambiguous RAD-tag sequence.

Note that the default rescue function (-r) corrects a single nucleotide error within the RAD tag and the inter-barcode distance (--barcode_dist_1) was set to four so that reads with fewer than four barcode sequencing errors could be rescued. After de-multiplexing and filtering, an average of 37,053,157 reads were retained per library—approximately 18% of the original data. Individuals retained a median of 427,849 reads each.

A draft version of the Siberian larch genome (*Larix sibirica*) was obtained in August 2017 from Dr. Konstantin Krutovsky, University of Göettingen, Germany (Kuzmin et al. 2019). The assembly was sparse and fragmented. The scaffold version of the draft genome contained 10,260,722 scaffolds and approximately 11.99 Gb of data, which was close to the expected genome size of 12.03 Gb (min. length: 201; max. length: 265,649; N50: 651,335). Approximately 8.00 Gb was derived from contiguous sequence (contig) data. It was assumed that the gaps in the scaffolds were largely attributable to highly repetitive regions. A search for the Sbf1 recognition site identified 43,003 potential cut sites. Note that the true number of restriction enzyme recognition sites was likely higher due to the 3.99 Gb of data missing from the scaffold assembly.

Because the larch genome was too large for the bowtie2 alignment program (max 3 Gb), fastq files were aligned to the Siberian larch genome using NextGenMap v. 0.5.5 (Sedlazeck et al. 2013), a flexible and highly sensitive short-read mapping tool. NextGenMap split the reference genome into overlapping Kmers against which Kmers extracted from short reads were compared to identify putatively matching genomic regions. If the number of matching Kmers passed a calculated threshold, regions were considered as candidates for mapping. Alignment scores were computed and the full alignment was reported for the candidate region(s) with the highest score. Note that

NextGenMapper only accepted 1,000 reference scaffolds. A custom Perl script written by Dr. Sam Yeaman at the University of Calgary was used to stitch together pseudoscaffolds using runs of 30 adenine nucleotides between scaffolds.

NextGenMap alignment was carried out separately for each individual. Output was converted from Sequence Alignment Map text format (SAM) to binary format (BAM) using Samtools v. 1.3.1 (Li et al. 2009). Samtools was also used to sort and index BAM files. Overall, 97.8% of retained larch sequences were successfully aligned to the Siberian larch draft genome. Unfortunately, 29% of aligned reads had a map quality score of zero, meaning they mapped to more than one location in the draft genome. This was likely due to the high proportion of repeat DNA present in conifer genomes. Another 13% of reads had scores below Phred 20 (i.e. below a 99% probability of correct alignment), and were also removed. After alignment and filtering based on mapping quality, individuals retained a median of 117,116 reads each.

Genotyping

PhiX control sequences were used to carry out base quality score recalibration (BQSR). BQSR methods model systemic base quality errors and adjust quality scores accordingly, providing more accurate base qualities overall. The Illumina control Enterobacteria phage phiX174 genome was uploaded from NCBI (NC_001422.1). Raw reads were aligned using NextGenMapper. PhiX accounted for an average of 10.23% of the total reads in the Sbf1 libraries. Picard v. 1.57 was used to generate a sequence dictionary for the PhiX reference genome (CreateSequenceDictionary.jar). GATK v. 3.6 HaplotypeCaller was used to call variable SNPs in each of the three libraries separately.

Five variant sites were identified. Sites of expected variation were masked during BQSR to avoid counting real variants as errors. BQSR was carried out for PhiX reads using the GATK BaseRecalibrator function. Base recalibration was rerun on the resulting output file, and the two files were compared using the GATK AnalyzeCovariates function. Graphical output showed that BQSR improved base quality across PhiX reads. BQSR was applied to larch bam files using the GATK PrintReads function.

Final larch genotypes were called with GATK UnifiedGenotyper function, which used a Bayesian likelihood model to estimate the most likely genotypes for a given locus across all individuals simultaneously. As per the default settings, a minimum base quality score of Phred 17 was required for a base call to contribute to a genotype and a minimum Phred score of 10 (90% confidence) was required to call genotypes. Both variant and reference calls were output together in variant call format (VCF). VCF files were sorted using Picard SortVCF. Sites with non-reference alternate alleles were subset using the GATK SelectVariants function. SNPs were subset in VCFtools (Danecek et al. 2011) by selecting for loci with a minor allele count of at least one (--mac 1). In total over 21 million nucleotides were genotyped, including 173,843 SNPs (0.8%).

Low quality SNPs were removed by filtering (Table 4). Because data were generated from a non-model organism, GATK variant quality score recalibration (VQSR) was not an option. Therefore hard filters were applied as recommended by the Broad Institute (Cambridge, MA, USA): QualByDepth (QD < 2.0), FisherStrand (FS > 60.0), StrandOddsRatio (SOR > 3.0), RMSMappingQuality (MQ < 40.0) and ReadPosRankSumTest (ReadPosRankSum < -8.0). QD was the variant confidence

Table 4. Filtering procedure for SNPs generated using restriction site associated DNA sequencing (RAD-seq) for 274 subalpine larch and ten western larch individuals.

Software	Function	Value	Filter	Process	No. Sites	% Kept
Picard	SortVcf	NA	NA	Sort VCF file	21,110,868	.
Gatk	SelectVariants	NA	sites	Remove invariant sites	857,840	.
VCFtools	mac	1	sites	Remove sites with < 1 copy of minor allele	173,843	100
Gatk	QD	< 2.0	sites	Normalize quality by depth and filter low quality	.	.
	FS	> 60.0	sites	Remove if strand bias present	.	.
	SOR	> 3.0	sites	Remove if strand bias present	.	.
	MQ	< 40.0	sites	Remove if mapping quality is low	.	.
	ReadPosRankSum	< -8.0	sites	Remove if ref/alt alleles map differently		
VCFtools	remove-filtered-all	NA	sites	Remove sites that did not pass GatK filters	62,816	36.1
	minGQ	< 20	genotypes	Remove genotypes called with < 99% confidence	.	.
	minDP	< 3	genotypes	Remove genotypes with < 3 reads	.	.
	exclude	Pop32_944	individuals	Remove individuals	62,816	36.1
	minQ	< 20	sites	Remove SNP called with < 99% certainty	61,783	35.5
	max-missing	> 0.50	sites	Remove SNPs absent in > 50% of individuals	9,177	5.3
BCFtools	filter -i 'AVG(FMT/DP)'	72.37494	sites	Remove sites with depth > (mean + 1.5*IQR)	8,073	4.6
VCFtools	maf	< 0.05	sites	Remove SNPs with minor allele frequency < 0.05	1,425	0.82
	min- & max-alleles	< 2 >	sites	Remove SNPs that are not biallelic	1,419	0.82
Plink	r2	< 0.50	sites	Remove SNPs in LD	1,077	0.62
VCFtools	thin	< 100	sites	Retain one SNP per 100 bases	809	0.47

divided by the unfiltered depth of non-homologous references samples. This filter took quality and depth into consideration, normalizing variant quality to avoid inflation associated with deep coverage. FS was the Phred-scale probability that there was strand bias at the site. Strand bias indicated whether the alternate allele was seen more or less often than the reference allele on the forward or reverse strand. SOR was another way to estimate strand bias using a test similar to the symmetric odds ratio test. MQ was the root mean square mapping quality over all the reads at the site. ReadPosRankSum compared the positions of the reference and alternate alleles to determine if they differed within reads. Note that hard filters cannot reliably separate true from false positives when coverage is shallow (< 10X).

Additional filtering was carried out in VCFtools. First, genotypes were filtered for minimum genotype quality of Phred 20 (--minGQ 20) and minimum depth of 3 reads (--minDP 3). Three reads per genotype is low but possible because GATK can confidently call genotypes with few reads when it assesses variants across all samples simultaneously. One individual from Ball Pass, AB, was excluded because it was found to have low depth and a high proportion of missing data (Pop32_944). SNPs were filtered for minimum quality of Phred 20 (--minQ 20) and a maximum of 50% missing data across individuals (--max-missing 0.50). SNPs with a mean depth greater than 1.5X the upper interquartile range were filtered using BCFtools in order to remove paralogs that may have mapped to a single site in the genome. SNPs with a minimum allele frequency less than 5% were filtered in VCFtools to remove low-frequency variants that may represent PCR or sequencing errors (--maf 0.05). Biallelic SNPs (--min-alleles 2, --max-alleles 2) were output in Plink format. Linkage disequilibrium (r^2) was calculated

between pairs of SNPs (--r2) in Plink. If a pair of SNPs had r^2 greater than 0.5, one SNP was removed until all pairs with high LD were broken up. This was a conservative filtering approach because SNPs found to be in statistical linkage across pseudoscaffolds were removed. This approach was preferred because pseudoscaffolds contain contigs that may or may not sit alongside each other in the genome. Finally, only a single SNP was retained per 100 bp (--thin 100), equivalent to one SNP per short-read sequence. After filtering, 809 SNPs were retained for further analysis. The final VCF file was converted into the file formats necessary for the analyses below using PGDspider v. 2.0.9.0 (Lischer and Excoffier 2012).

Data Analysis

Selectively neutral SNPs were identified using BayeScan v. 2.1 (Foll and Gaggiotti 2008). BayeScan uses Bayesian methods to assess differences in allele frequencies among populations and estimate population-specific F_{ST} coefficients. After running 100,000 iterations, SNPs with a posterior probability >0.95 were considered as outliers. In order to minimize false positives a prior odd of 10,000 was used, meaning the neutral model was 10,000 times more likely than the model with selection (Lotterhos and Whitlock 2014). F_{ST} outliers were removed using VCFtools, leaving 751 neutral SNPs for further analysis.

SNP data were loaded into R and missing genotypes were imputed using the *tab* function in the adegenet package with NA.method set to “mean” (Jombart 2008; Jombart and Ahmed 2011). A principal components analysis (PCA) was carried out using the

dudi.pca function in the ade4 package (Dray and Dufour 2007; Bougeard and Dray 2018).

Coordinate data were used to test for range-wide multivariate spatial structure. Because individual coordinate data were not available for the trees grafted *ex situ* at the Kalamalka Forestry Centre, orchard trees were assigned population-level coordinates. Western larch samples were removed along with a single individual from Indian Head, MT, which was found to cluster with the western larch populations (Pop43_1262). Both PCs and geographic distances were converted into Euclidian distances using the *dist* function and the correlation between genetic distance and geographic distance was tested for significance using the *mantel.randtest* function with 999 permutations.

Spatial PCA (sPCA) was carried out to determine if genetic clusters corresponded to geographic regions. Western larch samples were removed and genotype data were converted into allele counts by population using the *genind2genpop* function. Note that missing data were not imputed for this analysis. The sPCA analysis was carried out using the *spca* function with graph "type" set to 5, a minimum distance between any two neighbours of 1, and a maximum distance between any two neighbours of 2. The type 5 connection network defines neighbouring entities based on pairwise geographic distance. Both global and local spatial structures were tested for significance using Monte Carlo simulations in the *global.rtest* and *local.rtest* functions, respectively, with 999 permutations each. Principal components were interpolated using lagged principal scores in order to obtain a map of genetic clines. This was done using the *interp* function in the lazyeval package (Wickham 2017).

Individuals were assigned to genetic clusters using a discriminant analysis of principal components (DAPC), which generates synthetic variables (discriminant functions) that are used to assign individuals to genetic clusters with a given probability. First, the *find.clusters* function, a K-means clustering algorithm, was used to identify the optimum number of clusters based on the Bayesian Information Criterion (BIC). With 250 PCs included, BIC showed that four clusters should be retained. The *dapc* function was then used to assign individuals to these four groups. To avoid overfitting the model, 100 simulations of a DAPC model with 250 PCs and three discriminant functions were tested using the *optim.a.score* function, which suggested retaining 67 PCs (Appendix A). The final DAPC model was fit with 67 PCs and three discriminant functions. This model accounted for 54.5% of the variation in the data set. Other models (30 PCs and 100 PCs) were tested and were found to produce the same results, even though they accounted for different amounts of the total genetic variation.

DAPC posterior probabilities were used to assign individuals to regions for an analysis of molecular variation (AMOVA). Genind data were converted into a hierfstat dataframe using the *genind2hierfstat* function in the hierfstat package (Goudet and Jombart 2015). Missing data were replaced with the most common allele for each locus and hierarchical F_{ST} tests were carried out using the *varcomp.glob* function. A model was fit using species (two levels = western and subalpine larch), region (four levels = western larch, Cascades, Southern Rockies, Northern Rockies) and population as hierarchical factors and the *test.g* function was run with 999 permutations to test the significance of the variance component attributed to species. This process was repeated on a dataset containing only subalpine larch samples. The *test.g* function was run with 999

permutations to test the significance of the variance component attributed to region and the *test.between* function was run with 999 permutations to test the significance of the variance component attributed to populations within regions.

Assignment to genetic clusters was reassessed using the Bayesian clustering method implemented in STRUCTURE v. 2.3.4 (Pritchard et al. 2000; Falush et al. 2003). Default STRUCTURE parameters were used with a burn-in of 10,000 iterations and 100,000 MCMC steps. K-values (2 – 10) defining the number of putative populations were tested five times each. Optimum K was chosen using the delta K method (Evanno et al. 2005).

Population-level allele frequency data were used to generate a dendrogram using Provesti's genetic distance, which allows for missing data. Dendrogram topology and bootstrapped confidence estimates were calculated using the *aboot* function in the poppr package with 1000 samples (Kamvar et al. 2014; Kamvar et al. 2015).

Results

Genotyping

High molecular weight genomic DNA (>10 kb) was successfully recovered from silica-dried foliage after optimization of the DNA extraction protocol (Figure 4). There was some concern that DNA degradation may have occurred during the field season because DNA preservation is generally poorer in silica gel than in liquid nitrogen. However the silica gel successfully preserved intact DNA.

RAD-seq libraries were generally of high quality but did not pass all FastQC quality tests. Some failures were expected. First, data did not pass per-base sequence content tests due to higher-than-expected frequencies of certain nucleotides between positions 11 and 16. This corresponded to the shared Sbf1 RAD tag (5’ TGCAGG 3’). Second, data did not pass sequence duplication tests. A relatively low percentage of data remained after de-duplication (10 – 11%). This was because RAD-seq targeted and amplified the same sequences within and among individuals, meaning biological duplication of the same sequences was to be expected. PCR duplicates may also have been present but cannot be distinguished from biological duplicates in a single-end sequencing dataset such as this one. Finally, data did not pass Kmer content tests. This failure was predictable because high-frequency Kmers are present across the first 16 nucleotides of each sequence, corresponding to the ten-nucleotide barcode and the six-nucleotide RAD tag sequences.

Additional issues with specific libraries were identified by FastQC. First, library C446 contained six overrepresented sequences associated with the Floragenex control

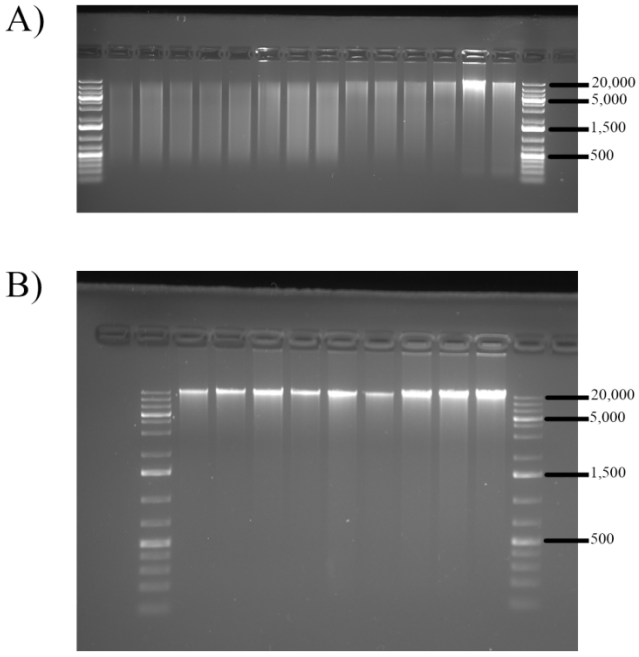

Figure 4. DNA extracted using (A) standard PL2 protocol and (B) optimized PL2 protocol visualized on 1.5% agarose gel. Note that (B) shows high molecular weight DNA (> 10 kb) in bright bands at the top of the gel.

generated from the *Saccharomyces bayanus* genome. FastQC estimated that these six sequences accounted for 0.66% of the total sequence data in this library. In total, control-derived sequences accounted for approximately 11% of the total reads in the C446 library. This was more than double the number of control sequences obtained from libraries C447 and C448, where control sequences accounted for approximately 4% of total reads. It was not clear why control sequences were so overrepresented in C446 but the simple explanation was that a higher concentration of yeast DNA was added during library preparation. Second, library C447 did not pass the FastQC per-tile sequence quality test. A subset of tiles on the flowcell showed very low quality around nucleotide position 18. A corresponding spike in uncalled bases (N content) was recorded at base 18 in the per-base N content plot. Conventional base calls were replaced with N values when the sequencer could not call a base with sufficient confidence. N values tend to appear when a library contains very biased sequence composition, confusing the base callers, but that should not be an issue at position 18, which was beyond the barcode and RAD tag sequences. The most common reason for N-calls is a general loss of quality. It seems most likely that this loss of quality was the result of a localized technical error that occurred during sequencing. Fortunately, this issue affected less than 10% of total sequences in library C447.

An average of 152 million raw sequence reads were obtained from each library (Table 3). After bioinformatics processing and alignment to a draft version of the Siberian larch genome, an average of 24 million reads were retained per library for calling genotypes. The largest read loss in all libraries (49%) occurred due to the presence of ambiguous RAD tags. It is not clear why such a large proportion of sequence

data had RAD-tag errors given that FastQC analysis showed high sequence quality between positions 11 and 16, the location of the restriction enzyme cut site overhang sequence. At least 9,000 different RAD tag sequences were observed in each library and, of these, only one sequence was correct (34% of reads). An additional 24 sequences contained a single error and could therefore be rescued (9% of reads), meaning 43% of reads passed the RAD-tag filter and 57% of reads did not. In practice only 49% of reads were discarded due to the presence of an ambiguous RAD tag, likely because some reads had already been discarded due to issues with read quality and/or errors in the barcode sequence. One known source of RAD-tag error was the presence of PhiX bacteriophage DNA sequences, which did not have barcodes or RAD tags. However PhiX only accounted for 10% of sequences per library, meaning the vast majority of RAD-tag errors could not be explained. In all three libraries the most common RAD-tag error was 'TGCACT', where the last two bases were incorrect (GG). This error accounted for approximately 2.5% of the total sequences in each library.

In total 21,110,868 nucleotides were genotyped, including 173,843 single nucleotide polymorphisms (0.8%). After filtering, 809 SNPs remained (Table 4). After removing loci that may be under selection, 751 SNPs were retained for further analysis. Individuals had an average of 545 SNPs. SNPs had a mean depth of 44 reads with 44% missing data across individuals. During SNP filtering, a large number of SNPs were removed due to a high proportion of missing data (> 50%). This was most likely attributable to under-sampling of the larch genome. When the number of SNPs per individual (prior to filtering) was plotted against the number of reads per individual after de-multiplexing, there was no saturation in the curve (Figure 5). More reads generated

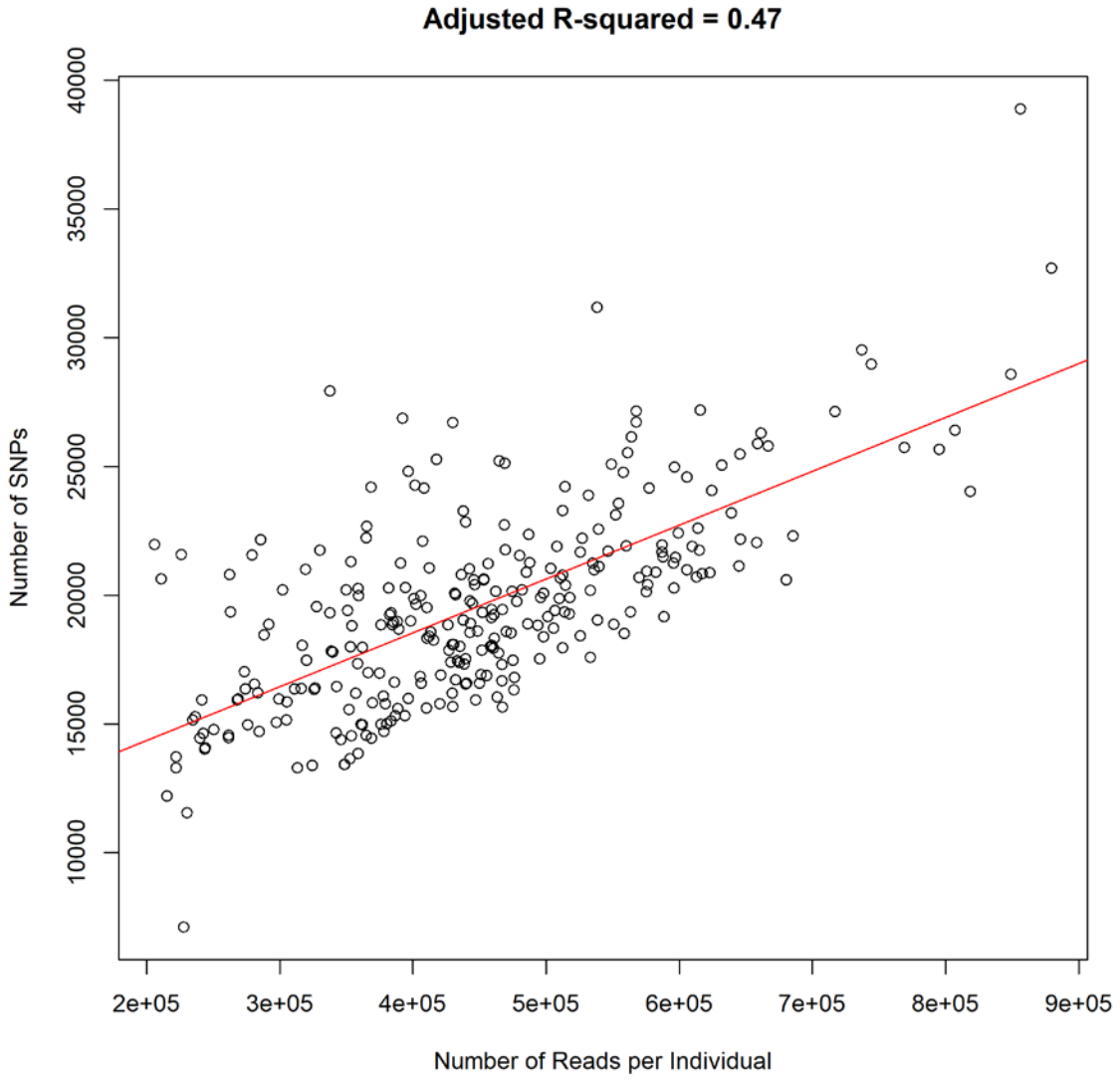

Figure 5. Number of reads per individual after de-multiplexing plotted against number of variant sites prior to filtering suggests that the *Larix lyallii* genome was undersampled, given that increased sequencing effort leads to a higher number of variant sites genotyped.

more SNPs, meaning increased sequencing effort per individual led to increased genotyping output. By sequencing a random portion of each individual's genome, many SNPs were not sequenced across individuals, and a high proportion of missing data was introduced into the data set.

Samples were not randomized across libraries. Significant differences in read depth, SNP count and missing data were detected among libraries. Individuals in libraries C447 and C448 had significantly higher read depths than individuals in library C446 ($p < 0.001$; $F = 2816$), with an average of 17 and 15 additional reads per SNP, respectively (Figure 6A). Individuals in library C448 had an average of 64 more SNPs ($p < 0.001$; $F = 47.48$; Figure 6B) and 8.6% less missing data ($p < 0.001$; $F = 47.78$; Figure 6C) compared to individuals in libraries C446 and C447. Despite these differences, coverage was reasonably consistent and missing data did not cluster within particular samples or SNPs (Figure 7). Furthermore, populations from the same geographic regions that were sequenced in separate libraries still clustered together in the analysis of genetic structure, as described below.

Data Analysis

Four genetic groups were identified by principal components analysis (Figure 8). When the first principal component (PC1), which accounted for 5.3% of the total variation in the dataset, was plotted against PC2 (3.6%) or PC3 (2.5%), dramatic differentiation between western larch and subalpine larch was observed. Within subalpine larch, three clusters corresponded to three broad geographic regions: the Cascade Range, the southern Rocky Mountains and the northern Rocky Mountains. While the first three

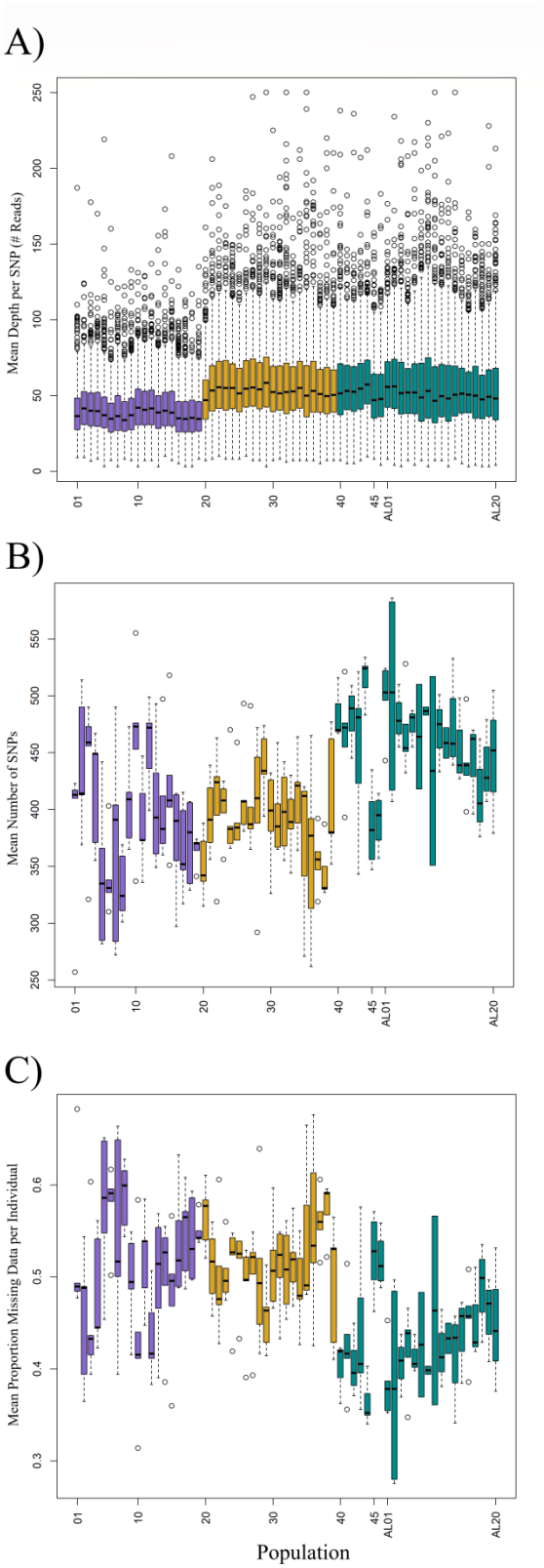

Figure 6. Differences among Sbf1 libraries (C446 in purple; C447 in yellow; C448 in green) in (A) mean depth per SNP, (B) SNP count per individual and (C) proportion missing data per individual.

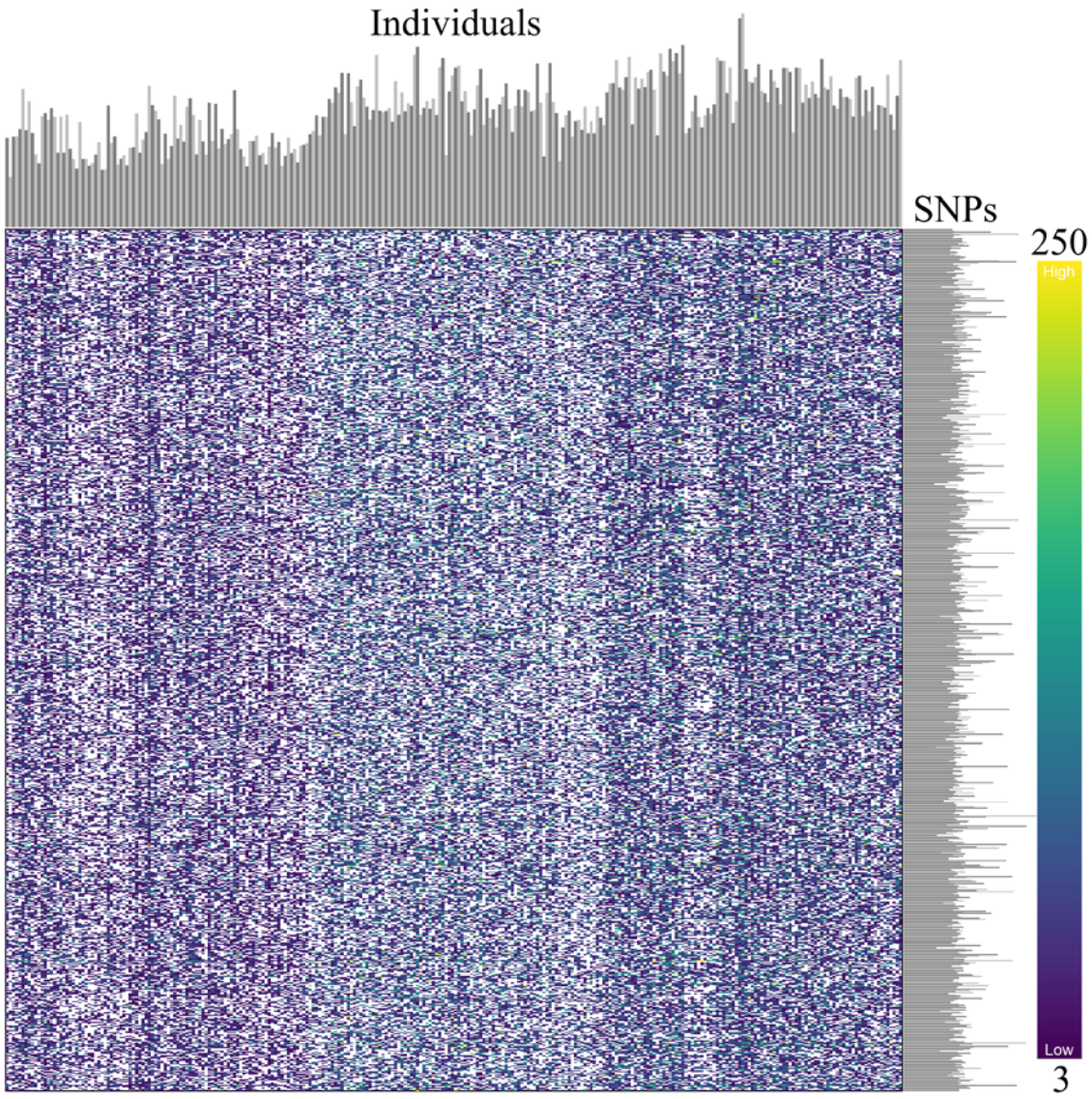

Figure 7. Heat map showing presence/absence (colour/white) and depth (colour scale) of 751 SNPs for 284 individuals of subalpine larch (*Larix lyallii*).

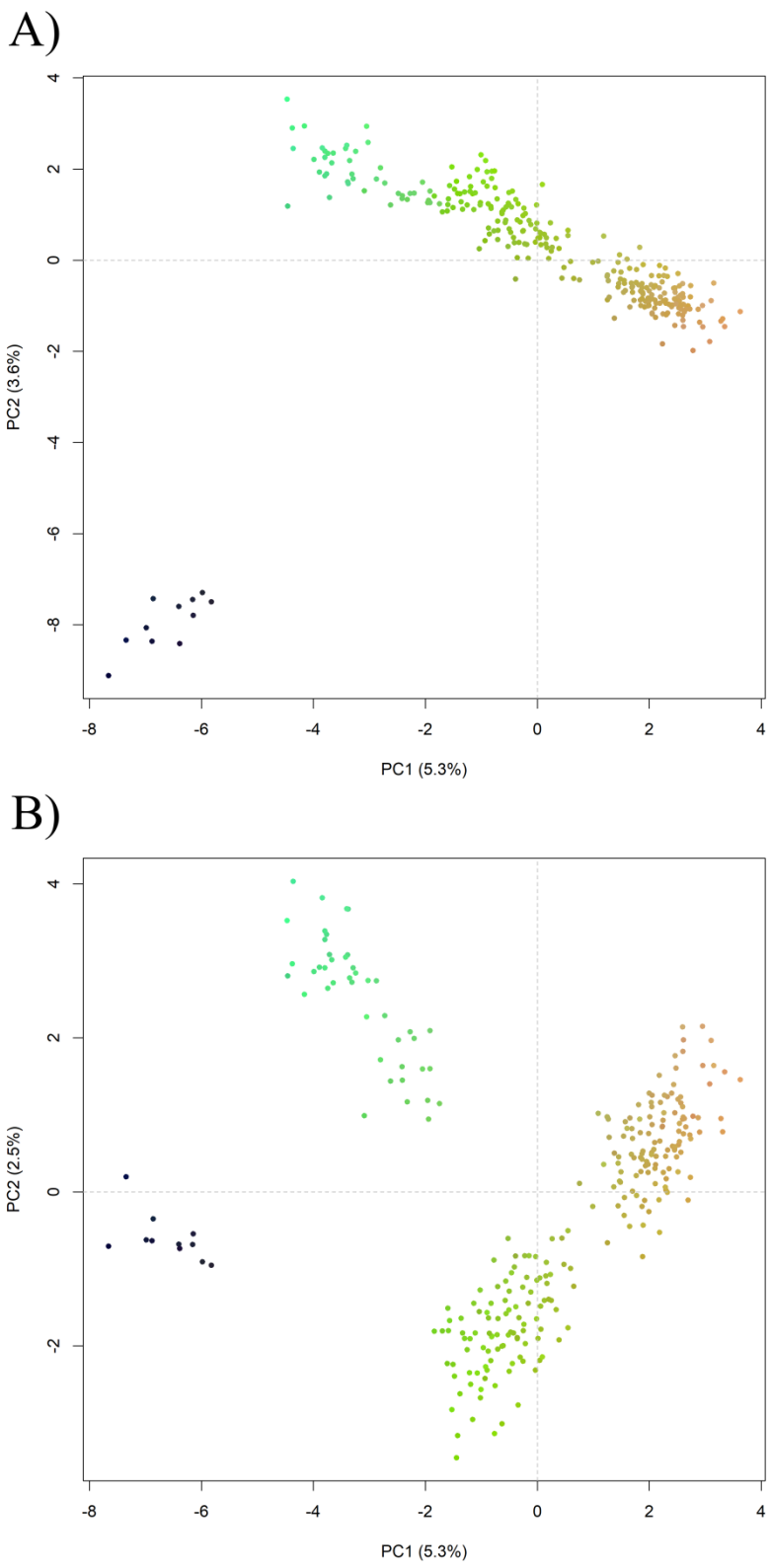

Figure 8. Colourplots for the first three principal components (PCs) associated with genetic variation in in 61 populations of subalpine larch (green/red) and two populations of western larch (blue).

PCs only represented 11.4% of the total variation, this was reasonable given the highly dimensional nature of the dataset.

One individual sampled at Indian Head, Montana, clustered with western larch and was removed from subsequent analyses of subalpine larch (Pop43_1262). This was not entirely unexpected given that western larch and subalpine larch overlapped in elevation at this site and the individual that was removed was sampled at the lowest elevation within the population. Tomentum on new stem growth, a defining morphological characteristic of subalpine larch, was observed on this individual. Nevertheless, it can be extremely difficult to separate subalpine larch and western larch based on morphological characteristics alone. Although western larch and subalpine larch had previously been observed to hybridize where their ranges overlap, PCA did not detect any hybrids intermediate between western larch and subalpine larch clusters.

Two additional analyses identified the same four genetic groups. First, K-means clustering identified four clusters. A discriminant analysis of principal components (DAPC) based on 67 PCs (54.5% of total variation) and three discriminant functions was used to assign individuals to these clusters. After assignment, the same genetic groups were identified: the western larch outgroup and subalpine larch in the Cascade Range, the southern Rocky Mountains and the northern Rocky Mountains (Figure 9). DAPC loadings showed that this pattern was driven by many alleles that each accounted for less than 1% of the total variation versus few alleles of large effect (Figure 10). The same four genetic groups were identified by STRUCTURE (Figure 11). Bayesian analysis was also able to successfully identify the western larch individual at Indian Head as well as one

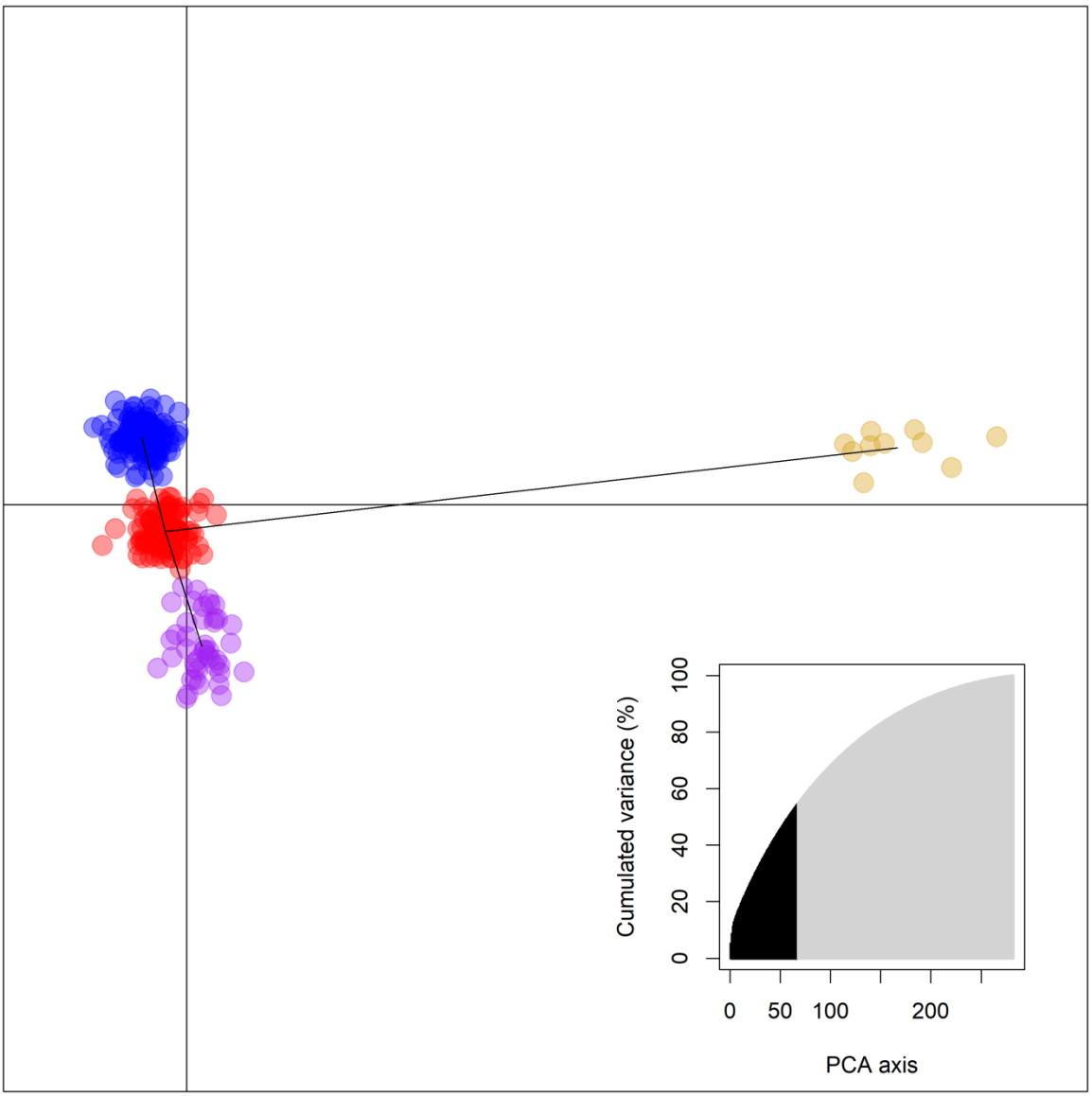

Figure 9. Three discriminant functions and 67 PCs (representing 54.5% of the total cumulative genetic variation as displayed in the inset) identify four genetically distinct clusters: western larch (yellow), subalpine larch in the northern Rocky Mountains (blue), subalpine larch in the southern Rocky Mountains (red) and subalpine larch in the Cascade Range (purple).

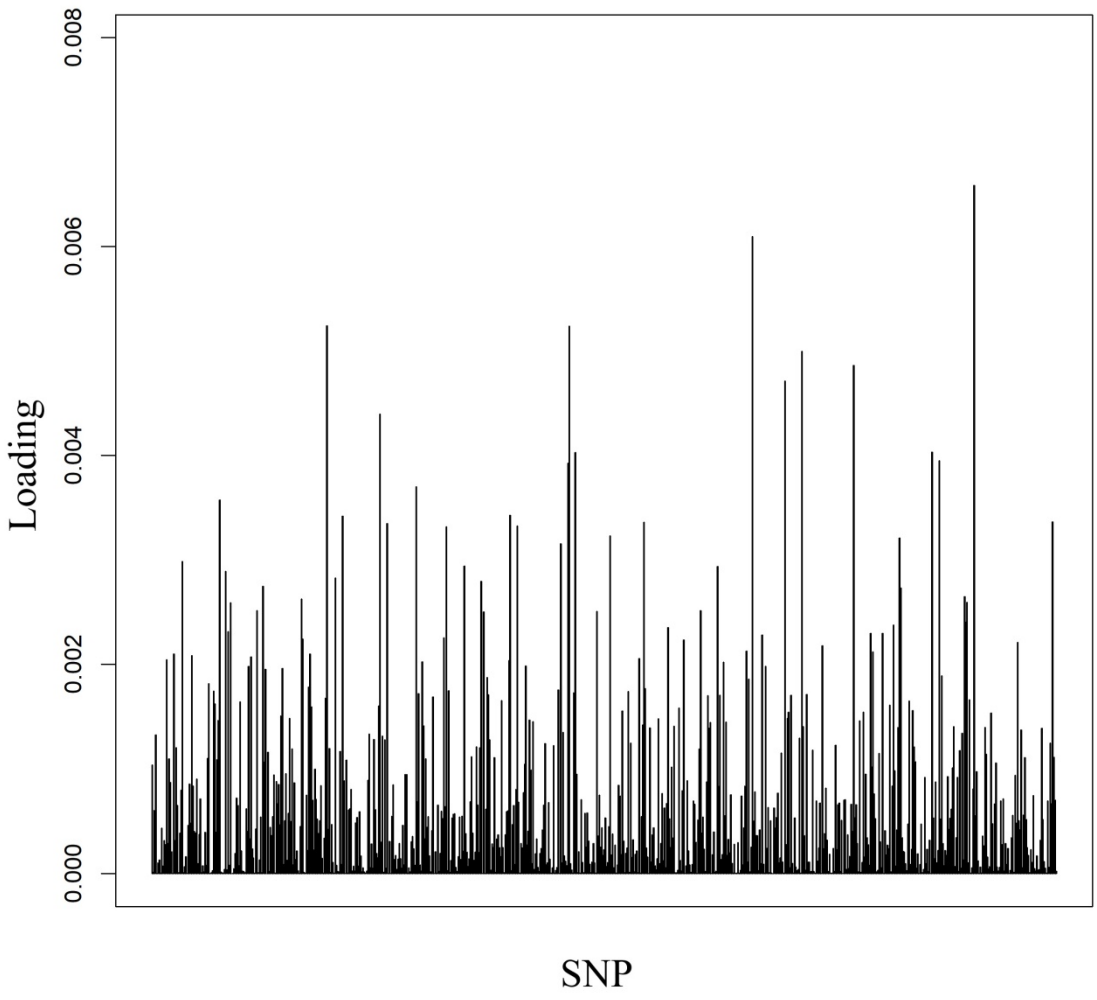

Figure 10. DAPC loading plot demonstrates that individual SNPs contribute small amounts of variation (< 0.8%) to principal components.

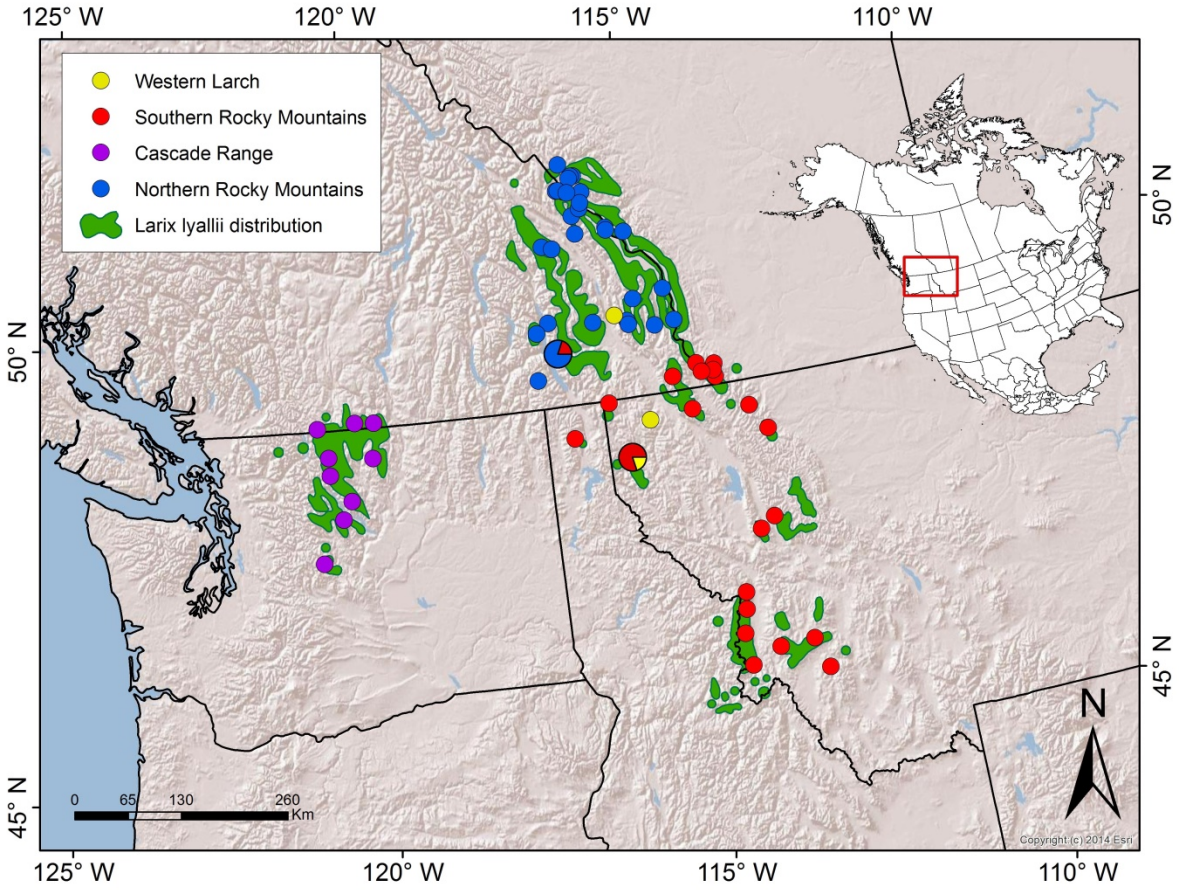

Figure 11. A discriminant analysis of principal components (DAPC) and a Bayesian STRUCTURE analysis identified four genetic clusters on the landscape: a western larch outgroup (yellow), a subalpine larch cluster in the Cascade Range (purple), a subalpine larch cluster in the southern Rocky Mountains (red) and a subalpine larch cluster in the northern Rocky Mountains (blue). One western larch tree was sampled at Indian Head, Montana. At Gray Peak Pass (Pop20) DAPC analysis identified two individuals from the southern Rockies while STRUCTURE analysis identified only one (pictured above).

individual in the northern Rockies that clustered with the southern Rockies. The proximity of this individual to the southern cluster means that it could be a migrant that arrived via long-distance dispersal.

Genetic structure reflected isolation by distance across the species range. A multivariate Mantel test found a significant correlation between genetic and geographic distance among subalpine larch samples ($p = 0.003$). This was expected given the major geographic disjunction present between the Cascade Range and the Rocky Mountains. A spatial analysis of principal components (sPCA) found large eigenvalues associated with global structure (Figure 12). Global structure was indeed significant ($p = 0.001$) while local structure was not ($p = 1.0$). A spatial map of genetic clines indicated that individuals from the northern Rocky Mountains were genetically differentiated from both the southern Rocky Mountains and the Cascade Range (Figure 13).

A dendrogram of Provesti's genetic distance was used to elucidate patterns of relatedness. The most ancient division occurred between subalpine larch and western larch (Figure 14). Next, subalpine larch populations in the Cascades split from subalpine larch populations in the Rocky Mountains. Within the Rockies, there was strong support for a split between the Northern Rockies and the Southern Rockies. Populations in the central Rocky Mountains formed a monophyletic group within the southern clade but this cluster had weak support (49.9%), indicating that this split was likely based on relatively few variants. Within the Cascades, Pop08, on the southern range margin, was basal relative to the other populations. Within the southern Rockies, Pop17, the farthest north, was identified as basal. Remaining outgroups were less geographically sensible. Within the northern region, Pop54, from the central part of the cluster, was identified as being

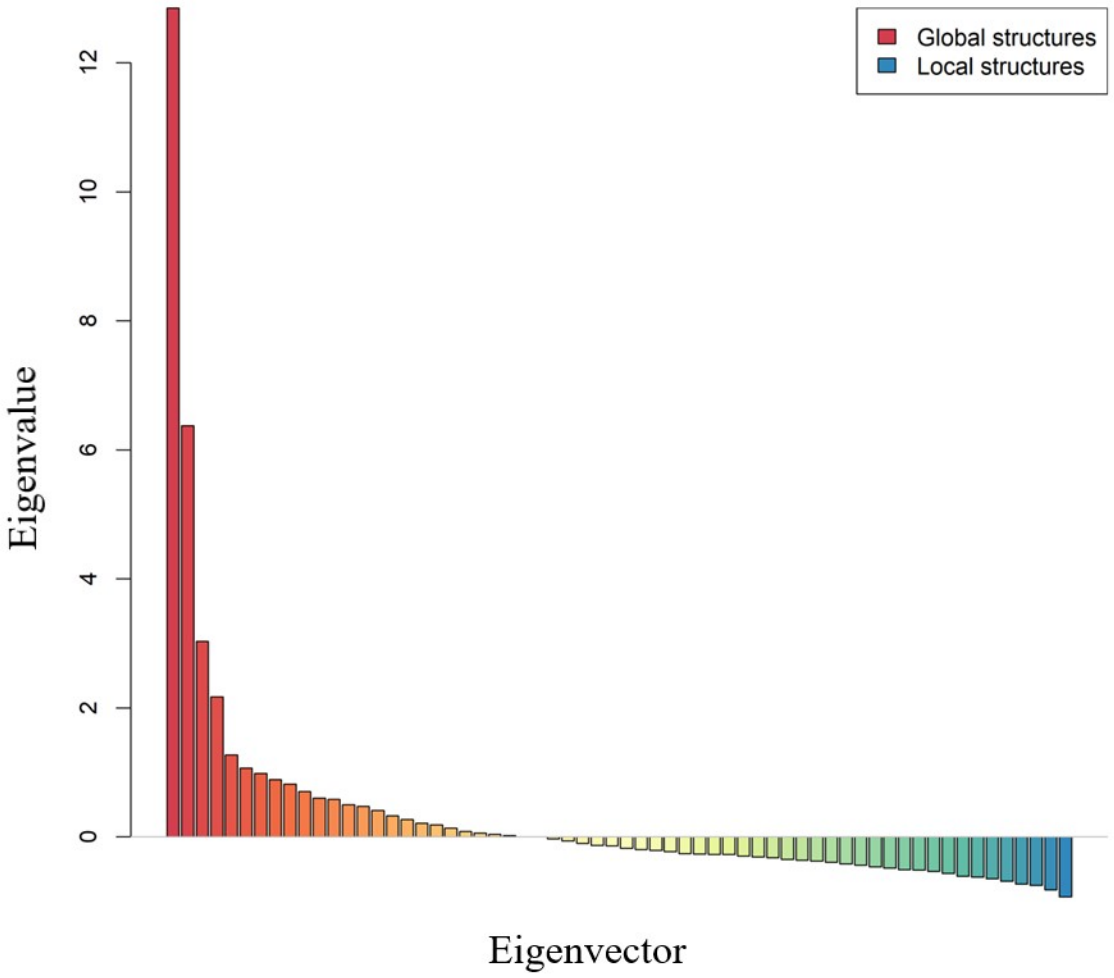

Figure 12. Eigenvalues derived from a spatial analysis of principal components indicate that there is global structure (large positive values) but no local structure (negative values) present in the dataset

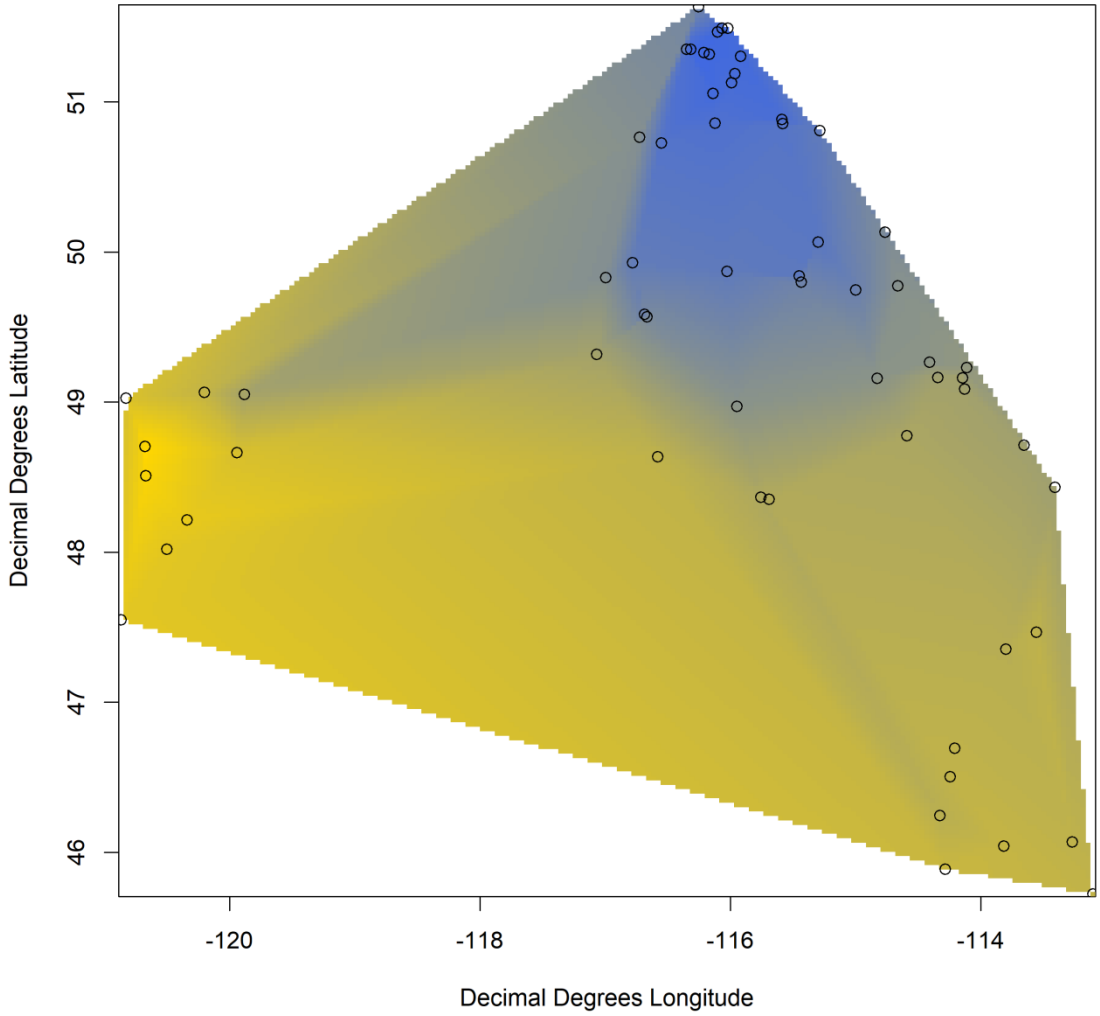

Figure 13. Genetic clines across the range of subalpine larch interpolated using lagged principal components from spatial PCA analysis and represented by color gradients.

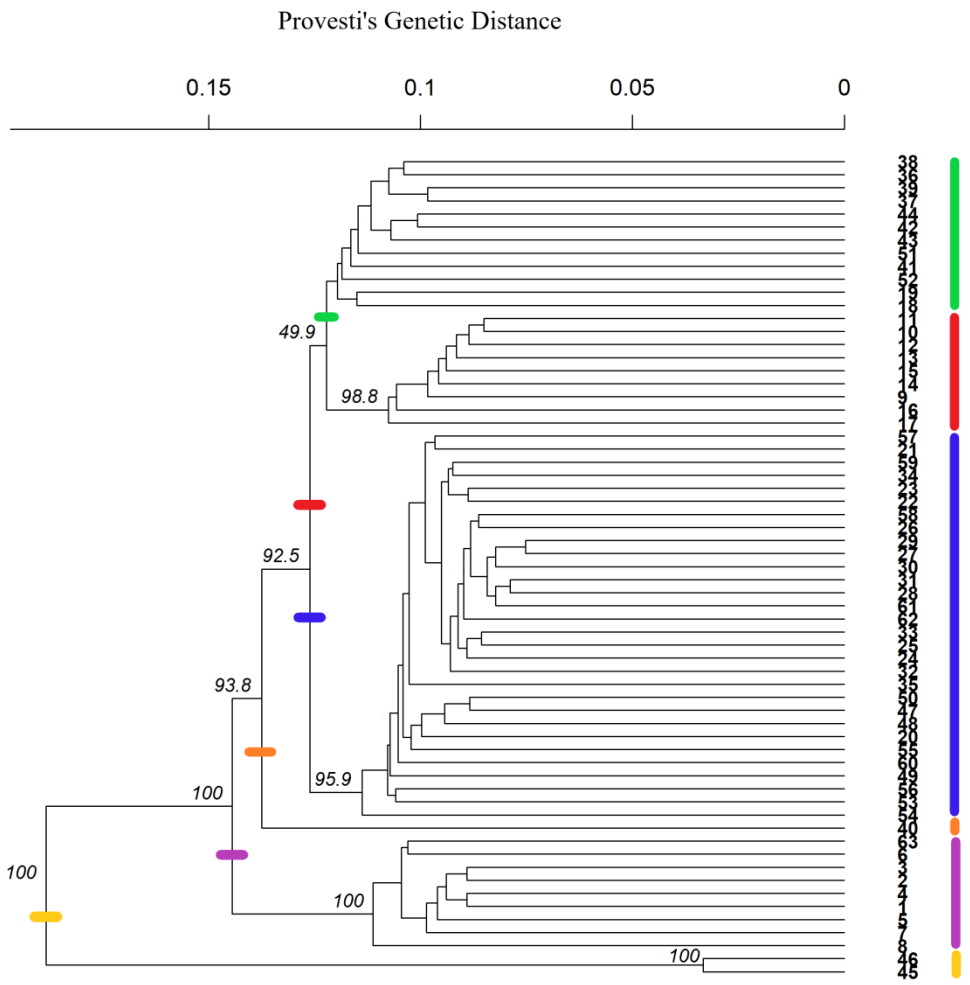

Figure 14. Dendrogram of Provesti's genetic distance with bootstrap support for two populations of western larch (yellow) and 61 populations of subalpine larch divided into three geographically sensible clusters: the Cascade Range (purple), the northern Rocky Mountains (blue), the southern Rocky Mountains (red) and one genetically distinct sub-cluster in the central Rockies (green). Population 40, from Glacier National Park, appears as an outgroup to all other populations in the Rocky Mountains (orange).

basal to the remaining populations. Finally, Pop40, a small fragment in Glacier National Park, was identified as basal to all other populations in the Rocky Mountains.

An analysis of molecular variance (AMOVA), or hierarchical F_{ST} analysis, was used to examine variance components attributable to species, regions, populations and individuals. While the majority of variation resided within individuals, the initial analysis attributed 24 % of variation to differences between species ($p = 0.001$). A second AMOVA analysis that excluded western larch found significant variance components attributed to regions (7.4 %; $p = 0.001$) and populations within regions (4.9 %; $p = 0.001$).

Discussion

This study identified population genetic structure across the natural range of subalpine larch. A spatial analysis of principal components (sPCA) detected the presence of significant global structure. A multivariate Mantel test detected significant isolation by distance. Finally, an analysis of molecular variance (AMOVA) found significant differentiation between regions and populations within regions. Together, these results provide the first evidence of range-wide population genetic structure in subalpine larch.

Four genetic clusters were identified by both PCA and Bayesian STRUCTURE analysis: a western larch outgroup, a subalpine larch cluster in the Cascades, a subalpine larch cluster in the southern Rocky Mountains and a subalpine larch cluster in the northern Rocky Mountains. Populations in the Cascade Range did not cluster with populations in south-central BC, as described previously, but this study did confirm the presence of a genetically differentiated sub-cluster in the central Rocky Mountains (Khasa et al. 2006). These groups are supported by a dendrogram of Provesti's genetic distance. In the dendrogram, the first split occurs between subalpine larch and western larch, as expected. This result is supported by AMOVA, which found significant differences in molecular variance between species. Subalpine larch and western larch can hybridize but do not often do so in nature. Although the two species occupy many of the same mountain ranges, they inhabit different altitudinal zones separated by 150 – 300 m of elevation. Occasional overlap occurs when subalpine larch moves into disturbed areas below its usual altitudinal limit, e.g. in avalanche chutes, on talus slopes or in burned areas. The two species are known to overlap and hybridize at 2000 m on Carlton Ridge in the Bitterroot Range. Carlson (1965) identified several natural hybrids using a

morphology-based hybridization index and later confirmed the presence of hybrids using foliar terpene profiles (Carlson et al. 1991). Artificial hybrids were produced via cross-pollination of the parental species (Carlson and Theroux 1993) and hybrids were observed to have faster germination rates, higher germination rates and greater stem diameter, providing some evidence for hybrid vigour (Carlson 1994). Growth was intermediate between the two parents but no outbreeding depression was observed. Based on these results, it seemed likely that introgression would be observed in this range-wide survey. However, out of the 274 subalpine larch trees included in this study, only one was identified as a western larch tree and that ID was definitive, not partial. Although that individual was observed to have tomentum on its new growth—normally a key identifying feature for subalpine larch—western larch trees were observed close by and this particular individual was sampled at the lowest elevation within the population at Indian Head, Montana. Thus it appears that species boundaries are maintained and hybrids remain rare across the natural range of subalpine larch.

Genetic clusters identified across the range of subalpine larch represent geographically sensible groupings. On the dendrogram, the first intraspecific split occurs between the Cascade Range and the Rocky Mountains. These two mountain chains are separated by a minimum of 200 km at their nearest point in southern British Columbia. Today, timberline populations in these two regions are reproductively isolated. It is possible that this isolation is ancient, dating back to the uplift of the Cascade Range, which ended during the Pliocene (McKee 1972). However it seems more likely that subalpine larch populations were better connected at the end of the Pleistocene (10 Ka – 1.8 Ma) than they are today. During the Pleistocene, the subalpine zone was lower,

probably between 900 and 1000 m lower than present, and covered a much larger area (Axelrod 1990). Although northern Washington, northern Idaho and northern Montana were covered by ice as recently as 13,500 years ago, the retreat of the Cordilleran ice sheet would have opened a subalpine corridor between the Cascade Range and the Rocky Mountains across north-central Washington (Dyke 2004). Evidence from whitebark pine, a common associate of subalpine larch, supports this hypothesis. Haploxylon pine pollen, most likely from whitebark pine, appeared in northwestern Montana and northern Idaho between 15,000 and 10,000 years ago (Mack et al. 1978, 1983) and a phylogenetic study of whitebark pine found a mitotype that moved west from Idaho via seed to colonize the northern Cascades (Richardson et al. 2002). Thus it is likely that subalpine larch enjoyed extended periods of connectivity between East and West during the Pleistocene. However this would have changed during the Holocene when climate shifted to become warmer and drier. Over the last 10,000 years, subalpine larch has likely retreated to higher elevations in both mountain chains in order to avoid abiotic stress and competition from faster-growing conifer species. Although the mechanism here may be somewhat different, reflecting recent altitudinal retreat versus fragmentation between isolated Pleistocene refugia, many conifer species in western North America show genetic differentiation between the eastern and western portions of their ranges (Critchfield 1984; Jaramillo-Correa et al. 2009).

The second major intraspecific split that can be observed in the dendrogram represents divergence between the southern and northern Rocky Mountains. It seems likely that the northern Rockies were colonized from the south given that the northern Rockies were covered by a continental ice sheet up until around 15,000 years ago, when

an unglaciated corridor began to open in the Rocky Mountain trench (Dyke 2004). Indeed, DAPC analysis found that individuals in the Cascades are more closely related to individuals in the southern Rockies than those in the northern Rockies. If populations in the East and West were connected at the end of the Pleistocene, the East was likely represented by the ancestors of individuals from the southern Rockies. When the ice sheet began to retreat, subalpine larch would have been able to migrate north. In many western conifers, northward expansion generated latitudinal clines in genetic variation due to bottlenecks resulting from founder events (Critchfield 1984; Jaramillo-Correa et al. 2009). Such patterns have been observed in some other western conifers, e.g. western white pine (Nadeau et al. 2015). Thus neutral processes like bottlenecks and genetic drift could be responsible for the genetic divergence observed between regions. An alternate hypothesis could be that subalpine larch populations were required to adapt in response to novel environmental conditions as they moved north, meaning genetic differentiation has been driven by natural selection. It is also possible that both neutral and adaptive processes have driven genetic divergence, as occurred when an invasive bank mole colonized Ireland (White et al. 2013). Regardless, step-wise expansion seems likely. The dendrogram shows a monophyletic sub-group within the southern clade that represents the central Rocky Mountains, providing additional support for differentiation that is strongly associated with latitude.

One result from the dendrogram of Provesti's genetic is difficult to explain. Population 40, which is located at Preston Park, just below Siyeh Pass in Glacier National Park, is identified as being genetically differentiated from all other populations in the Rocky Mountains. Given its position in the center of the species' latitudinal range, this is

difficult to explain. However it could be that this population is subject to unique pressures. Preston Park is now an isolated fragment and the population is very small – only 19 individuals could be sampled at this location instead of the usual 30. The average inter-tree sampling distance was 57 m based on measurements taken with the handheld GPS unit but the mean minimum inter-tree sampling distance calculated from the population distance matrix was only 32 m. Although this is a conservative estimate (as discussed in the Methods), it is still the second shortest mean minimum inter-tree sampling distance recorded in this study. Because the Preston Park population is so small, it could be experiencing strong genetic drift as well as high levels of inbreeding, leading to the random fixation of allelic variants, thus driving genetic differentiation. Knowledge of genetic diversity statistics would help clarify this issue. However inherent limitations in this dataset mean that answering this question is beyond the scope of the current study.

There are several issues that limit the inferences that can be made from this RAD-seq dataset. First, a maximum of five individuals were sequenced per population, making it inappropriate to calculate population-level diversity statistics. Second, because the larch genome was under-sampled, it is likely that genetic diversity was also under-sampled. Under-sampling also led to a high proportion of missing data, which reduced the number of SNPs available for analysis after filtration. Third, a low depth of coverage per locus makes it difficult to ascertain whether an individual is a truly a homozygote or, more simply, whether the second allele was missed during sampling. Average depth of coverage was high (44X) but individual genotypes were called with as few as three reads. Even for SNP loci with high coverage, diversity may be underestimated because the presence of PCR duplicates inflate estimates of homozygosity (Davey et al. 2011;

Parchman et al. 2018). Since this data set was generated using single-end sequencing, PCR duplicates cannot be identified and removed. Furthermore, RAD-seq datasets underestimate diversity due to the presence of mutations in the restriction enzyme recognition site. Such mutations prevent restriction enzymes from cutting the DNA strand, leading to allelic dropout, which again reduces genetic diversity (Gautier et al. 2013). Despite these challenges, the 751 markers obtained for this study provide high confidence that the patterns of spatial genetic variation identified in this study are reliable. The observed patterns are clear and geographically sensible.

The results of this analysis have high practical value and should be used to inform future management and conservation efforts. At present, BC's Genetic Conservation Technical Advisory Committee focuses primarily on assessing inventory on protected land to assess the conservation needs of native tree species. By this measure, subalpine larch is doing quite well, given the presence of two provincial parks in the northern Cascades (Manning Provincial Park, Cathedral Lakes Provincial Park) and numerous provincial and national parks in the Canadian Rockies (e.g. Banff, Kootenay, Yoho and Waterton Lakes National Parks). However this approach fails to account for the threat of *in situ* maladapation as a result of climate change. In order to provide effective conservation for subalpine larch, it is likely that assisted migration will have to be implemented in order to help this species to track its shifting climate niche. Natural rates of migration are extremely unlikely to keep pace with environmental change (McLachlan and Clark 2004; Gray and Hamann 2013). Alternatively, breeding populations could be established to safeguard the genetic diversity of the species (Yanchuk 2001). One such population already exists at the Kalmalka Forestry Centre in Vernon, BC. Indeed, the

grafted clones sampled for this study are part of that resource. However eight of the 18 populations archived at this site have fewer than 10 surviving individuals, the target number for provincial conservation. Furthermore, the Kalamalka site is extremely hot and dry, which could prove disastrous for subalpine larch if the irrigation system ever fails. Indeed, the international larch arboretum established at Hungry Horse, Montana, in 1992 was never irrigated and is now missing all of its high-elevation accessions. Additional reserves should consider more favorable locations for subalpine larch. All future conservation efforts should strive to preserve genetic variation from the genetically divergent groups identified across the species range.

To conclude, this dataset provides the first evidence for range-wide genetic structure in subalpine larch. While it is not yet clear whether these patterns were shaped by neutral processes acting during post-glacial expansion or by adaptive responses to natural selection, there is value to the conservation of both types of genetic variation. Conserving the first preserves the evolutionary history of a species. Conserving the second preserves functional divergence. Generally, the conservation of genetic diversity serves as a bulwark against future environmental change, providing the raw material necessary for adaptation.

CHAPTER 3: POPULATION GENOMICS

Introduction

Biogeography, the distribution of life across space and time, provides critical context for understanding the origin and evolutionary potential of species. For hundreds of years, biologists have collected and identified fossils and living organisms around the world in order to better understand life on Earth. Today, biogeography is a highly integrative discipline (Wen et al. 2013). Informed by geology, climatology, paleontology and molecular genetics, biogeography in turn informs our understanding of evolution, ecology and conservation. Phylogeography is a sub-discipline that utilizes molecular population genetics to examine intra-specific range dynamics over relatively shallow time scales. In this study, phylogeographic methods are used to elucidate the history and future prospects of a North American conifer species, subalpine larch (*Larix lyallii* Parl).

Modern patterns of genetic variation in North American conifers generally reflect climate change during the Quaternary Period, which includes both the Pleistocene (1.8 Ma – 10 kya) and Holocene Epochs (10 kya – Present). During the Pleistocene, cooling global temperatures led to the formation of continental ice sheets on both sides of the Rocky Mountains— the Laurentide Ice Sheet in the east and the Cordilleran Ice Sheet in the west. The Cordilleran Ice Sheet extended from the mountains of southeastern Alaska into northern Washington and northwestern Montana (Booth et al. 2003). During the Pleistocene the Cordilleran Ice Sheet underwent at least 16 large-scale advances, achieving its last glacial maximum (LGM) approximately 11,000 years ago. Major, repeated alterations of species' ranges during the Pleistocene led to a general loss of

genetic variation (Critchfield 1984). For western conifers, which all have ranges that extend south of the glacial boundary, multiple Pleistocene refugia were often established across climatically diverse topography. Genetic divergence arose between populations isolated in disjunct refugia. When the ice retreated at the end of the Pleistocene, conifers were once again able to expand their ranges north into present-day British Columbia. Rapid expansion creates distinctive patterns of spatial genetic variation: high-latitude populations tend to share recent ancestry with distant low-latitude populations whilst adjacent low-latitude populations are often genetically divergent as a result of enduring spatial disjunction (Jaramillo-Correa et al. 2009). These patterns have been observed in many conifer species, including tamarack (*Larix laricina*; Warren et al. 2016) and Douglas-fir (*Pseudotsuga menziesii*; Gugger et al. 2010; Wei et al. 2011).

At present, little is known about the biogeographic history of subalpine larch, a timberline conifer native to western North America. The earliest *Larix* macrofossils were found in the North-American High Arctic in a Middle Eocene formation (47 - 41 Ma) on Axel Heiberg Island (LePage and Basinger 1991; Jagels et al. 2001). Climate during the Eocene was much warmer than it is today. Conifers were dominant at high latitudes and at high elevations farther south (Axelrod 1990). The uplift of the Rocky Mountains, which began in the Late Cretaceous (70 Ma) and ended during the Late Eocene (37 Ma), provided plentiful high-elevation habitat (Elias 2002). At the end of the Eocene, global climate started to cool and conifers began to expand their ranges southward. Further expansions occurred during the Late Miocene and the Pliocene as mean annual temperature continued to drop and precipitation patterns became more seasonal (Graham 1998). The orogeny of the Sierras and the Cascades during the Pliocene provided

additional habitat of significant relief (McKee 1972). *Larix* macrofossils have been identified in a Miocene formation at Snake River Basin, Idaho (Axelrod 1965; Axelrod 1968), and a Pliocene formation at Birch Creek, Alaska (Miller and Ping 1994). Evidence from geology, climatology and paleontology thus suggests that *Larix* was broadly distributed in western North America prior to the major glaciation events of the Pleistocene.

Palynology and paleontology are often used to locate refugia and elucidate post-glacial migration routes (e.g. Gugger and Sugita 2010). Unfortunately, *Larix* not only shares a nonsaccate pollen morphology with its closest relative, *Pseudotsuga*, but the two pollen types are indistinguishable in the fossil record (Simak 1966). For example, the presence of *Larix*/*Pseudotsuga* pollen in lake sediments collected at a site near Nelson, BC, where a population of subalpine larch still grows today, suggests that subalpine larch may have succeeded in colonizing this site 8,200 years ago (Mustaphi and Pisaric 2014). However this evidence cannot be considered conclusive. Difficulty in distinguishing larch fossils is not restricted to pollen. Subalpine larch macrofossils cannot be distinguished from western larch macrofossils (*Larix occidentalis*) due to shared morphological characteristics such as exserted cone bracts. Molecular methods are necessary for elucidating the biogeographic history of subalpine larch. One previous study used seven microsatellite markers to assess the phylogeography of 19 populations of subalpine larch from the northern part of the species range (Khasa et al. 2006). The authors found low genetic diversity within populations and high genetic differentiation among populations in comparison to western larch, a closely related species with a more contiguous distribution at lower elevation. Together with deviations from mutation-drift

equilibrium in seven of the 19 populations, these results were used to support the interpretation that Holocene bottlenecks occurred via founder events and ensuing genetic drift in isolated populations. However, a focus on northern populations located on previously glaciated sites limited the inferences that could be made regarding the biogeographical history of the species.

In this study, next generation sequencing (NGS) methods were used to elucidate the range-wide phylogeography of subalpine larch. While early phylogeographic studies in plants were generally hampered by a lack of DNA sequences with sufficient intra-specific variation, NGS approaches can generate thousands of markers in non-model organisms, providing deep phylogenetic resolution (McCormack et al. 2013; Edwards et al. 2015). Additionally, sophisticated statistical phylogeographic methods allow for demographic hypothesis testing and parameter estimation.

Methods

Sample Collection

Fifteen populations of subalpine larch were selected to represent the natural range of the species (Figure 15; Table 1). Approximately 25 individuals were sequenced per population for a total of 366 samples. Note that these individuals form an independent sample; they were not sequenced in the previous chapter (p. 25).

Most trees were sampled at a minimum inter-tree distance of 50 m in order to avoid sampling close relatives. Based on the distance matrices calculated in Chapter 2 (p. 23), which provide conservative estimates of inter-tree distances, the average minimum inter-tree sampling distance was 57 m (Table 5). Four populations had an average minimum inter-tree sampling distance of less than 50 m but only one population, Holland Pass, had a distance of less than 45 m. Two populations (Holland Pass, MT; Molar Pass, AB) each had a pair of individuals spaced less than 20 m apart. While sampling nearby individuals increases the risk of sampling close relatives, it was necessary in very small populations.

Sequencing

DNA was extracted and quality criteria (QC) evaluated as previously described (Chapter 2, p. 24). Because three samples were successfully sequenced despite having Nanodrop 260/230 values below the initial QC cutoff (p. 25), that cutoff was lowered from 2.0 to 1.75. Based on this modified criterion, 80% of initial DNA extractions were considered successful. DNA from some individuals in target populations was individually re-extracted in order to meet QC.

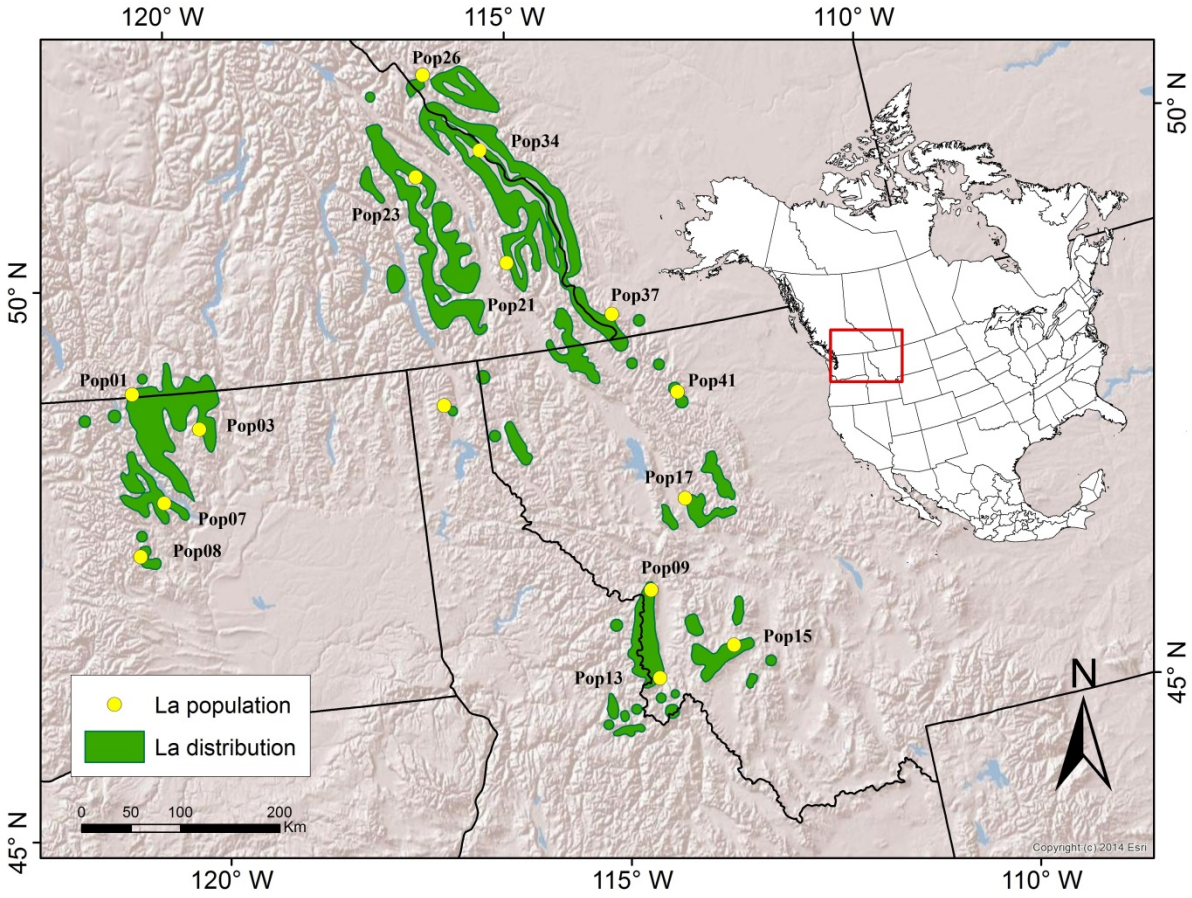

Figure 15. Subalpine larch (*Larix lyallii* Parl.) populations sampled for an analysis of population-level diversity statistics and demographic history.

Table 5. Populations of subalpine larch used for restriction enzyme associated DNA sequencing with the Pst1 restriction enzyme.

Pop.	Location	Latitude (decimal degrees)	Longitude (decimal degrees)	Elevation (m)	Number sequenced	Av. Min. Calculated Distance (m)
01	Frosty Mountain, E.C. Manning Provincial Park, BC	49.0265	-120.8275	2038	25	50
03	Tiffany Mountain, WA	48.6632	-119.9421	2247	25	51
07	Big Hill, WA	48.0183	-120.5016	2016	25	70
08	Windy Pass, WA	47.5494	-120.8689	2061	25	51
09	Carlton Ridge Research Natural Area, MT	46.6921	-114.2067	2483	25	72
13	Trapper Peak, MT	45.8859	-114.2824	2804	25	59
15	Storm Lake, MT	46.0670	-113.2661	2536	25	47
17	Holland Pass, MT	47.4655	-113.5552	2287	23	34
19	Roman Nose, ID	48.6345	-116.5804	1925	25	65
21	Sparkle Lake, Top of the World Provincial Park, BC	49.8386	-115.4481	2027	25	49
23	Tiger Pass, BC	50.7258	-116.5514	2187	25	54
26	Molar Pass, Banff National Park, AB	51.6327	-116.2538	2312	18	81
34	Wonder Pass, Banff National Park, AB	50.8833	-115.5891	2273	25	45
37	Bovin Lake, AB	49.2300	-114.1120	1978	25	85
41	Paradise Park, Glacier National Park, MT	48.4302	-113.4061	2024	25	50

After extraction, DNA was normalized to a final concentration of 20 ng/µL as previously described (Chapter 2, p. 26) and sent to Floragenex for RAD library preparation (Floragenex, Inc., Eugene, USA). Five multiplexed libraries (C577, C578, C579, C580, C581) were prepared as per standard Floragenex protocols (Chapter 2, p. 26) using the Pst1 restriction enzyme (target sequence 5' CTGCAG 3'). Pst1 targets a six-base subset of the Sbf1 recognition site, meaning it should cut all Sbf1 recognition sites plus additional Pst1 recognition sites. The draft Siberian larch genome has 43,003 Sbf1 sites and 888,022 Pst1 sites. Note that Pst1 is a methylation-sensitive restriction endonuclease. The practical implication of this sensitivity is that restriction endonuclease activity should target the non-methylated portion of the genome. For conifers, whose genomes can comprise up to 80% repetitive DNA sequences, including highly methylated transposable elements, using a methylation-sensitive enzyme should help focus sequencing effort on the expressed regions of the genome (Wegrzyn et al. 2014).

Agilent Bioanalyzer traces were used to validate library integrity. Average fragment size per library was 476 nucleotides (455 – 504). Libraries were sent to Genome Quebec (Montréal, Québec, Canada) for paired-end Illumina sequencing. In total, 475 individuals were sequenced on five lanes of Illumina Hi-Seq 2500 with 95 individuals pooled per library and one yeast control. Of these, 386 samples were pertinent to this experiment, including 366 subalpine larch samples and twenty controls that served as within- and among-library replicates (two of each type per library). Randomization was carried out by generating random numbers in R and ordering samples to assign them to specific plate and well positions. Samples were randomized within and among libraries prior to library preparation and sequencing.

Bioinformatics

Paired-end sequencing data were received from Genome Québec on July 16th, 2016, in fastqc format. Short-read sequences were 125 nucleotides in length. An average of 260 million read pairs was obtained per lane of sequencing, for a total of over 2.6 billion individual sequences.

Overall library quality was assessed using FastQC (Andrews 2010). All five libraries were generally of high quality and read quality stayed high until the end of both read1 and read2 sequences. However, four of the five libraries had quality scores below Phred 20 at the first position of read1. This was unexpected because quality usually starts high and decreases as errors accumulate over successive rounds of sequencing. It is therefore likely that position-specific quality problems are the result of a technical error that occurred during sequencing. In libraries C578 and C579, per tile sequence quality was flagged by FastQC as a possible concern, suggesting that quality problems were indeed localized on the sequencer flowcell. Overall, this drop in quality reflected only a subset of sequences. Mean quality remained above Phred 30 to the end of both read1 and read2 in all five lanes, meaning nucleotides were called with 99.9% certainty.

Libraries did not pass all FastQC quality tests. Some failures were expected due to the characteristics of the RAD-seq data, as previously described (Chapter 2, p. 41). Note that in this dataset, data did not pass per-base sequence content tests due to higher-than-expected frequencies of certain nucleotides between positions 11 and 15, corresponding to the shared Pst1 RAD tag (5’ TGCAG 3’). Note also that while RAD-seq data were not expected to pass Kmer content tests due to the presence of high-frequency Kmers in the

first 15 nucleotides of read1, corresponding to the ten-nucleotide barcode and the five-nucleotide RAD tag, Kmer spikes were also present at the beginning of read2 sequences (positions 1 – 21). Biases at the beginning of read2 may have reflected the presence of PCR duplicates with common read2 start points. Finally, a Kmer spike was observed in library C578 between positions 66 and 73 (GAGCGTCGTGTAG). Initially, the presence of this sequence raised concerns regarding possible contamination because the sequence corresponds to a 33-base Illumina read2 TruSeq adapter. Indeed, the full TruSeq adapter sequence was present in the raw data of all five libraries. However further exploration indicated that this sequence was more likely to represent a high-frequency repetitive element than adapter contamination. The P2 adapter sequence was not paired with the P1 Floragenex adapter sequence, read2 adapters were present in both read1 and read2 sequences, the corresponding Illumina read1 TruSeq adapter was not present, and the read2 adapter sequence start position was randomly distributed across reads. Thus no additional steps were taken to remove this sequence from the dataset.

Bioinformatics processing was carried out to clean and de-multiplex reads (Table 6). Reads were trimmed to remove P1 and P2 adapter sequences as described in the previous chapter (Chapter 2, p. 31). Adaptor sequences were identified and trimmed from the 3’ end of 8.9% of forward reads (read1) and 6.5% of reverse reads (read2). Cutadapt was also used to remove reads shorter than 50 nucleotides (2.4% of reads).

After adapters were removed, sequences were de-multiplexed and filtered using the *process_radtags* program in STACKS v 1.44 (Catchen et al. 2011; Catchen et al. 2013). Read quality was ensured via the implementation of the –c option, which removed

Table 6. Reads kept over successive stages of bioinformatics processing in five libraries generated using restriction site associated DNA sequencing (RAD-seq) with the restriction enzyme Pst1.

		Library							
Software	Process	C577	C578	C579	C580	C581	Av. # Reads Retained	% of raw reads retained	% lost during process
FastQC	Count raw reads	516,387,568	534,445,056	523,333,510	517,596,904	512,752,438	520,903,095	100	0.0
Cutadapt	Trim adaptors	42,191,053	31,171,422	46,558,138	42,816,497	37,680,934	520,903,095	100	0.0
	Remove if < 50 bp	501,456,606	531,945,148	504,573,900	501,968,592	502,205,526	508,429,954	97.6	2.4
STACKS process_radtags	Remove low quality reads	500,559,870	531,128,162	503,678,230	500,759,849	500,848,506	507,394,923	97.4	0.2
	Remove reads with ambiguous barcode	494,717,850	525,380,684	497,442,256	494,392,531	494,176,600	501,221,984	96.2	1.2
	Remove reads with ambiguous RAD tag	459,267,150	490,737,403	462,365,217	458,804,239	460,996,412	466,434,084	89.5	6.7
	Remove yeast reads	436,713,883	460,431,907	429,788,687	427,277,568	432,455,925	437,333,594	83.9	5.6
	Remove unprocessed	436,713,883	460,431,907	429,788,687	421,849,514	432,455,925	436,247,983	83.7	0.2
STACKS clone_filter	Remove unpaired remnants	401,239,948	425,760,250	394,655,938	386,015,252	398,963,152	401,326,908	77.0	6.7
	Remove PCR duplicates	146,162,286	148,882,552	169,619,828	129,065,782	131,976,102	145,141,310	27.8	49.2
NextGenMap	Remove reads that did not align to reference genome	139,425,269	142,711,200	162,495,333	123,107,501	125,129,728	138,573,806	26.6	1.3
GATK Unified Genotyper	Remove reads with mapping quality score = 0	101,521,549	104,039,636	118,494,876	89,647,102	90,998,024	100,940,237	19.4	7.2
	Remove reads failing bad mate filter	71,224,592	73,742,679	88,197,919	59,350,145	60,701,067	70,643,281	13.5	5.8

all reads with uncalled bases, and the –q option, which used a sliding-window approach to discard reads if average quality within a window dropped below Phred 10. On average, 0.2% of reads were discarded due to low quality, 1.2% of reads were discarded due to the presence of an ambiguous barcode sequence and 6.7% of reads were discarded due to the presence of an ambiguous RAD tag sequence. Note that the default rescue function (-r) corrected a single nucleotide error within the RAD tag. The inter-barcode distance (--barcode_dist_1) was set to four so that barcodes with fewer than four sequencing errors could be rescued. Barcodes were 10 bases long and differed by at least four nucleotides from all other barcodes. One individual (Pop01_22) had no reads remaining after Stacks processing and was therefore excluded from further analysis.

After filtering, unpaired reads were discarded. Paired-end sequencing data allows for the identification of PCR duplicates that share a common read2 start point. Random shearing generates random start points for biological duplicates, meaning reads with the same start point are most likely clones generated during PCR amplification. PCR duplicates were removed using the Stacks *clone_filter* module. On average, 49% of reads were identified as clones and removed from the data set. An average of 72,570,655 read pairs were retained per lane of sequencing after bioinformatics filtering—28% of the original data. Individuals retained a mean of 765,513 read pairs.

Reads were aligned to a draft of the Siberian larch genome (Kuzmin et al. 2019) as described in the previous chapter (Chapter 2, p. 32). Output was converted from Sequence Alignment Map text format (SAM) to binary format (BAM) using Samtools v. 1.3 (Li et al. 2009). Samtools was also used to sort and index BAM files. Overall, 96% of retained larch sequences were successfully aligned to the Siberian larch draft genome.

Unfortunately, 27% of aligned reads had a map quality score of zero, meaning they mapped to more than one location in the draft genome with equal probability. Aligned reads were also filtered out if read pairs did not map to the same pseudo-scaffold (bad mate filter). While it was likely that the highly fragmented Siberian larch draft genome was assembled into pseudo-scaffolds that did not reflect the true location of contigs within the genome, and that some sequences were indeed correctly aligned even though read pairs mapped to different pseudoscaffolds, filtering out these pairs reflected a conservative approach. Thus only 13.5% of the original read data were retained for calling genotypes. After alignment, individuals retained a mean of 1,461,749 reads.

Base quality score recalibration (BQSR) was carried out as described in the previous chapter (Chapter 2, p. 33). Raw reads were aligned to the PhiX genome using NextGenMapper v. 0.5.3 (Sedlazeck et al. 2013). PhiX reads accounted for an average of 7.7% of reads in Pst1 libraries, slightly below the expected 10%. However there were still over one billion PhiX nucleotides sequenced per library, meeting the requirement to proceed with BQSR.

Gentoyping

Genotypes were called using GatK UnifiedGenotyper as described in the previous chapter (Chapter 2, p. 34). Both variant and reference calls were output together in variant call format (VCF). VCF files were sorted using Picard SortVCF and sites with non-reference alternate alleles were subset using the GatK SelectVariants function. SNPs were subset in VCFtools (Danecek et al. 2011) by selecting for loci with a minor allele count of at least one (--mac 1). In total, almost 265 million nucleotides were genotyped

including 2.9 million SNPs (1.1%). Coverage was moderate with an average of 20 reads per SNP. Low-quality SNPs were removed by filtering as described in Chapter 2 (p. 34; Table 7). After filtering, 834 SNPs were retained for further analysis.

Genotyping error rates were assessed in the R statistical environment (R Core Team 2019) using 20 replicate pairs. Averaged across replicate pairs, 62% of genotypes were erroneous. The majority of these errors could be attributed to missing data in one of the two replicate samples (74% of errors). Indeed, the larch genome was under-sampled in this study, meaning individuals with more sequencing data had a higher number of genotyped SNPs (Figure 16). The remaining errors (26%) arose due to incorrect genotype calls. In most cases one sample was genotyped as a heterozygote while the other was genotyped as a homozygote (97% of genotyped errors).

Due to the high number of errors present in called genotypes, data were analyzed using ANGSD v. 0.916 (Korneliussen et al. 2014). ANGSD is useful for performing population genetic analyses on low-depth sequencing data. While standard NGS genotyping pipelines rely on a depth-based filter to exclude sequencing errors and accurately call heterozygotes versus homozygotes, the ANGSD software differs in that it uses a probabilistic approach to calculate genotype likelihoods, allowing SNPs derived from low-depth data to be retained along with a measure of genotype uncertainty. Genotype likelihood was estimated as the marginal probability of the sequencing data given a particular genotype.

Table 7. Filtering procedure for SNPs generated using restriction associated DNA sequencing (RAD-seq) with the Pst1 enzyme for 365 samples of subalpine larch representing 15 populations.

Software	Function	Value	Filter	Process	Pst1 SNPs	% Kept
Picard	SortVcf	NA	NA	Sort VCF file	264,864,302	
Gatk	SelectVariants	NA	sites	Remove reference sites	8,672,681	
VCFtools	mac	1	sites	Remove sites with < 1 copy of minor allele	2,910,758	100
Gatk	QD	< 2.0	sites	Normalize quality by depth and filter low quality		
	FS	> 60.0	sites	Remove if strand bias present		
	SOR	> 3.0	sites	Remove if strand bias present		
	MQ	< 40.0	sites	Remove if mapping quality is low		
	ReadPosRankSum	< -8.0	sites	Remove if ref/alt alleles map differently		
VCFtools	remove-filtered-all	NA	sites	Remove sites that did not pass GatK filters	2,187,604	75.2
	exclude	NA	individuals	Remove individuals		
	minGQ	< 20	genotypes	Remove genotypes called with < 99% confidence		
	minDP	< 3	genotypes	Remove genotypes with < 3 reads		
	minQ	< 20	sites	Remove SNP called with < 99% certainty	2,027,973	69.7
	max-missing	> 0.50	sites	Remove SNPs absent in > 50% of individuals	22,873	0.80
BCFtools	filter -i 'AVG(FMT/DP)'	69.11	sites	Remove sites with depth > (mean + 1.5 X IQR*)	21,511	0.70
VCFtools	maf	< 0.05	sites	Remove SNPs with minor allele frequency < 0.05	2,526	0.09
	min- & max-alleles	< 2 >	sites	Remove SNPs that are not biallelic	2,471	0.08
Plink	r2	< 0.50	sites	Remove SNPs in LD	1,973	0.07
VCFtools	thin	< 100	sites	Retain one SNP per 100 bases	832	0.03

* Interquartile range

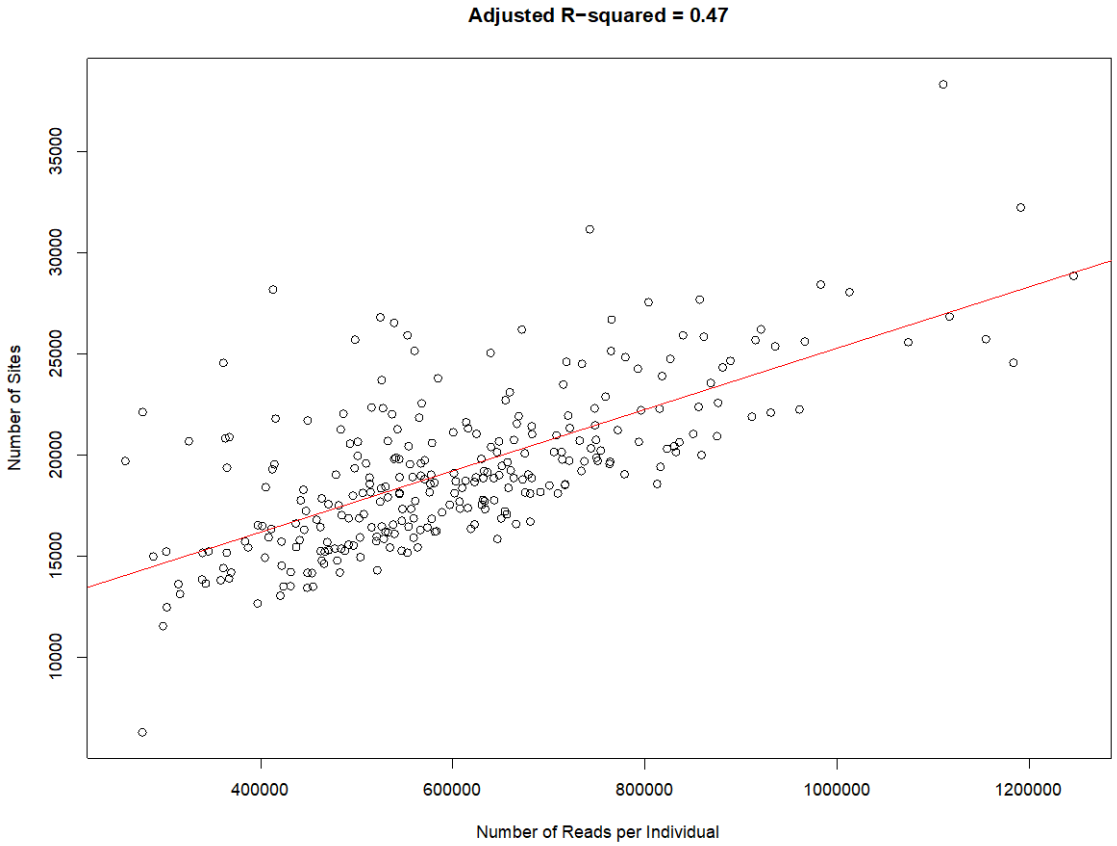

Figure 16. Number of reads per individual after de-multiplexing plotted against number of variant sites prior to filtering for 366 subalpine larch trees.

Data Analysis

In ANGSD, bam files (-bam) were used to calculate genotype likelihoods using the GATK UnifiedGenotyper algorithm (-GL 2). Base alignment quality (-baq 1), the probability of a read base being wrongly aligned, was calculated in order to reduce false SNP calls caused by misalignments around insertions and deletions (Li 2011). Per-sequence map quality scores were adjusted for excessive mismatches (-C 50). Reads with a mapping quality score below Phred 20 (-minMapQ 20) and bases with a quality score below Phred 20 (-minQ 20) were removed. Genotype likelihoods were used to count alleles (-doCounts 1), estimate allele frequencies (-doMaf 1) and infer major and minor alleles (-doMajorMinor 1). A likelihood ratio test statistic was used to filter out sites that were unlikely to be polymorphic (-SNP_pval 5e-2). Finally, SNPs with more than 50% missing data across individuals (-minInd N/2) and SNPs with more than two alleles were removed (-skipTriallelic 1). Minor allele frequency output files were uploaded into R and SNPs were sorted based on the number of individuals sequenced at each locus. SNPs with the largest number of individuals sequenced were preferentially retained while adjacent SNPs within 125 nucleotides were removed. Thus a single SNP was retained per RAD read.

1. Genetic structure

Genetic structure was assessed using two different methods: principal components analysis (PCA) and estimates of pairwise genetic distance. Range-wide samples were pooled and ANGSD was run as previously described with minor modifications (Appendix B). To calculate genotype probabilities, per-site allele frequency was used as a

prior (-doPost 1) with average individual inbreeding coefficients supplied to account for some deviation from Hardy-Weinberg equilibrium (-indF). SNPs were removed if they did not have a minimum minor allele frequency of at least 5% across all samples (-minMaf 0.05). SNPs that were genotyped in the largest number of individuals across all populations were preferentially retained when markers were thinned. In total 18,696 SNPs were retained for the analysis of genetic structure.

Genetic structure was assessed via PCA. In ANGSD, genotype probabilities were assigned at each site for each individual and output in binary format (-doGeno 32). The *ngsCovar* function in ngsTools (Fumagalli et al. 2014) was used to estimate the covariance matrix between individuals based on genotype probabilities. After eigenvalue decomposition of the matrix, the first two principal components were plotted in R.

Genetic structure was also assessed by plotting a dendrogram based on estimates of pairwise genetic distance. Genotype probabilities were output from ANGSD as posterior probabilities (-doGeno 8) and genetic distance was estimated using the *ngsDist* function in ngsTools. Estimates were bootstrapped by randomly sampling 20 blocks of SNPs with replacement 100 times. In R, mean individual pairwise genetic distance was calculated among populations and output in matrix format. Population mean pairwise genetic distances were converted into Newick tree format in FastME (LeFort et al. 2015) and plotted in R.

2. Heterozygosity

To estimate heterozygosity, the folded sample allele frequency (-fold 1) was calculated for each individual using the ANGSD settings described above with minor

modifications (Appendix B). For the unfolded SFS, the x-axis represented all possible allele frequencies (2N + 1 for N diploids) and the y-axis showed the frequency of loci with these allele frequencies. The folded SFS was used because an ancestral sequence was not available to polarize alleles as ancestral versus derived. The folded SFS had [N + 1] allele frequency values on the x-axis.

The 18,696 SNPs that were identified as being variable during the range-wide analysis of genetic structure were included in this analysis (-sites). SNPs were not filtered for polymorphism and fixed sites were retained. Individual SAF were used to estimate the individual SFS in ANGSD (realSFS). For a single individual, the folded SFS only has two values: the number of fixed sites and the number of heterozygous sites. For each individual, observed heterozygosity was calculated as the number of heterozygous loci divided by the total number of loci. Heterozygosity was analyzed in R using simple linear models. Region, population, latitude, longitude and elevation were tested individually as predictors of heterozygosity.

3. Inbreeding coefficients

Individual inbreeding coefficients were calculated within populations using the ANGSD settings described above with minor modifications (Appendix B). On average, 49,313 SNPs were retained per population. Output was saved as log genotype likelihoods in Beagle binary format (-doGlf 3) and inbreeding coefficients were calculated using the *ngsF* function in ngsTools. Reliable starting values were estimated by computing 20 initial searches. A deep search was performed to estimate the most likely inbreeding coefficient for each individual within the population. Individual inbreeding coefficients

were analyzed in R using simple linear models. Region, population, latitude, longitude and elevation were tested individually as predictors of inbreeding.

4. Tajima's D

Population SFS were used to test whether populations were evolving under mutation-drift equilibrium. Tajima's D is a summary statistic that compares the average number of pairwise differences (π) with the expected value (Θ) given the number of segregating (polymorphic) sites. Population SFS were calculated in ANGSD using the settings described above with minor modifications (Appendix B). The folded SFS was calculated with inbreeding coefficients included as priors (-indF). Posterior probabilities of per-site allele frequencies were thus calculated based on individual genotype likelihoods whilst accounting for individual inbreeding coefficients. SNPs identified within populations during the analysis of inbreeding were used to calculate the sample allele frequency (SAF) based on the posterior probabilities for each site (-sites). The SAF was used to estimate the SFS for each population (realSFS). Observed SFS were plotted using R. The population SFS was then used as a prior to compute allele frequency posterior probabilities (-pest) and the test statistic Θ (-doThetas). Output files were indexed (thetaStat do_stat) and a sliding window analysis was performed using a window length of 5000 bp (-win 5000) and a step of 1000 bp (-step 1000). One-sided t-tests were used to test whether estimated values were significantly different from zero across windows. Values were averaged across windows to obtain a mean value of Tajima's D for each population. Related statistics, Faye and Wu's H as well as Zeng's E, were also estimated during this analysis and were reported from the output. Faye and Wu's H is

similar to Tajima's D but focuses on alleles that have reached high frequency in the population and is thus helpful for detecting directional selection. Zeng's E is a sensitive estimator of both selective sweeps and population expansion.

5. Population differentiation (F_{ST})

Population SFS were calculated in ANGSD as described above with minor modifications (Appendix B). The unfolded SAF was specified and monomorphic sites were retained. For each pair of populations, the 2D joint SFS was generated in ANGSD (realSFS) and used as a prior (-sfs) to calculate the joint allele frequency probabilities for each locus. Pairwise F_{ST} values were calculated (-whichFST 1) after indexing the joint SFS (realSFS fst index). Output was used to calculate a weighted global estimate of F_{ST} (realSFS fst stats).

6. Demographic History

SFS calculated for F_{ST} analysis were used for single-population demographic simulations. After SFS were calculated, spectra were folded using a custom Python script written by Dr. David Marques, a post-doctoral researcher at the University of Bern, Switzerland. Folded spectra were used as input for fastsimcoal2 (Excoffier and Foll 2011), a software program that uses a continuous-time coalescent to simulate neutral molecular diversity for a given number of samples drawn from a population with a specified demographic history. Coalescent simulations were used to elucidate the evolutionary history of subalpine larch. For each population, the observed SFS was compared with the expected SFS generated under six different demographic scenarios

(Figure 17). Based on initial output, the ancestral population size was set to 30,000 alleles (i.e. 15,000 diploid individuals) for all populations and scenarios. The scenarios that were tested are as follows:

1. Scenario 1 (S1) described a population of constant size. Only one demographic parameter, the current effective population size, required estimation.
2. S2 described a population that underwent an instantaneous change in size, some number of generations before present. Two parameters required estimation: the current effective population size and the number of generations to resize.
3. S3 described a population that underwent two size changes. Three parameters required estimation: the current effective population size, the intermediate effective population size and the timing of the transition between the current and intermediate population sizes. Going backward, the length of the intermediate period lasted 10 generations.
4. S4 is the same as S3 except the intermediate period lasted 100 generations.
5. S5 is the same as S3 except the intermediate period lasted 500 generations.
6. S6 is the same as S3 except the intermediate period lasted 1,000 generations.
7. S7 is the same as S3 except the intermediate period lasted 2,000 generations.

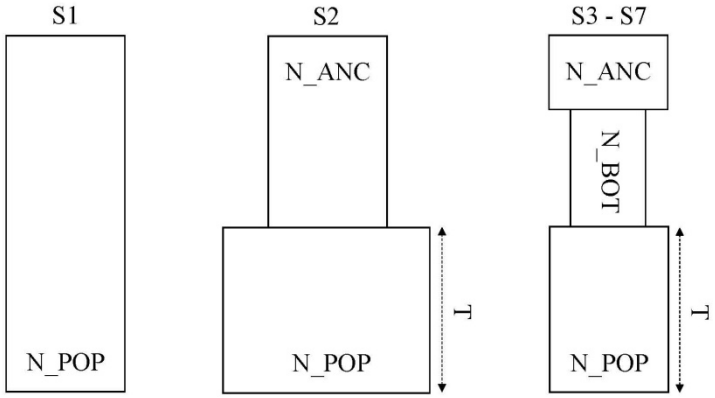

Figure 17. Seven demographic scenarios (S1 – S7) run for each population of subalpine larch required estimation of different demographic parameters (N_POP = current effective population size; N_BOT = intermediate effective population size; N_ANC = ancestral population size = 30,000 alleles; T = time of most recent resize in generations). Note that for scenarios S3 – S7, the period with an intermediate effective population size lasts a different number of generations (10, 100, 500, 1000, 2000).

Note that the length of a generation was not defined. If a generation was assumed to be 100 years, the intermediate population size in scenarios S3 – S7 lasted for 1,000 – 200,000 years. For all scenarios, sample size was defined as two times the number of diploid individuals (i.e. the number of alleles), the inbreeding coefficient was defined as the mean of individual inbreeding coefficients per population and the mutation rate was defined as 2.7e-8 (Hanlon 2018).

Demographic parameters were inferred from the observed SFS using a composite-likelihood method (-M). Initial values were randomly drawn from a defined search range and a conditional maximization algorithm (ECM) was used to estimate parameter values. The ECM algorithm maximizes each parameter in turn while holding the others at their last estimated value until estimates stabilize or the program reaches a maximum number of optimization cycles (-L 50). This process was repeated over 100,000 simulations (-n 100,000) to estimate the folded SFS (-m). In order to ensure that estimates did not reflect a local optimum associated with the initial search values, this process was repeated over 100 independent runs. The most likely parameter values were identified among runs for each population within each scenario. Model fit was used to identify the scenario that best explained the observed data.

To obtain confidence intervals for parameter estimates, population SFS were bootstrapped 100 times (realSFS –boostrap 100). Coalescent simulations were rerun for bootstrapped SFS. For each boostrapped SFS, ten independent runs were used to estimate parameter values starting with parameter estimates from the previously identified best run (--initValues).

Results

RAD-seq Data

Paired RAD-seq data were obtained for 366 individuals. One individual from Mount Frosty (Pop01_22) in Manning Provincial Park, BC, had no reads left after initial quality filtering and was therefore excluded from all further analyses. Overall, library quality was high. After filtering in Stacks, over 90% of sequence data were retained (Table 6). However, a high proportion of data was lost during bioinformatics processing. For example, 49% of data were ultimately identified as PCR duplicates and removed. Ultimately, 28% of data were retained for alignment to a draft of the Siberian larch genome.

After bioinformatics processing and successful alignment, individuals retained a mean of 1,461,749 reads for genotyping. The five libraries had significantly different numbers of reads per individual ($p < 0.001$; $F = 15.16$; Figure 18). Library C577 did not differ from library C578 ($p = 0.48$; t-value = 0.695) but had significantly fewer reads than library C579 (17%; $p < 0.001$; t-value = 4.171) and significantly more reads than C580 (11%; $p = 0.010$; t-value = -2.592) or C581 (10%; $p = 0.018$; t-value = -2.366). Differences should not impact inferences made from the data because samples were randomized within and across libraries.

Genetic Structure

ANGSD identified 38,927 potential SNPs across all populations. After thinning, 18,696 SNPs were retained for the analysis of genetic structure. While the minimum

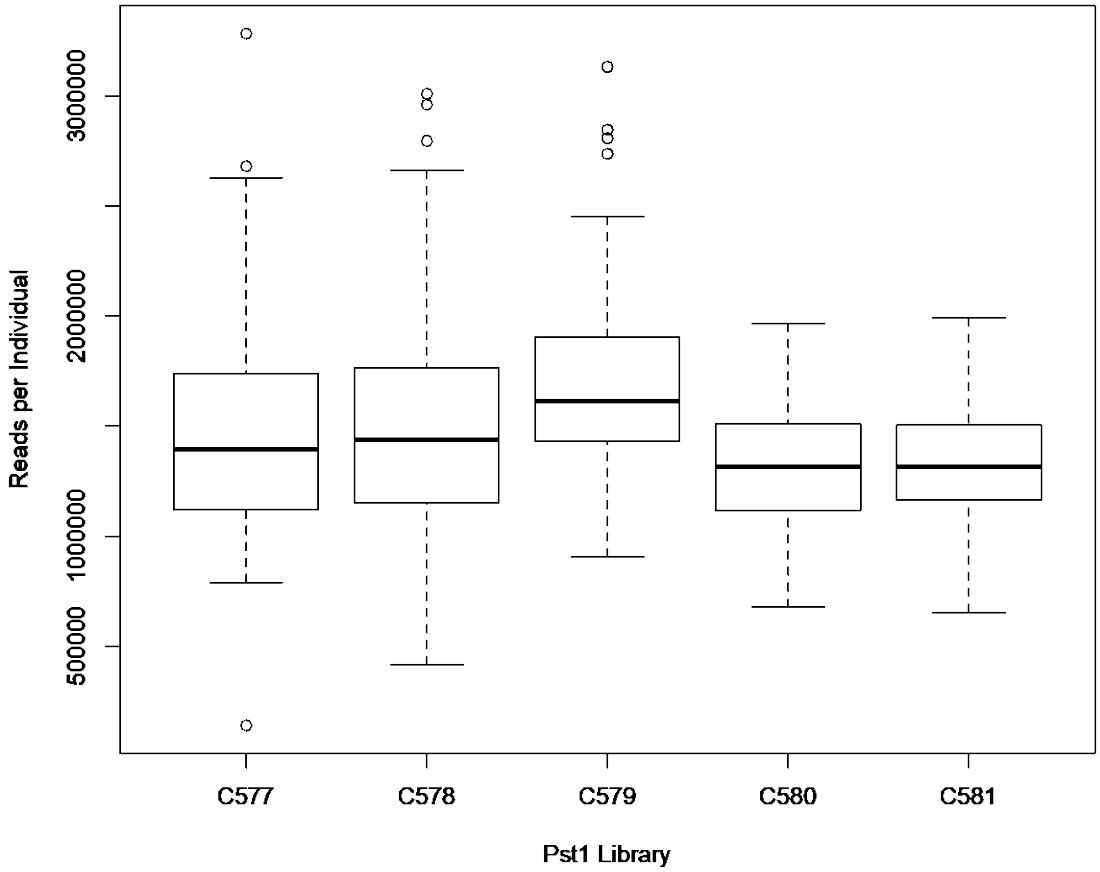

Figure 18. Mean number of reads per individual in the five Pst1 sequencing libraries after alignment.

inter-SNP distance was 125 nucleotides, the average distance between adjacent SNPs within scaffolds was generally much greater (median 311,161), suggesting minimal linkage between markers.

Genetic structure was examined using PCA. Four genetic clusters were observed: the Cascade Range, the southern Rocky Mountains, the central Rocky Mountains and the northern Rocky Mountains (Figure 19). Populations in the central Rockies fell between the southern and northern Rocky Mountain groups. The first principal component (PC) accounted for 3.01% of total genetic variation and the second PC accounted for 1.45% of total genetic variation.

A dendrogram was generated based on mean pairwise genetic distances between populations. Three genetic clusters were identified: the Cascades, the southern Rockies and the Northern Rockies (Figure 20). Populations in the central Rockies clustered with populations in the northern Rockies, although populations from the north did form a monophyletic sub-cluster within this group.

Heterozygosity

Observed heterozygosity (H_O) was calculated using range-wide SNPs identified during the analysis of genetic structure. SNPs were not sequenced in all individuals, however, so the number of SNPs used to calculate H_O varied among individuals (2,374 – 16,317) with a mean of 11,137.

Individual H_O varied between 0.06 and 0.38, with a mean of 0.17 across all individuals. Averaged by population, H_O ranged between 0.15 and 0.20 (Table 8). Region was not a significant predictor of H_O but population was ($F = 2.124$; $p = 0.0104$). When

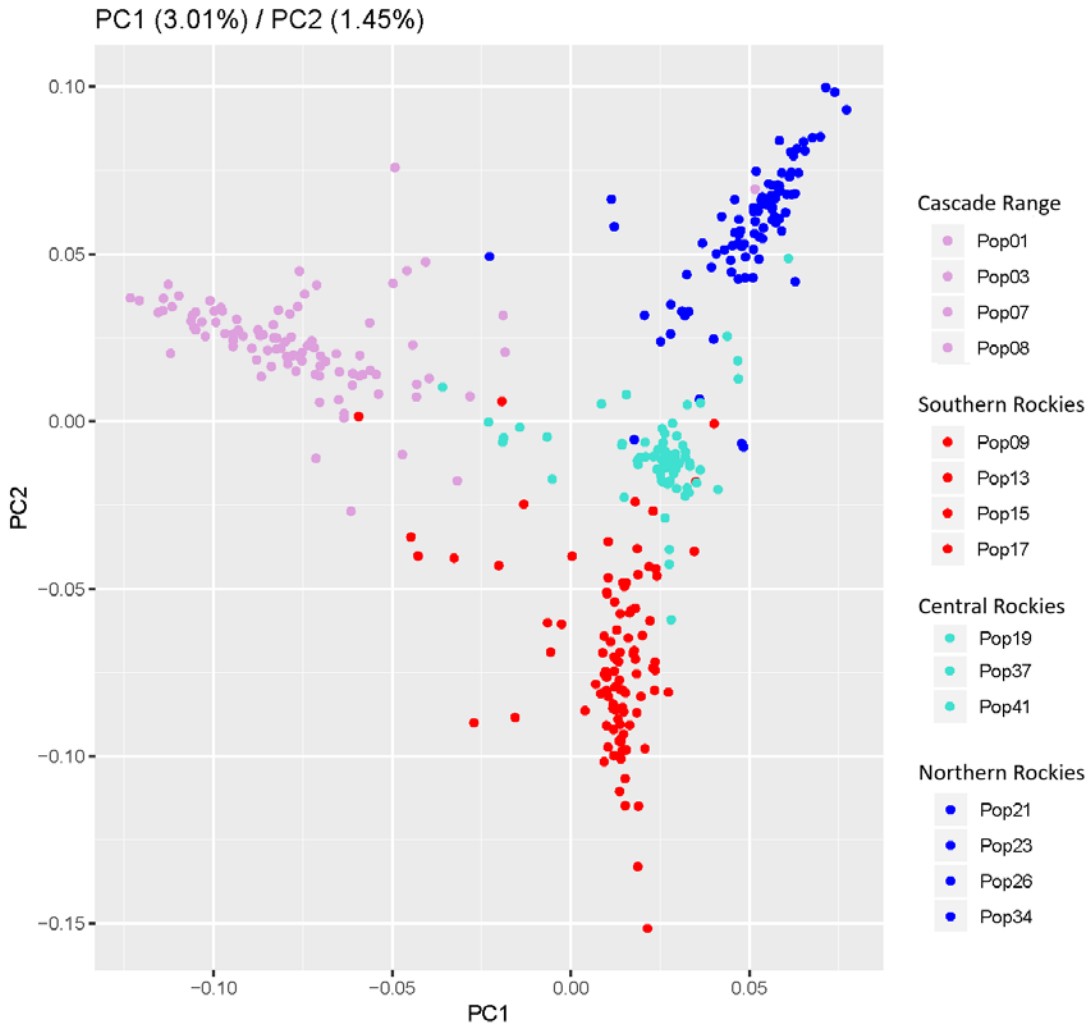

Figure 19. Principal components analysis of genetic variation for 365 subalpine larch trees sampled from 15 populations distributed across the species' natural range.

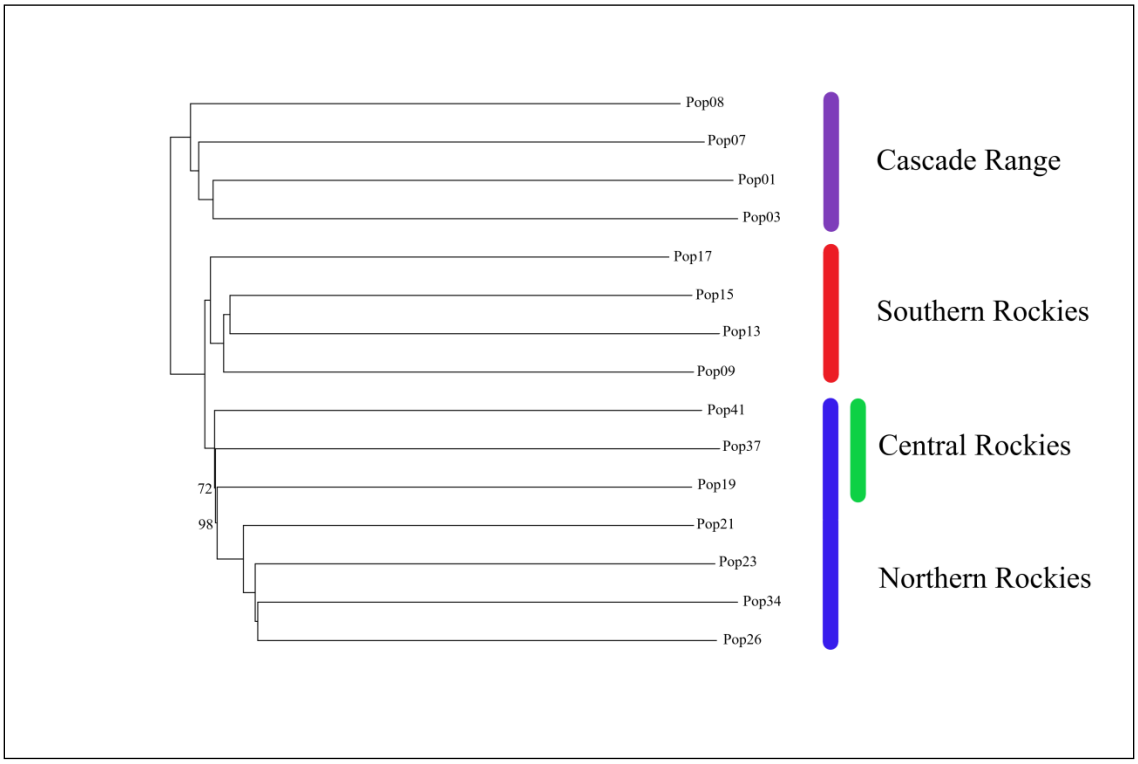

Figure 20. Dendrogram of mean pairwise genetic distances between 15 populations of subalpine larch representing the natural range of the species. Bootstrapping support for all divisions was 100% except for the two splits marked in the figure above.

Table 8. Inbreeding and observed heterozygosity for 15 populations of subalpine larch distributed across the species range.

Region	Population	N	SNPs*	F_{IS}	F_{IS} SD	H_O	H_O SD
Cascade Range	Pop01	24	53,809	0.16	0.08	0.18	0.05
	Pop03	25	63,473	0.20	0.12	0.19	0.08
	Pop07	25	51,980	0.20	0.11	0.20	0.06
	Pop08	25	33,739	0.25	0.12	0.15	0.06
South Rockies	Pop09	25	43,070	0.22	0.08	0.16	0.04
	Pop13	25	57,407	0.21	0.11	0.18	0.06
	Pop15	25	37,426	0.23	0.09	0.16	0.05
	Pop17	23	40,260	0.19	0.13	0.17	0.07
Central Rockies	Pop19	25	46,040	0.20	0.15	0.17	0.07
	Pop37	25	61,239	0.19	0.11	0.19	0.06
	Pop41	25	44,891	0.19	0.10	0.15	0.05
North Rockies	Pop21	25	48,645	0.17	0.12	0.17	0.06
	Pop23	25	44,313	0.20	0.10	0.15	0.05
	Pop26	18	48,179	0.17	0.10	0.15	0.06
	Pop34	25	65,213	0.15	0.09	0.18	0.05

*SNP counts for F_{IS} estimates, not individual H_O estimates

data were analyzed within mountain chains, population was a significant predictor of heterozygosity in the Cascades (F = 3.406; p = 0.0208) but not the Rockies (F = 1.556; p = 0.12). Individual heterozygosity was not predicted by latitude, longitude or elevation.

Inbreeding Coefficients

ANGSD identified an average of 118,747 SNPs per population for the analysis of inbreeding (85,788 – 153,392). After thinning, an average of 49,312 SNPs were retained per population (Table 8). While the minimum inter-SNP distance was 125 nucleotides, the average distance between adjacent SNPs within scaffolds was generally much greater (median 240,000), suggesting minimal linkage between markers.

Individual inbreeding coefficients varied between 0.00 and 0.66, with mean inbreeding across all individuals estimated at 0.19. Averaged by population, values ranged between 0.15 and 0.25 (Table 8). Observed heterozygosity showed a strong negative correlation with inbreeding coefficient (-0.89), as expected, despite having been calculated using different SNPs (i.e. SNPs identified in the range-wide analysis). Region was a significant predictor of inbreeding when the southern Rockies were compared to the northern Rockies if populations in the central Rockies were excluded (F = 6.366; p = 0.0125). Inbreeding was 3.7% higher in the southern Rockies: 21.2 % versus 17.5 % in the northern Rockies. Latitude was a significant predictor of inbreeding (F = 8.824; p = 0.003). When data were analyzed within mountain chains, latitude was a significant

predictor of inbreeding in both the Cascades (F = 6.819; p = 0.01) and the Rockies (F = 6.298; p = 0.0127). In both mountain chains, inbreeding decreased at higher latitudes. Note, however, that these models explained very little of the total variation in the dataset (adjusted R-squared = 0.02 – 0.03) and should thus be interpreted with caution. Inbreeding coefficients were not predicted by population, longitude or elevation.

Tajima's D

Tajima's D was calculated using polymorphic sites identified in the population-level analysis of inbreeding. When folded SFS generated from these SNPs were plotted, they appear to most closely match the model for stationarity, except for Population 26, which shows clear signs of recent expansion at the northern range margin in the Rocky Mountains (Appendix C).

Tajima's D assesses departures from a null model of neutral evolution. If the number of pairwise differences equals the number of segregating sites, then the population is evolving under mutation-drift equilibrium and Tajima's D equals zero. When Tajima's D is positive, indicating there are more pairwise differences than expected between individuals, balancing selection may be acting on the genome or the population may have undergone a recent contraction. When Tajima's D is negative, indicating there are fewer pairwise differences than expected, a selective sweep may have removed variation or the population may have undergone recent expansion. For all populations, average Tajima's D was negative (-0.85 to -0.22), indicating that populations have less variation than expected (Table 9). Since sites were initially filtered to remove SNPs that were not likely to be polymorphic as well as variants present at a

Table 9. Summary of diversity estimators for 15 populations of subalpine larch distributed across the species natural range.

Population	Windows	Tajima's D	t-value	p-value	Fay and Wu's H	t-value	p-value	Zeng's E	t-value	p-value
Pop01	87057	-0.30	-77.7	0.000	0.44	349.7	0.000	-0.47	-2220	0.000
Pop03	104047	-0.30	-85.3	0.000	0.45	388.2	0.000	-0.47	-2269	0.000
Pop07	94576	-0.36	-101.4	0.000	0.42	356.4	0.000	-0.46	-2250	0.000
Pop08	56848	-0.28	-60.7	0.000	0.44	282.7	0.000	-0.46	-1955	0.000
Pop09	73179	-0.30	-72.6	0.000	0.44	318.1	0.000	-0.46	-2114	0.000
Pop13	94056	-0.34	-93.2	0.000	0.43	356.1	0.000	-0.46	-2253	0.000
Pop15	64902	-0.29	-66.6	0.000	0.43	298.9	0.000	-0.46	-2094	0.000
Pop17	85163	-0.55	-147.9	0.000	0.37	303.0	0.000	-0.48	-2219	0.000
Pop19	84106	-0.45	-123.2	0.000	0.39	324.4	0.000	-0.46	-2223	0.000
Pop21	81907	-0.32	-79.9	0.000	0.43	326.1	0.000	-0.46	-2182	0.000
Pop23	67897	-0.24	-53.5	0.000	0.45	304.6	0.000	-0.46	-2065	0.000
Pop26	132830	-0.85	-303.5	0.000	0.34	373.1	0.000	-0.54	-2383	0.000
Pop34	96274	-0.22	-58.7	0.000	0.47	373.8	0.000	-0.47	-2193	0.000
Pop37	99162	-0.27	-77.6	0.000	0.45	385.4	0.000	-0.47	-2242	0.000
Pop41	71572	-0.24	-56.2	0.000	0.45	315.8	0.000	-0.46	-2107	0.000

frequency of less than 5% across the species range, it was unlikely that the observed excess of low-frequency variants represents sequencing or other error. T-tests across windows within populations confirmed that average Tajima's D values were significantly lower than zero. Values for Fay and Wu's H were positive in all populations (0.34 – 0.47) and values for Zeng's E were negative in all populations (-0.54 to -0.46), providing additional evidence that an excess of rare variants was present. When populations were compared qualitatively, Pop26 was identified as a clear outlier. It had the most negative values of Tajima's D and Zeng's E but the smallest value of Fay's & Wu's H. When the SFS for Pop26, on the northern range margin in the Rocky Mountains, was compared with the SFS for Pop01, on the northern range margin in the Cascade Range, the expansion signal was clearly visible (Figure 21).

Genetic Differentiation

Folded SFS generated using all SNPs appeared to most closely match the model for stationarity, including Population 26 (Appendix D). Mean pairwise F_{ST} was 0.18, indicating that there was significant genetic differentiation among populations of subalpine larch (Table 10). Mean F_{ST} values within each region were slightly lower: 0.10 in the Cascades, 0.09 in the southern Rockies and 0.16 in the central/northern Rockies. If the central and northern Rockies are separated, F_{ST} is 0.17 within the central Rockies and 0.09 in the northern Rockies. F_{ST} was slightly higher between regions, especially between populations from the Cascades and populations from the Rocky Mountains (average F_{ST} = 0.23). Populations from the southern Rockies and northern Rockies were differentiated

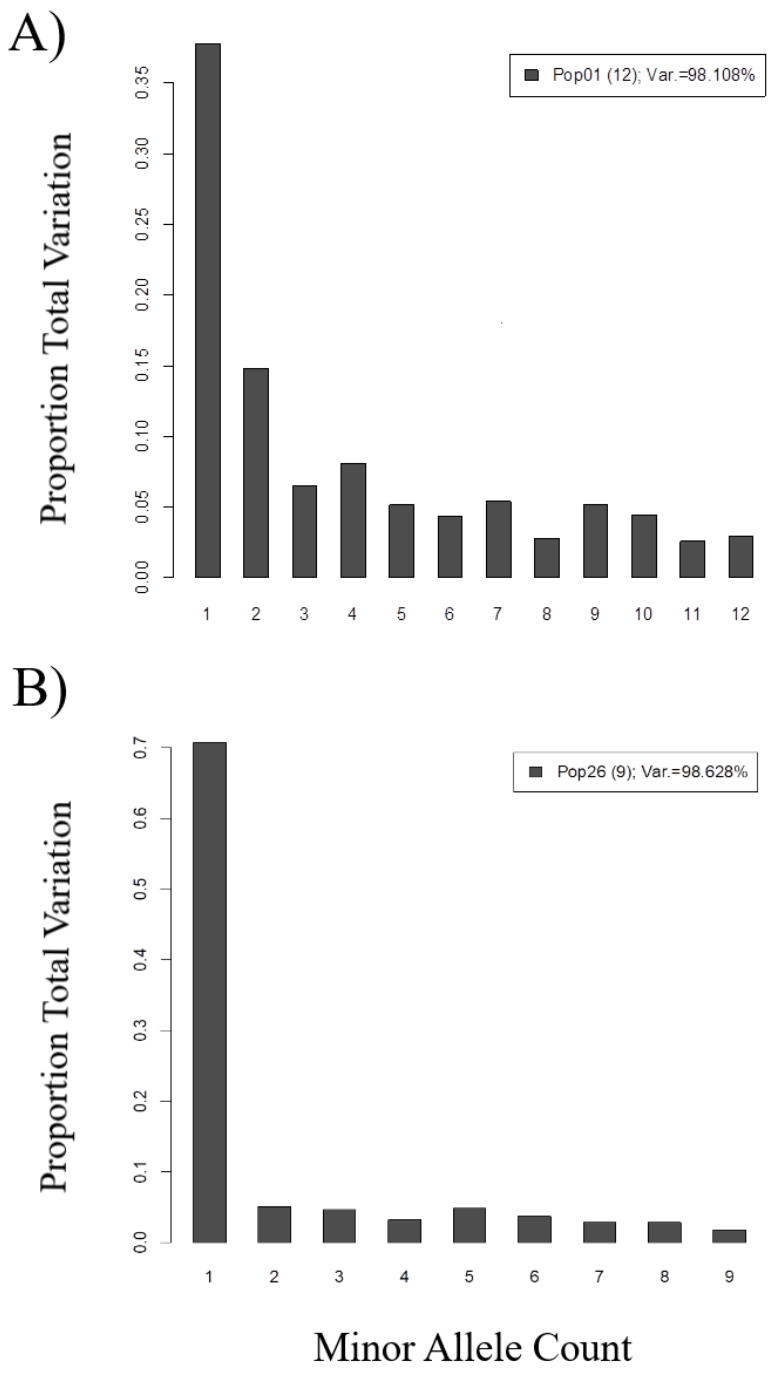

Figure 21. Population-level SFS for two populations of subalpine larch: Mount Frosty on the northern range margin of the Cascade Range (Pop01) and Molar Pass on the northern range margin of the Rocky Mountains (Pop26). Molar Pass has an excess of low-frequency variants, signalling ongoing expansion at the northern range margin.

Table 10. Pairwise global FST for 15 populations of subalpine larch representing the species natural range.

Region	Pop.	Pop01	Pop03	Pop07	Pop08	Pop09	Pop13	Pop15	Pop17	Pop19	Pop37	Pop41	Pop21	Pop23	Pop26
Cascades	Pop01	-													
	Pop03	0.11	-												
	Pop07	0.09	0.08	-											
	Pop08	0.13	0.12	0.09	-										
South Rockies	Pop09	0.22	0.21	0.17	0.20	-									
	Pop13	0.23	0.22	0.18	0.20	0.08	-								
	Pop15	0.22	0.20	0.18	0.20	0.07	0.06	-							
	Pop17	0.23	0.22	0.18	0.20	0.10	0.12	0.12	-						
Central Rockies	Pop19	0.25	0.23	0.20	0.23	0.16	0.17	0.16	0.16	-					
	Pop37	0.23	0.22	0.18	0.21	0.13	0.15	0.14	0.14	0.14	-				
	Pop41	0.27	0.26	0.22	0.24	0.18	0.19	0.18	0.18	0.19	0.17	-			
North Rockies	Pop21	0.25	0.24	0.20	0.23	0.16	0.17	0.16	0.16	0.17	0.15	0.19	-		
	Pop23	0.27	0.26	0.22	0.25	0.19	0.20	0.19	0.19	0.19	0.17	0.21	0.10	-	
	Pop26	0.28	0.27	0.23	0.25	0.19	0.21	0.19	0.19	0.20	0.18	0.22	0.11	0.11	-
	Pop34	0.28	0.27	0.23	0.25	0.19	0.20	0.18	0.19	0.19	0.17	0.21	0.08	0.07	0.09

by an average of 0.18. Populations in the central Rockies were slightly less differentiated from southern populations (0.16) than northern populations (0.18).

Demographic History

Of the seven demographic scenarios considered, a model with an intermediate effective population size (N_E) was preferred for all populations (Table 11). Support for the most likely demographic scenario was often weak, with small differences in AIC values (Appendix E). However changes in population size described by second-ranked models mirrored first-ranked models in 12 out of 15 cases (Appendix F). The demographic model with an intermediate N_E that endured for 2,000 generations was preferred for 10 out of 15 populations. Assuming a generation lasted 100 - 200 years, this equated to an intermediate population size that lasted for 200,000 - 400,000 years. For the remaining populations, the duration of the intermediate phase varied between 10 (Pop01, Pop21, Pop37), 500 (Pop26) and 1000 generations (Pop41). Populations of subalpine larch underwent significant size changes. The demographic scenario that described a constant population size (S1) was ranked as the least likely for 14 out of 15 populations (Appendix E).

Populations of subalpine larch appeared to have changed in size throughout the Pleistocene. The resize to the current N_E was estimated to have occurred between 181 and 2,148 generations ago or, if a generation time of 100 years is assumed, between 18,100 and 214,800 years ago (Table 11). Depending on the length of the intermediate phase, this corresponded to changes in population size that started between 26,100 and 414,800

Table 11. Parameter estimates with bootstrap confidence intervals (CI) for the preferred demographic scenarios of 15 populations of subalpine larch (S1 = constant population size; S2 = instant resize *t* generations ago; S3 = intermediate size change lasting 10 generations that ends *t* generations ago; S4 = intermediate size change lasting 100 generations that ends *t* generations ago; S5 = intermediate size change lasting 500 generations that ends *t* generations ago; S6 = intermediate size change lasting 1,000 generations that ends *t* generations ago; S7 = intermediate size change lasting 2,000 generations that ends *t* generations ago). ΔAIC is the difference between first- and second-ranked models.

Region	Pop.	Scenario	ΔAIC	Ancestral N_E* (CI)	Intermediate N_E* (CI)	Current N_E* (CI)	*t***	Description
Cascades	Pop01	S3	-11	15,000	380 (229 - 292,578)	276,496 (118,577 - 823,715)	830 (489 - 1,053)	bottleneck
	Pop03	S7	-6	15,000	17,245 (16,424 - 21,743)	174,534 (25,331 - 721,414)	258 (24 - 480)	expansion
	Pop07	S7	-31	15,000	34,631 (25,721 - 43,959)	21,394 (21,325 - 23,802)	2,148 (942 - 2,148)	retreat
	Pop08	S7	-5	15,000	32,457 (25,095 - 83,854)	24,017 (19,291 – 26,883)	1,701 (1,115 - 2,145)	retreat
South Rockies	Pop09	S7	-17	15,000	19,858 (17,786 - 27,369)	87,587 (14,065 - 651,153)	345 (624 - 920)	expansion
	Pop13	S7	-4	15,000	17,172 (15,451 - 21,182)	385,644 (79,387 - 826,249)	655 (507 - 814)	expansion
	Pop15	S7	-19	15,000	24,026 (19,068 - 27,023)	22,701 (5,696 - 608,115)	181 (14 - 751)	retreat
	Pop17	S7	-7	15,000	15,543 (15,024 - 20,048)	842,439 (29,564 - 842,439)	336 (94 - 501)	expansion
Central Rockies	Pop19	S7	-431	15,000	573,388 (375,796 - 1,145,201)	15,112 (11,110 - 18,552)	834 (503 - 1,260)	retreat
	Pop37	S3	-1	15,000	41,462 (495 - 268,187)	113,477 (23,230 - 786,215)	251 (203 - 839)	expansion
	Pop41	S6	-2	15,000	6,526 (5,638 - 8,603)	36,663 (27,233 - 351,980)	1,120 (550 - 1,471)	bottleneck
North Rockies	Pop21	S3	-3	15,000	154 (141 - 291)	64,561 (38,381 - 335,334)	967 (616 - 1,196)	bottleneck
	Pop23	S5	-2	15,000	4,483 (3,829 - 7,294)	108,117 (46,558 - 594,889)	886 (654 -1,108)	bottleneck
	Pop26	S7	-18	15,000	7,301 (6,184 - 9,591)	29,393 (25,398 - 505,089)	1,681 (682 - 2,154)	bottleneck
	Pop34	S7	-3	15,000	7,503 (6,977 - 8,347)	123,122 (71,476 - 827,294)	632 (493 -734)	bottleneck

*Effective population size (N_E) is the estimated number of diploid individuals; **Timing of resize to current N_E given as the number of generation

years ago. Although confidence intervals were large, spanning hundreds of generations, only three out of 15 populations had lower bounds that fell within the Holocene Epoch. Thus major changes in population size appeared to have occurred across the species range during the Pleistocene.

Different populations appeared to have experienced different types of population size changes. Bottlenecks were suggested in all four populations from the northern Rockies, one population fragment from the central Rockies (Pop41, Paradise Valley, Glacier National Park, MT) and one population from the northern Cascades (Pop01, Mount Frosty, BC). All of these populations were located on formerly glaciated sites. The average intermediate N_E for these populations was under 10,000 individuals but current N_E rebounded to an average of 120,000 individuals. Throughout the rest of the species range, populations appeared to have experienced either expansion or expansion followed by a more recent retreat. Expansion was observed in three populations from the southern Rockies, one population from the central Rockies (Pop37, Bovin Lake, AB) and one population from the northern Cascades (Pop03, Tiffany Mountain, WA). These populations had an average intermediate N_E of 22,000 individuals but expanded to an average current N_E of 321,000 individuals. Remaining populations underwent expansion followed by retreat. Retreat was observed in both populations from the southern Cascade Range, one population from the southern Rocky Mountains (Pop15, Storm Lake, MT) and one population fragment from the central Rocky Mountains (Pop19, Roman Nose, Idaho). Populations that retreated had an average intermediate N_E of 166,000 individuals but had now shrunk to an average current N_E of 21,000 individuals. Across the whole species range, the average current N_E was estimated to be 155,000 individuals.

Confidence intervals were large for estimates of N_E in most populations. The tightest confidence intervals were observed in three populations with the smallest estimates of current N_E, which were all populations undergoing retreat. Lower bounds around estimates of current N_E ranged between 11,110 (Pop19) and 21,325 (Pop07).

Discussion

Genetic Structure

This study confirms the presence of range-wide genetic structure in subalpine larch. Both a principal components analysis and a dendrogram of mean pairwise genetic distances identified three genetic clusters on the landscape: the Cascade Range, the southern Rocky Mountains and the northern Rocky Mountains. These genetic groupings are geographically sensible and support the results obtained in the previous chapter, despite the fact that genetic data were generated using a different restriction enzyme and different samples, and genotypes were inferred using a different methodology.

In this chapter, RADseq data were generated using the Pst1 restriction enzyme, which has a six-base recognition site and is a subset of the longer Sbf1 enzyme used in the previous chapter. Because Pst1 is shorter it occurs more frequently across the genome, meaning sequencing effort was distributed across a larger number of loci and fewer reads were obtained per locus. In general, low-depth datasets are considered less reliable due to uncertainty that arises from mapping and sequencing errors, as well as from the random sampling of haploid reads from diploid genotypes (Nielsen et al. 2011). Indeed, a very high average error rate (approx. 16%) was observed across 20 replicates due to haploid sampling of diploid genotypes when a filter-based genotyping approach was tested on the Pst1 dataset. It was thus preferable to skip filtering and calculate genotype likelihoods (GLs) based on aligned reads and their associated sequencing quality and mapping scores (Korneliussen et al. 2014). SNP variants and genotypes were called by combining GL information from multiple individuals with other priors, such as the inferred distribution of allele frequencies. Genotyping uncertainty was directly

incorporated into analyses of genetic structure. Thus despite the differences in the two approaches, consistency in clustering output provides high confidence that these results accurately reflect patterns of neutral genetic variation across the landscape.

The Pst1 dendrogram of genetic distances grouped populations in the central Rocky Mountains with populations in the North. The Sbf1 dendrogram from the previous chapter showed that populations in the central Rockies clustered most closely with populations in the South. There are several factors that could explain this difference. First, the Sbf1 dataset retained fewer than 800 SNPs while over 18,000 were retained in the Pst1 dataset. Random sampling of loci could have affected calculations of genetic distance in the Sbf1 dataset. In theory, the larger number of Pst1 variants should provide a more accurate estimate of genetic distance. Second, Sbf1-digested reads underwent single-end sequencing while Pst1-digested reads underwent paired-end sequencing, and could thus be filtered to remove PCR duplicates. During the PCR step of RAD-seq library preparation, stochastic processes can cause one allele to be amplified at a higher rate than the other allele for a given locus in an individual (Andrews et al. 2016). Genotyping errors result when PCR duplicates cause heterozygotes to appear as homozygotes or when alleles that contain PCR errors appear to be true alleles. In the Pst1 dataset, 49% of reads were identified as PCR duplicates and removed, demonstrating the importance of this issue. Although PCR duplicates should not systemically favour one allele over another, datasets with fewer loci are more likely to be biased. PCR duplicates in the Sbf1 dataset could not be removed and could therefore have biased allele frequencies. Finally, although loci with > 50% missing data were removed from both datasets, missing data were treated differently in the two calculations of genetic distance.

For the Sbf1 dataset, genetic distance was calculated in R using the *provesti.dist* function from the poppr package (Kamvar et al. 2014). This function uses a discrete dissimilarity matrix to return the number of observed differences divided by the number of possible differences. Missing data is not imputed for this analysis. For the Pst1 dataset, genetic distance was calculated using the *ngsDist* function in ngsTools (Fumagalli et al. 2016). This program was designed for low-depth NGS data and takes genotype uncertainty into account. Thus some genotypes would have had low likelihood but would not have been excluded. Given that the Pst1 dataset includes more loci, was filtered for PCR duplicates, and incorporates genotyping uncertainty into estimates of genetic distance, the relationships in the Pst1 dendrogram are more likely to be accurate. Regardless, the differences observed here are informative in their own way: populations in the central Rocky Mountains that are geographically intermediate between southern and northern clusters also show evidence of being genetically intermediate. This result fits within our current understanding of the biogeographic history of subalpine larch.

Biogeographic History

Subalpine larch has a biogeographic history similar to other conifers in western North America. During the Pleistocene, glacial advances forced a southward range contraction and fragmented the species distribution. Both the Sbf1 and Pst1 dendrograms of genetic distance identify the split between the Cascade Range and the Rocky Mountains as the oldest genetic division within the species. This result is supported by large F_{ST} values between populations from these two geographic regions (average 0.23). However it is unlikely that populations remained isolated in separate refugia throughout

the Pleistocene, as occurred for many other tree species (Critchfield 1984; Jaramillo-Correa et al. 2009). First, divisions within and between mountain chains all occur around the same time. The topology of the Pst1 dendrogram is shallow at the time of divergence and the split between the Cascades and the Rockies is closely followed by the split between the southern and northern Rockies. Second, average F_{ST} between the Cascade Range and the southern Rocky Mountains (0.20) is not much higher than average F_{ST} between the southern and northern Rocky Mountains (0.18), the central and northern Rocky Mountains (0.19) or even isolated fragments within the central Rockies [e.g. F_{ST} 0.19 between Pop19 (Roman Nose, Idaho) and Pop41 (Paradise Valley, Montana)]. This may indicate that divergence between the Cascade Range and the southern Rocky Mountains occurred shortly before subalpine larch began to expand its range northward, perhaps during the Vashon stade (18 – 13 kya).

As the climate warmed and the glaciers retreated, subalpine larch was able to migrate northward and to higher elevation. A climatically suitable refugium for three of subalpine larch's current associates—Engelmann spruce, subalpine fir and whitebark pine—likely existed on the Columbia Plateau 21,000 years ago (Roberts and Hamann 2015). If subalpine larch survived the Pleistocene in refugia on or around the Columbia Plateau, it would likely have been well-positioned to expand its range. Whitebark pine macrofossils discovered in early-Holocene sediment deposits at Castle Peak in the Coast Mountains suggests that this formerly glaciated site may have been invaded as early as 10,000 years ago (Clague and Mathewes 1989). Although the biogeographic history of subalpine larch differs somewhat from whitebark pine due to the fact that the latter species' seeds are dispersed by a bird, Clark's nutcracker, parallels can certainly be

drawn. Along with whitebark pine, cold-hardy subalpine larch could have been an early colonizer of the new subalpine expanses that opened at the end of the Pleistocene. Tajima's D and Zeng's E, two summary statistics that assess departures from neutrality, support a history of relatively recent expansion. Both statistics were negative in all populations of subalpine larch, meaning fewer pairwise differences were observed between individuals than were expected based on the number of segregating sites in the population. Selective sweeps can lead to a loss of genetic diversity but should be localized within the genome. Genome-wide loss of diversity is more likely the result of maintaining a small population size over a long period of time. Low-frequency variants are accumulated during subsequent population expansion. Finally, the preferred demographic scenarios obtained from coalescent simulations support a history of expansion for all populations. Six populations best fit a model that described a bottleneck followed by expansion, five populations best fit a model that described step-wise expansion and four populations appear to have experienced expansion followed by retreat.

Coalescent simulations revealed different demographic histories for populations across the species range. Six populations located on formerly glaciated sites in BC, Alberta and Montana had very small intermediate N_E values, averaging only 4,000 individuals, but rebounded to an average current N_E of 120,000 individuals. Bottlenecks have most likely the result of colonization by a relatively small number of founders. High current N_E likely reflects increased diversity due to migration from populations in the core of the range, as well as a high degree of connectivity between populations. However it is not clear that parameter estimates are accurate. It is possible that populations in the

northern Rockies are still connected but Pop41, in Glacier National Park, is now an isolated fragment. The fact that this isolation is not yet reflected in the estimate of current N_E could be attributable to the long lifespan of subalpine larch. Trees can live to be over 1,000 years old and will continue to contribute to the gene pool throughout their lives, meaning populations represent many overlapping generations. Indeed, the presence of overlapping generations may affect the accuracy of demographic parameter estimation in coalescent simulations. For example, bottlenecks were estimated to have ended between 632 and 1,681 generations ago. If a generation time of 100 years is assumed, current N_E was achieved between 63,200 and 168,100 years ago, tens of thousands of years before the Cordilleran ice sheet retreated and these populations were established in their current locations. Thus estimates for the timing of population size changes may not be reliable in this study. Simulations have found that the timing of recent expansion events can be somewhat reliably estimated from short-read sequencing data for simple demographic scenarios even when some degree of error is incorporated (Elleouet and Aitken 2018). However overlapping generations were not considered.

Migration pathways can be inferred from the lineages of modern populations. Given that Pop08, from Windy Pass, is basal on the dendrogram of genetic distance, it appears that the Cascade Range was colonized from a southern refugium. The story in the Rocky Mountains may be more complicated. Pop17, at Holland Pass, MT, and Pop41, in Paradise Valley, Glacier National Park, MT, are basal within the southern and northern clades, respectively. Pop17 is the most northerly population in the southern clade and Pop41 is the most southerly population in the northern clade, suggesting that populations in the Rocky Mountains may have been colonized out of a central refugium. Within all

three genetic groups, relationships between populations reflect a clear pattern of isolation by distance. While these patterns likely also reflect gene flow after expansion, they currently provide the best evidence for the post-glacial migration routes of subalpine larch.

Genetic Diversity

Theory predicts that rapid population growth and enhanced genetic drift at expanding range margins will reduce genetic variation, creating clines of high to low genetic diversity between core and range-front populations (Peischl et al. 2013). For many conifer species, northern populations established since the LGM tend to have less diversity than southern populations, with little geographic difference among populations (Critchfield 1984; Jaramillo-Correa et al. 2009). However such clines are not present in subalpine larch; heterozygosity does not vary significantly across the range. This has also been observed in eastern white pine (*Pinus strobus*) and was attributed to a gradually advancing expansion front and/or frequent long distance dispersal (LDD) events (Nadeau et al. 2015). Other empirical studies have also questioned the importance of founder effects in the range expansions of long-lived species. Norway spruce (Piotti et al. 2009) and European larch (Pluess 2011) show high levels of genetic variation, which are maintained at range fronts due to long-distance dispersal events, including wind pollination. Furthermore, young trees that succeed in establishing themselves at range margins will not reproduce immediately, meaning that source populations will contribute disproportionately to expansion fronts over long periods of time (Elleouet and Aitken 2019). Subalpine larch is likely to exhibit these features given that it is a wind-pollinated

species that requires at least a century to become sexually mature (Arno and Habeck 1972). Thus post-glacial expansion in subalpine larch most likely occurred as a slow, highly connected wave.

Genetic diversity is low in subalpine larch. Heterozygosity in populations of subalpine larch ranged between 0.15 and 0.20, indicating that this species may have a problem with overall low levels of genetic diversity. Heterozygosity is lower in this study than it is in most other studies of conifers, especially those that used isozymes (Hamrick et al. 1979; Hamrick et al. 1992) and microsatellites (White et al. 2007), including a previous study on 19 populations of subalpine larch (H_O = 0.389; Khasa et al. 2006). However this is most likely attributable to marker type. SNP-based estimates of heterozygosity in populations of *Pinus strobus* ranged between 0.26 and 0.30, which is higher than values observed in subalpine larch (Nadeau et al. 2015). Even *Pinus monticola*, which does exhibit clines in genetic diversity along south-north gradients in western North America, maintained heterozygosity levels above 0.22 in its most genetically depauperate northern populations (Nadeau et al. 2015).

Low diversity is most likely the result of genetic drift acting within isolated fragments. As climate warmed during the Holocene, subalpine larch was extirpated from lower-elevation peaks, creating discontinuities between patches of suitable habitat within its range. Within fragments, genetic drift removes additive genetic variation at a rate that is inversely proportional to a population's effective size, which can lead to genetic erosion in small populations (Aguilar et al. 2008). This likely explains the low levels of genetic diversity within populations of subalpine larch, as well as the high levels of genetic differentiation between them (average F_{ST} = 0.18). The combination of low or

absent gene flow with strong genetic drift acting in isolated populations leads to genetic divergence (Amos and Balmford 2001). Estimates of F_{ST} obtained in this study roughly align with the results of an earlier study on subalpine larch, which estimated an average F_{ST} of 0.15 between 19 populations from the northern portion of the species range (Khasa et al. 2006). High F_{ST} values are common in conifers with fragmented distributions, such as *Picea chihuahuana* (F_{ST} = 0.248; Ledig et al. 1997), *Pinus pinceana* (F_{ST} = 0.152; Ledig et al. 2001), *Pinus lagunae* (F_{ST} = 0.188; Molina-Freaner et al. 2001), *Pinus muricata* (F_{ST} 0.160; Molina-Freaner et al. 2001) and *Picea breweriana* (F_{ST} = 0.152; Ledig et al. 2005), as well as *Picea engelmanii*, which has a high degree of fragmentation in the southern portion of its range (F_{ST} = 0.147; Ledig et al. 2006). Thus high F_{ST} values between regions and populations within regions, especially isolated fragments in the central Rocky Mountains, likely reflect a history of fragmentation and drift.

Mating among close relatives is more common in small populations. In this study, populations had average inbreeding coefficients that ranged between 0.15 and 0.25, with the highest values in southern populations. An earlier study found that inbreeding within 19 populations of subalpine larch ranged between ranged between 0 and 0.21, averaging 0.07, with no clear geographic clines (Khasa et al. 2006). However these populations were all located in the northern Rockies, where the lowest inbreeding coefficients were detected in this study. Values obtained in this study are closer to those obtained from populations of *Larix gmelinii*, which averaged 0.27 (Larionova et al. 2004). It is also possible that estimates of inbreeding coefficients are somewhat inflated in this study. A high proportion of missing data (up to 50%) means that many alleles were called based on genotype likelihoods, skewing allele frequencies toward the most common allele at a

locus. However the use of over 18,000 SNPs may also provide a genome-wide assessment of inbreeding that the previous study, which relied on seven microsatellite markers, could not. Mating between close relatives is more likely in small populations and can lead to inbreeding depression. Most conifers rely on embryo abortion to deter selfing and avoid maladaptation due to inbreeding depression (Kärkkäinen and Savolainen 1993). However a high frequency of matings between close relatives can lead to the purging of deleterious alleles, which mitigates the negative effects of inbreeding depression. To assess whether or not subalpine larch is suffering from inbreeding depression, the fitness of outbred and inbred progeny would have to be compared.

In conclusion, subalpine larch has experienced a great deal of change over the Pleistocene and Holocene Epochs but is now facing a serious challenge. Range contraction during the Pleistocene may have led to an initial loss of genetic diversity in this species but the problem has likely been exacerbated by Holocene warming and altitudinal retreat. Coalescent simulations identified four populations that have experienced recent contraction: two populations from the southern Cascades, one population from the southern Rocky Mountains (Pop15, Storm Lake, MT) and one population fragment from the central Rocky Mountains (Pop19, Roman Nose, Idaho). These populations had an estimated average intermediate N_E of 166,000 individuals but have now shrunk to an estimated average current N_E of 21,000 individuals. As sustained, directional climate change proceeds, populations are likely to continue shrinking. Reductions in effective population size increase the strength of genetic drift, reduce the potential for natural selection to act, and increase the likelihood that deleterious alleles will reach fixation. Low genetic diversity further reduces a population's ability to

respond to selection. High levels of inbreeding may further increase subalpine larch's potential for maladaptation if they result in inbreeding depression. This could be an iterative process. Together, genetic drift and inbreeding depression in small populations contribute to an "extinction vortex", whereby small populations have reduced fitness, which leads to further reductions in population size, and so on until extinction (Gilpin and Soulé 1986). Low diversity, high inbreeding and restricted gene flow all call into question the future potential of subalpine larch for survival in a changing climate.

CHAPTER 4: LOCAL ADAPTATION FOR COLD TOLERANCE

Introduction

Spatial patterns of genetic variation reflect a complex history of dispersal, neutral evolutionary processes and natural selection. This complexity is difficult to capture when setting management and conservation priorities. Management strategies based on genetic markers typically seek to conserve the most genetically divergent groups and/or those groups with high levels of genetic diversity. Both criteria assume that neutral genetic variation reflects variation in ecologically important traits, and that the retention of adequate standing genetic variation will ensure future adaptability. However, these approaches ignore the fact that functional divergence may occur via small, localized changes in the genome. In salmon, a single mutation determines run-time, a complex physiological and behavioral phenotype that has ecosystem-level implications due to salmon's keystone status in the food web (Prince et al. 2017). In European crows, striking phenotypic differences are maintained by divergent selection acting on genes that represent less than 1% of the transcriptome (Poelstra et al. 2014). Such small genetic differences may be lost when conservation focus is on overall genetic diversity. In this study, cold tolerance was assessed in 18 populations of subalpine larch in order to identify local adaptation within a key trait for this timberline species.

Subalpine larch (*Larix lyallii* Parl.) is a deciduous conifer that grows within a restricted range in the North Cascade Range and Rocky Mountains of western North America (Arno and Habeck 1972). A poor competitor in mixed stands, subalpine larch has carved out a niche above the altitudinal limits of other tree species in the transitional

zone between the forest and alpine tundra biomes (1,500 – 3,000 m). This is a harsh environment. Average temperature at sites occupied by subalpine larch stays below freezing for more than half the year. The growing season, defined as the period over which average temperature is greater than 5.6 ˚C, lasts approximately 90 days (Arno and Habeck 1972). Even at the southern range margin, average temperature only exceeds 10 ˚C for two months of the year. At the northern range margin, average summer temperature does not usually exceed 10 ˚C. Although the exact mechanisms are not understood, subalpine larch is able to balance its growth budget at a much lower average temperature than other timberline trees, enabling it to survive in colder habitats than its competitors. Adaptation to extreme cold has allowed subalpine larch to occupy a habitat that provides access to plentiful light, a key resource for a deciduous species that must produce new needles every spring before achieving further carbon gains (Gower and Richards 1990).

Several physiological adaptations allow subalpine larch to thrive at timberline. First, deciduousness is thought to contribute to subalpine larch's overall hardiness. Larches differ strikingly from most other Pinaceae in that they have evolved a deciduous habit. Each autumn their needles senesce and fall, leaving their branches bare over winter. Deciduousness is adaptive in that it removes the possibility of damage to leaves by snow and ice, as well as winter water loss through stomata (Berg and Chapin 1994). Second, subalpine larch trees can avoid winter desiccation by isolating apical and lateral buds from the xylem tissue (Richards and Bliss 1986). Winter desiccation is a serious problem for high-elevation species that experience strong insolation while their roots remain frozen. Desiccation does not seem to affect subalpine larch as severely as it does

other species. At their study site in Marmot Valley, AB, Richards and Bliss (1986) found that subalpine larch suffered significantly less winter damage than three sympatric conifer species, sustaining two thirds less damage than the second hardiest species, subalpine fir (*Abies lasiocarpa*). Due to this adaptation, subalpine larch trees can maintain an upright growth form at sites where other conifers are only present as krummholz, which is only possible when buds remain undamaged. Finally, subalpine larch trees are able to tolerate extremely low water potential in their buds (Richards and Bliss 1986). This allows the buds to dehydrate over winter, thus avoiding cellular freezing. These and other adaptations allow subalpine larch to thrive at timberline.

At present little is known regarding local adaptation across populations of subalpine larch. However, it seems likely that local adaptation might exist for cold tolerance traits. Conifers in western North America exhibit clines in cold tolerance that allow them to alternate between active growth and winter dormancy in synchronization with the annual climatic cycle (Howe et al. 2003; Alberto et al. 2013). This growth cycle involves the cessation of growth followed by budset in late summer, cold acclimation in autumn, endodormancy and the attainment of maximum cold hardiness over winter, release from endodormancy, and deacclimation, growth initiation and bud flush in the spring. Subalpine larch begins growth when air temperatures rise to approximately 4 °C at the end of May (Arno and Habeck 1972; Worrall 1993). Cone and needle growth begin when branches are no longer covered by snow. Cone buds swell and clusters of needles start to emerge from short-shoot branches a few days later. Pollen cones mature as early as April although pollination does not occur until June, when female cones open (Carlson 1965). Needle elongation is usually finished by July. It is followed by a rapid burst of shoot

elongation that ceases by early August. Female cones mature at the end of August and drop their seed in early September. As photoperiod shortens in autumn, leaf abscission and bud dormancy are triggered. Needles begin to turn yellow and by the end of September trees will be completely golden. Cold injury is most likely to occur during phases of active growth. Therefore, most cold injury in natural stands occurs as a result of late spring frosts or early fall frosts (Timmis et al. 1994; Cannell and Smith 1984; Cannell et al. 1985a,b).

Adaptive differences in cold tolerance were examined using 18 populations of subalpine larch grafted *ex situ* at the Kalamalka Forestry Centre in Vernon, BC. This common garden planting is a valuable resource because it is extremely difficult to study phenology in natural populations of subalpine larch that are inaccessible for most of the year. Cold injury was measured in winter, spring and autumn. Correlations with environmental variables associated with parent trees and genetic variation were examined in order to elucidate the drivers of this important trait.

Methods

In this chapter, local adaption was examined using three different types of data: climate data, phenotypic data and genomic data. Cold tolerance phenotypes were assessed for 100 subalpine larch trees representing 18 populations from the northern portion of the species' range (Figure 22).

1. Climate data

Climate normals (1961 – 1990) were obtained from ClimateWNA (Wang et al. 2016) for 18 populations of subalpine larch based on the latitude, longitude and elevation of each population (Table 12). Nineteen annual climate variables and one seasonal climate variable were selected (Table 13) based on their biological relevance and their utility in previous studies of local adaptation in conifers (Yeaman et al. 2016; MacLachlan et al. 2017; Lotterhos et al. 2018). Climate variables were standardized in the R statistical environment by subtracting the mean and dividing by the standard deviation of each variable (R Core Team 2019).

A discriminant analysis of principal components (DAPC) was used to identify climatic regions across the landscape. A k-means clustering function from the adegenet package, *find.clusters*, identified the optimal number of climatic regions as two (Jombart 2008; Jombart and Ahmed 2011). Clusters were identified using four principal components (PCs) and one discriminant function, which minimized the Bayesian Information Criterion (BIC). To avoid overfitting, the *optim.a.score* function was run with 100 simulations. Two PCs and one discriminant function were retained to fit the final model using the *dapc* function. In total, 81% of environmental variation was

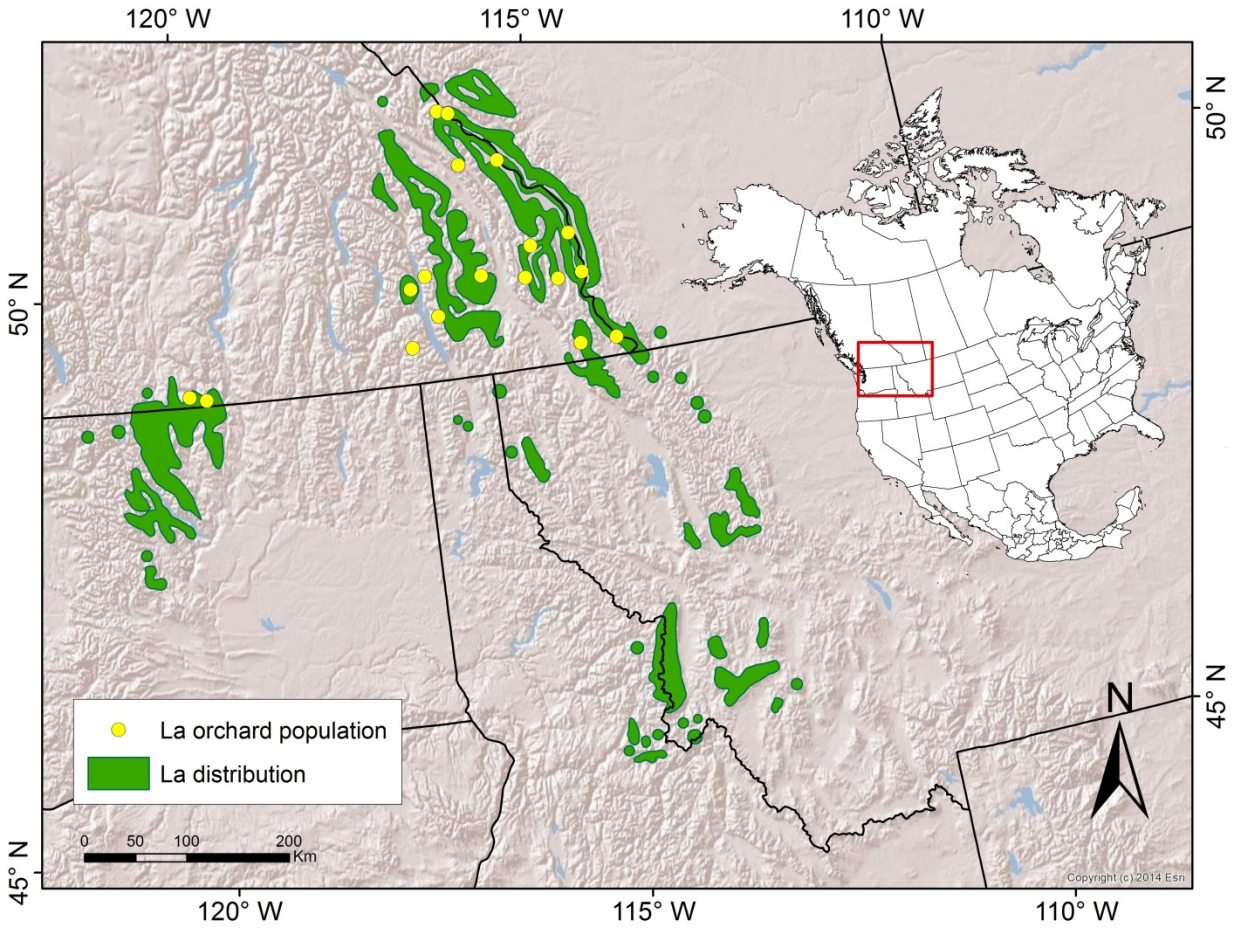

Figure 22. Origins of populations of subalpine larch (La) grafted *ex situ* at the Kalamalka Forestry Centre in Vernon, BC.

Table 12. Locations of 18 populations of subalpine larch from the Canadian portion of the species' range and the number of samples grafted *ex situ* at the Kalamalka Forestry Centre that were measured for cold tolerance (N).

Population	Location	Latitude	Longitude	Elevation	N
AL01	Baldy Mountain	49° 19'	117° 04'	1981	6
AL02	Burdett Peak - Gray Pass	49° 34'	116° 40'	2134	6
AL03	Mount Kaslo	49° 55' 683"	116° 46' 86"	2149	6
AL04	Fletcher Creek	49° 49' 803"	116° 59' 694"	2012	5
AL05	Inverted Ridge	49° 09' 433"	114° 49' 595"	2164	6
AL06	Sunkist Mountain	49° 09' 722"	114° 20' 572"	2210	5
AL07	Racehorse Pass	49° 46' 346"	114° 39' 644"	2210	5
AL08	Mount Kuleski	49° 44' 766"	114° 59' 759"	2179	6
AL09	Mount Dingley	49° 47' 834"	115° 25' 872"	2195	5
AL10	Mount Gass	50° 07' 849"	114° 45' 770"	2256	6
AL11	Mount Mike	50° 03' 962"	115° 17' 833"	2377	6
AL12	Luxor Pass - Mount Crook	50° 51' 454"	116° 07' 336"	2164	5
AL13	Mount Assiniboine	50° 51' 260"	115° 34' 936"	2210	6
AL14	Mount Bradford	49° 52' 224"	116° 01' 463"	2454	6
AL15	Twin Buttes	49° 05' 36"	120° 06' 45"	2270	6
AL16	Lake O'Hara	51° 21'	116° 19'	2200	6
AL17	Moraine Lake	51° 19'	116° 10'	2200	4
AL20	Harry Lake	49° 03'	119° 53'	2200	5

Table 13. Nineteen annual climate variables and one seasonal climate variable were used to study local adaptation in subalpine larch.

Environmental Variable (unit)	Abbreviation
Mean annual temperature (˚C)	MAT
Mean warmest month temperature (˚C)	MWMT
Mean coldest month temperature (˚C)	MCMT
Continentality (MWMT minus MCMT) (˚C)	TD
Minimum temperature in autumn (˚C)	Tmin_at
Mean annual precipitation (mm)	MAP
May to September precipitation (mm)	MSP
Annual heat-moisture index (MAT + 10)/(MAP/1000) (˚C/µm)	AHM
Summer heat-moisture index (MWMT/(MSP/1000) (˚C/µm)	SHM
Degree-days below 0˚C, chilling degree days	DD_0
Degree-days above 5˚C, growing-degree days	DD_5
Number of frost-free days (days)	NFFD
Frost-free period (days)	FFP
The day of the year on which FFP begins (Julian date)	bFFP
The day of the year on which FFP ends (Julian date)	eFFP
Precipitation as snow between August and July (mm)	PAS
Extreme minimum temperature over 30 years (˚C)	EMT
Extreme maximum temperature over 30 years (˚C)	EXT
Hargreaves reference evaporation (mm)	Eref
Hargreaves climatic moisture deficit (mm)	CMD

retained in the final model. DAPC posterior probabilities were used to assign populations to climatic regions for further analysis.

Individual climate variables were summarized by region. A simple linear model with region as a predictor was fit to each climate variable in order to elucidate climatic differences between regions.

2. Phenotypic data

2.1 Sampling

Cold tolerance in subalpine larch was assessed using 100 trees grafted *ex situ* at the Kalamalka Forestry Centre in Vernon, BC. Between four and six trees were sampled from each of 18 populations distributed across the Canadian portion of the species' range (Table 12; Figure 22). Sampling was carried out as previously described (Khasa et al. 2006). Cold tolerance was assessed in each of three seasons (winter, spring and autumn) over two consecutive years.

Winter sampling was carried out during the last week of December. Temperature data were obtained from a sensor (Hobo Micro Station, Onset Computer Corporation, Bourne, USA) located in the Assisted Migration and Adaptation Trial. Hourly data recorded by the sensors were used to identify daily maximum and minimum temperatures. In both years, night-time temperature first dropped below zero at the end of October (Figure 23). Maximum daily temperature stayed below 0 ˚C for 19 and 16 days prior to sampling in 2015 and 2016, respectively. Minimum temperature dropped below zero for 42 and 41 days prior to sampling in 2015 and 2016, respectively. Thus in both

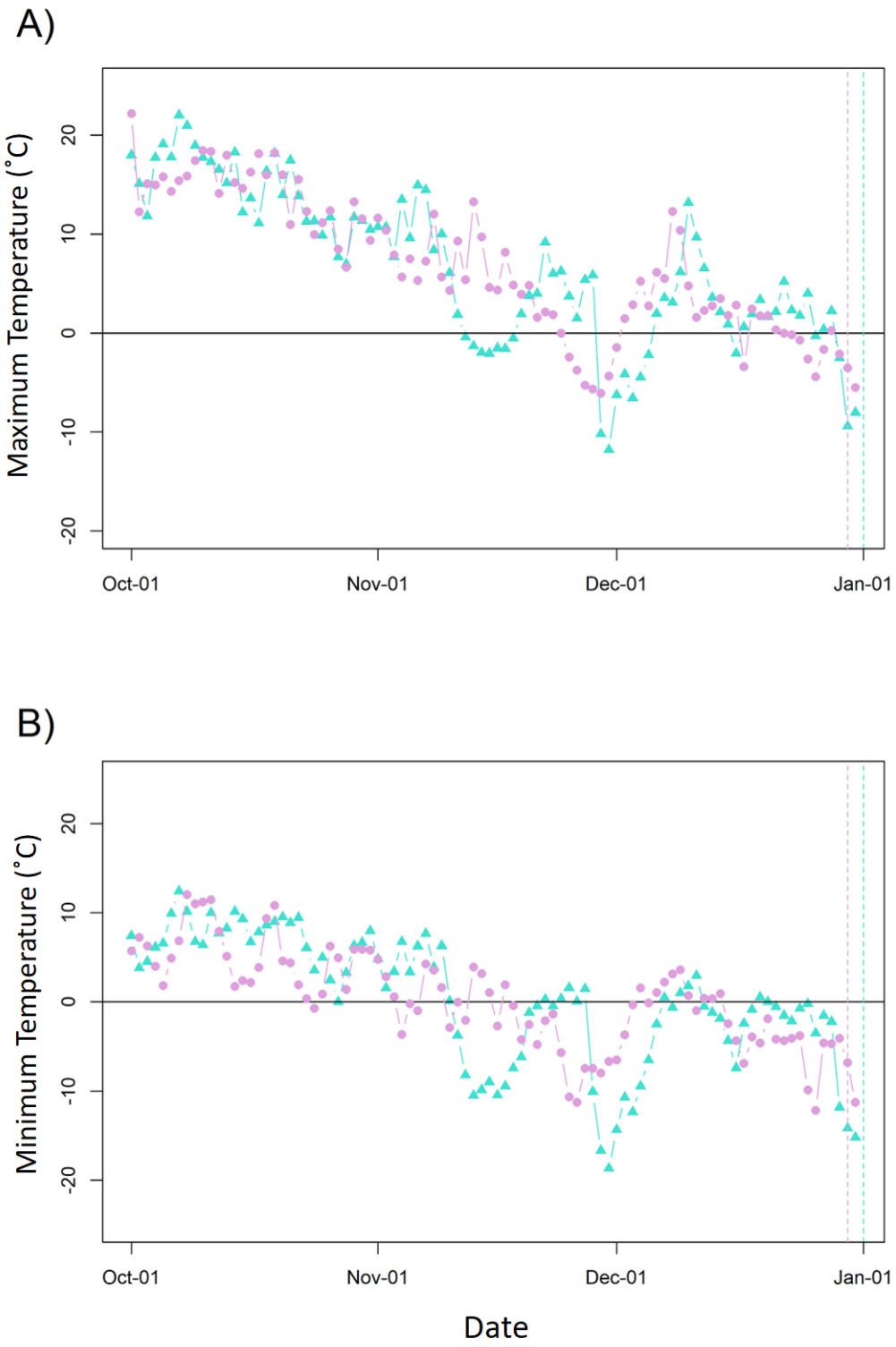

Figure 23. Maximum (A) and minimum (B) temperature at the Kalamalka Forestry Centre from October 1 until tissue was sampled from subalpine larch trees on December 30, 2014 (green), and December 29, 2015 (purple).

years trees were exposed to similar extended periods of cold prior to sampling and should have been fully hardened.

Spring sampling was carried out on March 30, 2015, and March 14, 2016. GDD at the Kalamalka Forestry Centre was calculated using mean daily temperature and a threshold of 1.46˚C [$\Sigma(T_{mean>1.46} - 1.46)$]. At the time of sampling, 215 and 115 GDD were recorded in 2015 and 2016, respectively (Figure 24). In the first year, sampling was late: three trees had already begun to flush. However spring cold tolerance is ultimately a measure of the relative timing of dehardening so all samples were retained for analysis.

Autumn sampling was carried out on October 19, 2015 and on October 17, 2016. In both years, temperatures dropped below 4 ˚C at least two weeks prior to sampling (Figure 25). The number of hardening degree days (HDD) below a threshold temperature was calculated using minimum daily temperature (Carles et al. 2012). Although the temperature threshold below which hardening occurs is not known in *Larix*, cold acclimation occurs below 14.5˚C in white spruce (*Picea glauca*), a North American conifer that is well adapted to survival at high latitudes (Greer et al. 2001). When a 14.5˚C threshold value was used, 323 and 353 HDD were accumulated between August 31st and the time of sampling (Figure 26). Spruce seedlings that accumulated 220 HDD were considered acclimatized for cold storage under operational settings (Carles et al. 2012), suggesting that *Larix* acquired some level of cold tolerance prior to autumn sampling.

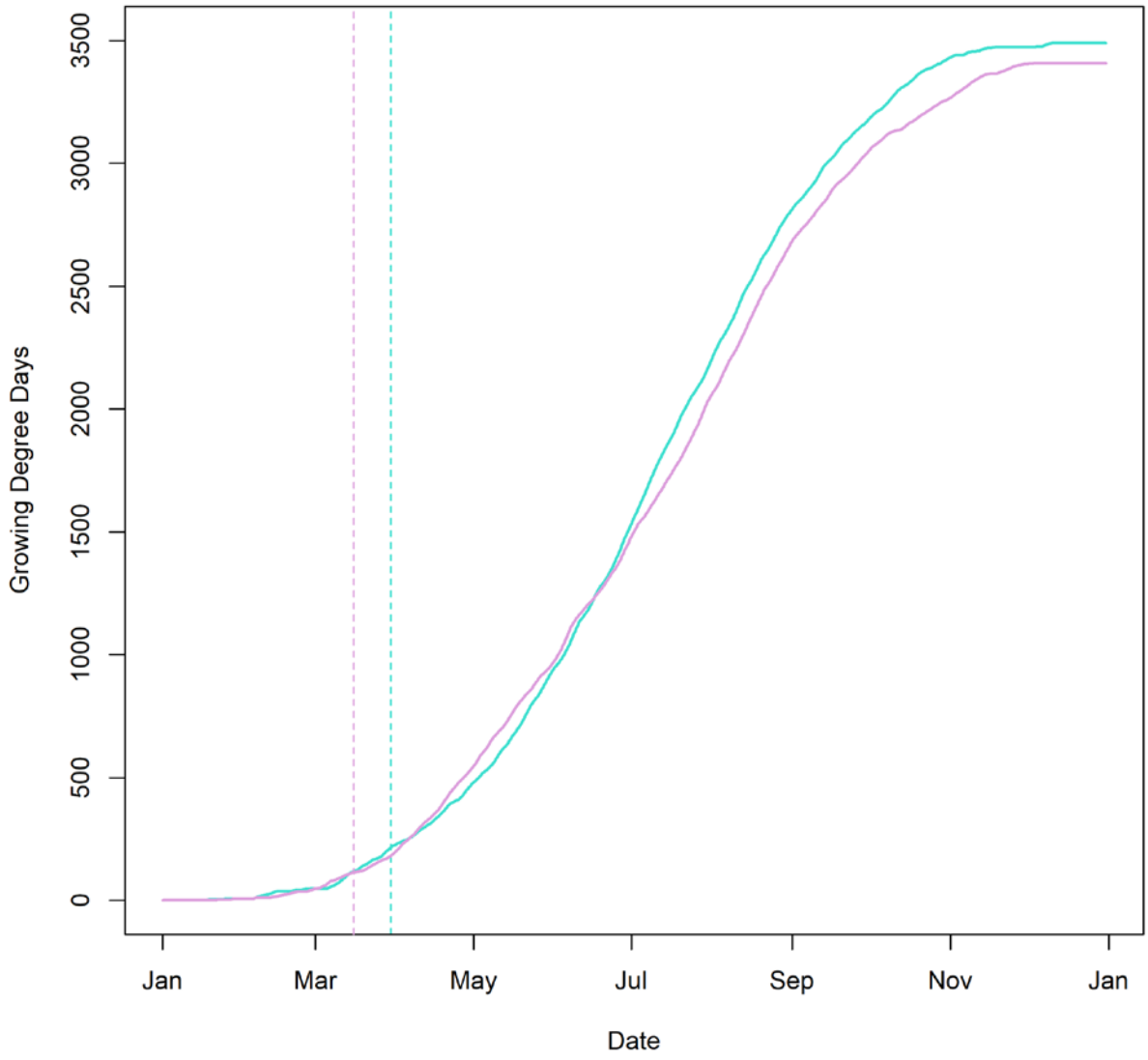

Figure 24. Growing degree days at the Kalmalka Forestry Centre prior to sampling subalpine larch stem tissue on March 30, 2015 (green), and March 14, 2016 (purple).

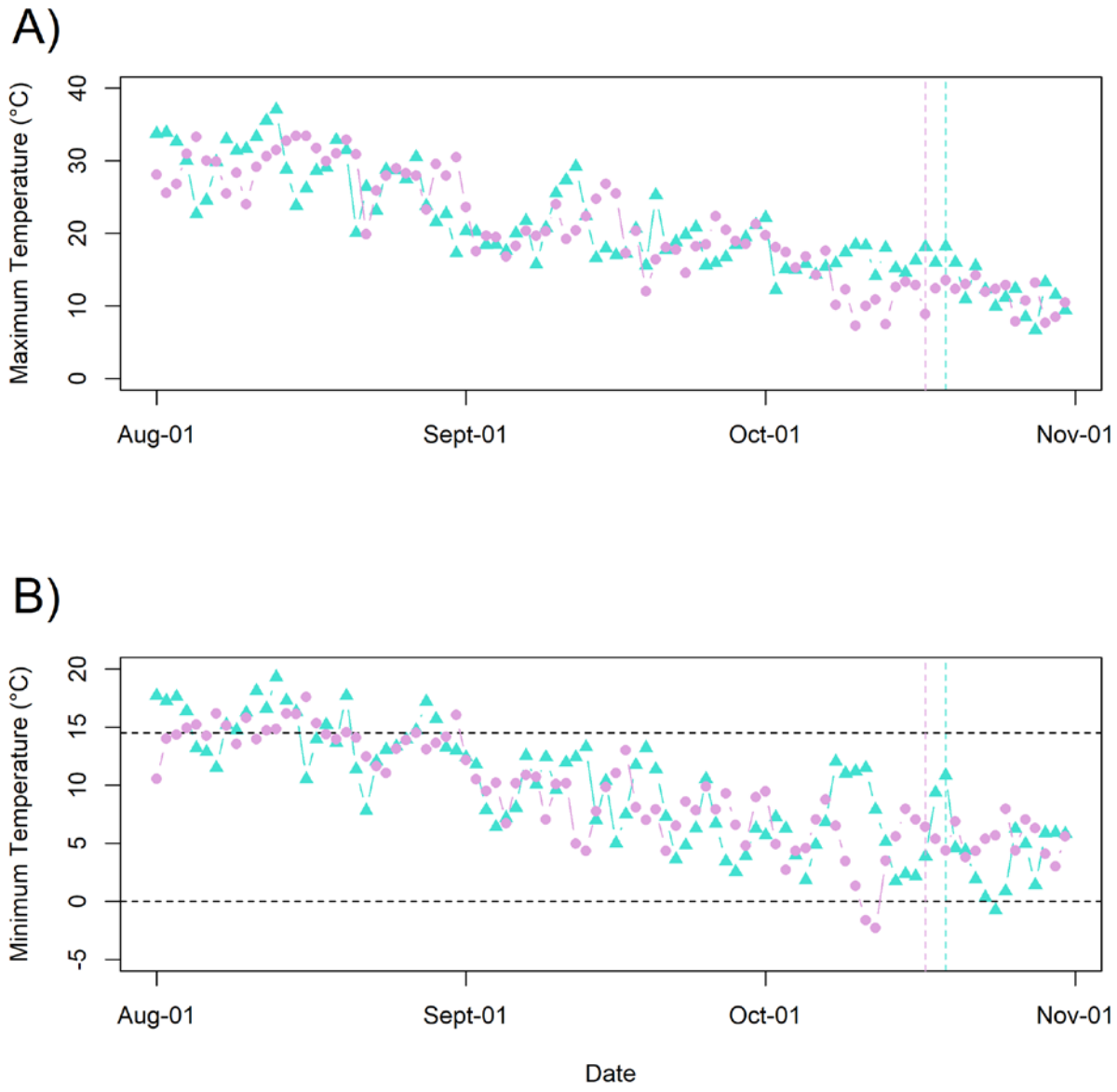

Figure 25. Maximum (A) and minimum (B) temperature at the Kalamalka Forestry Centre prior to autumn sampling of subalpine larch stem tissue on October 19, 2015 (green), and October 17, 2016 (purple).

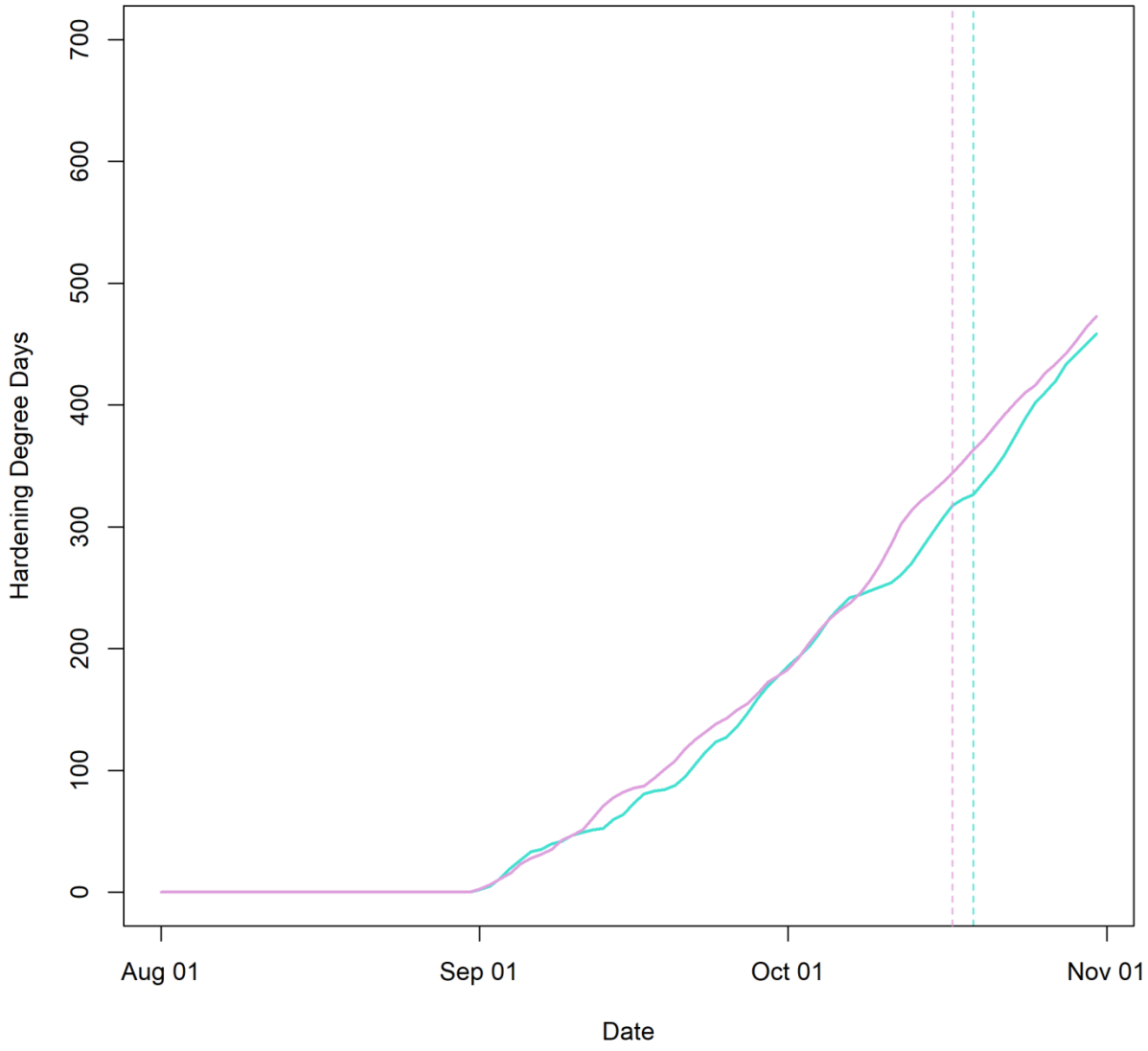

Figure 26. Accumulation of hardening degree days (HDD) at the Kalamalka Forestry Centre prior to autumn sampling of subalpine larch stem tissue on October 19, 2015 (green), and October 17, 2016 (purple).

2.2 Cold tolerance assessment

To assess cold tolerance, a sample of the most recent year's growth was collected from each of the 100 trees included in this study. If a single branch tip of suitable length was not available, up to three branch tips were collected. Samples were shipped overnight from Vernon to Victoria, BC. In the lab, terminal buds were removed and stem tissue was cut into 16 pieces, each 0.5 cm in length. If more than one branch tip was used, cuttings were mixed together before equal distribution into four vials. Five drops of distilled water were added to each vial to prevent desiccation. Vials were sealed and separated onto four trays, each representing a different freezing treatment, with vial position randomized within tray. In general, trays were moved to 4 ˚C overnight after setup. Note that in January 2015, trays spent two nights at 4 ˚C. In the spring of 2016, freezing treatments were applied directly after setup.

Cold tolerance was assessed after exposure to three different sub-zero temperatures: -20 ˚C, -30 ˚C and -40 ˚C. Three trays were placed in a programmable freezer (Caltec Scientific Ltd., Richmond, Canada) set to 0 ˚C and cooled at a rate of five degrees per hour. After reaching each experimental freezing temperature, temperature was held constant for one hour before one tray was removed. A fourth tray was kept at 4 ˚C to serve as an unfrozen control. After freezing, samples were moved to 4 ˚C to thaw overnight. Note that in the spring of 2015 the -40 ˚C treatment was dropped and a -10 ˚C treatment used instead because of concerns that trees had already started flushing and tissue would be more prone to injury.

After thawing, 10 mL of distilled water were added to each vial. Trays were kept at room temperature overnight on a shaker at 90 rpm. At the time of measurement, vials

were inverted to mix their contents. A conductivity probe (4020 conductivity meter, Jenway Ltd, Dunmow, UK) was used to measure conductivity due to electrolyte leakage. Trays were then transferred to an oven and heated until they reached 100 ˚C in order to ensure cellular rupture. Trays were removed from the oven and kept at room temperature overnight on the shaker. Conductivity was measured the following day. Relative conductivity (RC) was calculated as:

RC = (Electrical conductivity before boiling/Electrical conductivity after boiling)

The index of cold injury was calculated as:

$$CI = [(RC_{freezing\ test} - RC_{control})/(1 - RC_{control})]$$

Note that a higher index of cold injury indicates less cold tolerance.

Data were not successfully collected for all samples (Table 14). In January 2015, 28 lids cracked in the oven (out of 400) and these samples were excluded from further analysis. Additional exclusions were the result of cracked lids, lids that popped off in the oven and human error (e.g. spillage). In cases where individual controls were missing, the population average was used as the $RC_{control}$ value for the calculation of cold injury.

2.3 Analysis of phenotypic data

Cold injury data were first examined separately by sampling date. Negative cold injury values, which arise when more electrolyte leakage occurs in controls than in frozen tissue, were converted to zero values (Appendix G). Mean cold injury was plotted to confirm that damage increased at colder sub-zero temperatures.

Cold injury data were analyzed by season and freezing temperature. A simple linear model was fit to each dataset with population-of-origin included as a fixed effect.

Table 14. Number of subalpine larch samples assessed for cold tolerance across two years, three seasons and four treatments out of a maximum of 100 per cell.

		Control	Freezing Temperature (˚C)			
Year	Season	4 ˚C	-10	-20	-30	-40
2015	Winter	86	-	93	95	98
	Spring	100	100	100	99	-
	Autumn	97	-	94	99	88
2016	Winter	99	-	99	99	99
	Spring	96	-	96	97	97
	Autumn	99	-	85	99	99

Normality and homoscedasticity were assessed visually by plotting residuals and residuals versus fitted values, respectively. Because population explained less than 1% of variation in the -10˚C and -20˚C freezing tests, these data were excluded from further analysis. No further results will be reported from these freezing temperatures.

Cold injury data for the -30˚C and -40˚C treatments were analyzed separately by season. A simple linear model was fit to each dataset with climatic region (see above section 'Climate Data'), freezing temperature and year included as fixed effects. Model simplification led to the removal of non-significant predictors: year from the winter and spring analyses. Normality and homoscedasticity were assessed visually by plotting residuals and residuals versus fitted values, respectively. Autumn data were square-root transformed to better meet the assumptions of the model.

2.4 Phenotype-environment associations

Phenotype-environment associations were assessed for each season. Within seasons, cold injury at -40 ˚C (CI), the treatment with the highest index of cold injury, was averaged by tree and used to fit a simple linear model with each of the 20 climate variables as predictors. Spring samples were not averaged because they were only frozen at -40 ˚C in the second year of the experiment. Adjusted R-squared was assessed; variables that explained less than 5% of the total variation were removed due to a lack of linearity. Twelve variables were identified for the analysis of winter CI, 11 variables were identified for the analysis of autumn CI and seven variables were identified for the analysis of spring CI (Appendix H).

Phenotypic data were analyzed using redundancy analysis (RDA), a form of constrained ordination that determined how much variation was explained by a given set of variables. RDA is the multivariate analogue of simple linear regression. Selected climate variables were included in models fit to seasonal cold injury data using the *rda* function from the vegan package in R (Oksanen et al. 2018). Significance of climate predictors was determined using a permutation test as implemented by the *anova* function. Biplot loadings were used to identify the two most important climate variables shaping phenotypic variation in each season and a second model was fit using only these two variables. Reducing the number of predictors reduced variance inflation factors for constraints below 10 in all cases. To assess the importance of the selected climate variables, partial redundancy analysis was employed to control for geography (latitude and longitude). This accounts for neutral genetic structure across the landscape as well as the tendency of nearby locations to have similar climates.

Climate variables identified in the RDA analysis were used to assess phenotypic clines. CI was averaged by individual and population and plotted against climate. Mean CI data were fit with simple linear models to assess the strength of the linear relationship.

3. Genomic data

3.1 SNP data

Genotypic data were generated over two rounds of sequencing for the 100 trees with cold tolerance (CT) phenotypes (Table 15): 33 trees with the Sbf1 restriction

Table 15. Subalpine larch trees phenotyped for cold tolerance were genotyped over two rounds of RAD-seq utilizing two different restriction enzymes: Sbf1 and Pst1.

Population	Sbf1	Pst1	N
AL01	2	4	6
AL02	1	5	6
AL03	3	3	6
AL04	3	2	5
AL05	2	4	6
AL06	2	3	5
AL07	2	3	5
AL08	1	5	6
AL09	2	3	5
AL10	1	5	6
AL11	2	4	6
AL12	0	5	5
AL13	2	4	6
AL14	2	4	6
AL15	0	6	6
AL16	3	3	6
AL17	3	1	4
AL20	2	3	5
TOTAL:	33	67	100

enzyme (Ch. 2) and 67 trees with the Pst1 restriction enzyme (Ch.3). Pst1 is a six-base subset of the eight-base Sbf1 restriction enzyme. Trees sequenced using Sbf1 were not re-sequenced with Pst1 because the same loci were expected to be targeted by both enzymes. CT samples were pooled for analysis after sequences were aligned to the Siberian larch draft genome (Kuzmin et al. 2019) and base quality scores were recalibrated, as previously described (Chapter 2, p. 32; Chapter 3, p. 77).

Genotypes were called using two approaches. First, genotypes were called using the GATK UnifiedGenotyper software (Chapter 2, p. 34). SNP genotypes were filtered using VCFtools v. 0.1.14 (Chapter 2, p. 34). Two thresholds for missingness (0.5 and 0.3) were tested, as well as two thresholds for the minor allele frequency (0.05 and 0.02). Summary statistics were generated in VCFtools for mean depth per individual, mean number of SNPs per individual, mean number of missing loci per individual and individual genotype depths. This last file was reformatted in R prior to analysis. Output data were fit with simple linear models in R that included restriction enzyme as a fixed effect. The most stringent dataset (--maxmissing 0.50 --maf 0.05) and the most relaxed dataset (--maxmissing 0.30 --maf 0.02) were converted into STRUCTURE format using PGDspider (Lischer and Excoffier 2012) and loaded into R using the *read.structure* function from the adegenet package. Allelic data were converted to matrix format using the *tab* function, which replaces missing values with mean allele frequencies. Principal components analysis (PCA) was carried out using the *dudi.pca* function from the ade4 package (Bougeard and Dray 2018) and the first two PCs were plotted against each other. To test whether genetic data were geographically sensible, data were stratified by population and a dendrogram of Provesti's genetic distance was generated with bootstrap

support using the *aboot* function from the poppr package (Kamvar et al. 2014). Dendrograms were plotted using the *plot.phylo* function from the ape package (Paradis and Schliep 2018). Note that Provesti's genetic distance did not require the imputation of missing data.

Second, genotypes were called using the ANGSD software (Excoffier and Foll 2011; parameter settings in Appendix I). Within 125 nucleotides, the SNP with the highest number of individuals genotyped was preferentially retained. Three cut-offs for the minimum number of individuals required to call a genotype were tested: 30, 50 and 70 (out of 100). To test whether data from both rounds of sequencing were contributing to final genotype calls, PCA was carried out (Chapter 3, p. 83) and dendrograms of genetic distance were built using ngsTools (Fumagalli et al. 2014) and FastME (LeFort et al. 2015; Chapter 3, p. 83). ANGSD was also used to generate enzyme-specific datasets for Pst1 CT and Sbf1 CT samples. ANGSD genotypic data were formatted for further analysis using custom scripts in R.

Range-wide genomic datasets were generated for the identification of F_{ST} outliers and the examination of genotype-environment associations, as described below. The Sbf1 range-wide dataset was genotyped using GATK UnifiedGenotyper as described in Chapter 2 (p. 33) minus western larch samples. The Pst1 range-wide dataset was generated using ANGSD as described in Chapter 3 (p. 82) with minor modifications in parameter settings (Appendix I).

3.2 Genotype-phenotype associations

Random forest analysis (Breiman 2001) was used to identify loci that predict seasonal cold injury. Random forest (RF) is a machine-learning algorithm that can be used to identify the loci underlying polygenic traits (Brieuc et al. 2018). Its nonparametric framework allows for the detection of non-additive interactions between loci, meaning suites of loci that collectively explain phenotypic variation can be identified. RF analysis was carried out on the full CT dataset. ANGSD genotypes were used for RF analysis because regression does not allow for missing data.

Prior to analysis, phenotypes and genotypes were adjusted in order to avoid spurious associations caused by neutral population structure (Zhao et al. 2012). For each season, cold injury at -40 ˚C (CI) was fit with a simple linear model that included both population and year as categorical predictors. One exception was the spring analysis, where year was excluded because samples were only frozen at -40 ˚C in 2016. For the winter and autumn datasets, residuals were averaged by individual across years. Genotypic data were also adjusted for neutral population structure. For each locus, a generalized linear model was fit to a genotypic response (coded as 0, 1 or 2) with population as a predictor variable. Family was specified as Poisson and the *dispersiontest* function from the AER package (Kleiber and Zeileis 2008) was used to test for overdispersion. If overdispersion was detected, the model was rerun with family specified as quasipoisson. Residuals from each model were retained for RF analysis.

All RF analyses were performed using the randomForest package in R (Liaw and Wiener 2002). RF grows a “forest” of regression trees by sampling with replacement from a dataset. Approximately one third of the data is omitted from each run. A subset of

the total number of predictors (i.e., SNPs) is randomly sampled at each decision node (the mtry parameter) and the SNP with the greatest ability to improve the 'purity' of the node is used to split the node. Purity is maximized by minimizing error when the predictor is regressed against a continuous response variable. This process continues until the purity of the terminal nodes cannot be improved. It is repeated *t* times to create a forest with *t* trees (the ntree parameter). The power of an individual regression tree is determined by the proportion of variation explained by the tree. An individual SNP's importance can be evaluated via the average improvement in purity when that SNP is used to split a node.

To optimize parameter settings, ten RF analyses with increasing increments of 50,000 trees (50,000 – 500,000) were computed using six different values of the mtry parameter, i.e. a subset of the number of SNPs (*M*) to be assessed at each decision point ($\sqrt{M}$; 2*$\sqrt{M}$; 0.1**M*; 0.2**M*; *M*/3; *M*) and the proportion of phenotypic variance explained for each combination of parameters was plotted in R. Varying mtry did not improve the proportion of variation explained by the model. Thus, three RF analyses with 200,000 trees were computed using all loci as predictors (*M*) to assess the approximate importance of each locus. The correlation between SNP importance values across runs was assessed to ensure repeatability between forests. Marker importance values were used to subsample top-ranked loci (top 2 %, 3 %, 4 %, 5 %, 10 %, 20 % and 30 %) and determine the number of loci required to explain the maximum amount of phenotypic variation (Brieuc et al. 2018). Initial analyses found that the top 2% of loci predicted the maximum amount of observed phenotypic variance. A backward purging approach initiated with the top 4 % of SNPs was used to rank SNP importance across all three runs (Holliday et al. 2012). In these analyses, the least important loci were sequentially

removed, thus accommodating potential interactions between loci and accounting for the fact that initial importance values were approximations. Cross-validation was used to assess whether there was evidence of model overfitting. To do this, a bootstrapped regression cross-validation approach was implemented with 100 iterations using the *rf.crossValidation* function from the rfUtilities package (Evans and Cushman 2009; Evans and Murphy 2018).

The reference sequences surrounding predictor SNPs identified during RF analysis were searched in the BLASTn database (NCBI Resource Coordinators 2012). In total, 150 nucleotides on either side of each SNP were included in this search. Annotation was also explored using an annotation file for Siberian larch v 1.0 draft genome.

3.3 F_{ST} outliers

Genotypic data were analyzed using BayeScan v. 2.1 (Foll and Gaggiotti 2008) to identify F_{ST} outliers that are candidates for being under either balancing or directional selection. BayeScan estimates population-specific F_{ST} coefficients based on differences in allele frequencies between populations. Population size does not affect quality of output. The model was run with 100,000 iterations and prior odds of 10,000 in order to minimize false positives (Lotterhos and Whitlock 2014). BayeScan was used to identify F_{ST} outliers in four datasets:

1) CT ANGSD dataset
2) CT GATK UnifiedGenotyper (UG) dataset
3) Range-wide Pst1 ANGSD dataset
4) Range-wide Sbf1 UG dataset

SNPs identified as F_{ST} outliers were compared among datasets to identify overlap. The reference sequences surrounding SNPs that overlapped between both range-wide datasets (150 nucleotides on either side of each SNP) were used to search the BLASTn database. Annotation was also explored using an annotation file for Siberian larch v 1.0 draft genome.

3.4 Genotype-environment associations (GEA)

Correlations between genotypic and environmental gradients were assessed using bayenv2 (Coop et al. 2010; Günther and Coop 2013). This software uses a Bayesian approach to identify loci involved in local adaptation, either through unusual correlations between important ecological variables and allele frequencies or by extreme differences in allele frequencies between geographic regions. Two datasets were used to assess genotype-environment associations (GEA):

1) Range-wide Pst1 ANGSD dataset
2) Range-wide Sbf1 UG dataset

Empirical covariance in allele frequencies was estimated between populations using marker data from a subset of neutral SNPs identified using BayeScan. For the range-wide Sbf1 dataset with 789 SNPs, 400 neutral SNPs were used to assess neutral structure. For the range-wide Pst1 dataset with 18,992 SNPs, 5,000 neutral SNPs were used to assess neutral structure. Ten covariance matrices were calculated for each set of neutral SNPs using 100,000 MCMC iterations. The last draw from each of the ten runs was

output and a mean covariance matrix was calculated. Individual SNPs were tested against this null model, thus accounting for the neutral correlation of allele frequencies as well as differences in sample size.

For each SNP, output included a Bayes Factor (BF), which measures the strength of the linear relationship between allele frequency and environment, as well as Spearman's ρ, which gives the nonparametric correlation between allele frequency and environment. Because the stability of BFs is sensitive to the number of MCMC iterations (Blair et al. 2014), three replicate chains were run and BF values were compared across chains. For the range-wide Pst1 dataset, 100,000 MCMC iterations produced an average correlation of 0.996 between chains. For the range-wide Sbf1 dataset, implementing bayenv2 with 500,000 MCMC iterations produced an average correlation of 0.965 between chains. BF values and Spearman's ρ were averaged across runs for further analysis. Spearman's ρ was plotted using the heatmap.2 function from the gplots package (Warnes et al. 2019).

SNPs with at least moderate support for GEA (BF >= 3.0) were compared among datasets to identify overlap. The reference sequences surrounding SNPs that overlapped between both range-wide datasets (150 nucleotides on either side of each SNP) were used to search the BLASTn database. Annotation was also explored using an annotation file for the Siberian larch v 1.0 draft genome.

Results

Environmental Regions

A discriminant analysis of principal components on climate data from 18 populations of subalpine larch identified two climatically distinct regions in the northern portion of subalpine larch's range (Table 16). These correspond to one group of seven populations in the northern Rockies and another group of 11 populations in the southern Rockies and northern Cascades (Figure 27). Hereafter, these will be referred to as the northern and southern regions, respectively. Loadings for the two PCs used in this analysis indicated that the first PC was dominated by temperature-associated variables (e.g. minimum autumn temperature) while the second PC is dominated by precipitation-associated variables (e.g. mean annual precipitation). Overall the northern cluster was colder and wetter, especially in the summer (Table 16). Such environmental differentiation might have provided opportunities for local adaption across the landscape.

Analysis of Cold Injury

Mean cold injury was highest in the spring and lowest in the winter, while variation among individuals was greatest in the spring and least in the winter (Figure 28). Within seasons, freezing temperature and environmental region were significant predictors of cold injury. As expected, cold injury increased with increasingly severe freezing temperatures. When tissue was frozen at -40 ˚C versus -30 ˚C, cold injury increased by 8 % in winter ($p < 0.001$; $F = 142.78$), 16 % in spring ($p < 0.001$; $F = 70.77$) and by 10 % in autumn ($p < 0.001$; $F = 98.37$).

Table 16. Twenty climate variables were used to identify two environmental regions in the northern portion of subalpine larch's range. DAPC loadings for the first two principal components suggest the relative importance of each climate variable. Mean values of each variable are reported by region (north and south), with significant differences between regions indicated in bold.

Climate Variable	PC1 Loading	PC2 Loading	Northern Mean	Southern Mean	North-South p-value
MAT (˚C)	-0.29	-0.07	-1.7	0.2	< **0.001**
MWMT (˚C)	-0.23	-0.22	9.7	11.2	**0.009**
MCMT (˚C)	-0.26	0.15	-12.6	-9.7	< **0.001**
TD (˚C)	0.11	-0.30	22.4	20.8	0.069
Tmin_at (˚C)	-0.30	-0.01	-5.4	-3.1	< **0.001**
MAP (mm)	0.03	-0.36	1267	1245	0.876
MSP (mm)	0.20	-0.29	531	389	**0.008**
AHM (˚C/μm)	-0.16	0.32	6.7	8.7	0.053
SHM (˚C/μm)	-0.25	0.23	19.3	29.5	**0.001**
DD_0 (days)	0.29	-0.04	1753	1341	< **0.001**
DD_5 (days)	-0.23	-0.23	447	602	**0.007**
NFFD (days)	-0.27	-0.19	104	127	< **0.001**
FFP (days)	-0.26	-0.15	177	171	0.075
bFFP (date)	0.19	0.30	235	248	< **0.001**
eFFP (date)	-0.28	0.01	58	78	**0.001**
PAS (mm)	0.02	-0.29	801	812	0.913
EMT (˚C)	-0.26	0.12	-45.7	-40.7	< **0.001**
EXT (˚C)	-0.21	-0.22	27.4	28.6	0.056
Eref (mm)	-0.17	-0.17	348	381	0.105
CMD (mm)	-0.17	0.30	11	40	**0.012**

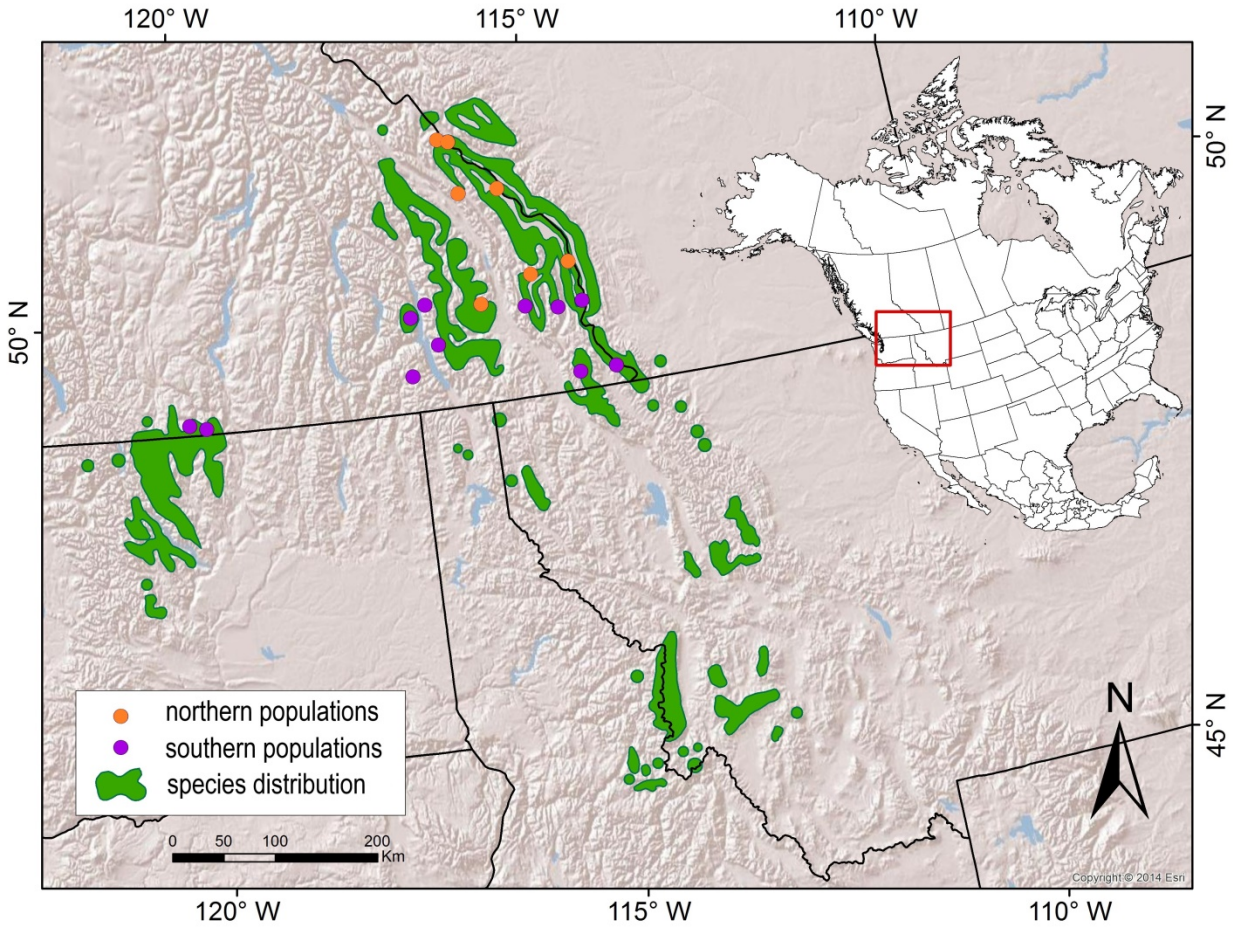

Figure 27. Eighteen populations of subalpine larch in the northern portion of the species range cluster into two distinct climatic regions (orange and purple) as per a discriminant analysis of principal components (DAPC) based on 20 climate variables.

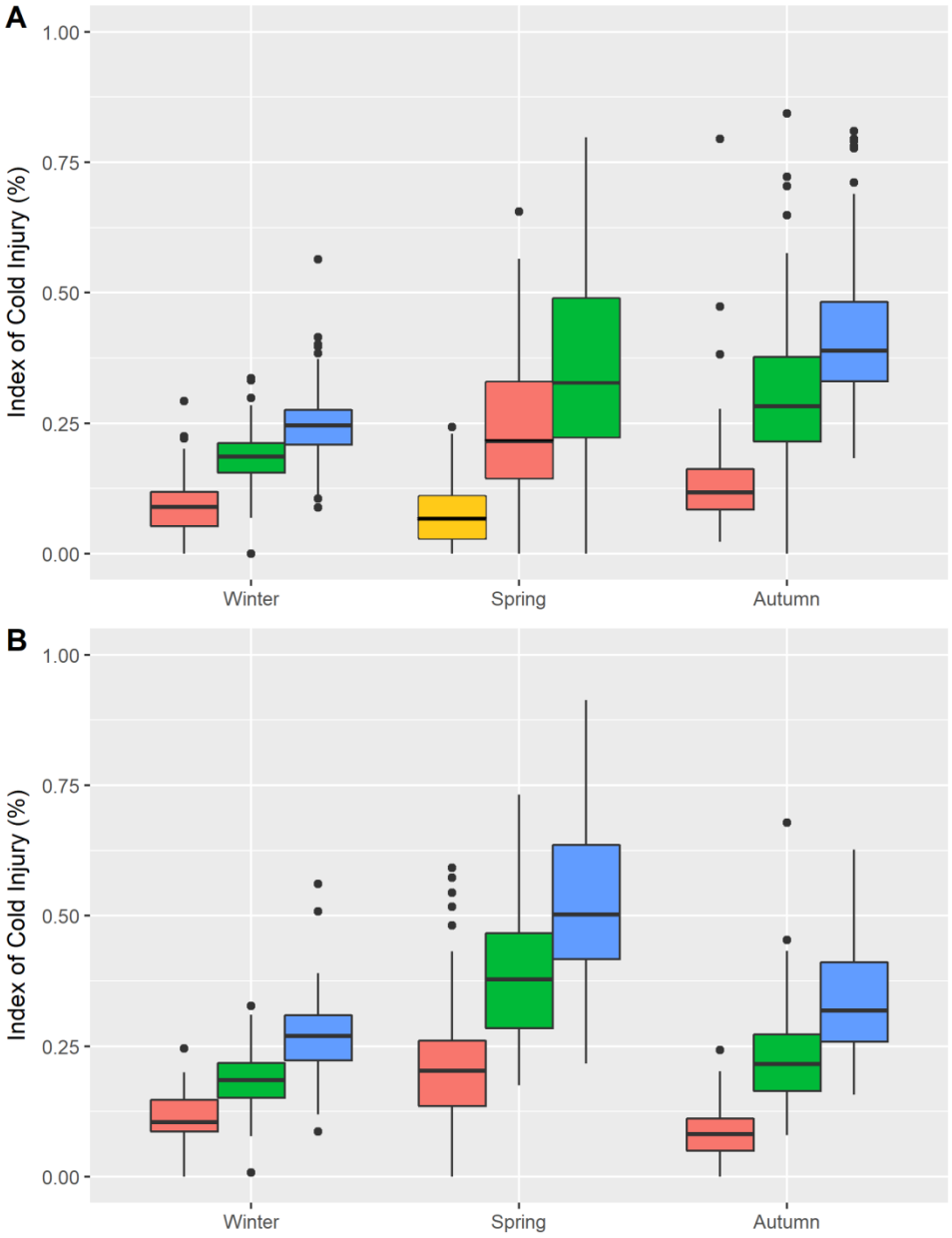

Figure 28. Mean index of cold injury for 100 subalpine larch trees grafted *ex situ* at the Kalamalka Forestry Centre, frozen at four different subzero temperatures (yellow = -10 ˚C; red = -20 ˚C; green = -30 ˚C; blue = -40 ˚C) in three different seasons in 2015 (A) and 2016 (B).

Individuals from the southern region incurred significantly more cold damage than individuals from the northern Rockies (Figure 29). The biggest difference between regions was observed in spring and the smallest was observed in winter. Damage was 4 % higher in trees from the southern region in winter ($p < 0.001$; $F = 29.62$), 8 % higher in spring ($p < 0.001$; $F = 19.58$) and 6 % higher in the autumn ($p < 0.001$; $F = 29.79$).

Finally, year was a significant predictor of autumn cold injury ($p < 0.001$; $F = 48.50$), with 7 % less damage in 2016. This may reflect the greater number of hardening degree days (+30) accumulated in the second year of sampling.

Phenotype-Environment Associations

Individual climate variables were significant predictors of cold injury at -40˚C (CI) in all three seasons. Redundancy analysis (RDA) identified degree days below zero (DD_0) and the length of the frost-free period (FFP) as the two most influential climate variables shaping variation in winter CI; continentality (TD) and mean coldest month temperature (MCMT) as the two most influential climate variables shaping variation in spring CI; and DD_0 and MCMT as the two most influential climate variables shaping variation in autumn CI. These pairs of climate variables were significant predictors of CI in winter ($p = 0.001$; $F = 14.013$), spring ($p = 0.002$; $F = 6.7414$) and autumn ($p = 0.001$; $F = 12.211$). When mean population cold injury was plotted against these climate variables, clear phenotypic clines emerged (Figure 30).

When RDA models were conditioned on geography (latitude and longitude), climate remained a significant predictor of winter CI ($p = 0.002$; $F = 9.5645$) but not spring CI ($p = 0.284$; $F = 1.3269$) or autumn CI ($p = 0.093$; $F = 2.504$). Partial

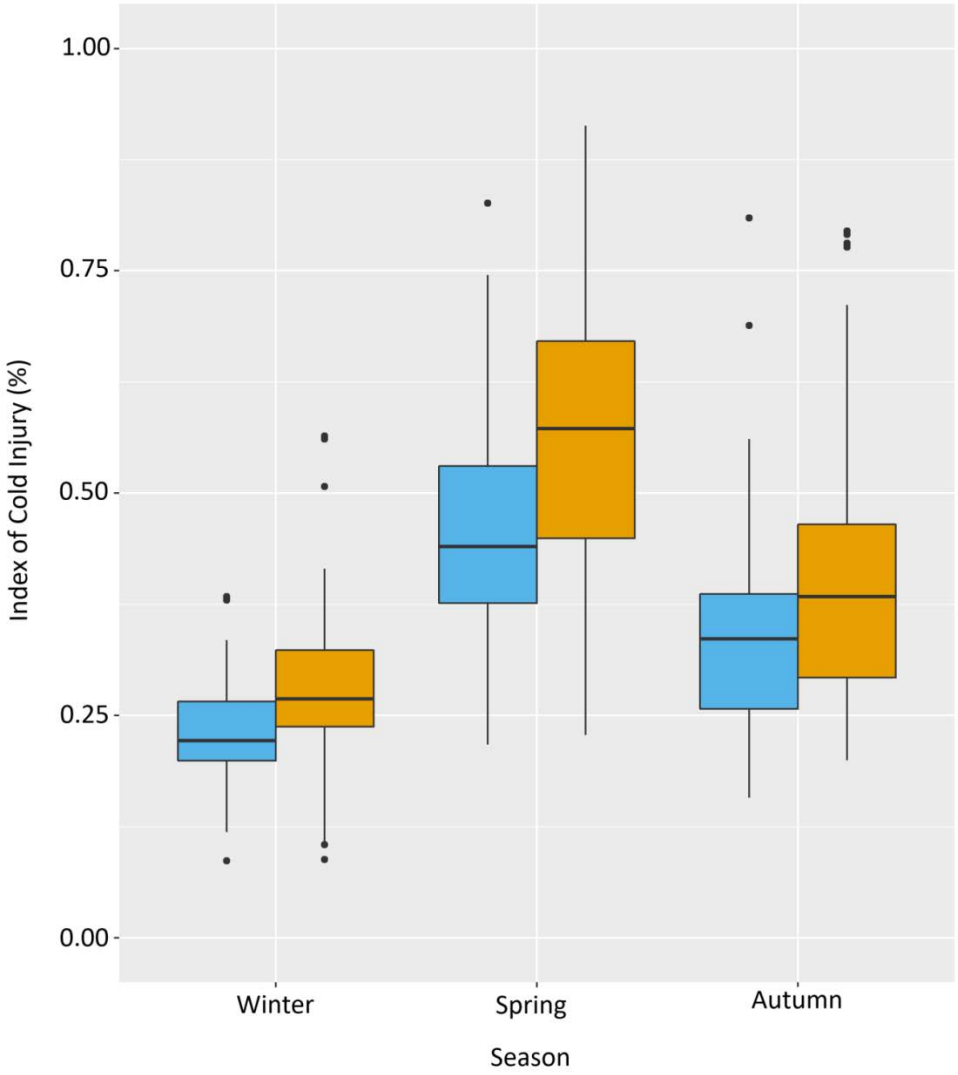

Figure 29. Cold injury is significantly higher for subalpine larch trees from the southern environmental region (orange) than the northern environmental region (blue) in all three seasons.

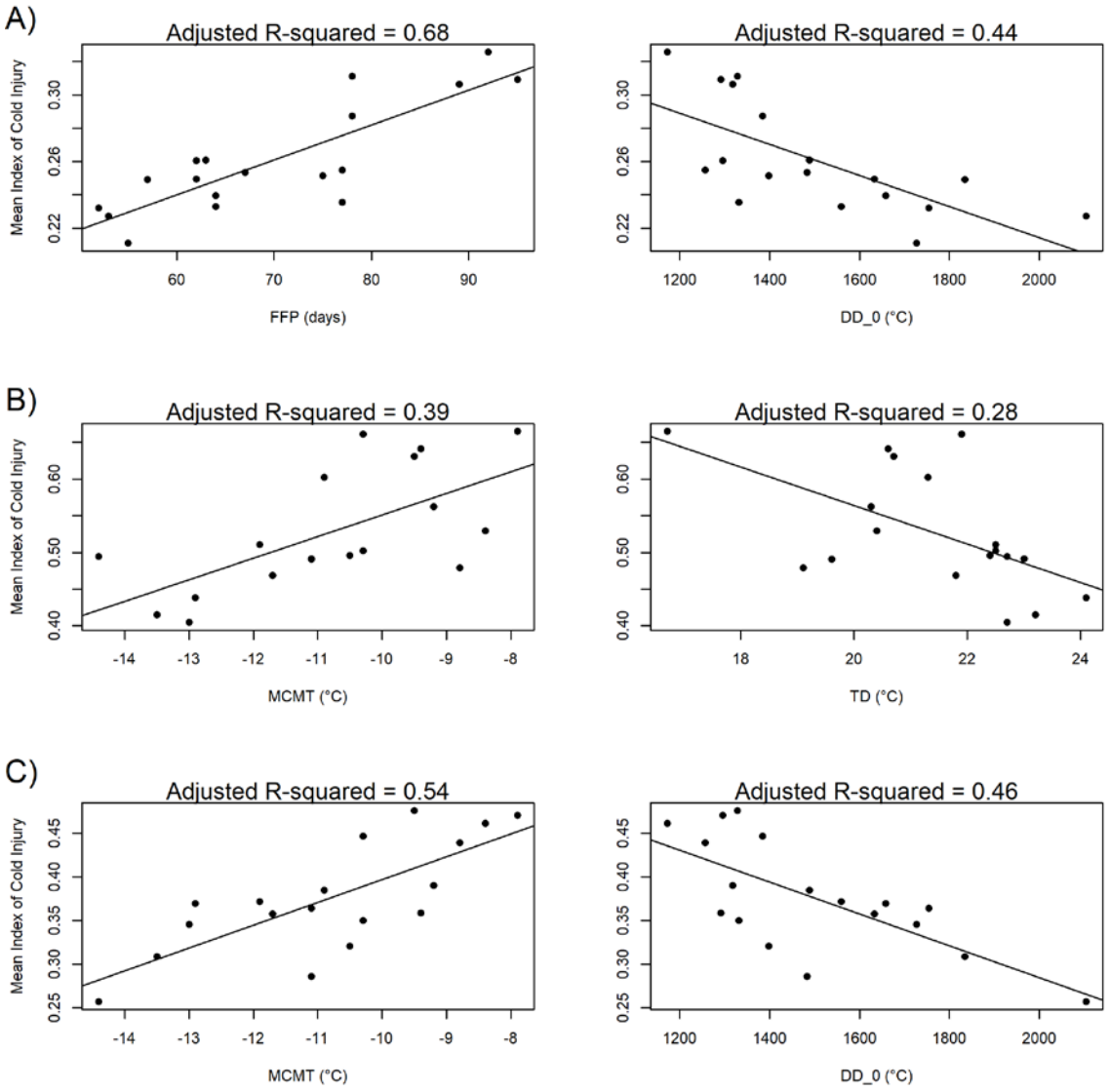

Figure 30. Subalpine larch mean population cold injury at -40˚C shows strong phenotypic clines along climatic gradients (A) in winter for frost-free period (FFP) and degree-days below zero (DD_0), (B) in spring for mean coldest month temperature (MCMT) and continentality (TD) and (C) in autumn for MCMT and DD_0.

redundancy analysis found that DD_0 and FFP accounted for 67 % of the explainable variance in winter CI while geography only accounted for 3 %; TD and MCMT accounted for 19 % of the explainable variance in spring CI while geography accounted for 5 %; and DD_0 and MCMT accounted for 20 % of the explainable variance in autumn CI while geography only accounted for 5 %. These results provide additional evidence for the importance of climate in shaping cold tolerance, especially in winter.

Genomic data

Sequencing for the 100 individuals phenotyped for cold tolerance (CT) was completed over two rounds after DNA digestion with two different restriction enzymes, Sbf1 and Pst1. In total, 33 samples were processed with the Sbf1 restriction enzyme and 67 samples were processed with the Pst1 restriction enzyme (Table 17).

Bioinformatics processing was completed separately for each round of sequencing (Table 18). Because CT samples were sequenced in libraries alongside non-CT individuals, it was not possible to report CT-specific summary statistics for all stages of processing (Table 17). Overall, Pst1 paired-end sequencing generated more reads per individual. Pst1 CT samples started with a median of over 4 million reads each, while Sbf1 CT samples started with less than a million reads each. After bioinformatics processing, Pst1 CT samples retained a median of 985,000 reads per individual while Sbf1 CT samples retained a median of 356,000 reads per individual.

CT samples were genotyped together using two different approaches. First, all genotypes were called using the GATK UnifiedGenotyper (UG) software. Across all sequenced nucleotides, less than 1 % were polymorphic. SNP variants were filtered to

Table 17. Individuals with cold tolerance phenotypes were sequenced over two rounds of sequencing (Sbf1 versus Pst1 restriction site associated DNA sequencing) in six separate libraries.

Library	Sbf1	Pst1
C446	0	.
C447	0	.
C448	33	.
C577	.	17
C578	.	15
C579	.	10
C580	.	7
C581	.	18

Table 18. Bioinformatics processing of individuals with cold tolerance (CT) phenotypes over two rounds of restriction associated DNA sequencing (RAD-seq) that utilized two different enzymes (Sbf1 and Pst1) on different numbers of individuals (33 and 67, respectively).

		Sbf1 Sequencing			Pst1 Sequencing		
Software	**Process**	**Total Sbf1**	**% Kept**	**Ind. Average**	**Total Pst1**	**% Kept**	**Ind. Average**
Cutadapt	Remove if < 50 bp	25,039,828	100.0	747,104	294,050,215	100.0	4,287,168
STACKS process_radtags	Remove low quality reads	*	*	*	*	*	*
	Remove reads with ambiguous barcode	*	*	*	*	*	*
	Remove reads with ambiguous RAD tag	*	*	*	*	*	*
	Output	20,619,571	82.3	246,696	*	*	*
STACKS clone_filter	Remove unpaired reads	NA	NA	NA	270,016,036	91.8	3,897,572
	Remove PCR duplicates	NA	NA	NA	98,141,988	33.4	1,439,238
NextGenMap	Remove reads that did not align to reference genome	16,925,934	67.6	502,872	93,596,182	31.8	1,359,103
UnifiedGenotyper	Remove reads with mapping quality score = 0	11,986,810	47.9	355,906	68,129,126	23.2	985,168

*Not possible to separate cold tolerance individuals from sequencing library

obtain reliable genotypes (Table 19). When the threshold for missing genotypes was relaxed from 0.5 to 0.3, the final number of SNPs increased from 465 to 1,337. Under this relaxed filter, mean missingness increased and mean depth per locus decreased. Decreasing the threshold for the minimum allele frequency from 0.05 to 0.02 (i.e. four alleles from the sample of 100 individuals) further increased the final number of SNPs to from 1,337 to 1,593 but had no further impact on mean missingness or mean depth per locus.

Pst1 and Sbf1 samples were not equally represented in final UG genotype calls (Table 20). Overall, Sbf1 samples had greater depth per sequenced locus but fewer loci per individual. Thus Sbf1 individuals had a much higher proportion of missing data across loci. When PCA data for the first two PCs were plotted, individuals processed with Sbf1 were tightly clustered relative to those processed with Pst1 in both the stringent (--maxmissing 0.50 --maf 0.05) and relaxed (--maxmissing 0.30 --maf 0.02) datasets (Appendix J Figure 1). Despite this, dendrograms of Provesti's genetic distance showed reasonable genetic structure (Appendix J Figure 2). Both the stringent dataset and the relaxed dataset identified populations from the Cascade Range (AL15 and AL20) as a monophyletic outgroup. Because relaxed filtering did not substantively improve the quality of the dataset (i.e. it did not remove the effect of enzyme), the stringent CT dataset with 465 SNPs was retained for further analysis with BayeScan, as per the methods implemented in Chapter 2.

CT samples were also genotyped using ANGSD. Three thresholds for the minimum number of individuals needed to retain a SNP were tested: 70, 50 and 30 out of 100. Under the stringent threshold (70), 122 loci were retained for PCA. Along the first

Table 19. Filtering of GATK UnifiedGenotyper SNPs for 100 subalpine larch individuals with cold tolerance phenotypes.

Software	Function	Value	Filter	Process	**Stringent**	**% Kept**	**Medium**	**% Kept**	**Relaxed**	**% Kept**
Picard	SortVcf	NA	NA	Sort VCF file	105,139,800	.	105,139,800	.	105,139,800	.
Gatk	SelectVariants	NA	sites	Remove invariant sites	**2,827,530**	.	2,827,530	.	2,827,530	.
VCFtools	mac	1	sites	Remove sites with < 1 copy of minor allele	**822,103**	100	822,103	100	822,103	100
Gatk	QD	< 2.0	sites	Normalize quality by depth and filter low quality	**.**		.		.	
	FS	> 60.0	sites	Remove if strand bias present	**.**		.		.	
	SOR	> 3.0	sites	Remove if strand bias present	**.**		.		.	
	MQ	< 40.0	sites	Remove if mapping quality is low	**.**		.		.	
	ReadPosRankSum	< -8.0	sites	Remove if ref/alt alleles map differently	**.**		.		.	
VCFtools	remove-filtered-all	NA	sites	Remove sites that did not pass GatK filters	**655,103**	79.7	655,103	79.7	655,103	79.7
	exclude	NA	individuals	Remove individuals	**.**		.		.	
	minGQ	< 20	genotypes	Remove genotypes called with < 99% confidence	**.**		.		.	
	minDP	< 3	genotypes	Remove genotypes with < 3 reads	**.**		.		.	
	minQ	< 20	sites	Remove SNP called with < 99% certainty	**606,055**	73.7	606,055	73.7	606,055	73.7
	max-missing	> 0.50; > 0.30; > 0.30	sites	Remove SNPs absent in > 50% of individuals	**4,741**	0.58	**11,478**	1.4	11,478	1.4
BCFtools	filter -i 'AVG(FMT/DP)'	60.893; 47.322; 47.322	sites	Remove sites with depth > (mean + 1.5*IQR)	**4,291**	0.52	**10,548**	1.3	10,548	1.3
VCFtools	maf	< 0.05; < 0.02; < 0.02	sites	Remove SNPs with minor allele frequency < 0.05	**1,328**	0.16	**4,694**	0.57	**6,298**	0.77
	min- & max-alleles	< 2 >	sites	Remove SNPs that are not biallelic	**1,295**	0.16	**4,568**	0.56	**6,013**	0.73
Plink	r2	< 0.50	sites	Remove SNPs in LD	**1,013**	0.12	**3,308**	0.40	**4,563**	0.56
VCFtools	thin	< 100	sites	Retain one SNP per 100 bases	**465**	0.06	**1,337**	0.16	**1,593**	0.19

Table 20. SNPs called using GATK UnifiedGenotyper with different filter settings for the proportion of genotypic data required (MaxMissing) and the minor allele frequency (MAF) showed significant differences for key summary statistics depending on whether sequence data had been generated with the Pst1 restriction enzyme (67 subalpine larch trees) or the Sbf1 restriction enzyme (33 subalpine larch trees).

Filter	MaxMissing	MAF	Trait	Pst1	Sbf1	F-value	P-value	R^2
Stringent	0.5	0.05	Mean depth per individual	26	91	851	< 0.001	0.9
			Mean number of SNPs per individual	457	32	62979	< 0.001	0.998
			Mean proportion of missing SNPs per individual	0.18	0.94	2698	< 0.001	0.96
			Mean depth per SNP	27	96	8978	< 0.001	0.23
Medium	0.3	0.05	Mean depth per individual	14	60	1537	< 0.001	0.94
			Mean number of SNPs per individual	1197	174	6209	< 0.001	0.98
			Mean proportion of missing SNPs per individual	0.44	0.89	595	< 0.001	0.86
			Mean depth per SNP	16	65	35452	< 0.001	0.32
Relaxed	0.3	0.02	Mean depth per individual	14	59	1527	< 0.001	0.94
			Mean number of SNPs per individual	1427	210	6382	< 0.001	0.98
			Mean proportion of missing SNPs per individual	0.44	0.89	587	< 0.001	0.86
			Mean depth per SNP	16	64	39303	< 0.001	0.30

two PCs, variation clearly reflected the enzyme that had been used to subsample the genome prior to sequencing (Appendix K Figure 1). The standard filter (50) generated more loci (2,197) but PCA showed that Sbf1 samples were tightly clustered relative to Pst1 samples (Appendix K Figure 2), likely as a result of missing Sbf1 genotypes being imputed as the most common Pst1 allele. Separation by enzyme is still clearly visible in an individual-based dendrogram of genetic distance (Appendix K Figure 3). Relaxed filtering (30) generated 26,253 loci but PCA once again found that Sbf1 samples were clustered relative to Pst1 samples (Appendix K Figure 4) and separation by enzyme is clearly visible in an individual-based dendrogram of genetic distance (Appendix K Figure 5). When an earlier version of the relaxed dataset was examined, where SNPs within 125 nucleotides were not filtered by preferentially retaining loci with the highest number of sequenced individuals, 52,825 SNPs were retained and the spread of the Sbf1 samples along the first two PC's was much improved (Appendix K Figure 6). A threshold of 30 individuals allowed the 33 individuals genotyped with Sbf1 to contribute more loci to the dataset. Without preferentially retaining loci with the most individuals sequenced, Sbf1 loci were more likely to be retained in the dataset. However an individual-based dendrogram of genetic distance once again showed clear separation by enzyme (Appendix K Figure 7). Less stringent filters did not substantively improve the quality of the dataset (i.e. remove the effect of enzyme) so the CT dataset generated using standard filters was retained for BayeScan and random forest analyses, as per the methods implemented in Chapter 3.

Range-wide genomic datasets were generally of high quality. The Sbf1 dataset, which was genotyped using GATK UnifiedGenotyper, included a total of 789 SNPs with

a mean of 534 SNPs per individual. Loci had a mean depth of 44 reads and a mean missingness of 0.44. The range-wide Pst1 ANGSD dataset retained 18,992 SNPs.

Genotype-Phenotype Associations

Parameter optimization runs did not detect significant improvements in the proportion of variation explained by different numbers of SNPs considered at each decision node (mtry) so initial SNP importance values were estimated by considering all SNPs at each node. The ntree parameter was set to 200,000 in order to ensure that genotype-phenotype associations were reliably identified. For each season, initial SNP importance values were highly correlated across three runs (winter > 99%; spring > 94%; autumn > 96%).

Subsets of top-ranked SNPs were used to fit new RF models. Initial RF models did not explain any of the observed variation in spring or autumn cold injury so these analyses were not pursued (Appendix L). However genotypic variation did explain a proportion of the phenotypic variation observed for winter cold injury (Figure 31). The proportion of phenotypic variance explained (PVE) was maximized by the top 2% of SNPs. After a backwards purging approach was applied to the top 4% of SNPs, less than 1% of loci were retained: eight SNPs predicted the maximum amount of variation in winter CI (Table 21). Cross-validation found that the eight SNPs associated with winter CI explained a median permuted variance of 48% with a median permuted MSE of only 0.001% (Appendix M).

Half of the SNPs associated with winter cold tolerance were located in known genic regions (Appendix N). In BLASTn, a 300-nucleotide sequence surrounding SNP

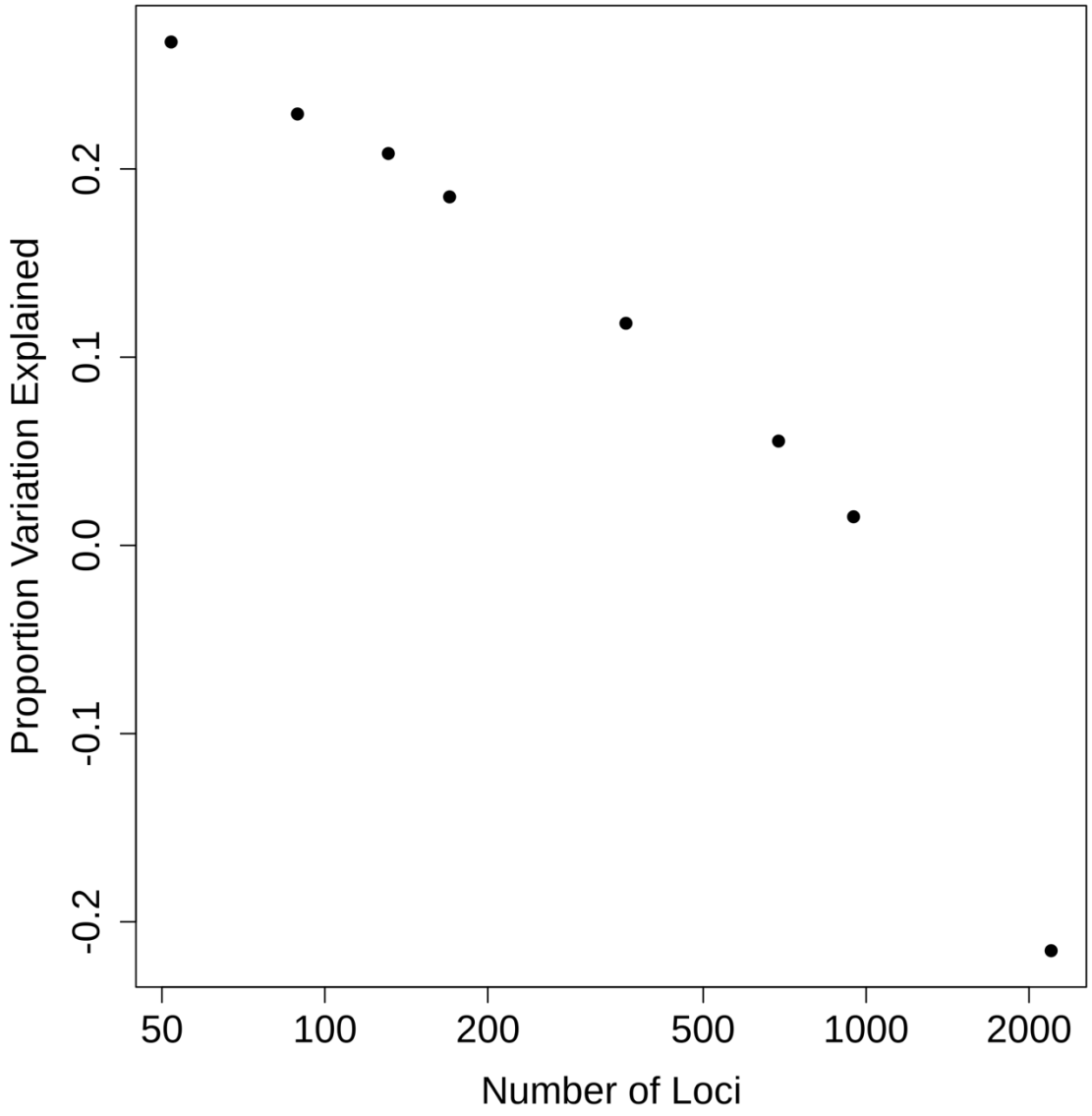

Figure 31. The top 2% of SNPs ranked based on initial importance value estimation predicted the highest proportion of observed variation in winter cold injury for 100 subalpine larch trees.

Table 21. SNPs that predict phenotypic variation in winter cold injury as per random forest analysis for 100 individuals of subalpine larch representing 18 populations from the northern portion of the range.

SNP	Major	Minor	BayeScan F_{ST} Outlier	bayenv2 environmental correlation	Correlated Climate Variables
Pseudo_151_6087320*	G	A	Pst1, CT ANGSD	-	-
Pseudo_225_5038052*	G	A	Pst1	-	-
Pseudo_247_15753552	T	A	-	Sbf1	Tmin_at, MAT, MWMT, MCMT, MSP, DD_0, DD5, NFFD, eFFP, EMT
Pseudo_271_24459634*	C	G	Pst1, CT ANGSD	Pst1	MSP
Pseudo_349_4095395*	G	A	-	-	-
Pseudo_405_917624	C	T	-	Sbf1	TD, MSP
Pseudo_709_6262212*	C	A	Pst1, CT ANGSD	-	-
Pseudo_734_3455909	T	G	Pst1, CT ANGSD	Sbf1	AHM

*Not genotyped in range-wide Sbf1 dataset thus no BayeScan or bayenv2 results possible

Pseudo_225_5038052 matched with 5.8S ribosomal RNA genes in *Larix decidua* (Accession MG216878.1; E-value = 5e-104), *Larix laricina* (Accession MF348954.1; E-value = 5e-104), *Larix lyallii* (Accession MG216886.1; E-value 4e-100), *Larix sibirica* (AF041345.1; E-value 4e-90) and others. Of the remaining seven SNPs, four could be mapped without error to an updated version of the genome and its associated annotation file. Three of these four SNPs (Pseudo_151_6087320, Pseudo_349_4095395, Pseudo_405_917624) were in genic regions identified as producing mRNA.

F_{ST} Outliers

Four datasets with different sample sizes and numbers of SNPs were tested for F_{ST} outliers (Table 22). Only one SNP was identified as an F_{ST} outlier in the CT UG dataset but 470 were detected when the same dataset was genotyped using ANGSD. For range-wide data, only 54 outlier loci were identified in the range-wide Sbf1 UG dataset but over 6,000 outliers were identified in the range-wide Pst1 ANGSD dataset.

There was some overlap between outliers detected in the different datasets. Although the two CT datasets shared 143 SNPs, none of these were F_{ST} outliers. Similarly no overlap in F_{ST} outliers was identified between the CT UG dataset and the range-wide Pst1 and Sbf1 datasets. However outlier loci from the CT ANGSD dataset had significant overlap with both the range-wide Pst1 dataset (369 SNPs) and the range-wide Sbf1 dataset (148 SNPs). Overlap between the two range-wide datasets was minimal: although 175 SNPs were shared between datasets, only eight were F_{ST} outliers (Table 23). Five of these outliers overlapped with the CT ANGSD dataset.

Table 22. Subalpine larch datasets used to test for F_{ST} outliers using BayeScan have different numbers of samples and SNPs.

Dataset	Genotyping Software	N	Loci	P >= 0.95
Cold tolerance (CT)	ANGSD	100	2,197	470
Cold tolerance (CT)	GATK UnifiedGenotyper	100	465	1
Range-wide Pst1	ANGSD	364	18,992	6,086
Range-wide Sbf1	GATK UnifiedGenotyper	275	789	54

Table 23. BayeScan F_{ST} outliers that overlap between range-wide Pst1 dataset genotyped with ANGSD and range-wide Sbf1dataset genotyped with GATK UnifiedGenotyper and their associated.

SNP	CT ANGSD BayeScan FST Outlier	bayenv2 environmental correlation	Correlated Climate Variables
Pseudo_222_20364504	-	-	-
Pseudo_248_20818947	Yes	-	-
Pseudo_272_3301043	Yes	-	-
Pseudo_334_391595	Yes	-	-
Pseudo_446_532154	-	Pst1	Tmin_at, MAT, MAP, AHM, PAS
Pseudo_638_8142505	Yes	Pst1	MCMT, TD, MSP, SHM, bFFP, CMD
Pseudo_705_2951648	Yes	-	-
Pseudo_820_4506024	-	Pst1, Sbf1	Tmin_at, MAT, MWMT, MCMT, DD_0, DD5, NFFD, bFFP , eFFP , FFP, EMT, EXT, eREF*, CMD*

*Correlated climate variables identified via bayenv2 analysis on Pst1 dataset versus Sbf1 dataset

Half of the eight SNPs that were identified as F_{ST} outliers in both range-wide datasets were located in known genic regions (Appendix O). Three hits were obtained from BLASTn. A 300-nucleotide sequence around SNP Pseudo_248_20818947 matched with mRNA sequences in *Picea sitchensis* (Accession EF678132.1; E-value 3e-76) and *Picea glauca* (BT102158.1; E-value 1e-74). Second, SNP Pseudo_272_3301043 matched with mRNA sequences in *Picea glauca* (Accession BT116120.1; E-value 1e-30) and *Picea sitchensis* (Accession HM197891.1; E-value 5e-24). The mRNA in *Picea sitchensis* was characterized as a glycosyl hydrolase-like protein (Holliday et al. 2010). Note, however that Pseudo_272_3301043 SNP was located between two regions of direct alignment, meaning it may be in tight linkage with the gene versus in the genic region itself. Third, SNP Pseudo_705_2951648 matched with hypothetical proteins in *Pinus taeda* (Accession JQ020377.1 and JQ18933.1; E-value = 7e-58) and *Pinus lambertiana* (Accession JQ020383.1; E-value 3e-56). Of the remaining five SNPs, two could be mapped without error to an updated version of the Siberian larch genome and its associated annotation file. SNP Pseudo_446_532154 was in a genic region identified as producing mRNA.

Genotype-Environment Associations

For the range-wide Sbf1 and Pst1 datasets, bayenv2 neutral covariance matrices accurately captured neutral genetic structure (Appendix P). Tested against this null, 1,159 genotype-environment relationships with at least moderate support (Bayes Factor >= 3.0) were identified in the range-wide Sbf1 UG dataset and 28,397 relationships were identified in the Pst1 ANGSD dataset (Table 24). There was very little overlap between

Table 24. Number of SNPs correlated with 20 environmental variables in bayenv2 for subalpine larch range-wide Sbf1 data genotyped with GATK UnifiedGenotyper and range-wide Pst1 data genotyped with ANGSD and their overlap.

Bayes Factor	Climate Variable																			
	AHM	**bFFP**	**CMD**	**DD_0**	**DD5**	**eFFP**	**EMT**	**Eref**	**EXT**	**FFP**	**MAP**	**MAT**	**MCMT**	**MSP**	**MWMT**	**NFFD**	**PAS**	**SHM**	**TD**	**Tmin_at**
Sbf1 UG Dataset																				
< 3.0	726	730	749	733	721	718	732	735	728	725	732	730	740	755	716	715	728	745	747	716
3.0 - 9.9	57	46	37	42	56	57	44	47	50	45	41	45	37	27	61	57	49	40	39	55
10.0 - 19.9	3	8	1	11	7	10	9	4	8	14	11	11	10	5	8	10	7	3	2	12
20.0 - 29.9	3	0	2	3	3	0	2	2	3	2	4	2	2	1	3	2	2	1	0	4
30.0 - 99.9	0	5	0	0	2	4	2	1	0	2	0	1	0	1	1	5	2	0	1	2
100 - 149.9	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0
> 150	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pst1 ANGSD																				
< 3.0	18303	17368	17846	17302	17168	17666	17104	17538	17483	17521	18174	17257	17095	17860	17070	17521	18330	17883	17566	17348
3.0 - 9.9	470	863	671	996	980	783	1093	874	873	815	532	969	1085	727	1012	803	406	649	844	940
10.0 - 19.9	87	259	191	275	308	212	297	236	262	237	112	278	335	148	309	259	110	184	223	258
20.0 - 29.9	28	119	56	88	124	86	133	86	98	99	47	139	118	55	136	99	43	56	70	116
30.0 - 99.9	64	210	117	184	228	133	191	139	179	151	65	191	177	107	249	167	61	105	124	184
100 - 149.9	9	30	23	28	41	36	37	30	28	42	13	31	42	16	48	36	8	19	19	30
> 150	29	141	86	117	141	74	135	87	67	125	47	125	138	77	166	105	32	94	144	114
Overlap																				
< 3.0	155	144	149	150	143	148	147	146	146	151	156	149	147	157	141	148	155	153	152	148
3.0 - 9.9	0	1	0	1	1	0	0	0	0	0	1	1	1	1	1	0	1	2	1	0
10.0 - 150	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
> 150	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

the datasets generated by the different enzymes: within climate variables, only 14 relationships were identified with at least moderate support between genotype and environment in both datasets, representing 12 unique SNPs. Across climate variables, 30 SNPs were identified with at least moderate support for at least one climate variable in both datasets (Table 25). This lack of overlap reflected the fact that there was very little overlap in general, with only 175 SNPs shared between datasets.

A subset of SNPs that overlap between range-wide bayenv2 datasets were located in known genic regions (Appendix Q). Three matches were identified in BLASTn. First, a 300-nucleotide sequence surrounding SNP Pseudo_198_22285064 overlapped somewhat (16% not including SNP location) with mRNA from *Picea glauca* (Accession BT110550.1; E-value 2e-13) and *Picea sitchensis* (Accession EF087801.1; E-value 2e-13). Second, SNP Pseudo_377_15881506 matched with mRNA from *Picea glauca* (Accession BT108636.1; E-value 1e-94). Third, SNP Pseudo_404_2501874 overlapped somewhat (48% not including SNP location) with mRNA from *Picea glauca* (Accession BT119855.1; E-value 1e-34). Of the remaining 27 SNPs, 18 could be mapped without error to an updated version of the genome and its associated annotation file. Seven of these 18 SNPs (Pseudo_252_21191933, Pseudo_254_22353006, Pseudo_297_6014296, Pseudo_508_9553361, Pseudo_530_3964246, Pseudo_655_5097907 and Pseudo_940_3109473) were in genic regions identified as mRNA, although exact function remains unknown.

Within datasets, there was a great deal of overlap between SNPs across climate variables (Appendix R). The 1,159 genotype-environment relationships identified in the Sbf1 dataset represented 304 unique SNPs and the 28,397 relationships in the Pst1 dataset

Table 25. SNPs identified by bayenv2 as being correlated with environmental gradients in both the range-wide Pst1 dataset genotyped using ANGSD and the range-wide Sbf1 dataset genotyped using GATK UnifiedGenotyper.

SNP	CT ANGSD BayeScan FST Outlier	Pst1 Correlated Climate Variables	Sbf1 Correlated Climate Variables	Overlappping Climate Variables
Pseudo_47_1721469	YES	bFFP	bFFP	bFFP
Pseudo_112_5519243	YES	MCMT	Tmin_at, MAT, DD5, NFFD, bFFP, eFFP, FFP	-
Pseudo_198_22285064	-	MWMT, AHM	MAP, MSP, AHM, PAS	AHM
Pseudo_236_2642487	YES	TD	AHM	-
Pseudo_252_21191933	-	TD, SHM	bFFP, eFFP, FFP	-
Pseudo_254_22353006	YES	MCMT	MAP, AHM, PAS	-
Pseudo_259_19308369	-	MCMT, eFFP	Eref, CMD	-
Pseudo_297_6014296	YES	MSP, SHM, NFFD, bFFP, FFP, CMD	Tmin_at, MSP, SHM, eFFP, EMT	MSP, SHM
Pseudo_330_14254778	YES	eREF	MAP, AHM, eFFP	-
Pseudo_351_6175164	-	SHM, CMD	MWMT, DD5, EXT	-
Pseudo_369_4132364	YES	MWMT, EXT, Eref	Tmin_at, MCMT, eFFP, EMT	-
Pseudo_377_15881506	YES	Tmin_at, MCMT, SHM, EMT	bFFP, CMD	-
Pseudo_404_2501874	YES	eREF	Tmin_at, NFFD, bFFP, eFFP, FFP	-
Pseudo_408_7240070	YES	bFFP	SHM	-
Pseudo_508_9553361	YES	MCMT, TD, MSP, SHM, bFFP, Eref, CMD	SHM, CMD	SHM, CMD
Pseudo_530_3964246	YES	MWMT, DD5, EXT	CMD	-
Pseudo_561_10943560	-	MWMT, Eref	eFFP, FFP	-
Pseudo_596_1954353	-	TD	TD, bFFP	TD

SNP	CT ANGSD BayeScan FST Outlier	Pst1 Correlated Climate Variables	Sbf1 Correlated Climate Variables	Overlappping Climate Variables
Pseudo_602_8643462	-	MAP, AHM, PAS	MAP, PAS	MAP, PAS
Pseudo_603_10102834	YES	bFFP	EXT, Eref	-
Pseudo_655_5097907	YES	TD	MAP, PAS	-
Pseudo_678_759073	-	Tmin_at, MAT, MWMT, MCMT, MSP, DD_0, DD5, NFFD, eFFP, FFP, EMT, EXT, Eref	SHM, CMD	-
Pseudo_680_3876454	YES	bFFP	Tmin_at, MAT, MWMT, DD5, CMD	-
Pseudo_696_4766119	YES	MCMT, TD, MSP, SHM, bFFP, CMD	Tmin_at, MAT, MCMT, DD_0, EMT	MCMT
Pseudo_707_5269576	YES	MAT, MWMT, DD_0, DD5, EXT	Tmin_at, MAT, MWMT, MCMT, DD_0, DD5, NFFD, bFFP, eFFP, FFP, EMT	MAT, MWMT, DD_0, DD5
Pseudo_741_262116	-	MAP, AHM	NFFP	-
Pseudo_817_2066018	YES	SHM	NFFD, bFFP	-
Pseudo_819_1710072	YES	TD	MAT, MWMT, DD5, NFFD, bFFP, EXT	-
Pseudo_820_4506024	YES	Eref, CMD	Tmin_at, MAT, MWMT, MCMT, DD_0, DD5, NFFD, bFFP, eFFP, FFP, EMT, EXT	-
Pseudo_940_3109473	-	NFFD, bFFP, FFP	Tmin_at, MAT, MWMT, MCMT, DD_0, DD5, EMT, EXT, Eref, CMD	-

represented 7,195 unique SNPs. When Spearman's ρ was plotted for each climate variable, clear relationships between groups of SNPs emerged. Clusters of SNPs are correlated with groups of correlated temperature-associated climate variables, such as DD_0, minimum autumn temperature (Tmin_at), end of the frost-free period (eFFP), extreme minimum temperature (EMT) and MCMT (Figure 32). SNPs generated from Pst1 data and Sbf1 data exhibit clear overlap across temperature-associated climate variables (e.g. EMT, DD_0 and MCMT). These results provided support for the polygenic nature of cold tolerance traits involved in local adaptation to climate among populations of subalpine larch.

Overlap Between Analyses

SNPs that predicted phenotypic variation in winter cold tolerance overlap with F_{ST} outliers and SNPs correlated with environmental gradients (Table 21). Five SNPs were identified as F_{ST} outliers by BayeScan. Four SNPs have at least moderate support for relationships with environmental gradients, including temperature-associated variables (Appendix S). Two SNPs are both BayeScan outliers and correlated with environmental gradients. These results connected SNPs that explained phenotype directly to F_{ST} and clines in environmental variation.

Eight F_{ST} outliers identified by BayeScan overlapped between range-wide Pst1 and Sbf1 datasets (Table 23). Three of these SNPs were also correlated with climate variables as per bayenv2 analysis (Appendix T).

There was significant overlap between F_{ST} outliers identified by BayeScan and SNPs correlated with environmental gradients identified by bayenv2. Within the range-

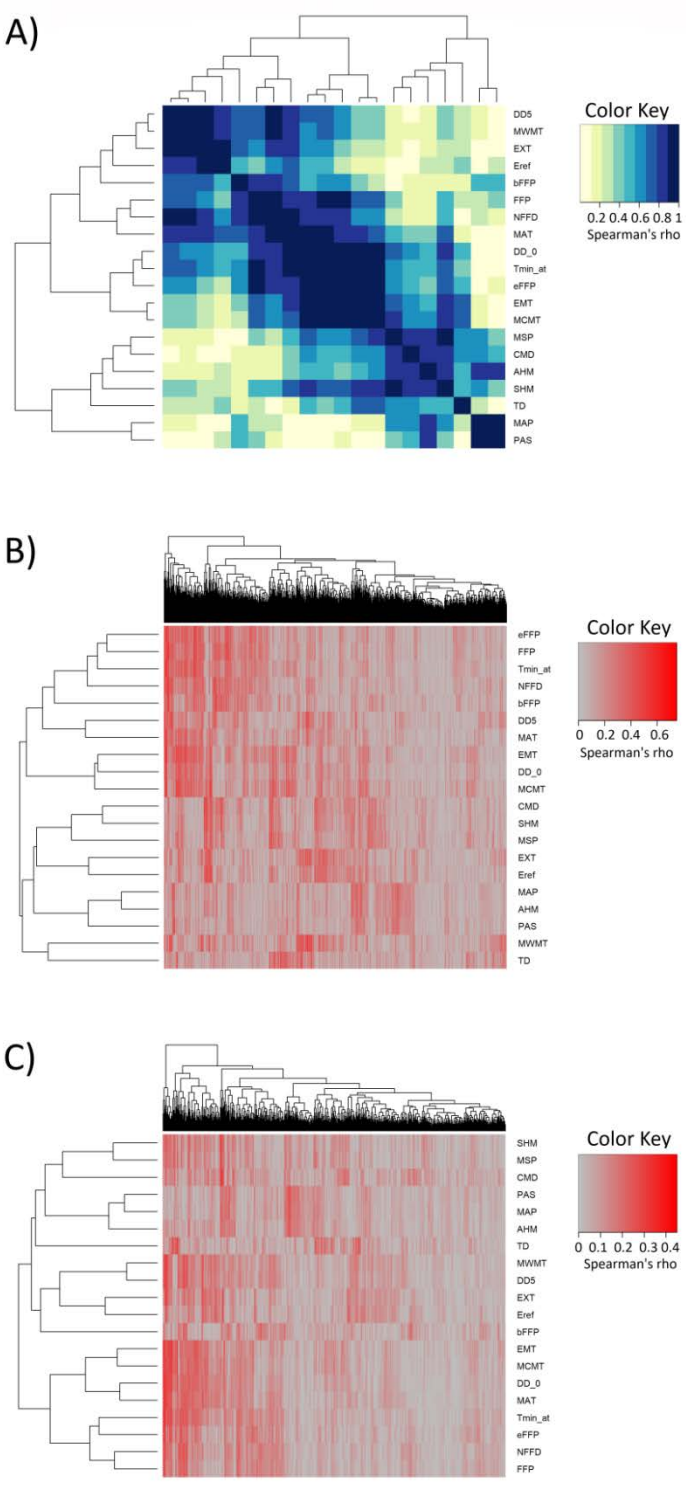

Figure 32. Correlations between climate variables on the y-axis and (A) climate variables, (B) Pst1 SNPs and (C) Sbf1 SNPs on the x-axis. Note that SNPs were generated for 100 subalpine larch trees representing 18 populations from the northern portion of the species range.

wide Sbf1 dataset, which includes 789 SNPs, 13 SNPs were identified that had both a high probability of being an F_{ST} outlier according to BayeScan (>= 0.95) and at least moderate support (BF >= 3.0) for correlation with at least one environmental variable. Within the range-wide Pst1 dataset, which includes 18,992 SNPs, 2,370 SNPs were identified that had both a high probability of being an F_{ST} outlier and at least moderate support for correlation with one or more environmental variable. A subset of 276 SNPs had very strong support (BF >= 150). Across datasets and climate variables, 30 SNPs with at least moderate support for an environmental correlation (BF >= 3.0) overlapped between the range-wide Pst1 and range-wide Sbf1 datasets (Table 25; Appendix U). All 30 of these SNPs were also F_{ST} outliers according to BayeScan analyses for the range-wide Pst1 and Sbf1 datasets. Twenty of these SNPs were also identified as F_{ST} outliers during BayeScan analysis of the CT ANGSD data.

Discussion

Subalpine larch exhibits local adaptation for cold tolerance within the northern portion of its range. Although timberline is often described as a relatively consistent environment (Richardson and Friedland 2009), a discriminant analysis of principal components found that 18 populations of subalpine larch occupy two climatically differentiated regions within a restricted latitudinal range (49.05 – 51.35˚). The northern region, comprising seven populations from the northern Rockies, was significantly colder and wetter than the southern region, which includes nine populations from the southern Rockies and two populations from the northern Cascades. Individuals from the north had significantly higher cold tolerance in all three seasons (winter, spring, autumn) compared to individuals from the south, providing strong evidence that environmental differentiation has shaped phenotypic variation. Partial redundancy analysis confirmed that temperature-associated climate variables were significant predictors of winter and autumn cold tolerance even when models were conditioned on geography (latitude and longitude). This was particularly evident for winter cold tolerance, for which the length of the frost-free period (FFP) and the number of degree-days below zero (DD_0) predicted 65% of the explainable variation in winter cold injury, while geography only accounted for 3%. Trees from environments with a shorter FFP and more DD_0 have higher winter cold tolerance. When cold tolerance measurements were averaged by population and plotted against climate predictors, clear phenotypic clines emerged. Winter cold tolerance exhibited the strongest cline in relation to FFP (R^2 = 68%) but autumn cold tolerance also showed a strong relationship with mean coldest month

temperature (R^2 = 54%). These results provide the first evidence of local adaptation in subalpine larch.

Cold tolerance is an important adaptive trait in conifers. Around the world, cold tolerance has allowed conifers to dominate temperate and boreal ecosystems. Adaptive differences among species have long been understood to play a key role in defining range boundaries (Sakai and Weiser 1973). Phenotypic clines associated with latitude and elevation have been observed in many different species (Howe et al. 2003). For most conifers in western North America, latitudinal clines generally arise over large distances (e.g. Liepe et al. 2016). Because conifers are wind-pollinated, pollen dispersal from the core of the range can easily swamp out local adaptation on the margins (Kawecki 2008). Off the eastern coast of the United States, viable pollen from loblolly pine pollen has been observed 41 km from shore (Williams 2010). In the Canadian Arctic, an unusual atmospheric event led to the deposition of pine and spruce pollen thought to have originated almost 3,000 km away (Campbell et al. 1999). However subalpine larch demonstrates phenotypic differentiation over a relatively small geographic scale. Local adaptation evolves when the strength of selection is greater than the rate of migration (Haldane 1930). At timberline it seems likely that selection for cold tolerance is strong. Indeed, western larch (*Larix occidentalis*) populations separated by 500 m in elevation exhibit genetic differentiation for spring frost damage (Rehfeldt 1995). It is also possible that local adaptation has been facilitated by restricted gene flow in the Rocky Mountains. Prevailing winds from the southwest in combination with rugged topography may limit south-north pollen dispersal in the spring. Divergence in phenological traits correlated with cold tolerance, such as the timing of bud flush, encourages assortative mating within

regions (Soularue and Kremer 2014). A similar pattern of north-south divergence for freezing resistance in Japanese larch (*Larix kaempferi*) was also observed over a similarly small area (35 – 37˚N) in the mountains of central Honshu (Scheumann and Schonbach 1968; Okada et al. 1971).

It is likely that climate has played an important role in shaping variation in spring and autumn cold tolerance even though partial redundancy analysis (pRDA) did not find that climate was a significant predictor of cold injury in these seasons. Previous work on Douglas-fir found that spring and autumn cold tolerance were not genetically correlated so results were not expected to be the same across seasons (O'Neill et al. 2000, 2001). Still, late spring frosts and early autumn frosts are the most likely to cause damage—as conifers undergo deacclimation and acclimation, respectively—and these traits are generally understood to be under strong selection. Several factors could be obfuscating the relationship between climate and cold injury. First, spring sampling may have been late. Spring cold injury values were high and variation among individuals was also high, suggesting that deacclimation may have progressed too far at the time of sampling. Worrall (1993) found that subalpine larch trees growing at Blackwall Mountain, B.C., have a threshold temperature requirement of only 1.46˚C and can initiate bud burst after 67 ± 4.5 growing degree days (GDD). In Worrall's study, bud burst was considered to have occurred when leaves extended 0.25 cm beyond bud scales and half of all buds that would eventually flush had flushed. In this study, trees were sampled after 215 and 115 GDD across the two years. While these values are considerably higher than Worrall's estimate of the GDD required for flushing, only three trees had visible needles in 2015 and no trees showed signs of flushing in 2016 so it is not clear how big a role timing

played. Second, climate gradients in western North America are highly correlated with geographic gradients so accounting for geography could mask the effects of climate. For example, the two climate variables identified during redundancy analysis of spring cold injury, MCMT and TD, are strongly correlated with latitude and longitude, respectively (> 0.80). Accounting for neutral genetic structure thus weakens the explanatory power of the model. However it is important to note that significant differences in spring and autumn cold tolerance were observed between northern and southern regions, and climate variables were significant predictors of cold tolerance when the model was not conditioned on geography. Indeed, although the relationships were not as strong as those for winter cold tolerance, phenotypic clines were observed for all seasons.

Subalpine larch demonstrated high levels of cold tolerance. After freezing at -40 ˚C, mean winter and autumn damage stayed below 50 %. This was expected given the results of previous cold-tolerance testing on closely related species. Sakai and Weiser (1973) found that bud and twig tissues collected mid-winter from *Larix laricina* demonstrated freezing resistance at -80 ˚C in a freezer and -196 ˚C in liquid nitrogen. *Larix occidentalis* twigs collected in Idaho also showed freezing resistance up to -50 ˚C. A more recent study found that relative conductivity (i.e. electrolyte leakage) for *Larix occidentalis* seedlings frozen at -18 ˚C dropped below 25 % in late October (L'Hirondelle et al. 2006). By comparison, mean damage for samples frozen at -20 ˚C was never above 25 %, regardless of season. Although the exact physiological mechanisms for cold tolerance in subalpine larch are not yet understood, the relatively high levels of tolerance observed in this study demonstrate that subalpine larch is well adapted to life at timberline, where extreme minimum temperatures can reach -50 ˚C. Future research

should seek to clarify the specific mechanisms that allow subalpine larch to thrive in its high-elevation habitat.

The Genetic Basis of Local Adaptation

This study identified loci associated with winter cold tolerance in subalpine larch. Random forest analysis identified eight SNPs as predictors of winter cold injury. Together, they explain a median of 48% of observed phenotypic variation. This is in line with results obtained from previous studies (Howe et al. 2003). In Douglas-fir (*Pseudotsuga menziesii*), 11 QTLs each explained an average of 4.1% of the phenotypic variance in fall cold hardiness (Jermstad et al. 2001). In Sitka spruce (*Picea sitchensis*), 19 genic SNPs identified via association analysis explained 28.1% of variation in winter cold hardiness (Holliday et al. 2010). Five of the SNPs identified in this study were also identified as F_{ST} outliers by BayeScan in the range-wide Pst1 dataset, providing additional evidence that these loci are subject to selection. Furthermore, two F_{ST} outliers are correlated with environmental gradients: SNP Pseudo_271_24459634 is correlated with variation in mean summer precipitation and SNP Pseudo_734_3455909 is correlated with variation in annual heat-moisture index. Although temperature-associated variables are known to be important drivers of cold adaptation in the conifers in western North America and SNP Pseudo_247_15753552 is indeed strongly correlated with gradients for MAT, MCMT, DD_0 and EMT, correlations with MSP are more common in this dataset. It is possible that genetic variation for cold tolerance may be associated with other stress responses. One phytohormone, abscisic acid, is known to induce gene expression in response to both cold and drought stress (Bray 1993). Cellular water deficits are common

to both stress conditions. In this respect, subalpine larch's ability to isolate the apical and lateral buds from the xylem tissue overwinter would appear to play a key role in survival in these extreme environments (Richards and Bliss 1986). Subalpine larch trees cope with extremely low water potential in their buds in order to reduce cellular freezing and winter damage. Although this ability to cope with low water potential during winter is a cold-associated trait, it is possible that it could impart some benefit with respect to drought stress and variation in MSP. Thus it is possible that local adaptation for cold tolerance may maintain standing genetic variation with adaptive value for other traits. Such variation may prove especially valuable for facilitating future adaptation given that drought events are predicted to increase with climate change (IPCC 2013).

Elucidating the genomic basis of local adaptation in subalpine larch was not an easy task. A recent study in lodgepole pine was able to explain roughly 70% of variation in autumn cold injury using three PC's generated from 196 phenotype-associated SNPs identified using randomForest (Mahony et al. 2019). However, that study had much larger sample sizes, both in terms of the number of individuals phenotyped (1,594 trees) and the number of SNPs used (31,634 SNPs were obtained after filtering from an Affymetrix Axiom 50K SNP Array). With only 100 individuals and 2,197 SNPs, the power to detect genotype-phenotype associations was limited in this study. Furthermore, the limited genomic resources available for poorly studied non-model organisms present inherent limitations. RAD-seq does not include a variant discovery process that generates targeted probes. Instead, the whole genome is subsampled using a restriction enzyme. In this study, two challenges arose as a result of this process. First, missing data was rife. Although genotypes were called for all individuals using a likelihood-based approach in

ANGSD, up to 50% of genotypes were missing at each locus. This limits the precision and accuracy of association analysis, especially when sample size is small. Second, samples were processed using two different restriction enzymes, which produced unexpectedly different results. Although the Pst1 restriction enzyme cut site sequence is a subset of the longer Sbf1 cut site sequence, the same loci were not cut and amplified in both datasets, making it more difficult to merge the datasets for analysis. The fact that BayeScan analysis and bayenv2 analysis produced SNPs of interest that overlapped across datasets speaks partly to luck (i.e. that enough data was generated in both datasets) and partly to the strong signals coming from these particular loci. Future work should seek to increase sample size and provide consistency in terms of genotyping effort to further explore the loci identified in this study. Finally, the ability to ascribe function when SNPs of interest are identified was limited in this study due to limited genomic resources. Because reads were aligned to a highly fragmented version of the Siberian larch genome, and relatively few resources exist for conifers in terms of functional information on genes, it is not possible to tell whether or not all SNPs are in coding regions and what impact they may have on the biology of the species. Although the exact function of the sequences where SNPs in this study were found is not known, it is possible that they lie within genic regions, within regulatory elements, or linked non-coding DNA. If function can be elucidated, this may eventually provide an understanding of the mechanisms driving genetic differentiation for cold tolerance in populations of subalpine larch.

Future Perspectives

Although marker-based management strategies often focus on the conservation of genetically divergent units and/or the preservation of genetic diversity, local adaptation may result from genetic divergence at relatively few loci. In this study, eight loci were identified that predict a significant portion of the variation in one trait, winter cold tolerance. This confirms that this complex physiological trait is polygenic and suggests that more comprehensive sequencing would likely reveal additional markers. However the fact that nearly half of the variation in this trait can be explained by eight markers indicates that these markers should be prioritized for conservation efforts. Marker-based selection could target warm-adapted alleles if conservation is to occur *in situ*. If an assisted migration strategy is considered, then cold-adapted populations should be prioritized for northward transfer. Finally, the association between SNPs that predict cold tolerance and mean summer precipitation indicates that assessing adaptation in a single trait is likely to provide a limited view of the species biology. Future work on local adaptation in subalpine larch should incorporate populations from the southern portion of the species range, which may show stronger drought adaptation, and consider additional traits linked to drought adaptation.

CHAPTER 5: CONCLUSIONS AND FUTURE PERSPECTIVES

This research was the first to describe intraspecific patterns of genetic variation in the timberline conifer, subalpine larch. Genetic structure across the species range was assessed using 61 populations of subalpine larch. Three major genetic clusters were identified using principal components analysis (PCA), a discriminant analysis of principal components (DAPC) and Bayesian STRUCTURE analysis. A dendrogram of Provesti's genetic distance identified a fourth monophyletic clade in the central Rocky Mountains, near the US-Canadian border. However, bootstrap support for this cluster was low, indicating that its genetic differentiation is dependent on relatively few loci. The genetic structure of subalpine larch was confirmed using sequencing data generated by a different restriction enzyme (Pst1 versus Sbf1). SNP data were generated with a different genotyping approach (likelihood-based ANGSD approach versus a traditional SNP filtering pipeline). Observed heterozygosity was found to be low and inbreeding coefficients were found to be high within populations. F_{ST} estimates of genetic differentiation between populations were also high, suggesting that subdivided populations are experiencing reproductive isolation. Negative values of Tajima's D point to a demographic history of expansion. Demographic parameters estimated from fastsimcoal2 coalescent modelling found relatively high estimates of current effective population size but also indicated that populations may be shrinking, especially in the south. Adaptive genetic variation in the cold tolerance trait was observed. Two environmental regions were identified within the northern portion of the species range based on a DAPC analysis of 20 climate variables. In all seasons, populations from the northern Rocky Mountains had significantly higher cold tolerance than populations from

the southern Canadian Rocky Mountains and the northern Cascades. Redundancy analysis indicated that climate accounted for the largest proportion of explainable variance in winter cold tolerance. Strong clines for winter cold injury were observed in relation to the frost-free period, as well as degree days below zero. Phenotypes were correlated with genotypes using random forest analysis in R. Eight SNPs were identified that explained a significant portion of the phenotypic variance in winter cold tolerance. Genotypes were also correlated with environmental gradients using bayenv2. Many SNPs had significant environmental associations, even after correcting for neutral genetic structure. This confirmed that adaptive traits in subalpine larch are polygenic. Altogether, this study provided valuable insights into the biogeographic history and future prospects of subalpine larch, a species that is likely to face serious challenges as climate change progresses.

Long-lived tree species living at high elevation are thought to be extremely vulnerable to climate change. The fact that these species cannot track their shifting climate niches means that they will be forced to adapt *in situ* to relatively rapid environmental change. A long generation time means that adaptation is likely to lag behind the rate of environmental change. Increasingly, populations will become maladapted to their environments. They will be more likely to suffer both increased mortality and reduced reproductive success. Local extirpation of severely maladapted populations may disproportionately reduce the effective size of the population. Small populations are more likely to be affected by genetic drift, potentially leading to the random fixation of deleterious alleles. Small populations may also experience higher levels of inbreeding, increasing the potential for inbreeding depression. Inbreeding can

thus further reduce reproductive success, initiating an extinction vortex. As directional climate change proceeds in the mountains of western North America, subalpine larch populations will face abiotic challenges due to late-summer drought events and biotic competition as other species shift their ranges to higher elevation in order to track their own shifting climate niches. Given the low genetic diversity, the high inbreeding coefficients, and the high F_{ST} values found in this study, it appears that subalpine larch has already experienced increased fragmentation and population subdivision as a result of Holocene warming and climate change. Such patterns have been observed in other conifers with fragmented ranges. Taken together, these patterns indicate that subalpine larch is poorly positioned to respond effectively to environmental change and to sustain itself into the future without assistance.

Some of the results of this study do not bode well for future persistence of the species under predicted climate change scenarios. To this end, management and conservation in subalpine larch may involve different strategies. It has been suggested that a combination of *in situ*, *inter-situ* and/or *ex situ* strategies are appropriate for the management of forest trees. In British Columbia, the Forest Genetics Council has, to date, focused on *in situ* conservation within protected areas. However, given that climate change is creating significant challenges within the current range of subalpine larch, *inter situ* and *ex situ* options should also be explored. At present, there is only one *inter situ* resource for subalpine larch, which is the breeding arboretum at the Kalamalka Forestry Centre in Vernon, BC. However this collection may not provide the best opportunity for long-term conservation for the following reasons. First, the collection has a small sample size within populations. A median of 10 individuals per population remain out of the 30

that were originally collected; some populations are represented by as few as four individuals. Second, the site is too hot and too dry for high-elevation species. Successfully grafted subalpine larch trees have experienced significant mortality at the Kalamalka site. Mean annual temperature (MAT) at Kalamalka is 8 ˚C while MAT in natural populations, including the southern portion of the species range, is much lower (-0.3 ˚C). Furthermore, summer temperatures in the Okanagan Valley regularly exceed 30 ˚C and occasionally surpass 40 ˚C. The trees in this particular *inter situ* collection are therefore reliant on a functional irrigation system and could face severe stress if that system were to fail. In my opinion, the trees in this collection should be re-grafted and planted at a second location with a cooler climate. Finally, *ex situ* seed collections could be expanded. To date, Alberta, Idaho and Montana have none. Washington State has seed collections from two populations. The Province of British Columbia only has five seed lots of subalpine larch at the Surrey Tree Seed Centre and these are generally considered low quality. Expanding *inter situ* and *ex situ* resources would ensure that the Province of British Columbia is prepared to engage in active management if populations of subalpine larch suffer ill effects due to climate change.

Building seed collections would allow for more active management strategies to be implemented, such as assisted migration. Because populations are unlikely to track their shifting climate niches, it may be appropriate to move them into areas that are likely to be climatically suitable in the future. For subalpine larch, that would mean moving trees north. Species distribution models could be used to identify appropriate sites for planting. Subalpine larch could thrive on these sites. Often, a species' fundamental niche is not the same as its realized niche. It seems likely that subalpine larch could survive and

thrive in more northern habitats that it simply didn't reach due to slow migration rates. Local adaptation for cold tolerance should be considered when deciding which populations to move north. Common-garden provenance or progeny tests should be established to study adaptive traits such as growth, disease resistance and drought tolerance. Assisted migration could thus provide multiple benefits to buffer against predicted climate change and possible maladaptation within subalpine larch.

Drought tolerance is a trait that is likely to become increasingly important, especially along the southern range margin. This study identified several allelic clines associated with drought-associated climate variables. As the climate continues to get warmer, it is very likely that subalpine larch will be required to adapt to drier summer conditions in order to persist within its current range. Currently, drought screening protocols are being developed at the Kalamalka Forestry Centre for commercially important tree species, which are also experiencing drought events that prevent establishment during reforestation. These same protocols could be applied to subalpine larch if seed becomes available for testing.

The markers developed in this study have potential for future use. They could be used to reveal mechanisms of cold tolerance subalpine larch. Such studies depend on more conifer genomic resources becoming available in the future, which may not be simple. Conifer genomes are difficult to assemble because of both their large size and their high proportion of repetitive elements. Recent progress has been made to assemble longer contiguous sequences for loblolly pine and *Sequoiadendron giganteum*. Using Hi-C technology, researchers have now assembled a version of the sequoia genome that includes a single chromosome a billion nucleotides in length. There are also genome

assemblies for Norway spruce, white spruce and Siberian larch. Having contiguous sequence to align to would allow for the identification of selective sweeps that signal divergence as well as architectural features such as inversions that may play an important role in local adaptation. Unfortunately, these cannot be identified from the highly fragmented genome used in this study. Transcriptomics approaches may also be helpful for elucidating adaptive differences between individuals and populations. Future studies could therefore continue to explore the molecular basis of adaptive genetic variation in subalpine larch.

Future work should seek to expand on the results of this study. Alignment to the recently-published Siberian larch mitochondrial genome and chloroplast genome would provide additional insight into the biogeographic history of subalpine larch. *Larix*'s three genomes have different modes of inheritance: the chloroplast genome is paternally inherited, the mitochondrial genome is maternally inherited and the nuclear genome is biparentally inherited. Fine-grained genetic structure could be explored with organelle-specific markers. Thinking beyond subalpine larch, the markers developed in this study could be compared to other high-elevation larch species to look for convergent evolution. Comparisons with low-elevation larch species may provide insights into the ecological differentiation of these species. Finally, additional sequence data could be used to resolve, once and for all, the phylogeny of the North America larches.

In conclusion, this study provided a new perspective on subalpine larch, an iconic species in the mountain ecosystems of western North America. The information obtained from this study provides a solid foundation for all future management and conservation efforts. It also raises new avenues of inquiry that could be explored in order to ensure that

subalpine larch remains a viable component of the forest ecosystem, and that future generations can enjoy the beauty of this high-elevation conifer species.

Bibliography

Aguilar, R., Quesada, M., Ashworth, L., Herrerias-Diego, Y., and Lobo, J. (2008). Genetic consequences of habitat fragmentation in plant populations: susceptible signals in plant traits and methodological approaches. *Molecular Ecology*, 17: 5177-5188.

Aitken, S.N., Yeaman, S., Holliday, J.A., Wang, T., and Curtis-McLane, S. (2008). Adaptation, migration or extirpation: climate change outcomes for tree populations. *Evolutionary Applications*: doi:10.1111/j.1752-4571.2007.00013.x

Alberto, F.J., Aitken, S.N., Alía, R., Gonzáles-Martínez, S.C., Hänninen, H., Kremer, A., Lefèvre, F., Lenormand, T., Yeaman, S., Whetten, R., and Savolainen, S. (2013). Potential for evolutionary response to climate change – evidence from tree populations. *Global Change Biology*, 19: 1645-1661.

Alvarez, A.D. (1994). Analisis historico-ecologico de los bosques de *Pseudotsuga* en Mexico. Folieto Tecnico No. 24. Secretaria de Agricultura y Recursos Hidraulicos, Instituto Nacional de Investigaciones Forestales y Agropecurias, Centro de Investigacion Regional del Golfo Centro Campo Experimental el Palmar, Ver. Mexico.

Amos, W., and Balmford, A. (2001). When does conservation genetics matter? *Heredity*, 87: 257-265.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Andrews, K.R., Good, J.M., Miller, M.R., Luikart, G., and Hohenlohe, P.A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, 17: 81-92.

Armenise, L., Simeone, M.C., Piredda, R., and Schirone, B. (2012). Validation of DNA barcoding as an efficient tool for taxon identification and detection of species diversity in Italian conifers. *European Journal of Forest Research*, 131: 1337-1353.

Arno, S.F. (1990). Alpine Larch. *In* Silviculture of North America, Volume I, Conifers Agricultural Handbook. *Edited by* Burns, R.M., and Honkala, B.H. USDA Forest Service. Washington, DC, USA. Pp. 152-159.

Arno, S.F., and Habeck, J.R. (1972). Ecology of alpine larch (*Larix lyallii* Parl.) in the Pacific Northwest. *Ecological Monographs*, 42(4): 417-450.

Axelrod, D.I. (1965). The Miocene Trapper Creek flora of southern Idaho. *Journal of Paleontology*, 39(3): 510-511.

Axelrod, D.I. (1968). Tertiary floras and topographic history of the Snake River Basin, Idaho. *Geological Society of America Bulletin*, 79(6): 713-734.

Axelrod, D.I. (1990). Age and origin of subalpine forest zone. *Paleobiology*, 16(3): 360-369.

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A., and Johnson, E.A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, 3(10): e3376 (7 pp.).

Bentz, B.J., Régnière, J., Fettig, C.J., Hansen, E.M., Hayes, J.L., Hicke, J.A., Kelsey, R.G., Negrón, J.F., and Seybold, S.J. (2010). Climate change and bark beetles in the western United States and Canada: direct and indirect effects. *BioScience*, 60(8): 602-613.

Berg, E.E., and Chapin, F.S. III. (1994). Needle loss as a mechanism of winter drought avoidance in boreal conifers. *Canadian Journal of Forest Research*, 24: 1144-1148.

Blair, L.M., Granka, J.M., and Feldman, M.W. (2014). On the stability of the bayenv method in assessing human SNP-environment associations. *Human Genomics*, 8: 1 (13 pp.).

Booth, D.B., Troost, K.G., Clague, J.J., and Waitt, R.B. (2003). The cordilleran ice sheet. *Developments in Quaternary Science*, 1: 17-43.

Bougeard, S., and Dray, S. (2018). Supervised multiblock analysis in R with the ade4 package. *Journal of Statistical Software*, 86(1): 1-17.

Bray, E.A. (1993). Molecular responses to water deficit. *Plant Physiology*, 103: 1035-1040.

Breiman, L. (2001). Statistical modelling: the two cultures. *Statistical Science*, 16(3): 199-215.

Brieuc, M.S.O., Waters, C.D., Drinan, D.P., and Naish, K.A. (2018). A practical introduction to Random Forest for genetic association studies in ecology and evolution. *Molecular Ecology Resources*, 18: 755-766.

Cannell, M.G.R., and Smith, R.I. (1984). Spring frost damage on young *Picea sitchensis*. 2. Predicted dates of budburst and probability of frost damage. *Forestry*, 57: 177-197.

Cannell, M.G.R., Murray, M.B., and Sheppard, L.J. (1985a). Frost avoidance by selection for late budburst in *Picea sitchensis*. *Journal of Applied Ecology*, 22: 931-941.

Cannell, M.G.R., Sheppard, L.J., Smith, R.I., and Murray, M.B. (1985b). Autumn frost damage on young *Picea sitchensis.* 2. Shoot frost hardening, and the probability of frost damage in Scotland. *Forestry*, 58: 145-166.

Campbell, I.D., McDonald, K., Flannigan, M.D., and Kringayark, J. (1999). Long-distance transport of pollen into the Arctic. *Nature*, 399: 29-30.

Carles, S., Lamhamedi, M.S., Stowe, D.C., Veilleux, L., and Margolis, H.A. (2012). An operational method for estimating cold tolerance thresholds of white spruce seedlings in forest nurseries. *The Forestry Chronicle*, 88(4): 448-456.

Carlson, C. (1965). Interspecific hybridization of *Larix occidentalis* and *Larix lyallii*. MSc Dissertation, University of Montana. 52 pp.

Carlson, C.E. (1994). Germination and early growth of western larch (*Larix occidentalis*), alpine larch (*Larix lyallii*), and their reciprocal hybrids. *Canadian Journal of Forest Research,* 24: 911-916.

Carlson, C.E., and Theroux, L.J. (1993). Cone and seed morphology of western larch (*Larix occidentalis*), alpine larch (*Larix lyallii*), and their hybrids. *Canadian Journal of Forest Research*, 23: 1264-1269.

Carlson, C.E., Cates, R.G., and Spencer, S.C. (1991). Foliar terpenes of a putative hybrid swarm (*Larix occidentalis* x *Larix lyallii*) in western Montana. *Canadian Journal of Forest Research*, 21: 876-881.

Catchen, J.M., Amores, A., Hohenlohe, P., Cresko, W., and Postlethwait, J.H. (2011). Stacks: building and genotyping loci *de novo* from short-read sequences. *G3: Genes, Genomes, Genetics*, 1: 171-182.

Catchen, J., Hohenlohe, P.A., Bassham, S., Amores, A., and Cresko, W.A. (2013). Stacks: an analysis tool set for population genomics. *Molecular Ecology*, 22: 3124-3140.

Clague, J.J., and Mathewes, R.W. (1989). Early Holocene thermal maximum in western North America: new evidence from Castle Peak, British Columbia. *Geology*, 17: 277-280.

Coop, G., Witonsky, D., Di Rienzo, A., and Pritchard, J.K. (2010). Using environmental correlations to identify loci underlying local adaptation. *Genetics*, 185: 1411-1423.

Critchfield, W.B. (1984). Impact of the Pleistocene on the genetic structure of North American conifers. *In* Proceedings of the Eight North American Forest Biology Workshop. *Edited by* Lanner, R.M. Utah State University, Logan, Utah, USA. Pp. 70-118.

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., McVean, G., Durbin, R., and 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics, 27*(15): 2156-2158.

Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M., and Blaxter, M.L. (2011). RADSeq: next-generation population genetics. *Nature Reviews Genetics*, 12: 499-510.

Doyle, J. (1918). Observations on the morphology of *Larix leptolepis*. *Scientific Proceedings Royal Dublin Society*, 15: 310-327 + 2 pls.

Dray, S., and Dufour, A. (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software*, 22(4): 1-20.

Dyke, A.S. (2004). An outline of North American deglaciation with emphasis on central and northern Canada. *Developments in Quaternary Sciences*, 2(B): 373-424.

Edwards, S.V., Shultz, A.J., and Campbell-Station, S.C. (2015). Next-generation sequencing and the expanding domain of phylogeography. *Folia Zoologica*, 64(3): 187-206.

Elias, S.A. (2002). Rocky Mountains. Smithsonian Institution Press, Washington, DC, USA. 164 pp.

Elleouet, J.S., and Aitken, S.N. (2018). Exploring Approximate Bayesian Computation for inferring recent demographic history with genomic markers in nonmodel species. *Molecular Ecology Resources*, 18: 525-540.

Elleouet, J.S., and Aitken, S.N. (2019). Long-distance pollen dispersal during recent colonization favors a rapid but partial recovery of genetic diversity in *Picea sitchensis*. *New Phytologist*, 222: 1088-1100.

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simualation study. *Molecular Ecology*, 14: 2611-2620.

Evans, J.S., and Cushman, S.A. (2009). Gradient modeling of conifer species using random forest. *Landscape Ecology*, 24: 673-683.

Evans, J.S., and Murphy, M.A. (2018). rfUtilities. Available online at: https://cran.r-project.org/package=rfUtilities

Farjon, A. (1990). Pinaceae. Drawings and descriptions of the genera *Abies, Cedrus, Pseudolarix, Keteleeria, Nothotsuga, Tsuga, Cathaya, Pseudotsuga, Larix* and *Picea*. Koeltz Scientific Books, Königstein, Germany. Pp. 330.

Falush, D., Stephens, M., and Pritchard, J.K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164: 1567-1587.

Florin, R. (1963). The distributin of conifer and taxad genera in time and space. *Acta Horti Bergiani*, 20(4): 121-312.

Foll, M., and Gaggiotti, O.E. (2008). A genome scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics*, 180: 977-983.

Fumagalli, M., Viera, F.G., Linderoth, T., and Nielsen, R. (2014). ngsTools: methods for population genetic analyses from next-generation sequencing data. *Bioinformatics*, 30(10): 1486-1487.

Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdelhué, C., Pudlo, P. Cornuet, J.-M., and Estoup, A. (2013). The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology*, 22: 3165-3178.

Gernandt, D.S., and Liston, A. (1999). Internal transcribed spacer region evolution in *Larix* and *Pseudotsuga* (Pinaceae). *American Journal of Botany*, 86(5): 711-723.

Gilpin, M.E., and Soulé, M.E. (1986). Minimum viable populations: the process of population extinction. *In* Conservation biology: the science of scarcity and diversity. *Edited by* Soulé, M.E. Sinauer Associates, Sunderland, Massachusetts, USA. 584 pp.

Gower, S.T., and Richards, J.H. (1990). Larches: deciduous conifers in an evergreen world. *BioScience*, 40(1): 818-826.

Goudet, J., and Jombart, T. (2015). hierfstat: estimation and tests of hierarchical f-statistics. Available online at: https://CRAN.R-project.org/package=hierfstat

Graham, A. (1998). Late Cretaceous and Cenozoic History of North American Vegetation: North of Mexico. Oxford University Press, Oxford, UK. 370 pp.

Gray, L.K., and Hamann, A. (2013). Tracking suitable habitat for tree populations under climate change in western North America. *Climatic Change*, 117(1-2): 289 – 303.

Greer, D.H., Leinonen, I., and Repo, T. (2001). Modelling cold hardiness development and loss in conifers. *In* Conifer Cold Hardiness. *Edited by* Bigras, F.J., and Colombo, S.J. Kluwer Academic Publishers, Dordrecht, Netherlands, pp. 437-460.

Gros-Louis, M.-C., Bousquet, J., Pâques, and Isabel, N. (2005). Species-diagnostic markers in *Larix spp.* based on RAPDs and nuclear, cpDNA and mtDNA gene sequences, and their phylogenetic implications. *Tree Genetics and Genomes*, 1: 50-63.

Gugger, P.F., and Sugita, S. (2010). Glacial populations and postglacial migration of Douglas-fir based on fossil pollen and macrofossil evidence. *Quaternary Science Reviews*, 29: 2052-2070.

Gugger, P.R., Sugita, S., and Cavender-Bares, J. (2010). Phylogeography of Douglas-fir based on mitochondrial and chloroplast DNA sequences: testing hypotheses from the fossil record. *Molecular Ecology*, 19: 1877-1897.

Günther, T., and Coop, G. (2013). Robust identification of local adaptation from allele frequencies. *Genetics*, 195(1): 205-220.

Haldane, J.B.S. (1930). A mathematical theory of natural and artificial selection (Part VI, Isolation). *Mathematical Proceedings of the Cambridge Philosophical Society*, 26: 220-230.

Hamann, A., and Wang, T. (2006). Potential effects of climate change on ecosystem and tree species distribution in British Columbia. *Ecology*, 87(11): 2773-2786.

Hamrick, J.L., Linhart, Y.B., and Mitton, J.B. (1979). Relationships between life history characteristics and electrophoretically detectable genetic variation in plants. *Annual Review of Ecology and Systematics*, 10: 173-200.

Hamrick, J.L., Godt, M.J.W., and Sherman-Broyles, S.L. (1992). Factors influencing levels of genetic diversity in woody plant species. *New Forests*, 6: 95-124.

Hanlon, V.C.T. (2018). Heritable somatic mutations accumulate slowly in Sitka spruce but increase the per-generation mutation rate considerably. MSc Dissertation, University of British Columbia. 52 pp.

Holliday, J.A., Yuen, M., Ritland, K., and Aitken, S.N. (2010). Postglacial history of a widespread conifer produces inverse clines in selective neutrality tests. *Molecular Ecology*, 19: 3857-3864.

Holliday, J.A., Suren, H., and Aitken, S.N. (2012). Divergent selection and heterogeneous migration rates across the range of Sitka spruce (*Picea sitchensis*). *Proceedings of the Royal Society B*, 279(1734): 1675-1683.

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

IPCC. (2013). Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovern- mental Panel on Climate Change. *Edited by:* Stocker, T.F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P.M. Cambridge University Press, Cambridge, United Kingdom. 1535 pp.

IUCN. (2017). Guidelines for Species Conservation Planning Version 1.0. IUCN Species Survival Commission Species Conservation Planning Sub-Committee. IUCN, Gland, Switzerland. 114 pp. doi: https://doi.org/ 10.2305/IUCN.CH.2017.18.en

Jagels, R., LePage, B.A., and Jiang, M. (2001). Definitive identification of *Larix* (Pinaceae) wood based on anatomy from the middle Eocene, Axel Heiberg Island, Canadian High Arctic. *IAWA Journal*, 22(1): 73 – 83.

Jaramillo-Correa, J.P., Beaulieu, J., Khasa, D.P., and Bousquet, J. (2009). Inferring the past from the present phylogeographic structure of North American forest trees: seeing the forest for the genes. *Canadian Journal of Forest Research*, 39: 286-307.

Jeffries, D.L., Copp, G.H., Lawson Handley, L., Olsén, K.H., Sayer, C.D., and Hänfling, K.H. (2016). Comparing RADseq and microsatellites to infer complex phylogeographic patterns, an empirical perspective in the Crucian carp, *Carassius carassius*, L. *Molecular Ecology*, 25: 2997-3018.

Jermstad, K.D., Bassoni, D.L., Wheeler, N.C., Anekonda, T.S., Aitken, S.N., Adams, N.T., and Neale, D.B. (2001). Mapping of quantitative trait loci controlling adaptive traits in coastal Douglas-fir. II. Spring and fall cold-hardiness. *Theoretical and Applied Genetics*, 102: 1152-1158.

Jombart, T. (2008). adegenet: an R package for the multivariate analysis of genetic markers. *Bioinformatics*, 24(11): 1403-1405.

Jombart, T., and Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, 27(21): doi: 10.1093/bioinformatics/btr521

Kaiser, S.A., Taylor, S.A., Chen, N., Sillett, T.S., Bondra, E.R., and Webster, M.S. (2017). A comparative assessment of SNP and microsatellite markers for assigning parentage in a socially monogamous bird. *Molecular Ecology Resources*, 17: 183-193.

Kamvar, Z.N., Tabima, J.F., and Grünwald, N.J. (2014). Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, 2: e281. doi:10.7717/peerj.281

Kamvar, Z.N., Brooks, J.C., and Grünwald, N.J. (2015). Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality. *Frontiers in Genetics*, 6: 208. doi:10.3389/fgene.2015.00208

Kärkkäinen, K., and Savolainen, O. (1993). The degree of early inbreeding depression determines the selfing rate at the seed stage: model and results from *Pinus sylvestris* (Scots pine). *Heredity*, 71: 160-168.

Kawecki, T.J. (2008). Adaptation to marginal habitats. *Annual Review of Ecology, Evolution and Systematics*, 39: 321-342.

Kelly, A.E., and Goulden, M.L. (2008). Rapid shifts in plant distribution with recent climate change. *Proceedings of the National Academy of Sciences*, 105(33): 11823 – 11826.

Kelley, D., and Richards, C. (2018). oce: analysis of oceanographic data. Available online at: https://CRAN.R-project.org/package=oce

Khasa, D.P., Jaramillo-Correa, J.P., Jaquish, B. and Bousquet, J. (2006). Contrasting microsatellite variation between subalpine and western larch, two closely related species with different distribution patterns. *Molecular Ecology*, 15(13): 3907 – 3918.

Kleiber, C., and Zeileis, A. (2008). Applied econometrics with R. Springer-Verlag, New York, USA. 222 pp.

Korneliussen, T.S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: analysis of next-generation sequencing data. *BMC Bioinformatics*, 15(356): 1-13.

Kuzmin, D.A., Feranchuck, S.I., Sharov, V.V., Cybin, A.N., Makolov, S.V., Putintseva, Y.A., Oreshkova, N.V., and Krutovsky, K.V. (2019). Stepwise large genome assembly approach: a case of Siberian larch (Larix sibirica Ledeb). *BMC Bioinformatics*, 20(Suppl. 1): 35-46.

Larionova, A.Y., Yakhneva, N.V., and Abaimov, A.P. (2004). Genetic diversity and differentiation of Gmelin larch *Larix gmelinii* populations from Evenkia (central Siberia). *Russian Journal of Genetics*, 40(10): 1127-1133.

Ledig, F.T., Jacob-Cervantes, V., Hodgskiss, P.D., and Eguiluz-Piedra, T. (1997). Recent evolution and divergence among populations of a rare Mexican endemic,

Chihuahua spruce, following Holocene climatic warming. *Evolution*, 51(6): 1815-1827.

Ledig, F.T., Capó-Arteaga, M.A., Hodgskiss, P.D., Sbay, H., Flores-López, C., Conkle, M.T., and Bermejo-Velázquez, B. (2001). Genetic diversity and the mating system of a rare Mexican piñon, *Pinus pinceana*, and a comparison with *Pinus maximartinezii* (Pinaceae). *American Journal of Botany*, 88(11): 1977-1987.

Ledig, F.T., Hodgskiss, P.D., and Johnson, D.R. (2005). Genetic diversity, genetic structure and mating system of Brewer spruce (Pinaceae), a relict of the Arcto-Tertiary forest. *American Journal of Botany*, 92(12): 1975-1986.

Ledig, F.T., Hodgskiss, P.D., and Johnson, D.R. (2006). The structure of genetic diversity in Engelmann spruce and a comparison with blue spruce. *Canadian Journal of Botany*, 84(12): 1806-1828.

LeFort, V., Desper, R., and Gascuel, O. (2015). FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Molecular Biology and Evolution*, 32(10): 2798-2800.

LePage, B.A., and Basinger, J.F. (1991). A new species of *Larix* (Pinaceae) from the early Tertiary of Axel Heiberg Island, Arctic Canada. *Review of Palaeobotany and Palynology*, 70: 89-111.

Le Roux, J.J., Strasberg, D., Rouget, M., Morden, C.W., Koordom, M., and Richardson, D.M. (2014). Relatedness defies biogeography: the tale of two island endemics (*Acacia heterophylla* and *A. koa*). *New Phytologist*, 204(1): 230-242.

L'Hirondelle, S.J., Simpson, D.G., and Binder, W.D. (2006). Overwinter storability of conifer planting stock: operational testing of fall frost hardiness. *New Forests*, 32: 307-321.

Li, H. (2011). Improving SNP discovery by base alignment quality. *Bioinformatics*, 27(8): 1157-1158.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup.

(2009). The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, 25(16): 2078-2079.

Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2(3): 18-22.

Liepe, K.J., Hamann, A., Smets, P., Fitzpatrick, C.R., and Aitken, S.N. (2016). Adaptation of lodgepole pine and interior spruce to climate: Implications for reforestation in a warming world. *Evolutionary Applications*, 9(2): 409-419.

Lin, C.-P., Huang, J.-P., Wu, C.-S., Hsu, C.-Y., and Chaw, S.-M. (2010). Comparative chloroplast genomics reveals the evolution of Pinaceae genera and subfamilies. *Genome Biology and Evolution*, 2: 504-517.

Lischer, H.E.L., and Excoffier, L. (2012). PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, 28(2): 298-299.

Lotterhos, K.E., and Whitlock, M.C. (2014). Evaluation of demographic history and neutral parameterization on the performance of F_{ST} outlier tests. *Molecular Ecology*, 23: 2178-2192.

Lotterhos, K.E., Yeaman, S., Degner, J.C., Aitken, S.A., and Hodgins, K.A. (2018). Modularity of genes involved in local adaptation to climate change despite physical linkage. *Genome Biology*, 19: 157 (24 pp).

Lynch, M., and Lande, R. (1993). Evolution and extinction in response to environmental change. *In* Biotic Interactions and Global Change. *Edited by* Kareiva, P., Kingsolver, J., and Huey R. Sinauer Assoc., Inc., Sunderland Mass. Pp. 234 -250.

Mack, R.N., Rutter, N.W., Bryant, V.M., and Valastro, S. (1978). Reexamination of postglacial vegetation history in northern Idaho: Hager Pond, Bonner, Co. *Quaternary Research*, 10(2): 241-255.

Mack, R.N., Rutter, N.W., and Valastro, S. (1983). Holocene vegetation history of the Kootenai River Valley, Montana. *Quaternary Research*, 20(2): 177-193.

MacLachlan, I.R., Wang, T., Hamann, A., Smets, P, and Aitken, S.N. (2017). Selective breeding of lodgepole pine increases growth and maintains climatic adaptation. *Forest Ecology and Management*, 391: 404-416.

Mahony, C.R., MacLachlan, I.R., Lind, B.M., Yoder, J.B., Wang, T., and Aitken, S.N. (2019). Evaluating genomic data for management of local adaptation in a changing climate: a lodgepole pine case study. BioRxiv, 568725.

Mamet, S.D., Brown, C.D., Trant, A.J., and Laroque, C.P. (2019). Shifting global *Larix* distributions: northern expansion and southern retraction as species respond to changing climate. *Journal of Biogeography*, 46: 30-44.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBNet.journal, 17*(1): 10-12. doi: https://doi.org/10.14806/ej.17.1.200

McCormack, J.E., Hird, S.M., Zellmer, A.J., Carstens, B.C., and Brumfield, R.T. (2013). Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, 66: 526-538.

McKee, B. (1972). Cascadia: The Geologic Evolution of the Pacific Northwest. McGraw-Hill, Inc., USA. 352 pp.

McLachlan, J.S., and Clark, J.S. (2004). Reconstructing historical ranges with fossil data at continental scales. *Forest Ecology and Management*, 197: 139-147.

Mimura, M., and Aitken, S.N. (2007). Adaptive gradients and isolation-by-distance with postglacial migration in Picea sitchensis. *Heredity*, 99: 224-232.

Miller, C.N. Jr., and Ping, L. (1994). Structurally preserved larch and spruce cones from the Pliocene of Alaska. *Quaternary International*, 22/23: 207-214.

Miller, M.R., Dunham, J.P., Amores, A., Cresko, W.A., and Johnson, E.A. (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, 17: 240-248.

Molina-Freaner, F., Delgado, P., Piñero, D. Perez-Nassar, N., and Alvarez-Buylla, E. (2001). Do rare pines need different conservation strategies? Evidence from three Mexican species. *Canadian Journal of Botany*, 79: 131-138.

Mustaphi, C.J.C., and Pisaric, M.F.J. (2014). Holocene climate-fire-vegetation interactions at a subalpine watershed in southeastern British Columbia, Canada. *Quaternary Research*, 81: 228-239.

Nadeau, S., Godbout, J., Lamothe, M., Gros-Louis, M.-C., Isabel, N., and Ritland, K. (2015). Contrasting patterns of genetic diversity across the ranges of *Pinus monticola* and *Pinus strobus*: a comparison between eastern and western North American postglacial colonization histories. *American Journal of Botany*, 102(8): 1342-1355.

NCBI Resource Coordinators. (2012). Database resources of the National Centre for Biotechnology Information. *Nucleic Acids Research*, 41: D8-D20.

Nielsen, R., Paul, J.S., Albrechtsen, A., and Song, Y.S. (2011). Genotyping and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12: 443-451.

O'Connell, L.M., Ritland, K., and Thompson, S.L. (2008). Patterns of post-glacial colonization by western redcedar (*Thuja plicata*, Cupressaceae). *Botany*, 86: 194-203.

Okada, S., Kurahasi, A., and Sakai, A. (1971). Differences in freezing resistance in winter of the Japanese larch seedlings from natural forests in 20 different localities. *Journal of Japanese Forestry Science*, 52: 377-379.

Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Szoecs, E., and Wagner, H. (2018). vegan: community ecology package. Available online at: https://CRAN.R-project.org/package=vegan

O'Neill, G.A., Aitken, S.N., and Adams, W.T. (2000). Genetic selection for cold hardiness in coastal Douglas-fir seedlings and saplings. *Canadian Journal of Forest Research*, 30: 1799-1807.

O'Neill, G.A., Adams, W.T., and Aitken, S.N. (2001). Quantitative genetics of spring and fall cold hardiness in seedlings from two Oregon populations of coastal Douglas-fir. *Forest Ecology and Management*, 149: 305-318.

Ostenfeld, C.H., and Larsen, C.S. (1930). The species of the genus *Larix* and their geographical distribution. *Kongelige Danske Videnskabernes-Selskabs Biologiske Meddelelser*, 9(2): 1-107.

Paradis, E., and Schliep, K. (2018). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35(3): 526-528.

Parchman, T.L., Jahner, J.P., Uckele, K.A., Galland, L.M., and Eckert, A.J. (2018). RADseq approaches and applications for forest tree genetics. *Tree Genetics & Genomes*, 14(39): 1-25.

Pauli, H., Gottfried, M., Dullinger, S., Abdaladze, O., Akhalkatsi, M., Alonso, J.L.B., Coldea, G., Dick, J., Erschbamer, B., Calzado, R.F., Ghosn, D., Holten, J.I., Kanka, R., Kazakis, G., Kollár, J., Larsson, P., Moiseev, P., Moiseev, D., Molau, U., Mesa, J.M., Nagy, L., Pelino, G., Puşcaş, M., Rossi, G., Stanisci, A., Syverhuset, A.O., Theurillat, J.-P., Tomaselli, M., Unterluggauer, P., Villar, L., Vittoz, P., and Grabherr, G. (2012). Recent plant diversity changes on Europe's mountain summits. *Science*, 336: 353 – 355.

Peischl, S., Dupanloup, I., Kirkpatrick, M., and Excoffier, L. (2013). On the accumulation of deleterious mutations during range expansions. *Molecular Ecology*, 22(24): 5972 – 5982.

Piotti, A., Leonardi, S., Piovani, P., Scalfi, M., and Menozzi, P. (2009). Spruce colonization at treeline: where do those seeds come from? *Heredity*, 103: 136-145.

Pluess, A.R. (2011). Pursuing glacier retreat: genetic structure of a rapidly expanding *Larix decidua* population. *Molecular Ecology*, 20: 473-485.

Poelstra, J.W., Vijay, N., Bossu, C.M., Lantz, H., Ryll, B., Müller, I., Baglione, V., Unneberg, P., Wikelski, M., Grabherr, M.G., and Wolf, J.B.W. (2014). The

genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science*, 344: 1410-1414.

Price, R.A., Olsen-Stojkovich, J., and Lowenstein, J.M. (1987). Relationship among the genera of Pinaceae: an immunological comparison. *Systematic Botany*, 12(1): 91-97.

Prince, D.J., O'Rourke, S.M., Thompson, T.Q., Ali, O.A., Lyman, H.S., Saglam, I.K., Hotaling, T.J., Spidle, A.P., and Miller, M.R. (2017). The evolutionary basis of premature migration in Pacific salmon highlights the utility of genomics for informing conservation. *Science Advances*, 3: e160319 (11 pp.).

Pritchard, J.K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155: 945-959.

R Core Team. (2019). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available online at: https://www.R-project.org/

Ran, J.-H., Shen, T.-T., Wu, H., Gong, X., and Wang, X.-Q. (2018). Phylogeny and evolutionary history of Pinaceae updated by transcriptomic analysis. *Molecular Phylogenetics and Evolution*, 129: 106-116.

Rehfeldt, G.E. (1995). Genetic variation, climate models and the ecological genetics of *Larix occidentalis*. *Forest Ecology and Management*, 78: 21-37.

Richards, J.H., and Bliss, L.C. (1986). Winter water relations of a deciduous timberline conifer, *Larix lyallii* Parl. *Oecologia*, 69(1): 16-24.

Richardson, A.D., and Friedland, A.J. (2009). A review of the theories to explain arctic and alpine treelines around the world. *Journal of Sustainable Forestry*, 28: 218-242.

Richardson, B.A., Brunsfeld, S.J., and Klopfenstein, N.B. (2002). DNA from bird-dispersed seed and wind-disseminated pollen provides insights into postglacial colonization and population genetic structure of whitebark pine (*Pinus albicaulis*). *Molecular Ecology*, 11: 215-227.

Roberts, D.R., and Hamann, A. (2015). Glacial refugia and modern genetic diversity of 22 western North American tree species. *Proceedings of the Royal Society B,* 282: 20142903 (9 pp.).

Sakai, A., and Weiser, C.J. (1973). Freezing resistance of trees in North America with reference to tree regions. *Ecology*, 54(1): 118-126.

Scheumann, W., and Schonbach, H. (1968). Die Prilfung der Frost Resistenz von 25 *Larix leptolepis* Herkunften eines internationalen Provenienzversuches mit Hilfe von Labor- Prufverfahren. *Archiv fur Forstwesen*, 17: 597-611.

Schorn, H.E. (1994). A preliminary discussion of fossil larches (*Larix*, Pinaceae) from the Arctic. *Quaternary International*, 22/23: 173-183.

Sedlazeck, F.J., Rescheneder, P., and von Haeseler, A. (2013). NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics*, 29(21): 2790-2791.

Semerikov, V.L, and Lascoux, M. (2003). Nuclear and cytoplasmic variation within and between Eurasian *Larix* (Pinaceae) species. *American Journal of Botany*, 90(8): 1113-1123.

Semerikov, V.L., Zhang, H., Sun, M., and Lascoux, M. (2003). Conflicting phylogenies of *Larix* (Pinaceae) based on cytoplasmic and nuclear DNA. *Molecular Phylogenetics and Evolution*, 27: 173-184.

Semerikov, V.L., Semerikova, S.A., and Polezhaeva, M.A. (2013). Nucleotide diversity and linkage disequilibrium of adaptive significant genes in *Larix* (Pinaceae). *Russian Journal of Genetics*, 49(9): 915-923.

Simak, M. (1966). Karyotype analysis of *Larix griffithiana* Carr. *Hereditas*, 56(1): 137 – 141.

Soularue, J.-P., and Kremer, A. (2014). Evolutionary responses of tree phenology to the combined effects of assortative mating, gene flow and divergent selection. *Heredity*, 113: 485-494.

Timmis, R., Flewelling, J., and Talbert, C. (1994). Frost injury prediction model for Douglas-fir seedlings in the Pacific Northwest. *Tree Physiology*, 14(7/8/9): 855-869.

Wang, X.-Q., and Ran, J.-H. (2014). Evolution and biogeography of gymnosperms. *Molecular Phylogenetics and Evolution*, 75: 24-40.

Wang, T., Hamann, A., Spittlehouse, D.L., and Murdock, T.Q. (2016). ClimateWNA—high-resolution spatial climate data for western North America. *Journal of Applied Meteorology and Climatology*, 51: 16-29.

Warnes, G.R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., and Venables, B. (2019). gplots: various R programming tools for plotting data. Available online at: https://CRAN.R-project.org/package=gplots

Warren, E., de LaFontaine, G., Gérardi, S., Senneville, S., Beaulieu, J., Perron, M., Jaramillo-Correa, J.-P., and Bousquet, J. (2016). Joint inferences from cytoplasmic DNA and fossil data provide evidence for glacial vicariance and contrasted post-glacial dynamics in tamarack, a transcontinental conifer. *Journal of Biogeography*, 43: 1227-1241.

Wei, X.-X., and Wang, X.-Q. (2004a). Evolution of 4-coumarate:coenzyme A ligase (4Cl) gene and divergence of *Larix* (Pinaceae). *Molecular Phylogenetics and Evolution*, 31(2): 542-553.

Wei, X.-X., and Wang, X.-Q. (2004b). Recolonization and radiation in *Larix* (Pinaceae): evidence from nuclear ribosomal DNA paralogues. *Molecular Ecology*, 13(10): 3115-3123.

Wei, X.-X., Yang, Z.-Y., Li, Y., and Wang, X.-Q. (2010). Molecular phylogeny and biogeography of Pseudotsuga (Pinaceae): insights into the floristic relationship between Taiwan and adjacent areas. *Molecular Phylogenetics and Evolution*, 55: 776-785.

Wei, X.-X., Beaulieu, J., Khasa, D.P., Vargas-Hernández, J., López-Upton, J., Jaquish, B., and Bousquet, J. (2011). Range-wide chloroplast and mitochondrial DNA imprints reveal multiple lineages and complex biogeographic history for Douglas-fir. *Tree Genetics & Genomes*, 7: 1025-1040.

Wegrzyn, J.L., Liechty, J.D., Stevens, K.A., Wu, L.-S., Loopstra, C.A., Vasquez-Gross, H.A., Dougherty, W.M., Lin, B.Y., Zieve, J.J., Martínez-García, P.J., Holt, C., Yandell, M., Zimin, A.V., Yorke, J.A., Crepeau, M.W., Puiu, D., Salzberg, S.L., de Jong, P.J., Mockaitis, K., Main, D., Langley, C.H., and Neale, D.B. (2014). Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. *Genetics*, 196: 891-909.

Wen, J., Ree, R.H., Ickert-Bond, S.M., Nie, Z., and Funk, V. (2013). Biogeography: where do we go from here? *Taxon*, 62(5): 912-927.

Westerling, A.L., Hidalgo, H.G., Cayan, D.R., and Swetnam, T.W. (2006). Warming and earlier spring increase western U.S. forest wildfire activity. *Science*, 313: 940 – 943.

White, T.L., Adams, W.T., and Neale, D.B. (2007). Forest Genetics. CABI, Wallingford, UK. 682 pp.

White, T.A., Perkins, S.E., Heckel, G., and Searle, J.B. (2013). Adaptive evolution during ongoing range expansion: the invasive bank mole (*Myodes glareolus*) in Ireland. *Molecular Ecology*, 22: 2971-2985.

Wickham, H. (2017). lazyeval: Lazy (non-standard) evaluation. Available online at: https://CRAN.R-project.org/package=lazyeval

Williams, C.G. (2010). Long-distance pine pollen still germinates after meso-scale dispersal. *American Journal of Botany*, 97(5): 846-855.

Williamson, T.B., Colombo, S.J., Duinker, P.N., Gray, P.A., Hennessey, R.J., Houle, D., Johnston, M.H., Ogden, A.E., and Spittlehouse, D.L. (2009). Climate Change and Canada's Forests: From Impacts to Adaptation. Sustainable Forest Management Network and Natural Resources Canada, Canadian Forest Service, Northern Forestry Centre, Edmonton, AB, Canada. 104 pp.

Worrall, J. (1993). Temperature effects on bud-burst and leaf-fall in subalpine larch. *Journal of Sustainable Forestry*, 1(2): 1-19.

Yanchuk, A.D. (2001). A quantitative framework for breeding and the conservation of forest tree genetic resources in British Columbia. *Canadian Journal of Forest Research*, 31: 566-576.

Yeaman, S., Hodgins, K.A., Lotterhos, K.E., Suren, H., Nadeau, S., Degner, J.C., Nurkowski, K.A., Smets, P., Wang, T., Gray, L.K., Liepe, K.J., Hamann, A., Holliday, J.A., Whitlock, M.C., Riesberg, L.H., and Aitken, S.N. (2016). Convergent local adaptation to climate in distantly related conifers. *Science*, 353(6306): 1431-4133.

Zhao, Y., Chen, F., Zhai, R., Lin, X., Wang, Z., Su, L., and Christiani, D.C. (2012). Correction for population stratification in random forest analysis. *International Journal of Epidemiology*, 41: 1798-1806.

Appendix A: A-score Optimisation

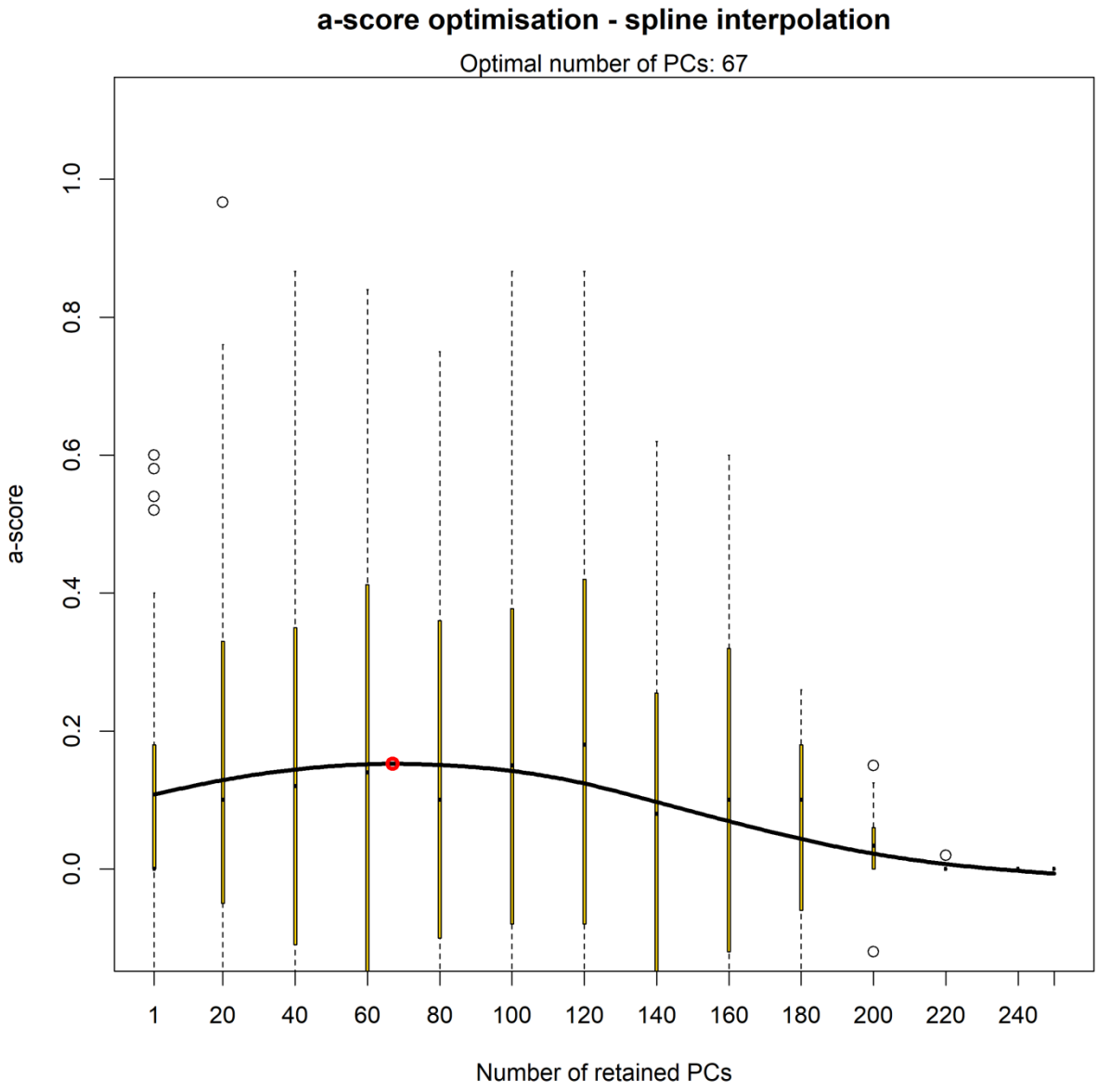

Figure 1. One hundred simulations were run using the *optim.a.score* function and 67 was identified as the optimal number of principal components to include in a discriminant analysis of principal components (DAPC) in order to avoid overfitting.

Appendix B: ANGSD Parameter Settings

Table 1. Parameter values for subalpine larch population genomics analyses run using the software ANGSD.

	PCA Prep.[1]	PCA	Genetic Distance	Inbreeding Prep.[2]	Inbreeding (F_{IS})	Population SFS	Tajima Prep.[3]	Tajima's D	Heterozygosity
-bam	YES	YES	YES	YES	YES	YES	YES	YES	YES
-ref	*L. sibirica*	*L. sibirica*	*L. sibirica*	*L. sibirica*	*L. sibirica*	*L. sibirica*	*L. sibirica*	*L. sibirica*	*L. sibirica*
-anc	-	-	-	-	-	*L. sibirica*	*L. sibirica*	*L. sibirica*	*L. sibirica*
-GL	2	2	2	2	2	2	2	2	2
-baq	1	1	1	1	1	1	1	1	1
-C	50	50	50	50	50	50	50	50	50
-minMapQ	20	20	20	20	20	20	20	20	20
-minQ	20	20	20	20	20	20	20	20	20
-doCounts	1	1	1	1	1	1	1	1	1
-doMaf	1	1	1	1	1	1	1	1	1
-doMajorMinor	1	1	1	1	1	1	1	1	1
-sites	-	PCA Prep.	PCA Prep.	-	F_{IS} Prep.	-	F_{IS} Prep.	F_{IS} Prep.	PCA Prep.
-SNP_pval	0.05	0.05	0.05	0.05	0.05	-	0.05	-	-
-minMaf	0.05	0.05	0.05	-	-	-	0.05	0.05	-
-minInd	183	183	183	N/2	N/2	N/2	N/2	N/2	-
-skipTriallelic	1	1	1	1	1	1	1	1	1
-indF	F_{IS}	F_{IS}	F_{IS}	-	-	-	F_{IS}	F_{IS}	-
-doGeno	32	32	8	-	-	-	-	-	-
-doGlf	-	-	-	3	3	-	-	-	-
-doPost	1	1	1	-	-	-	-	-	-
-doSaf	-	-	-	-	-	2	2	2	1
-fold	-	-	-	-	-	-	1	1	1
-doThetas	-	-	-	-	-	-	-	1	-
-pest	-	-	-	-	-	-	-	From Prep.	-

Preparatory analyses were run in order to identify [1]range-wide variants separated by 125 nucleotides, [2]population-level variants separated by at least 125 nucleotides and [3]to generate population-level SFS needed as input for further analyses

Appendix C: Population Site Frequency Spectra

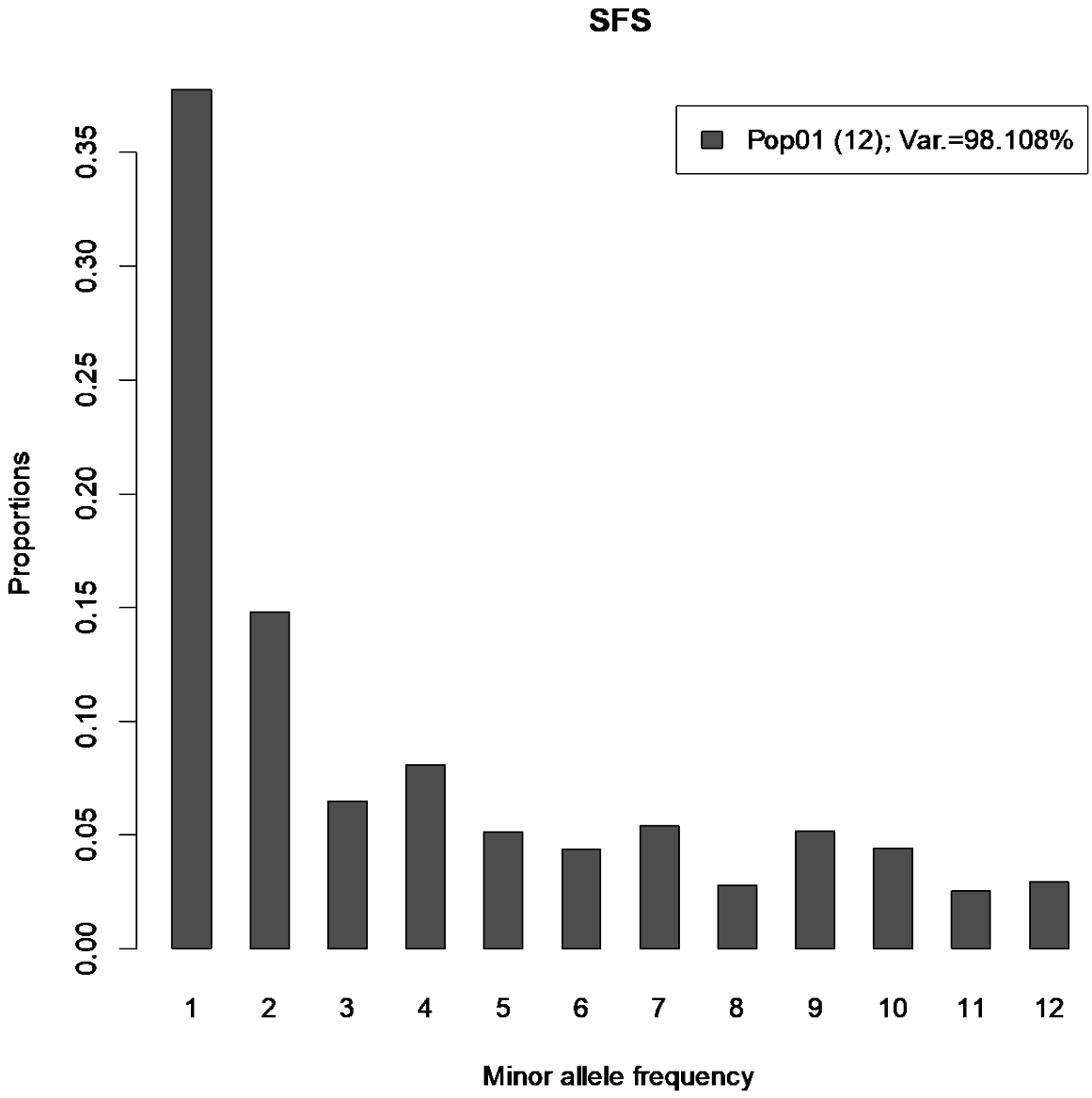

Figure 1. Site frequency spectrum for Pop01, Frosty Mountain, BC, generated by filtering for polymorphic sites separated by at least 125 nucleotides.

Appendix C Population Site Frequency Spectra

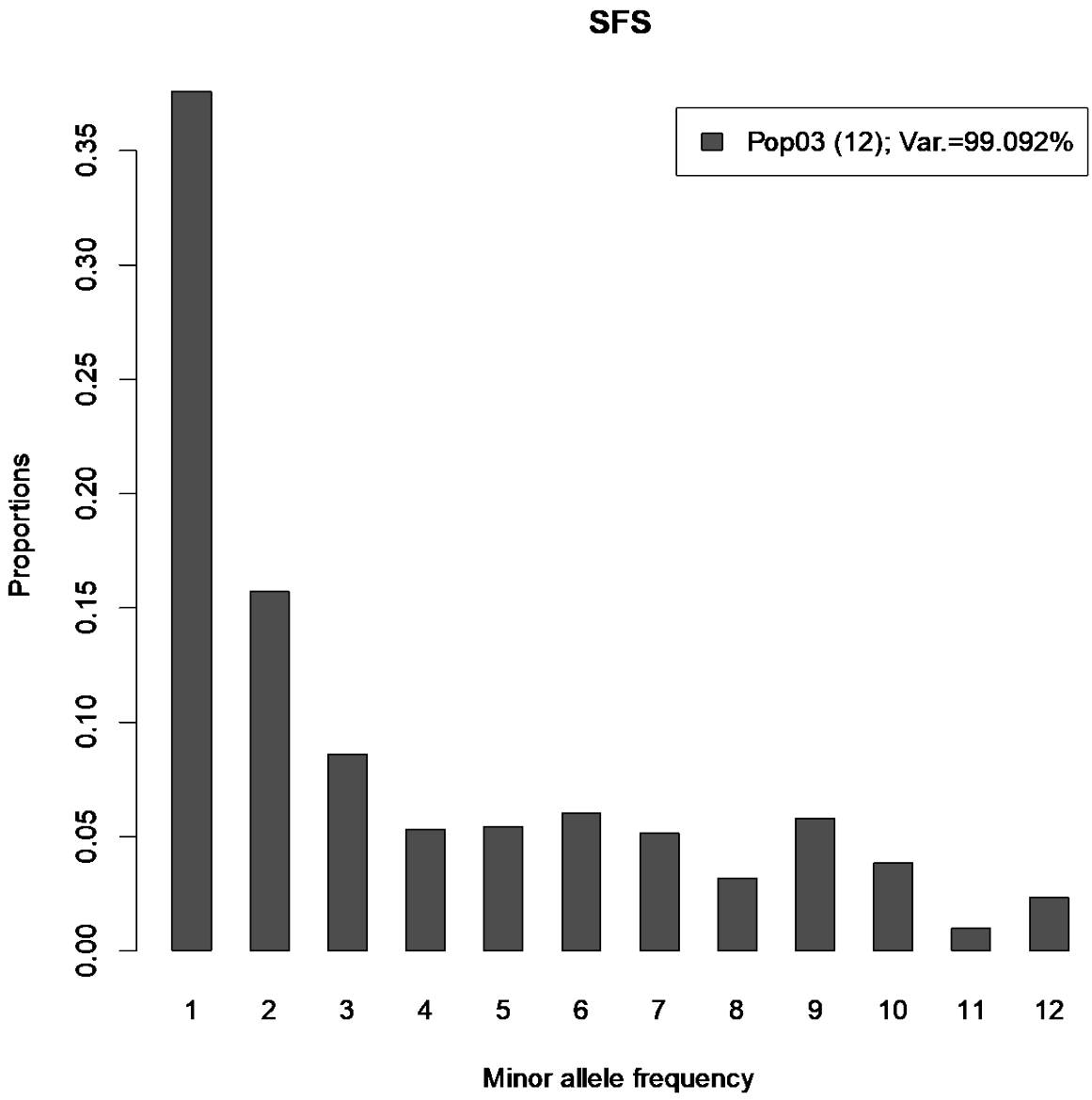

Figure 2. Site frequency spectrum for Pop03, Tiffany Mountain, WA, generated by filtering for polymorphic sites separated by at least 125 nucleotides.

Appendix C Population Site Frequency Spectra

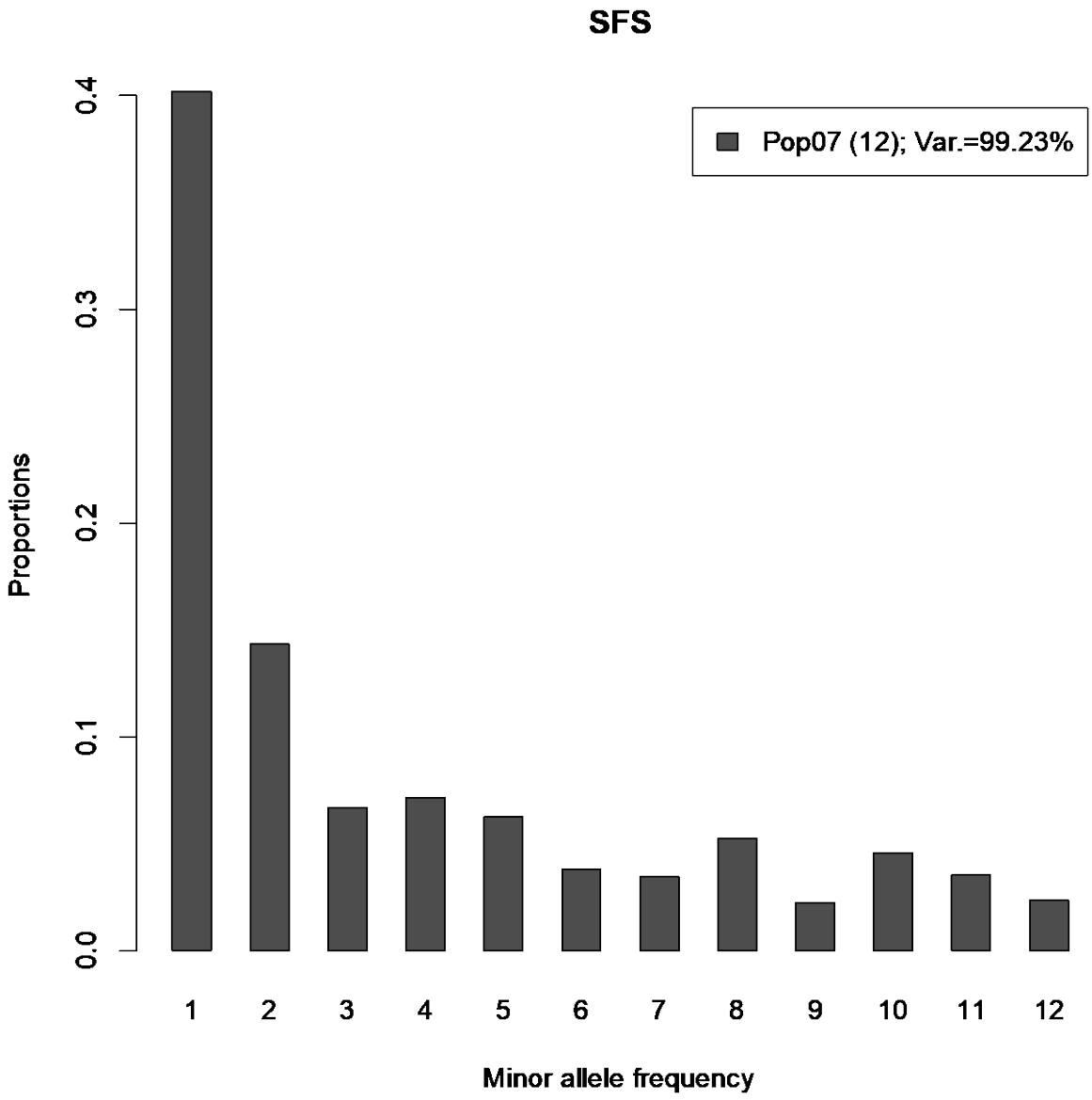

Figure 3. Site frequency spectrum for Pop07, Big Hill, WA, generated by filtering for polymorphic sites separated by at least 125 nucleotides.

Appendix C Population Site Frequency Spectra

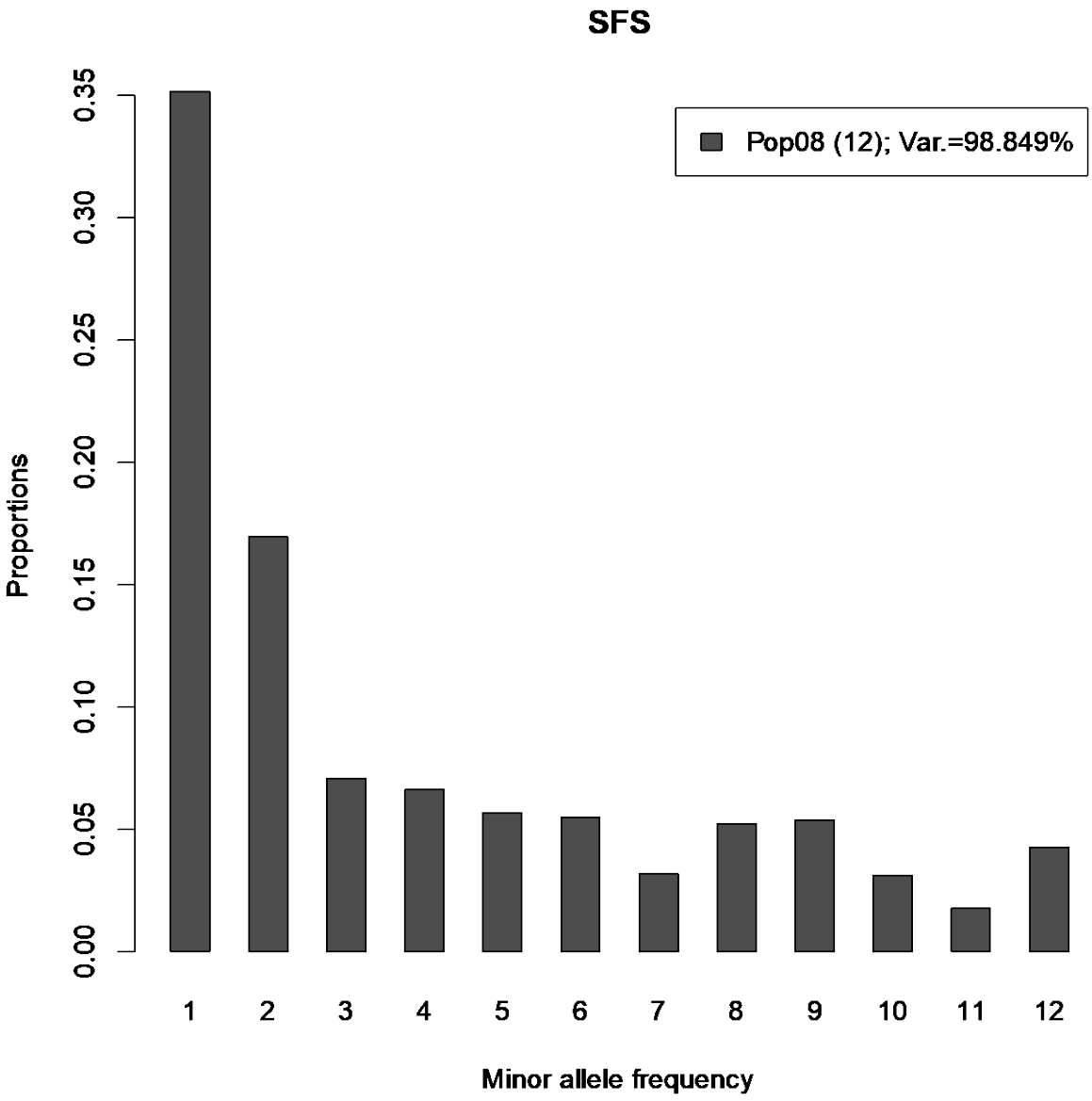

Figure 4. Site frequency spectrum for Pop08, Windy Pass, WA, generated by filtering for polymorphic sites separated by at least 125 nucleotides.

Appendix C Population Site Frequency Spectra

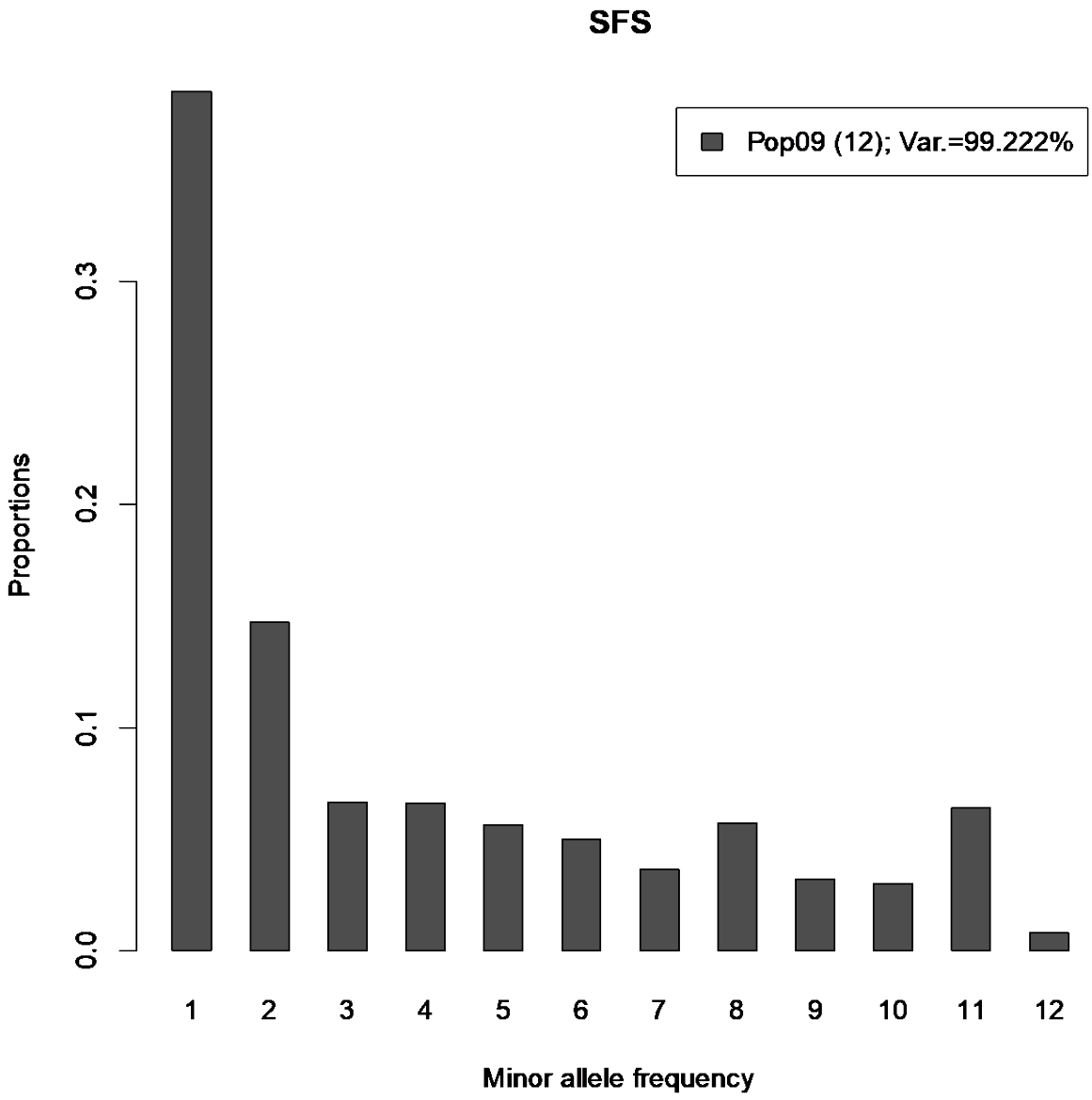

Figure 5. Site frequency spectrum for Pop09, Carlton Ridge, MT, generated by filtering for polymorphic sites separated by at least 125 nucleotides.

Appendix C Population Site Frequency Spectra

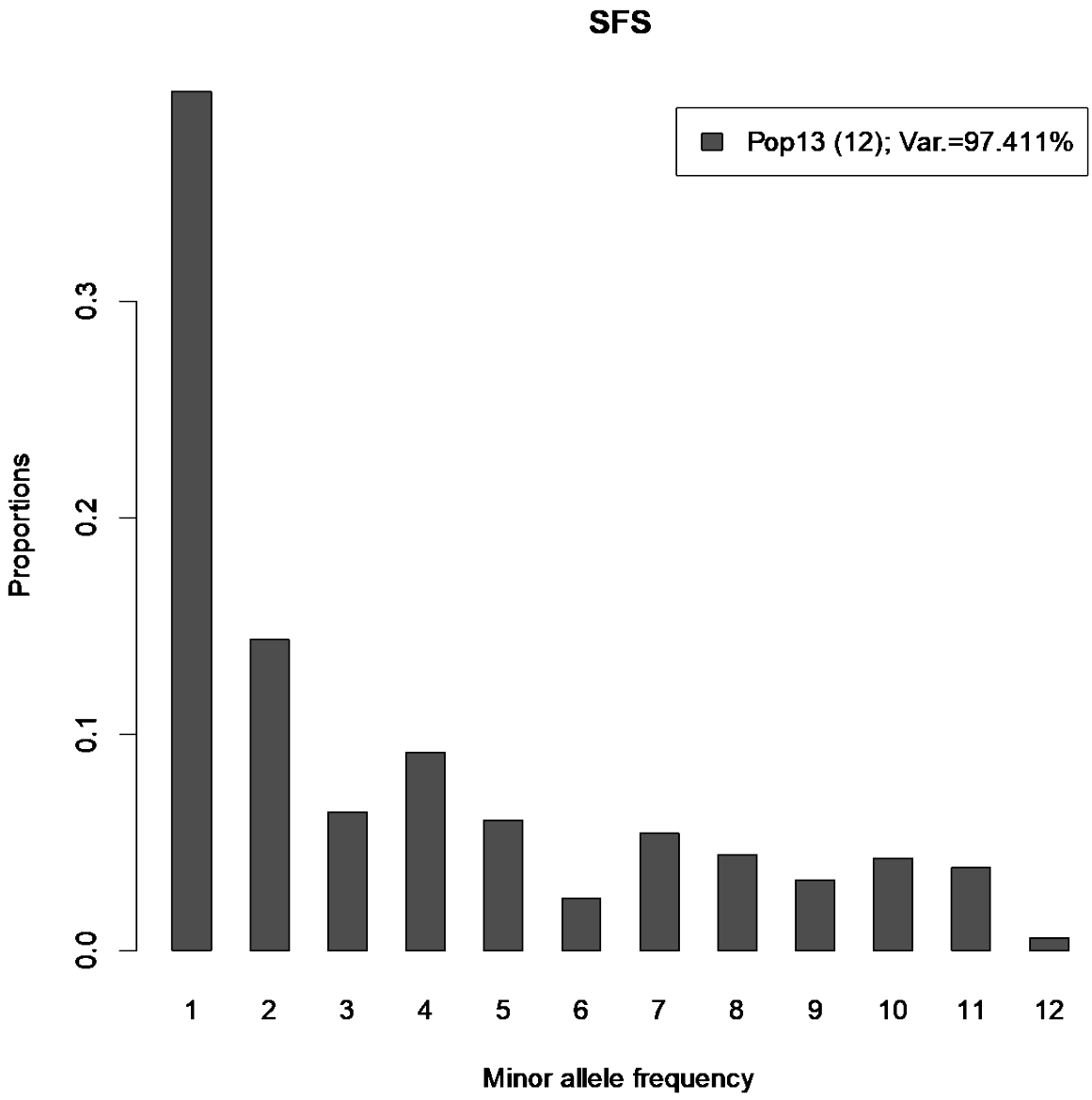

Figure 6. Site frequency spectrum for Pop13, Trapper Peak, MT, generated by filtering for polymorphic sites separated by at least 125 nucleotides.

Appendix C Population Site Frequency Spectra

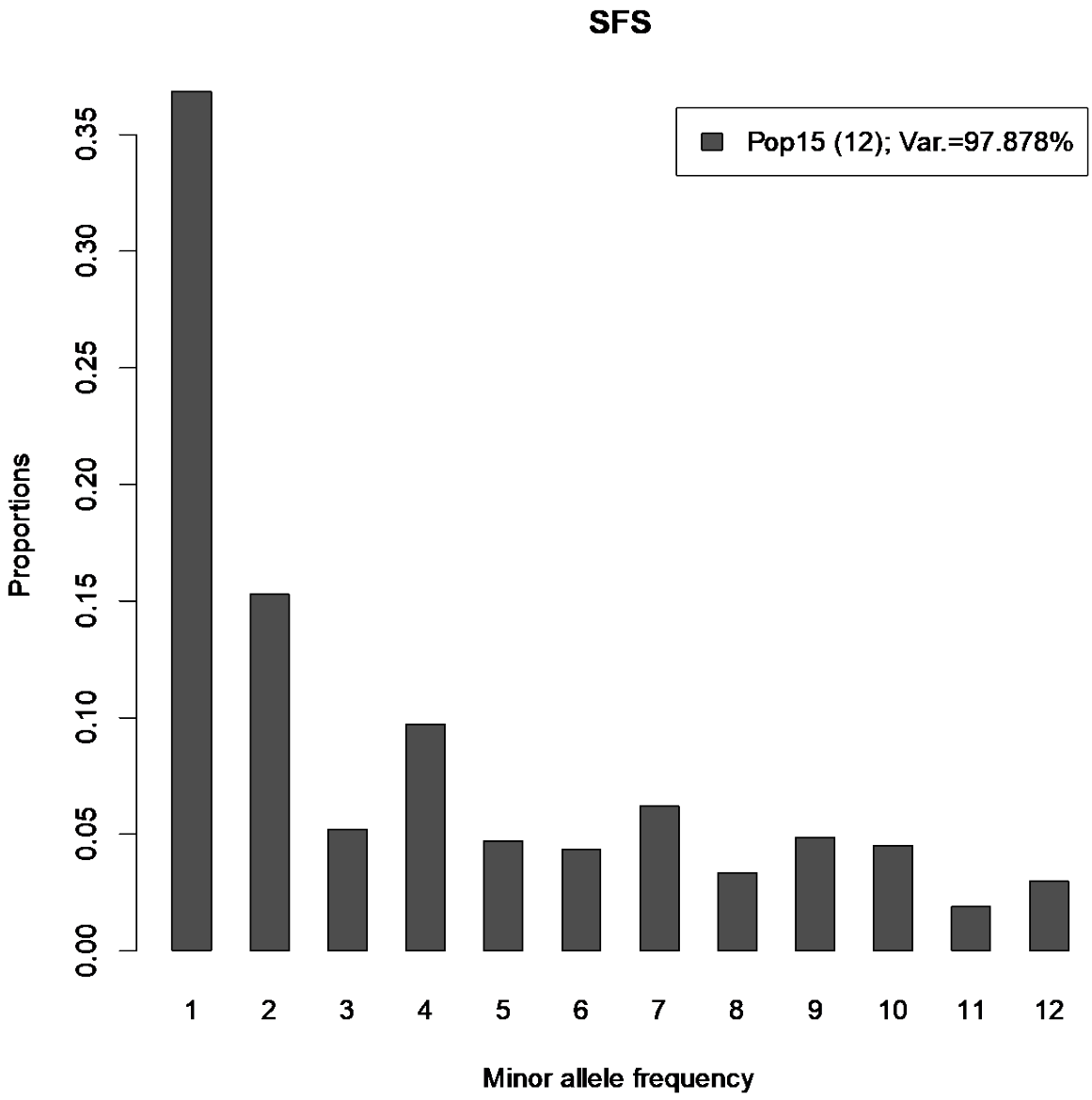

Figure 7. Site frequency spectrum for Pop15, Storm Lake, MT, generated by filtering for polymorphic sites separated by at least 125 nucleotides.

Appendix C Population Site Frequency Spectra

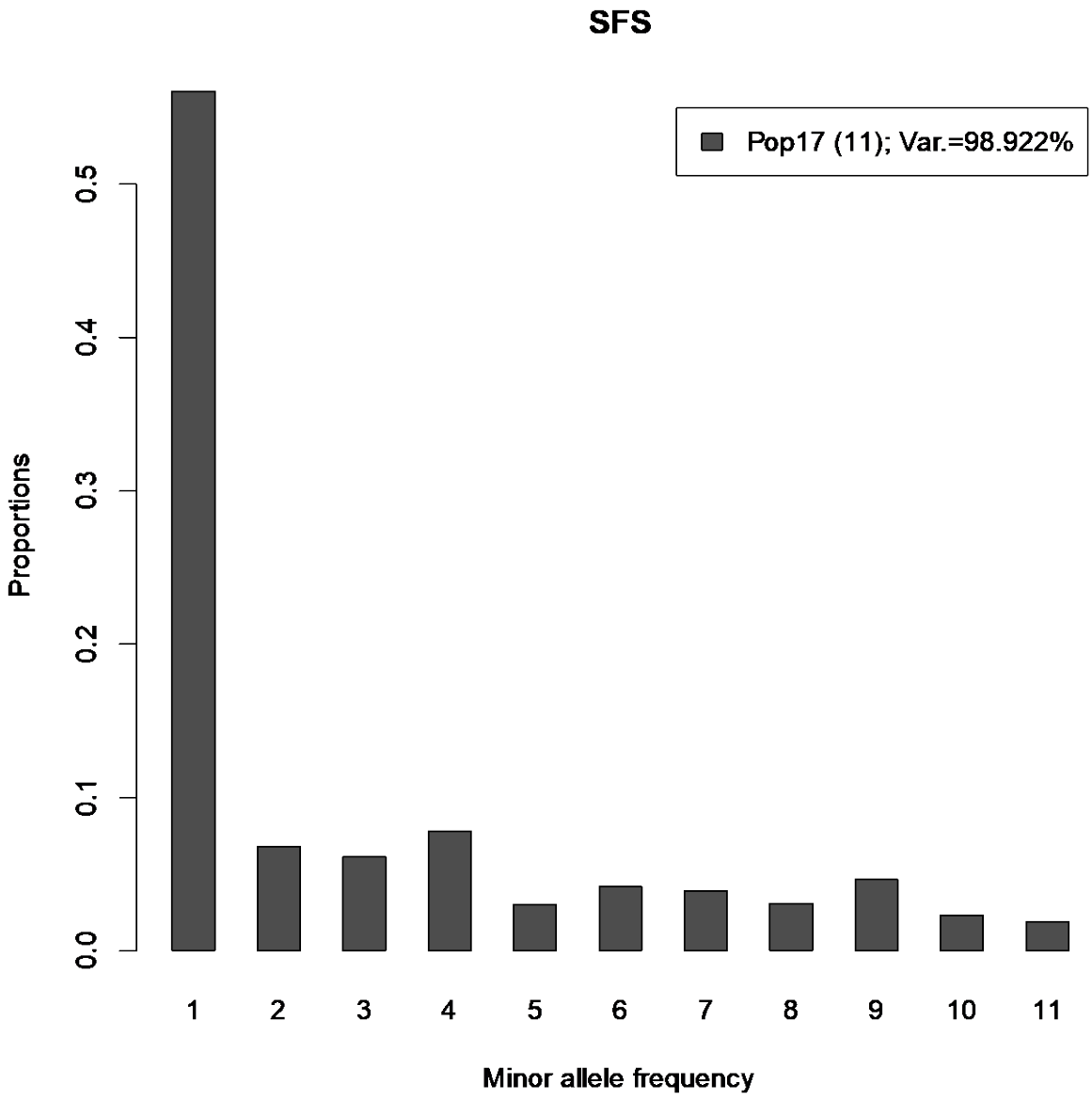

Figure 8. Site frequency spectrum for Pop17, Holland Pass, MT, generated by filtering for polymorphic sites separated by at least 125 nucleotides.

Appendix C Population Site Frequency Spectra

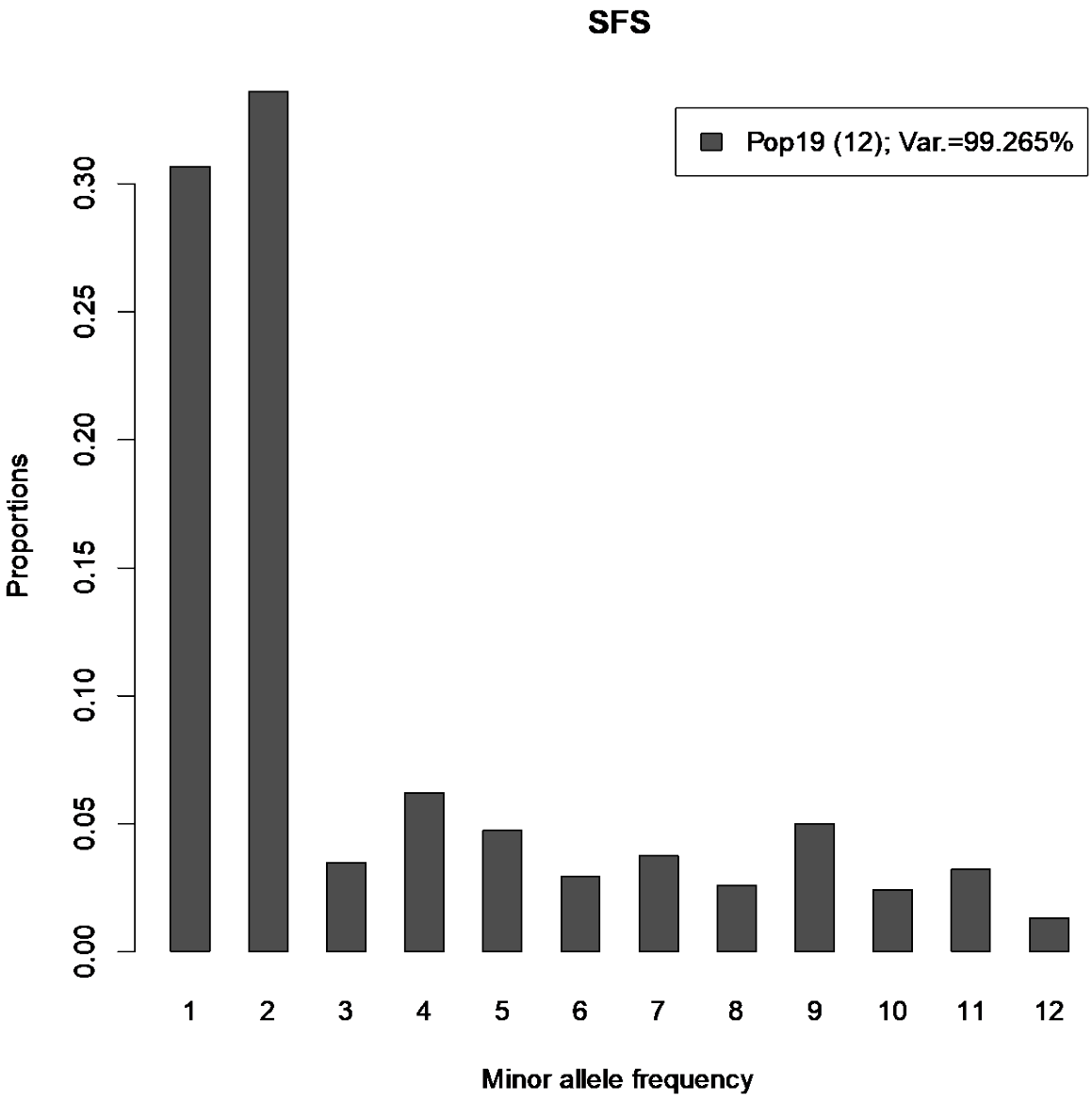

Figure 9. Site frequency spectrum for Pop19, Roman Nose, ID, generated by filtering for polymorphic sites separated by at least 125 nucleotides.

Appendix C Population Site Frequency Spectra

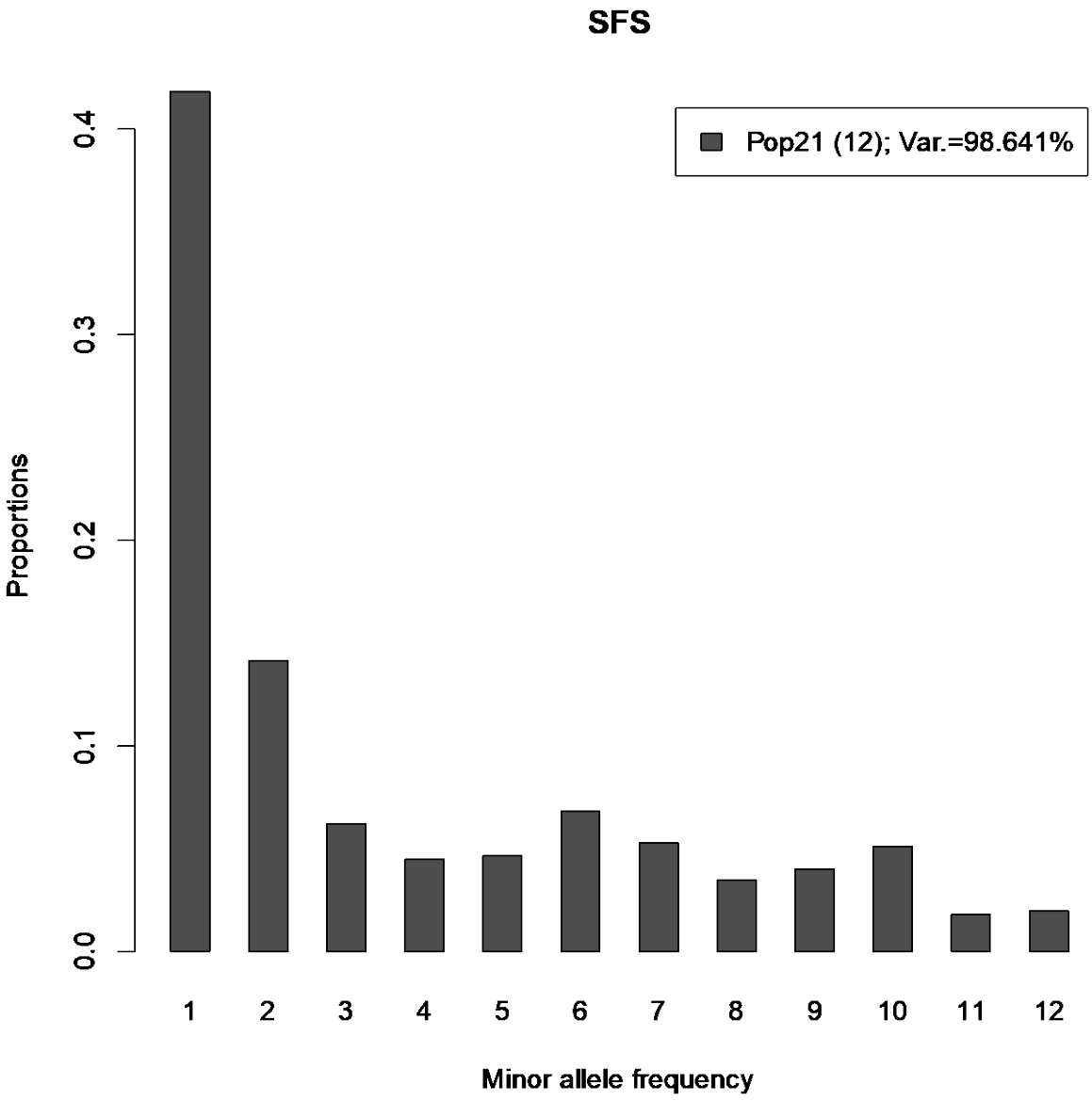

Figure 10. Site frequency spectrum for Pop21, Sparkle Lake, BC, generated by filtering for polymorphic sites separated by at least 125 nucleotides.

Appendix C Population Site Frequency Spectra

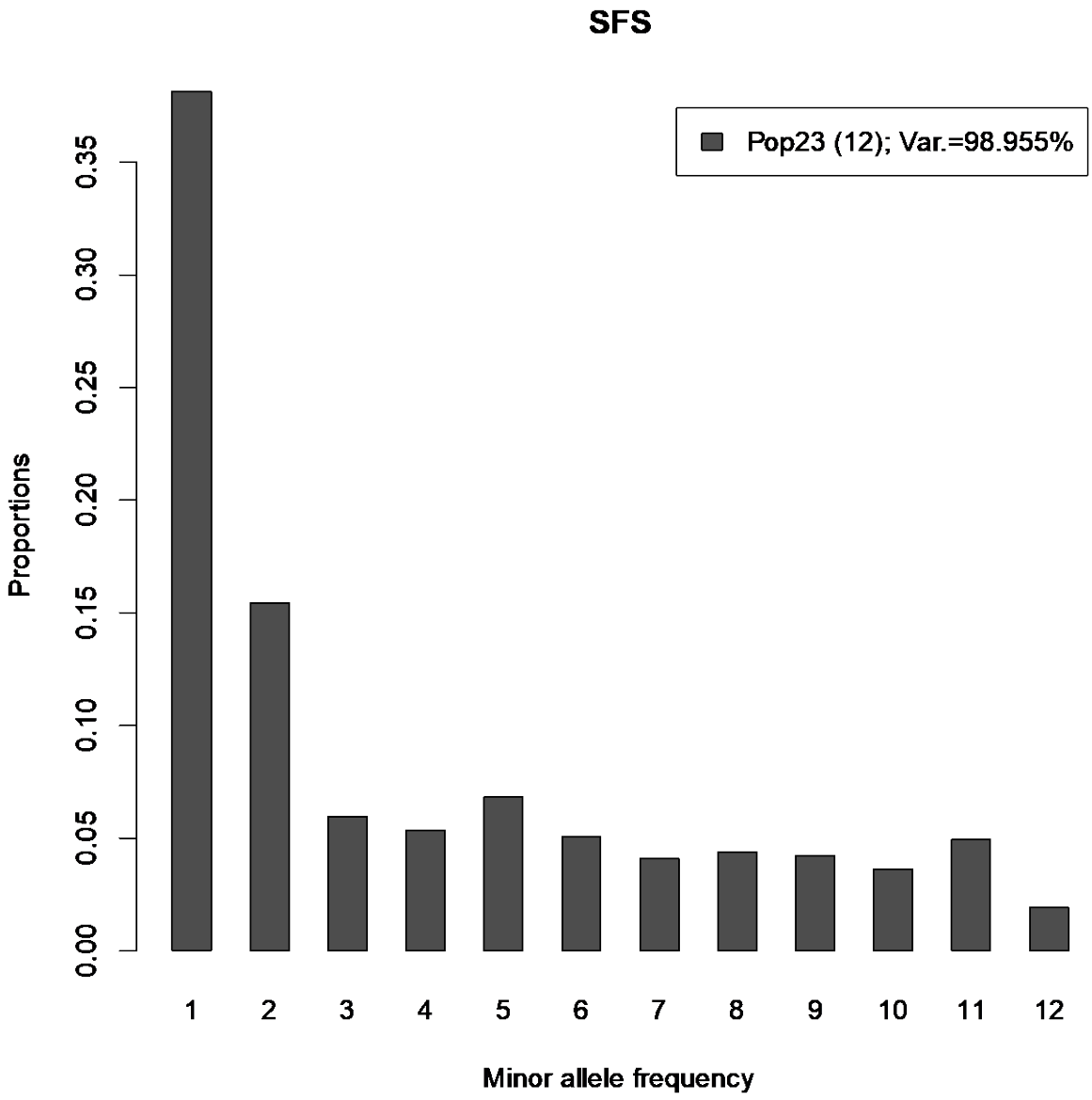

Figure 11. Site frequency spectrum for Pop23, Tiger Pass, BC, generated by filtering for polymorphic sites separated by at least 125 nucleotides.

Appendix C Population Site Frequency Spectra

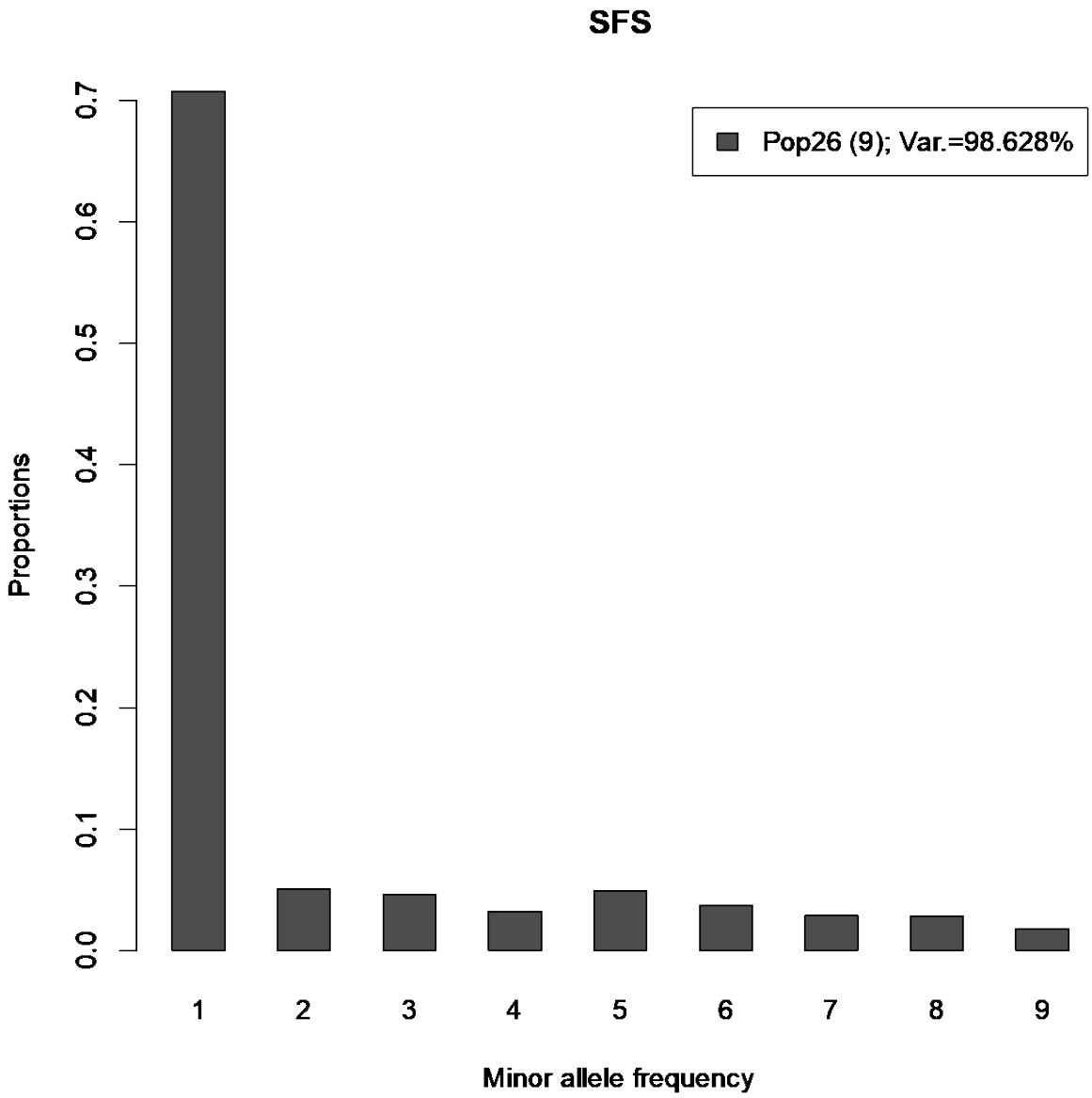

Figure 12. Site frequency spectrum for Pop26, Molar Pass, AB, generated by filtering for polymorphic sites separated by at least 125 nucleotides.

Appendix C Population Site Frequency Spectra

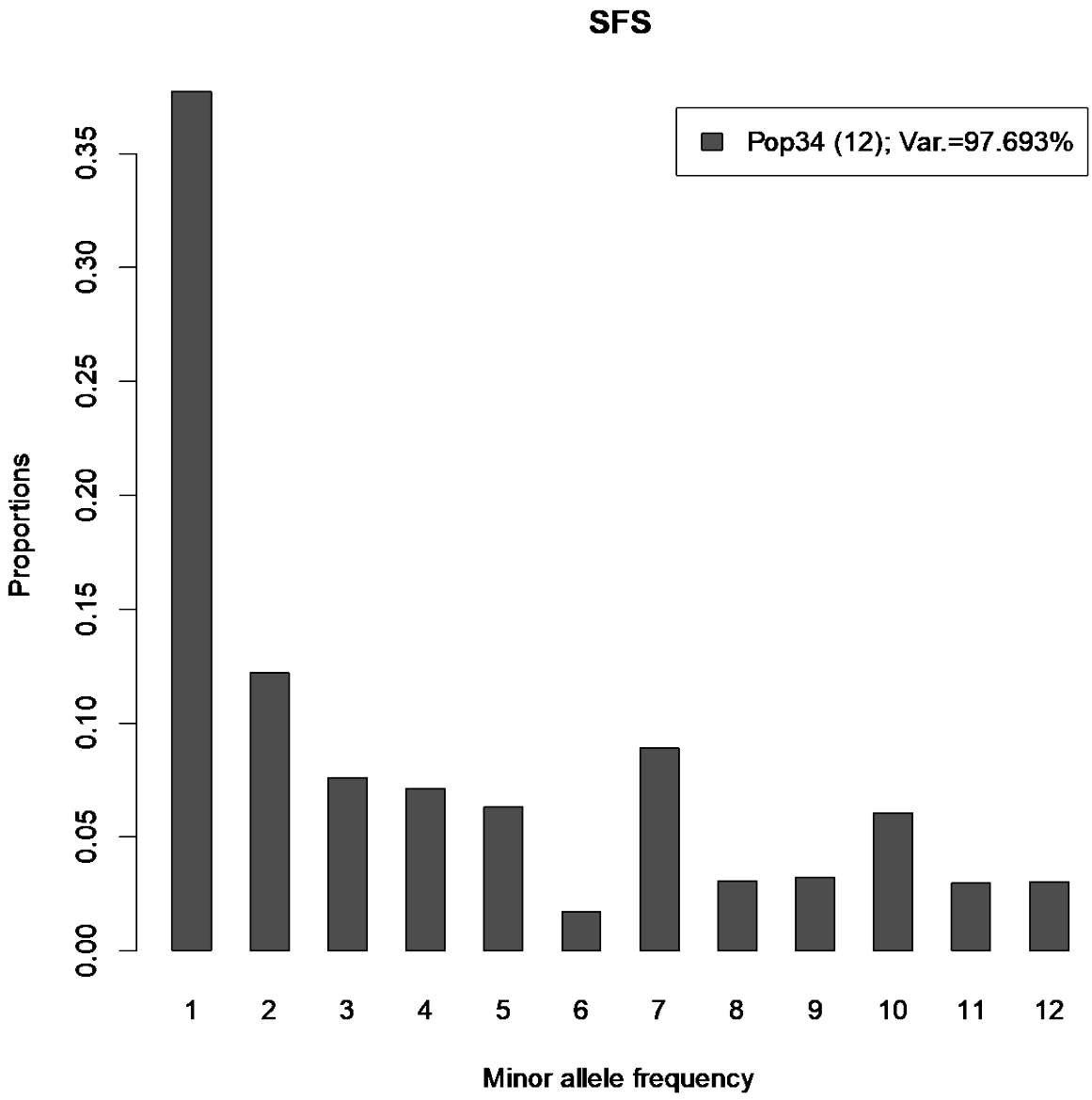

Figure 13. Site frequency spectrum for Pop34, Wonder Pass, AB, generated by filtering for polymorphic sites separated by at least 125 nucleotides.

Appendix C Population Site Frequency Spectra

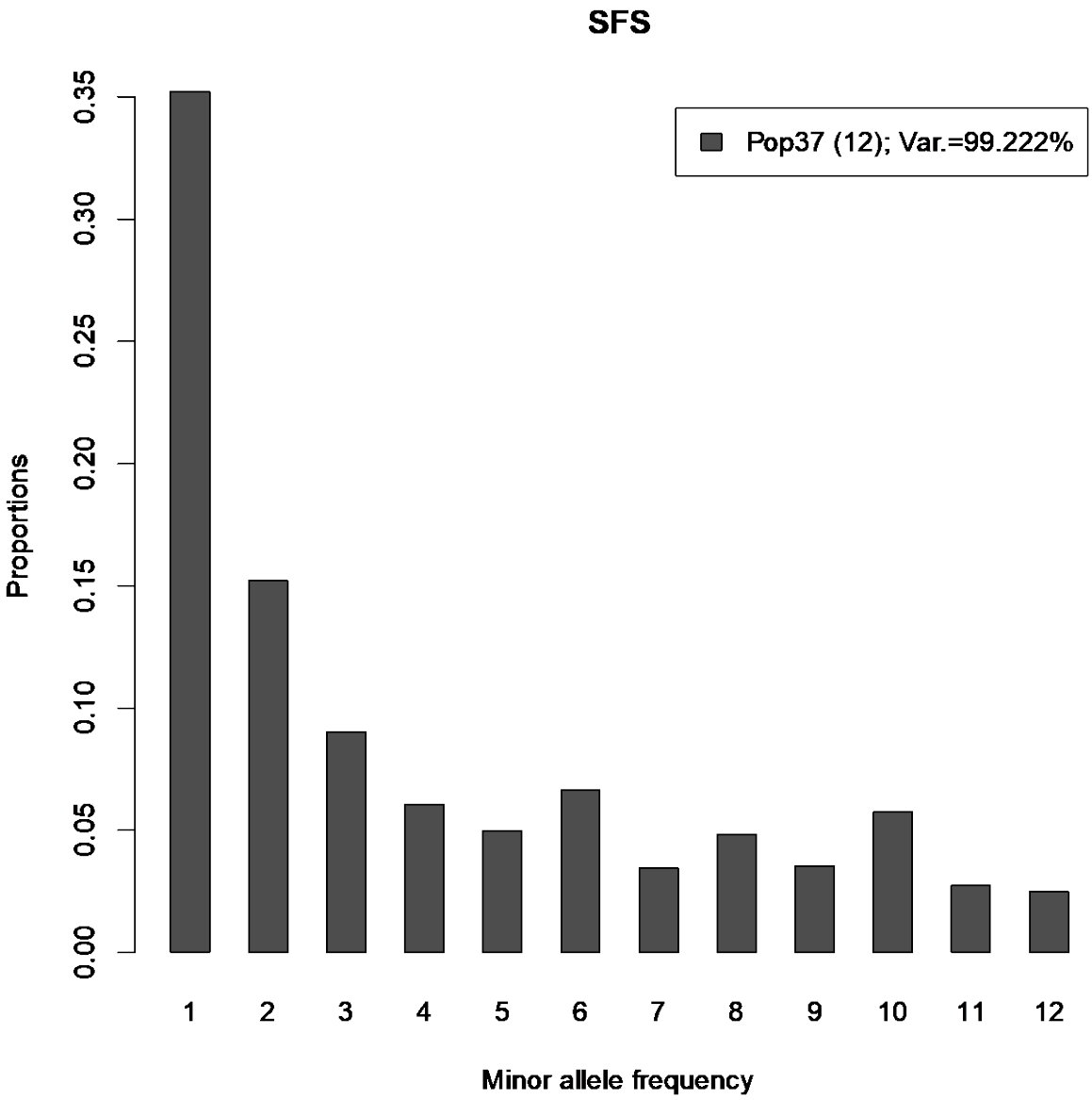

Figure 14. Site frequency spectrum for Pop37, Bovin Lake, AB, generated by filtering for polymorphic sites separated by at least 125 nucleotides.

Appendix C Population Site Frequency Spectra

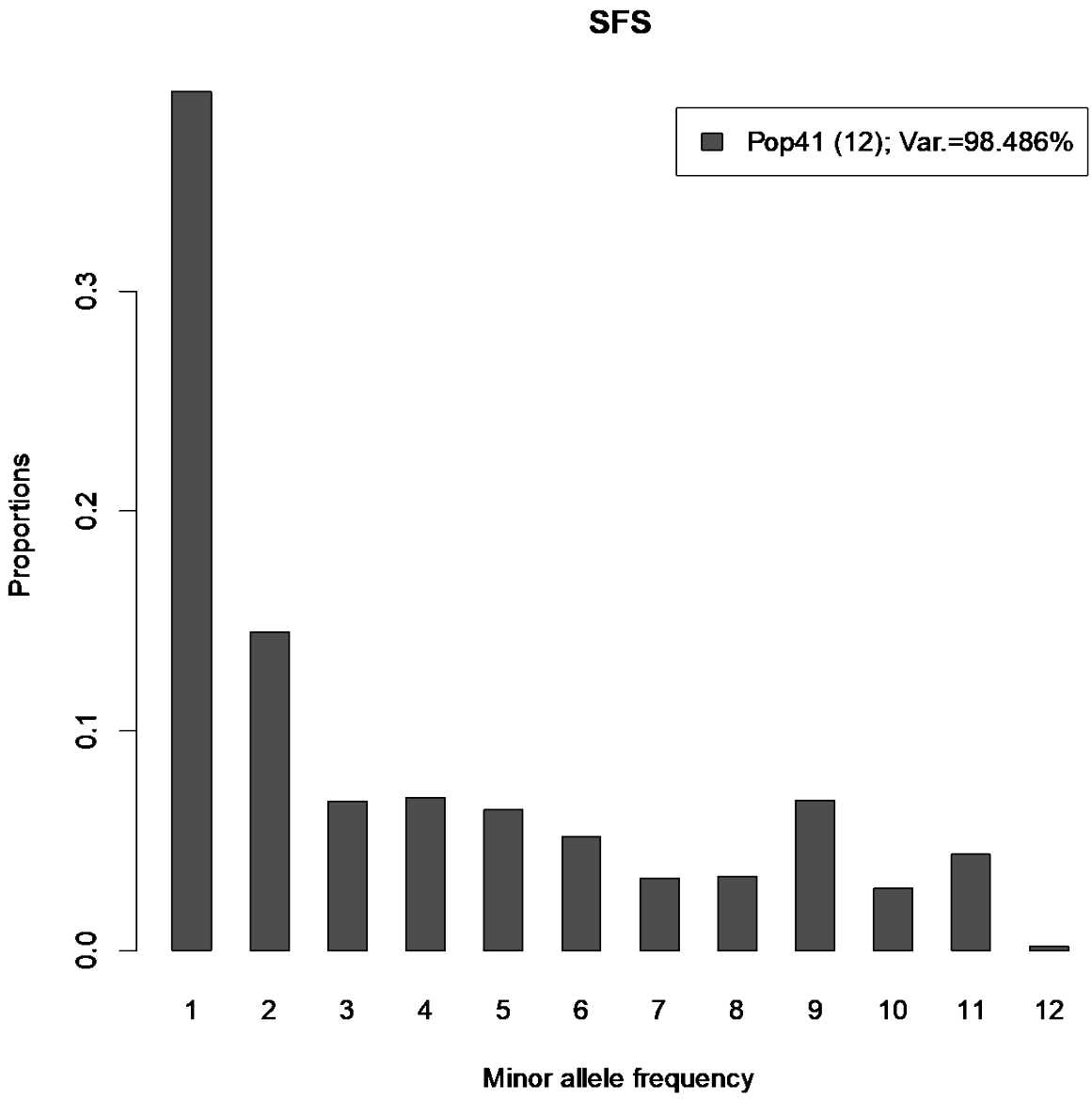

Figure 15. Site frequency spectrum for Pop41, Paradise Park, MT, generated by filtering for polymorphic sites separated by at least 125 nucleotides.

Appendix D: Population Site Frequency Spectra (All Sites)

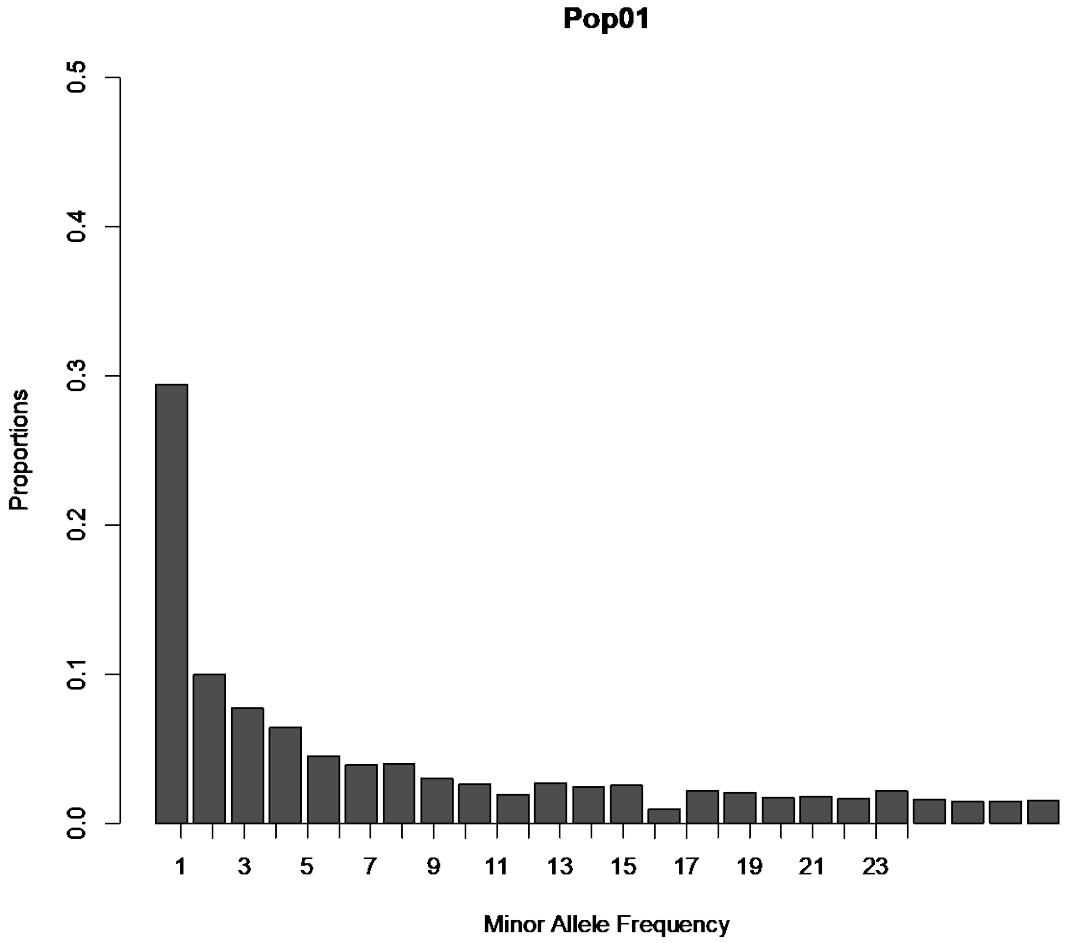

Figure 1. Site frequency spectrum for Pop01, Frosty Mountain, BC, generated using all sites

Appendix D Population Site Frequency Spectra (All Sites)

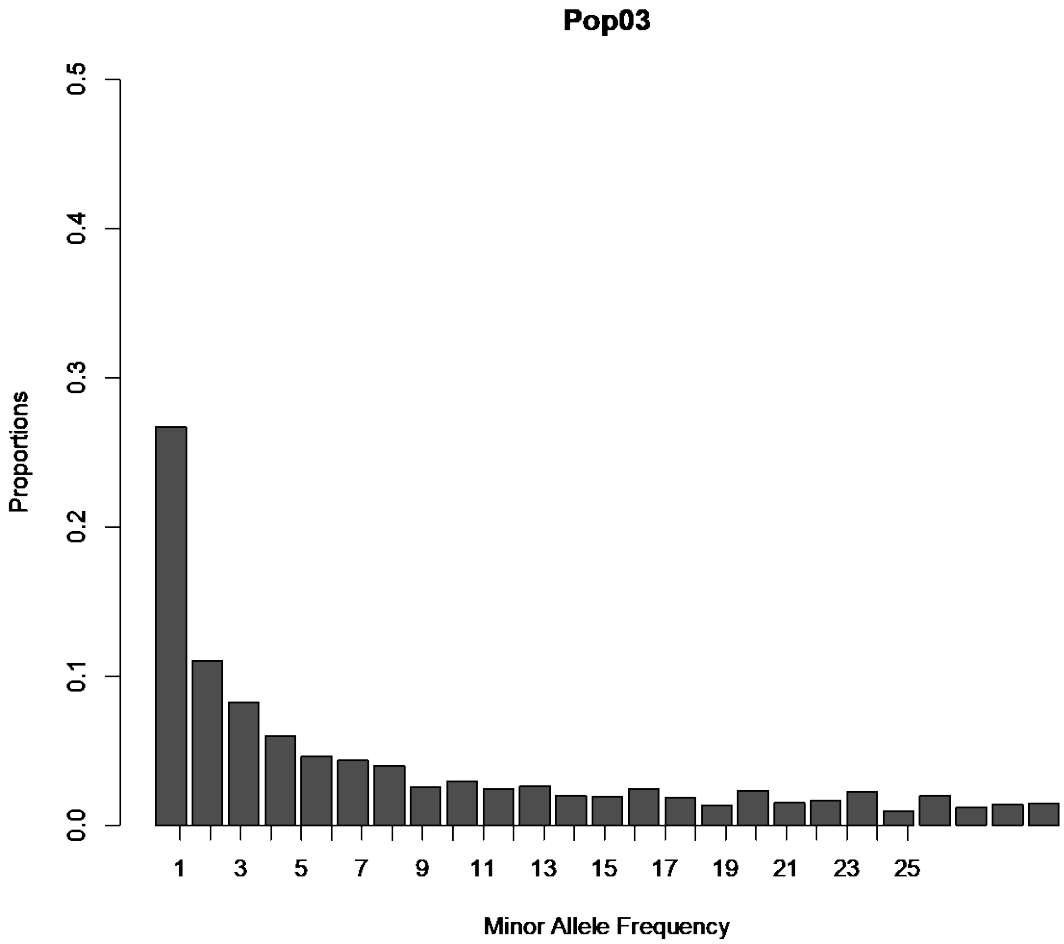

Figure 2. Site frequency spectrum for Pop03, Tiffany Mountain, WA, generated using all sites

Appendix D Population Site Frequency Spectra (All Sites)

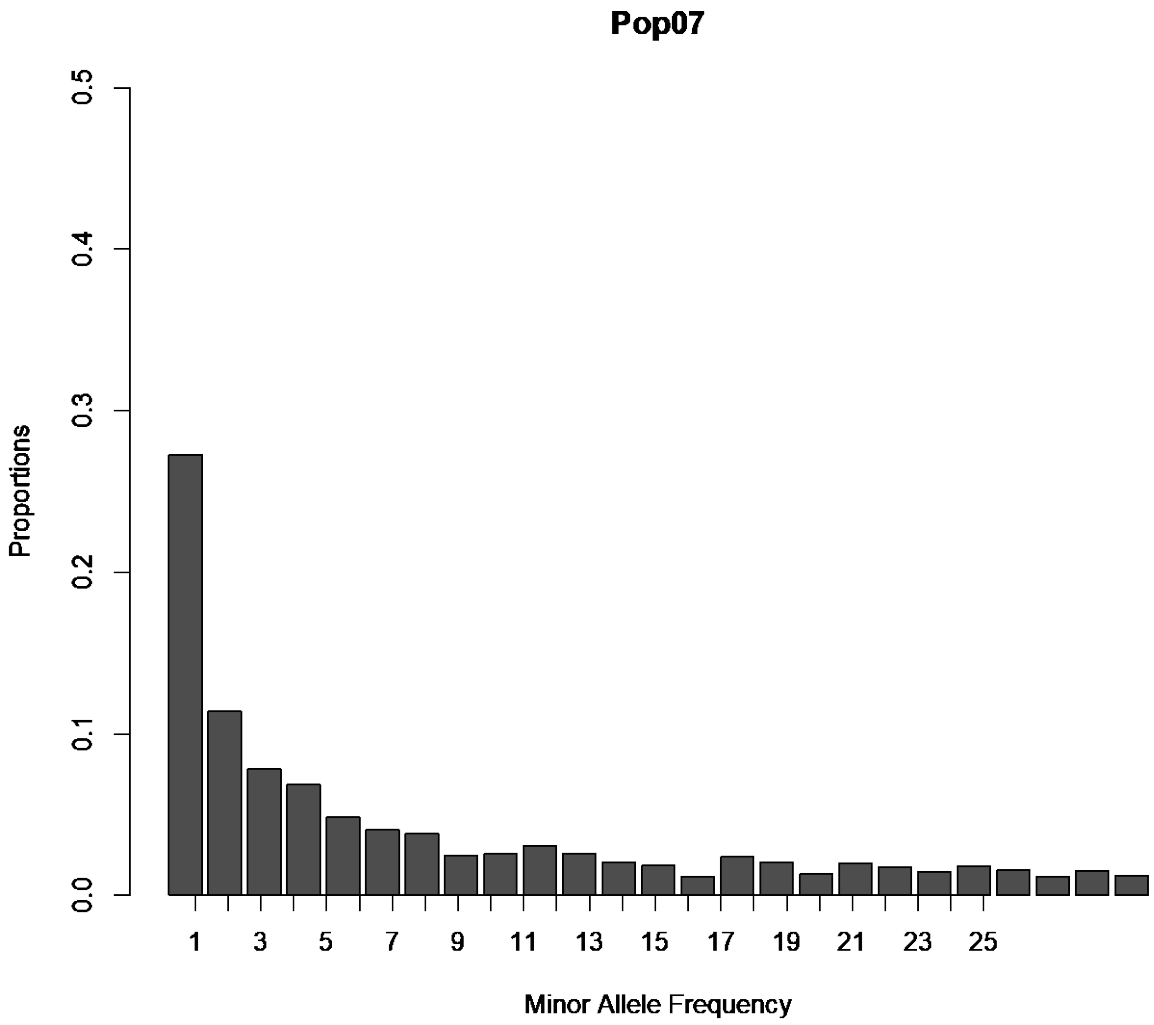

Figure 3. Site frequency spectrum for Pop07, Big Hill, WA, generated using all sites

Appendix D Population Site Frequency Spectra (All Sites)

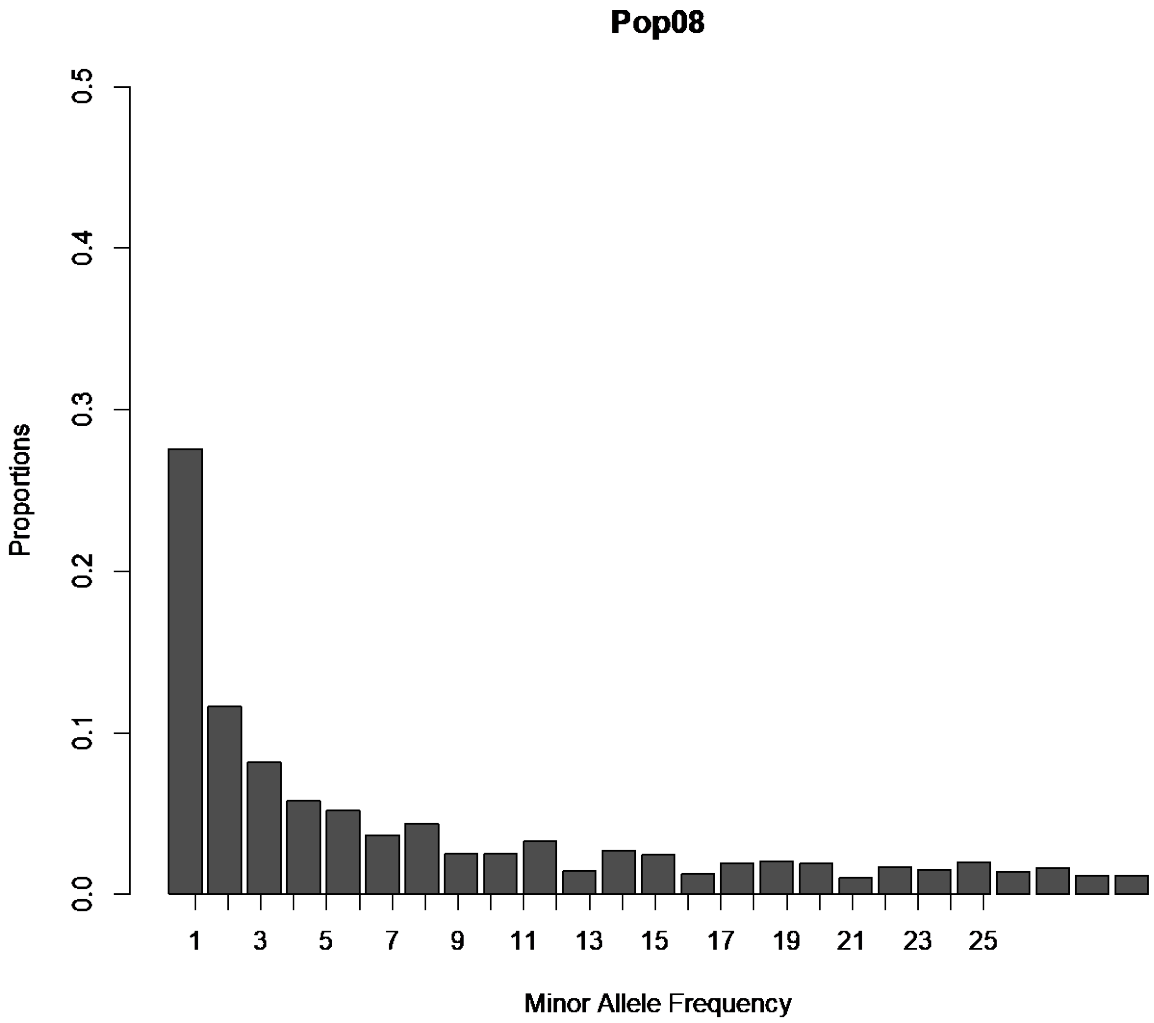

Figure 4. Site frequency spectrum for Pop08, Windy Pass, WA, generated using all sites

Appendix D Population Site Frequency Spectra (All Sites)

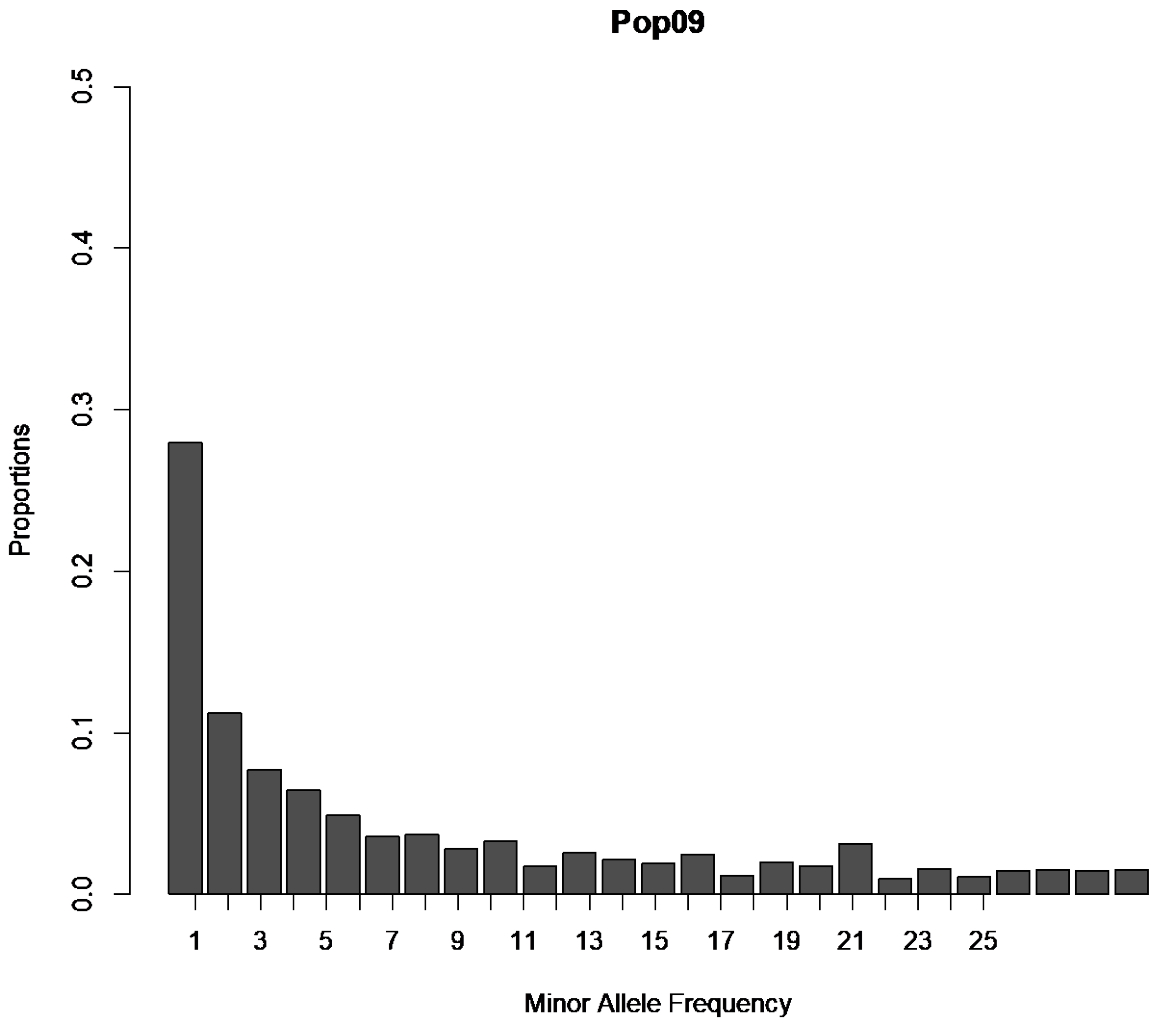

Figure 5. Site frequency spectrum for Pop09, Carlton Ridge, MT, generated using all sites

Appendix D Population Site Frequency Spectra (All Sites)

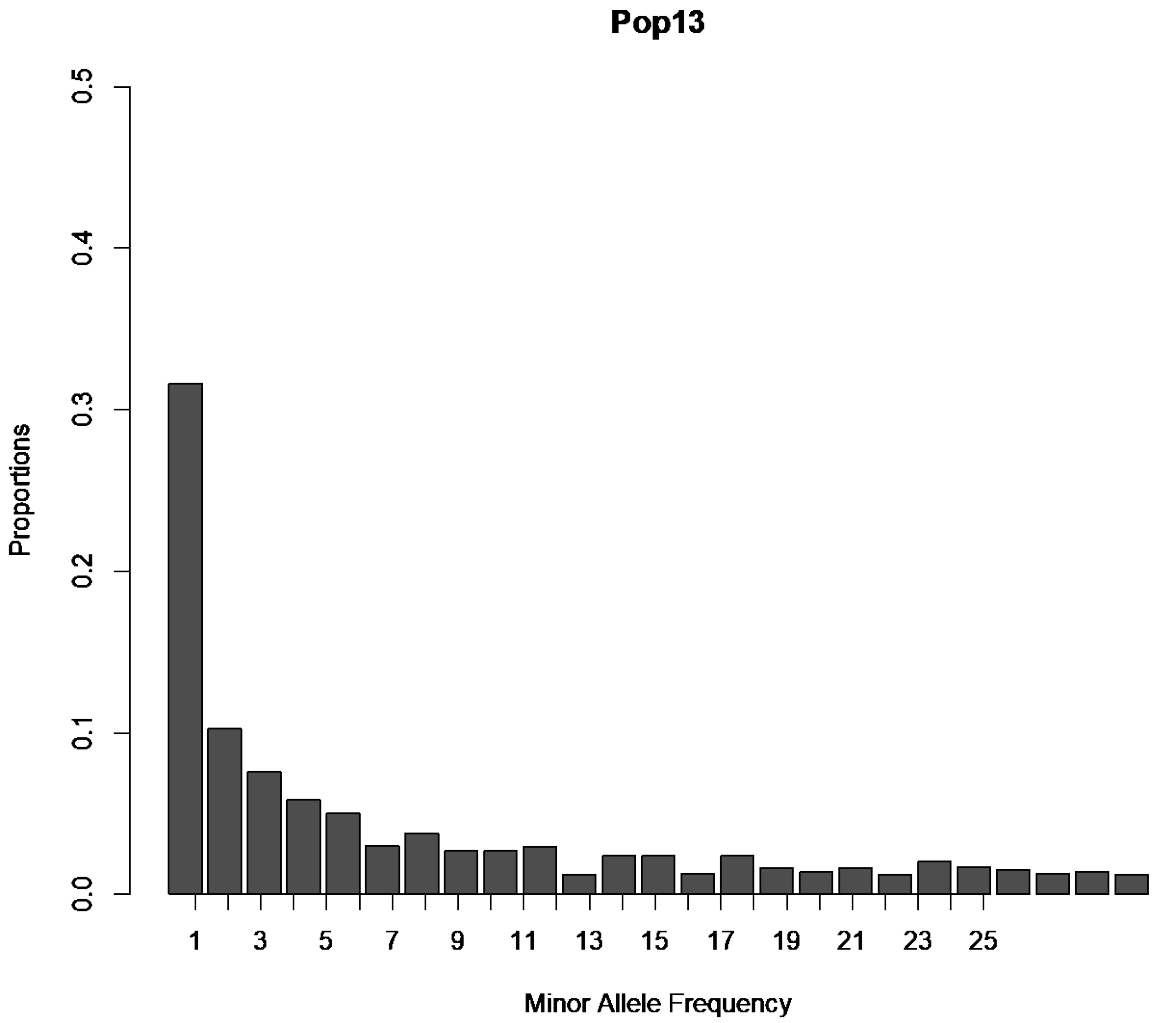

Figure 6. Site frequency spectrum for Pop13, Trapper Peak, MT, generated using all sites

Appendix D Population Site Frequency Spectra (All Sites)

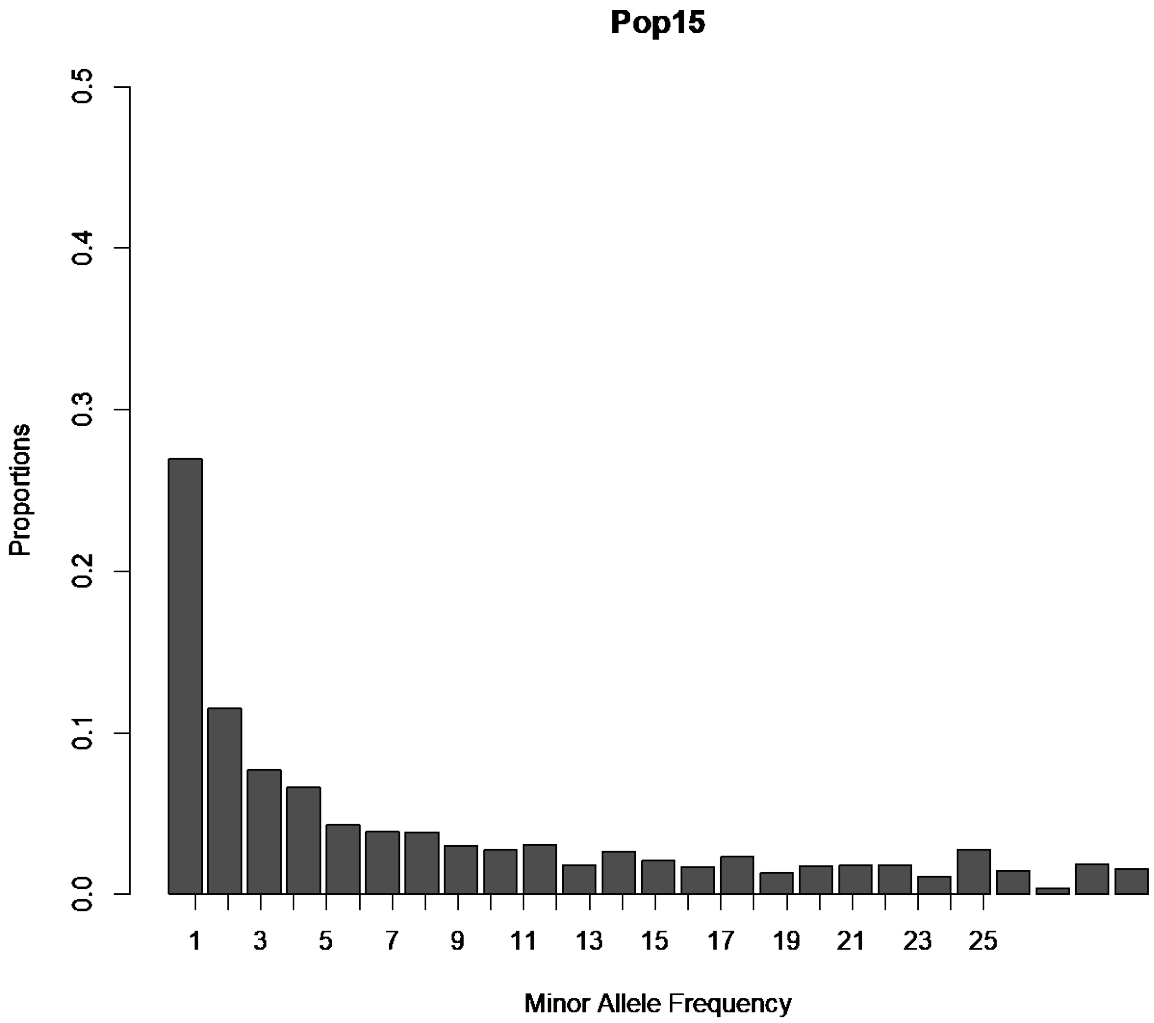

Figure 7. Site frequency spectrum for Pop15, Storm Lake, MT, generated using all sites

Appendix D Population Site Frequency Spectra (All Sites)

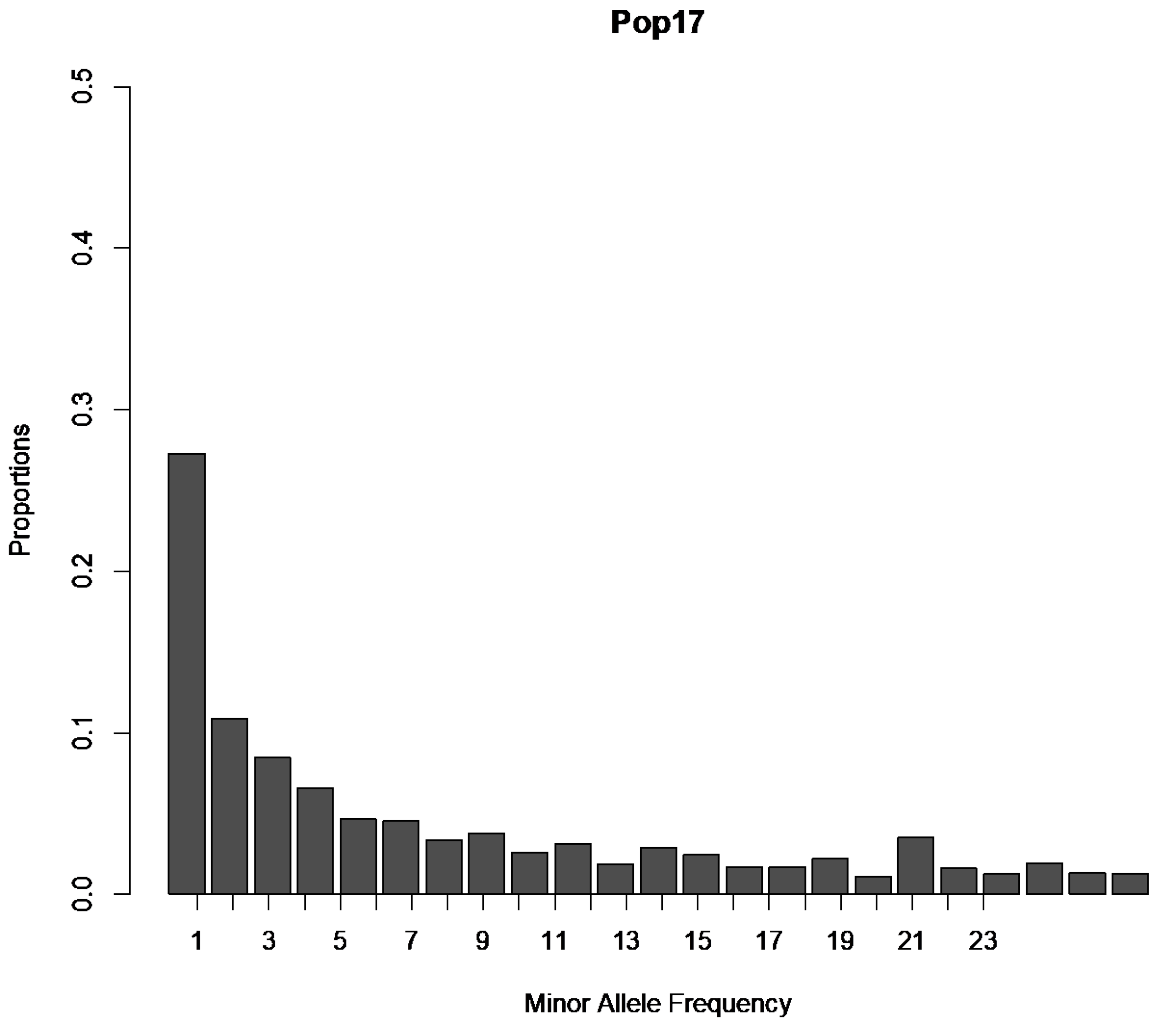

Figure 8. Site frequency spectrum for Pop17, Holland Pass, MT, generated using all sites

Appendix D Population Site Frequency Spectra (All Sites)

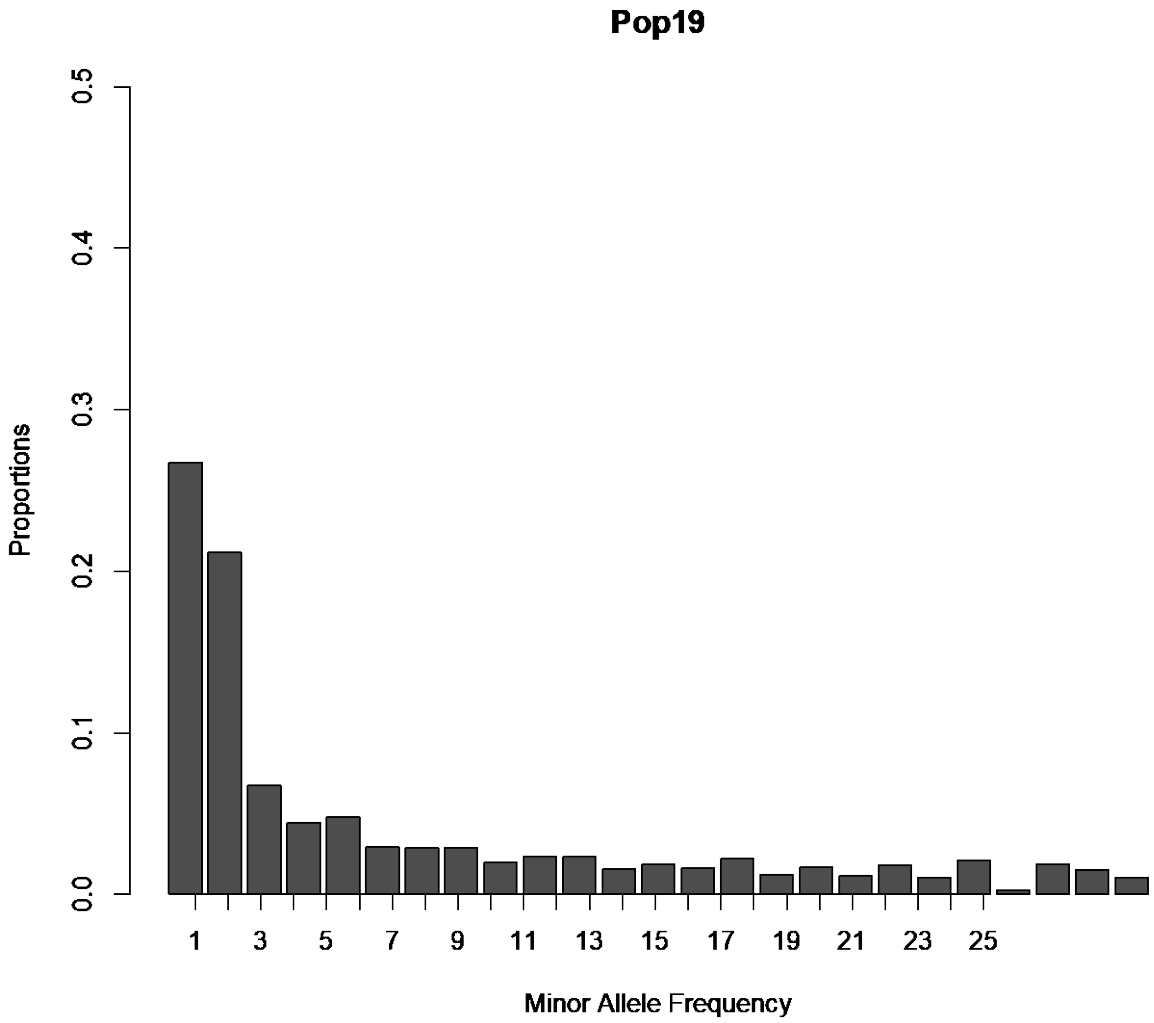

Figure 9. Site frequency spectrum for Pop19, Roman Nose, ID, generated using all sites

Appendix D Population Site Frequency Spectra (All Sites)

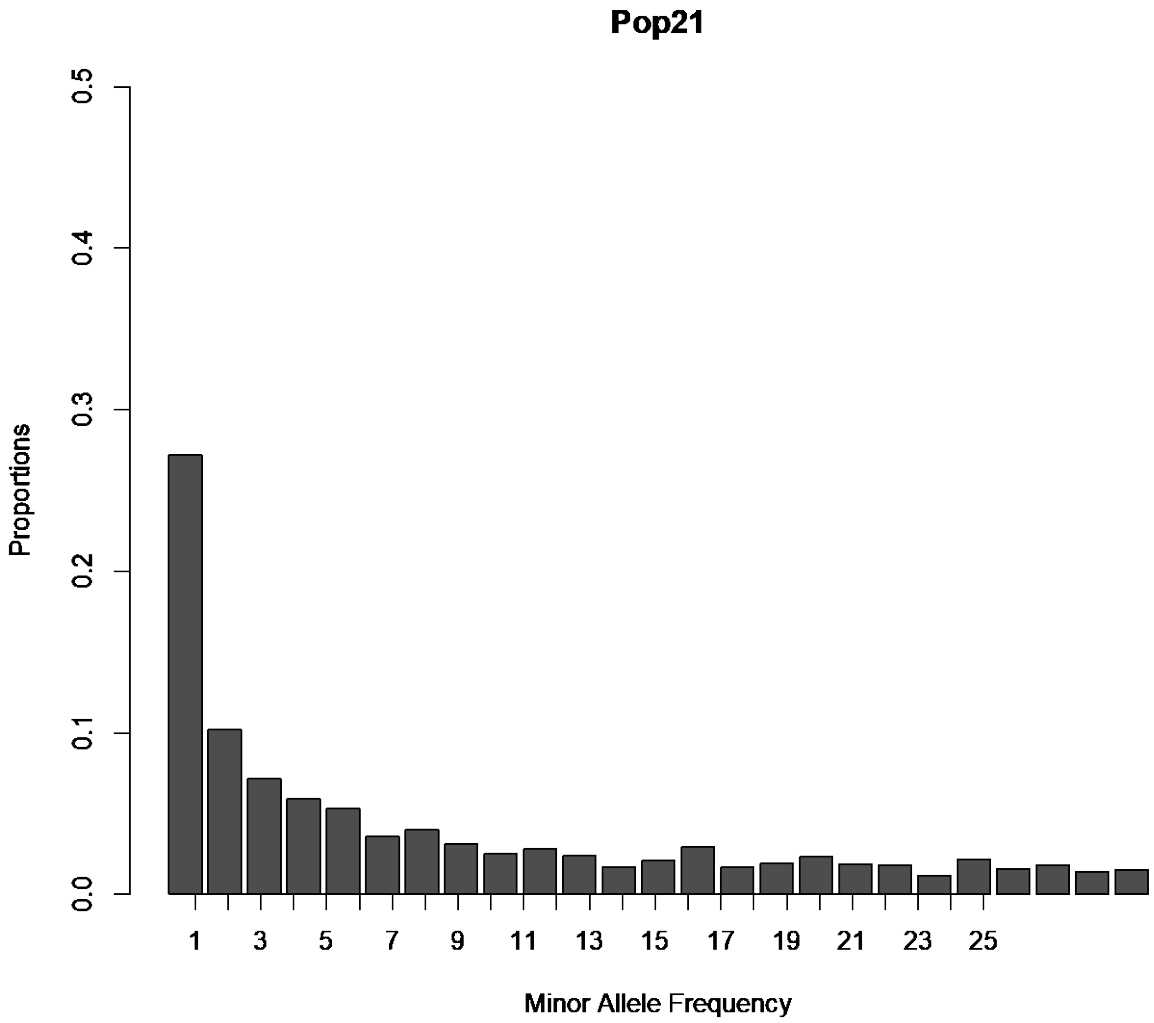

Figure 10. Site frequency spectrum for Pop21, Sparkle Lake, BC, generated using all sites

Appendix D Population Site Frequency Spectra (All Sites)

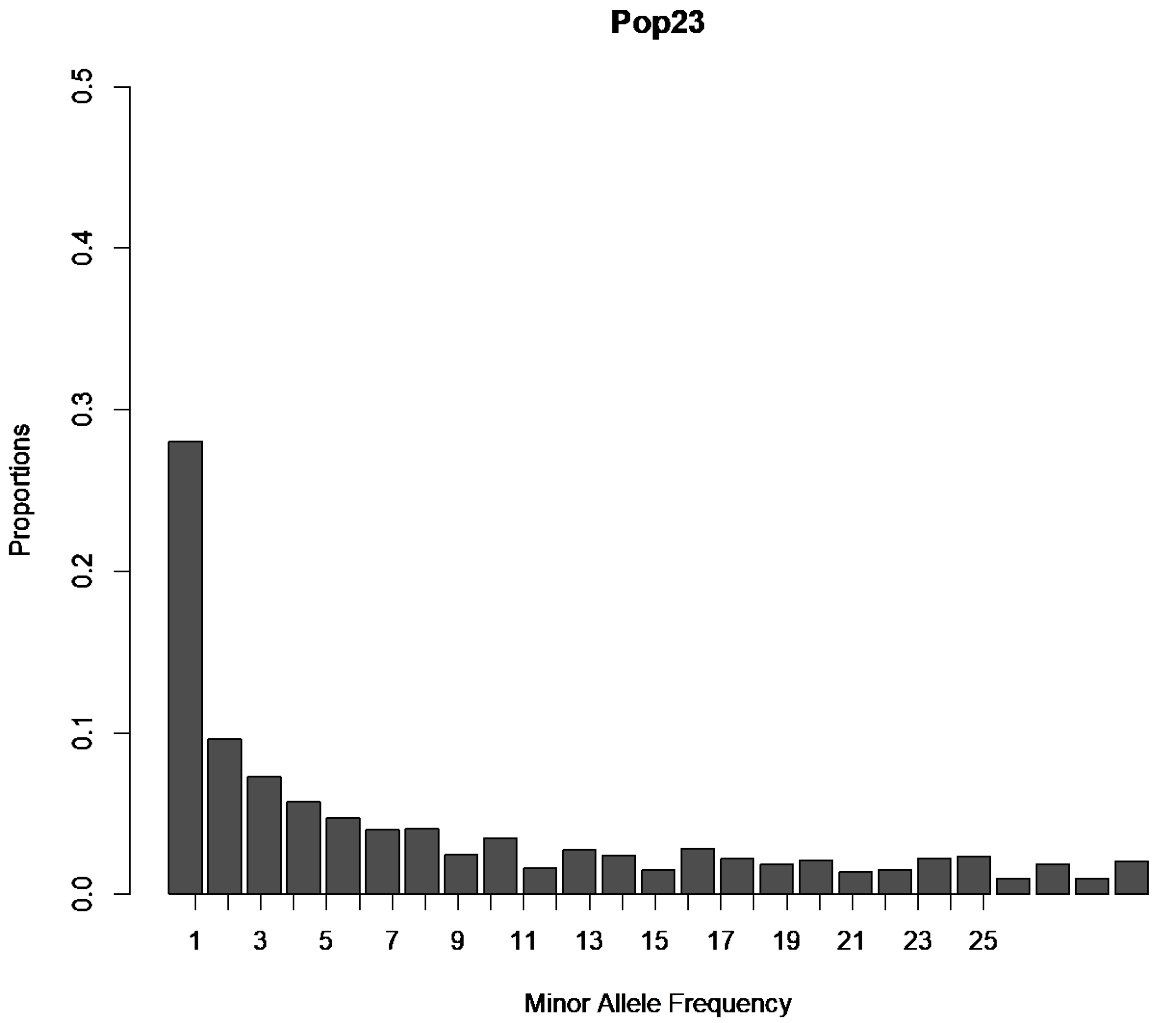

Figure 11. Site frequency spectrum for Pop23, Tiger Pass, BC, generated using all sites

Appendix D Population Site Frequency Spectra (All Sites)

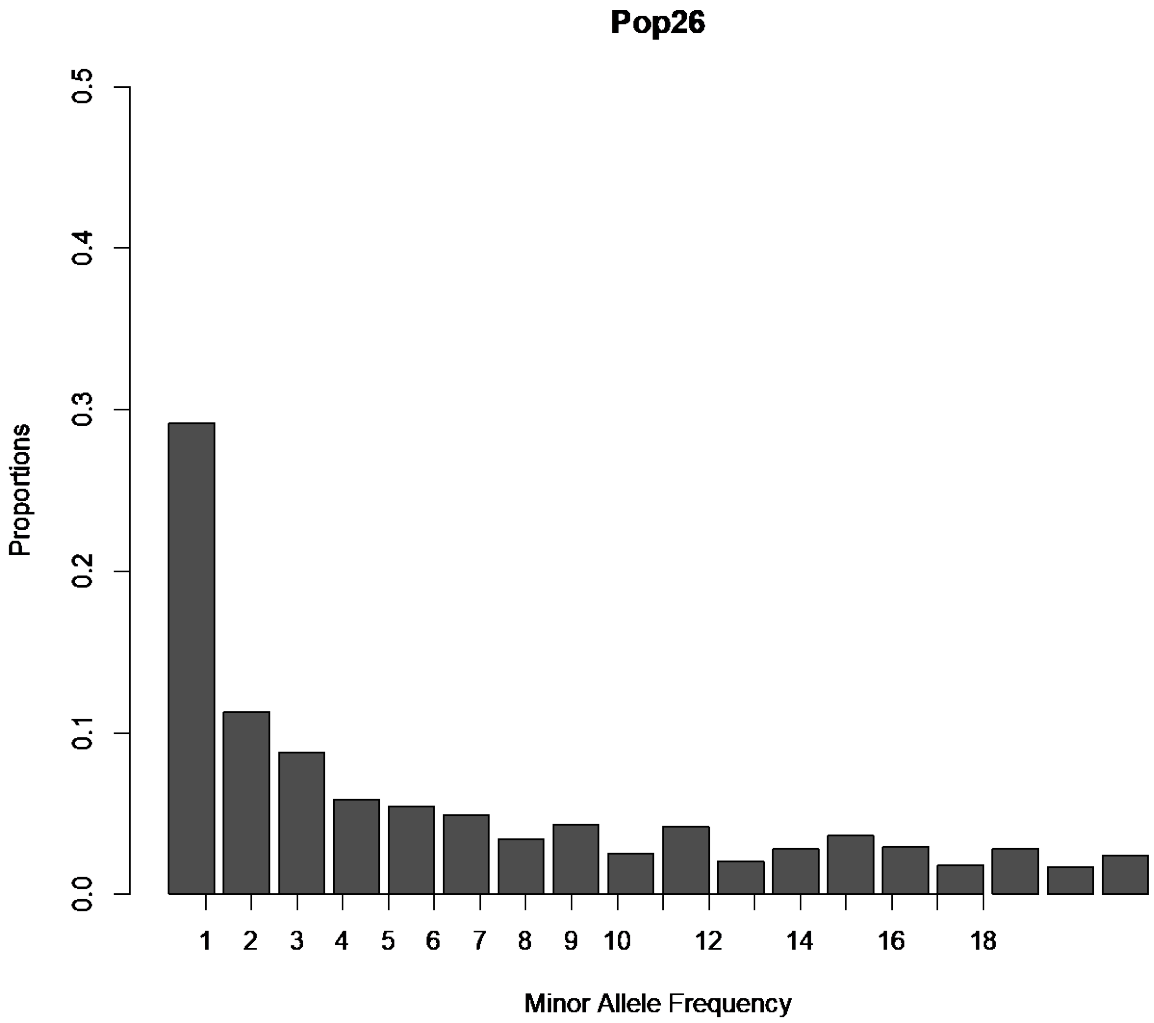

Figure 12. Site frequency spectrum for Pop26, Molar Pass, AB, generated using all sites

Appendix D Population Site Frequency Spectra (All Sites)

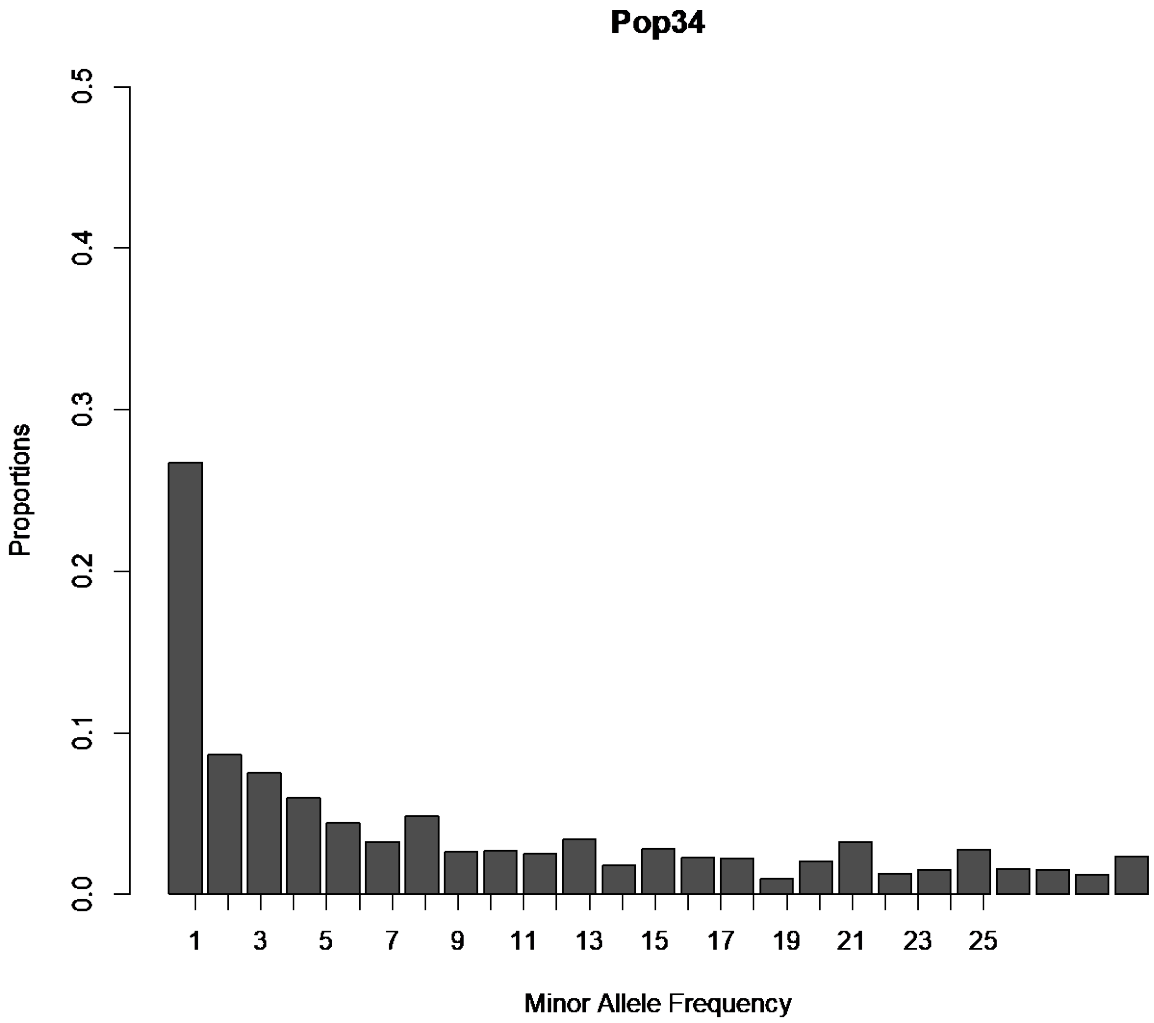

Figure 13. Site frequency spectrum for Pop34, Wonder Pass, AB, generated using all sites

Appendix D Population Site Frequency Spectra (All Sites)

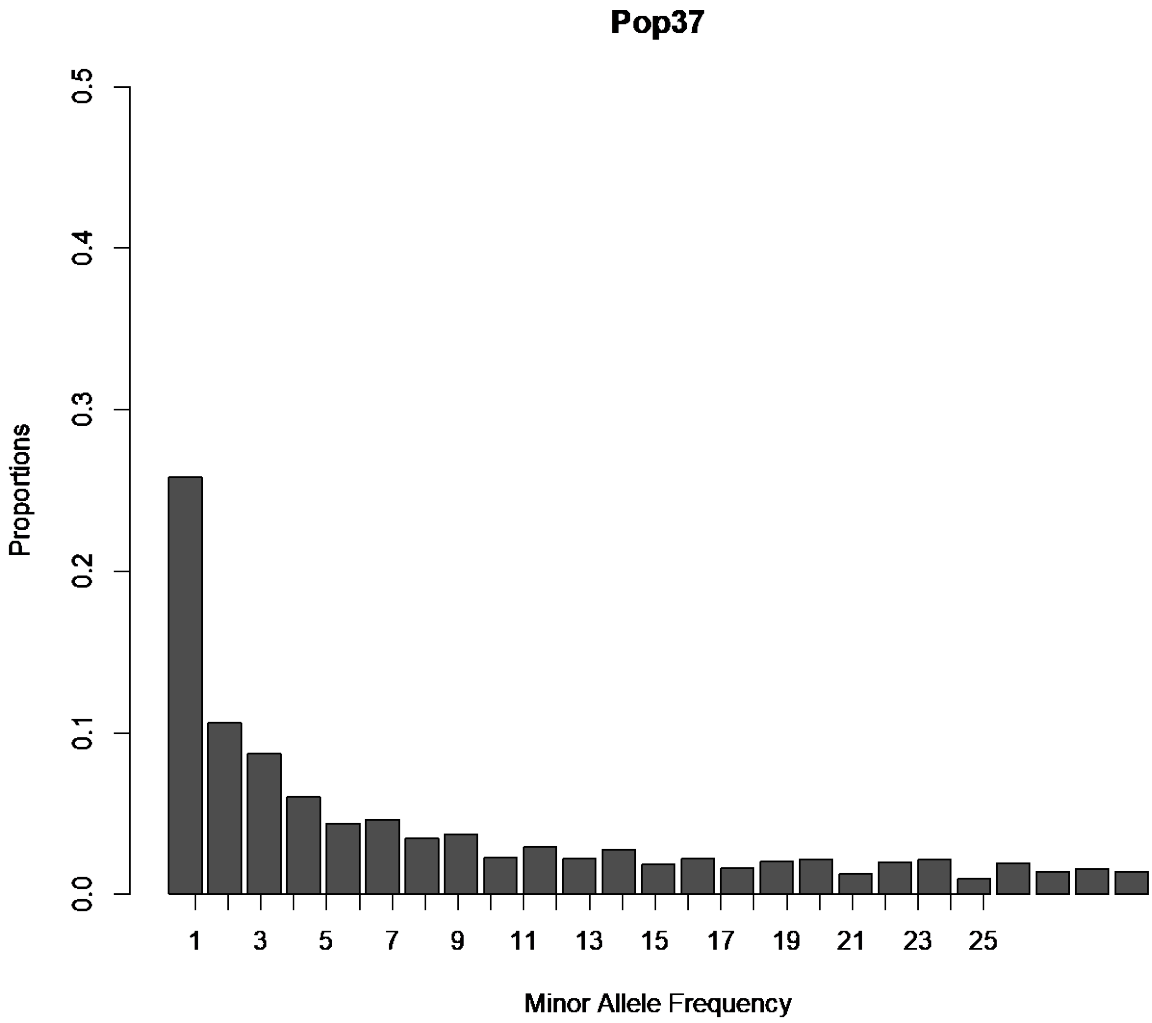

Figure 14. Site frequency spectrum for Pop37, Bovin Lake, AB, generated using all sites

Appendix D Population Site Frequency Spectra (All Sites)

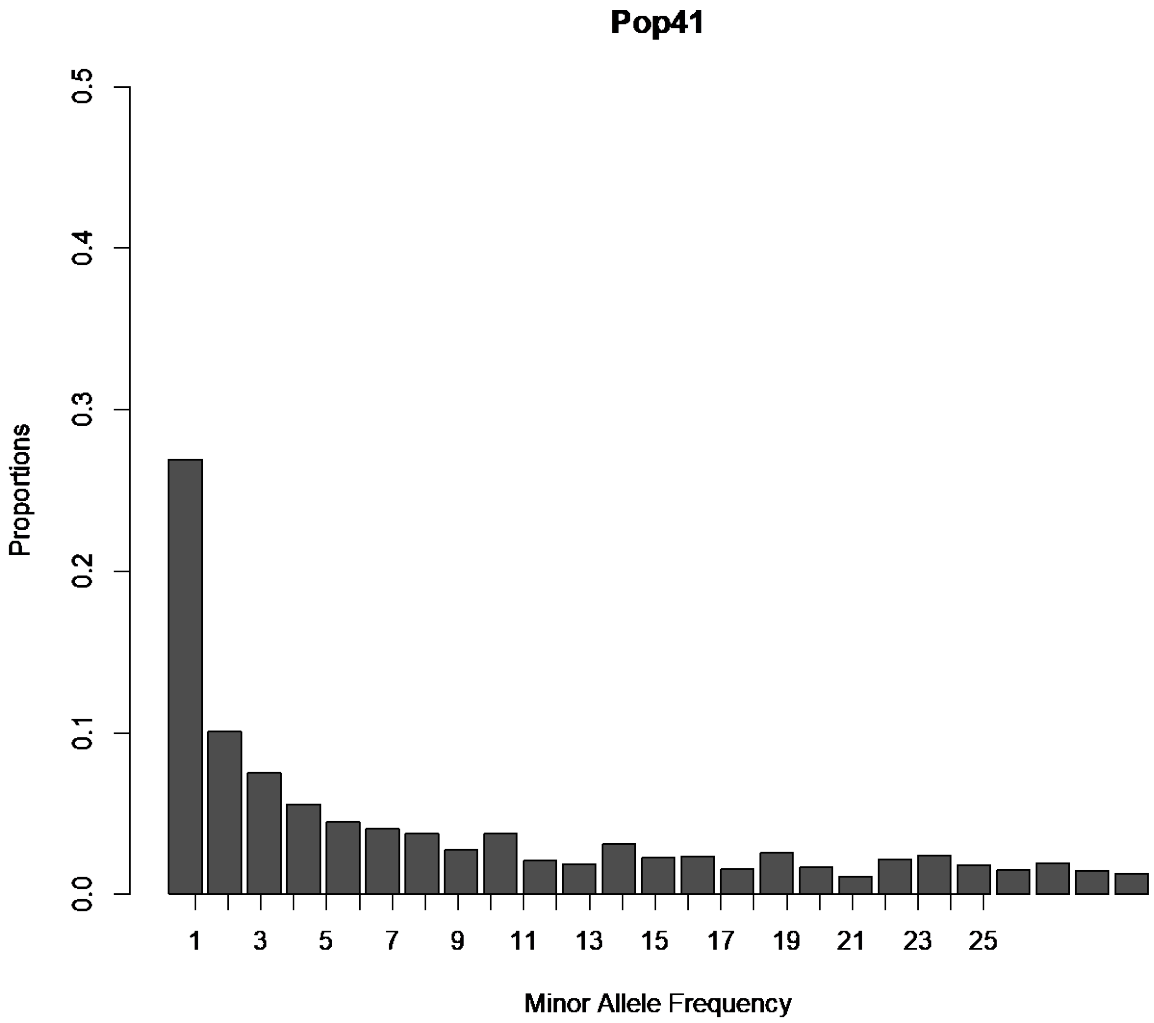

Figure 15. Site frequency spectrum for Pop41, Paradise Park, MT, generated using all sites

Appendix E: AIC-Based Demographic Scenario Rankings

Table 1. Coalescent simulations were run 100 times for seven single-population demographic scenarios (S1 = constant population size; S2 = instant resize; S3 = intermediate size change lasting 10 generations; S4 = intermediate size change lasting 100 generations; S5 = intermediate size change lasting 500 generations; S6 = intermediate size change lasting 1,000 generations; S7 = intermediate size change lasting 2,000 generations) and an AIC-based model selection process was used to select the most likely scenario. Differences in AIC values between scenarios are reported here for 15 populations of subalpine larch representing the species natural range.

Pop.	ΔL	AIC	scenario 1	ΔAIC1	scenario 2	ΔAIC2	scenario 3	ΔAIC3	scenario 4	ΔAIC4	scenario 5	ΔAIC5	scenario 6	ΔAIC6	scenario 7
Pop01	255	1,068,874	S3	-11	S7	0	S5	0	S4	-4	S2	-15	S6	-1359	S1
Pop03	299	1,285,347	S7	-6	S6	-11	S4	-3	S5	-13	S2	-3	S3	-603	S1
Pop07	249	1,112,418	S7	-31	S6	-41	S5	-75	S4	-13	S3	-7	S2	-519	S1
Pop08	246	738,561	S7	-5	S6	-32	S5	-66	S4	-18	S2	-8	S3	-353	S1
Pop09	346	894,305	S7	-17	S5	-5	S6	-9	S4	-7	S2	0	S3	-754	S1
Pop13	326	1,188,174	S7	-4	S4	-5	S3	-1	S5	-1	S2	-4	S6	-2,878	S1
Pop15	352	788,251	S7	-19	S2	-9	S5	0	S4	0	S3	-3	S6	-384	S1
Pop17	302	814,174	S7	-7	S3	-1	S5	-2	S6	-1	S2	-8	S4	-396	S1
Pop19	891	970,839	S7	-431	S6	-319	S5	-397	S4	-101	S3	-8	S2	-2,139	S1
Pop21	196	945,510	S3	-3	S7	-3	S6	-5	S5	-11	S4	-118	S2	-461	S1
Pop23	404	848,205	S5	-2	S7	-9	S6	-12	S4	0	S3	-280	S2	-433	S1
Pop26	339	868,243	S7	-18	S6	-17	S4	-6	S5	-26	S3	-238	S2	-72	S1
Pop34	802	1,226,333	S7	-3	S6	-44	S5	-31	S4	-10	S3	-888	S1	-150	S2
Pop37	291	1,231,485	S3	-1	S5	-2	S4	-2	S7	-2	S2	-10	S6	-365	S1
Pop41	221	860,986	S6	-2	S7	-19	S5	0	S3	-14	S4	-201	S2	-236	S1

Appendix F: Parameter Estimates for Second-Ranked Demography Scenario

Table 1. Parameter estimates generated for the second-most preferred demographic scenario (S1 = constant population size; S2 = instant resize *t* generations ago; S3 = intermediate size change lasting 10 generations that ends *t* generations ago; S4 = intermediate size change lasting 100 generations that ends *t* generations ago; S5 = intermediate size change lasting 500 generations that ends *t* generations ago; S6 = intermediate size change lasting 1,000 generations that ends *t* generations ago; S7 = intermediate size change lasting 2,000 generations that ends *t* generations ago) for single-population coalescent simulation models representing 15 populations of subalpine larch from across the species range

Region	Population	Secondary Scenario	Ancestral N_E*	Intermediate N_E*	Current N_E*	*t***	Description	Difference
Cascades	Pop01	S7	15,000	14,549	493,310	548	bottleneck	
	Pop03	S6	15,000	20,606	172,165	207	expansion	
	Pop07	S6	15,000	113,508	22,588	2,027	retreat	
	Pop08	S6	15,000	203,660	22,449	2,151	retreat	
Southern Rockies	Pop09	S5	15,000	38,823	28,372	51	retreat	✓
	Pop13	S4	15,000	46,707	152,942	815	expansion	
	Pop15	S2	15,000	775	27,153	1,549	bottleneck	✓
	Pop17	S3	15,000	1,390	133,621	508	bottleneck	✓
Central Rockies	Pop19	S6	15,000	539,921	27,491	2,063	retreat	
	Pop37	S7	15,000	16,990	278,249	146	expansion	
	Pop41	S7	15,000	10,310	78,485	543	bottleneck	
Northern Rockies	Pop21	S6	15,000	8,586	464,357	551	bottleneck	
	Pop23	S6	15,000	6,861	136,889	755	bottleneck	
	Pop26	S5	15,000	5,777	57,798	1,146	bottleneck	
	Pop34	S7	15,000	7,503	123,122	632	bottleneck	

*Current effective population size (NE) given in the number of diploid individuals

**Timing of most recent population resize given as the number of generation

Appendix G: Samples with Negative Cold Injury Values

Table 1. Number of individuals with higher electrolyte leakage in controls than in frozen samples for cold injury assessment of 100 subalpine larch trees frozen at different sub-zero temperatures

	Winter	Spring	Autumn
Tray -10˚C	0	10	0
Tray -20˚C	6	4	2
Tray -30˚C	1	1	1
Tray -40˚C	0	0	0

Appendix H: Climate Variables Retained For Redundancy Analysis

Table 1. Climate variables with linear relationships > 5% identified for inclusion in initial redundancy analysis

Winter climate variables	Spring climate variables	Autumn climate variables
Tmin_at	Tmin_at	Tmin_at
MAT	MCMT	MAT
MWMT	TD	MCMT
MCMT	DD_0	TD
SHM	eFFP	MSP
DD_0	FFP	SHM
DD5	EMT	DD_0
NFFD		eFFP
bFFP		FFP
eFFP		EMT
FFP		CMD
EMT		

Appendix I: Parameter settings for ANGSD

Table 1. Parameter settings for ANGSD genotyping, principal components analysis (PCA) and a dendrogram of genetic distance (Gdist). Genotyping was carried out on 100 subalpine larch samples phenotyped for cold tolerance. PCA and Gdist analyses were carried out using range-wide data generated using restriction enzyme associated DNA sequencing (RAD-seq) with the Pst1 restriction enzyme.

Parameter	Genotyping	PCA	Gdist
-bam	yes	yes	yes
-ref	*L. sibirica*	*L. sibirica*	*L. sibirica*
-anc	*L. sibirica*	*L. sibirica*	*L. sibirica*
-GL	2	2	2
-baq	1	1	1
-C	50	50	50
-minMapQ	20	20	20
-minQ	20	20	20
-doCounts	1	1	1
-doMaf	1	1	1
-doMajorMinor	1	1	1
-SNP_pval	5.00E-02	5.00E-02	5.00E-02
-skipTriallelic	1	1	1
-sites	yes	yes	yes/no*
-doGlf	.	no	no
-doGeno	4	32	32
-doPost	1	1	1
-indF	.	no	no
-minMaf	0.05	0.05	0.05
-doSaf	.	no	no
-fold	.	no	no
-minInd	N/2	70/50/30**	70/50/30**

*Different values were tested

**Different values (out of 100) were tested

Appendix J: Cold Tolerance UnifiedGenotyper Genotypes

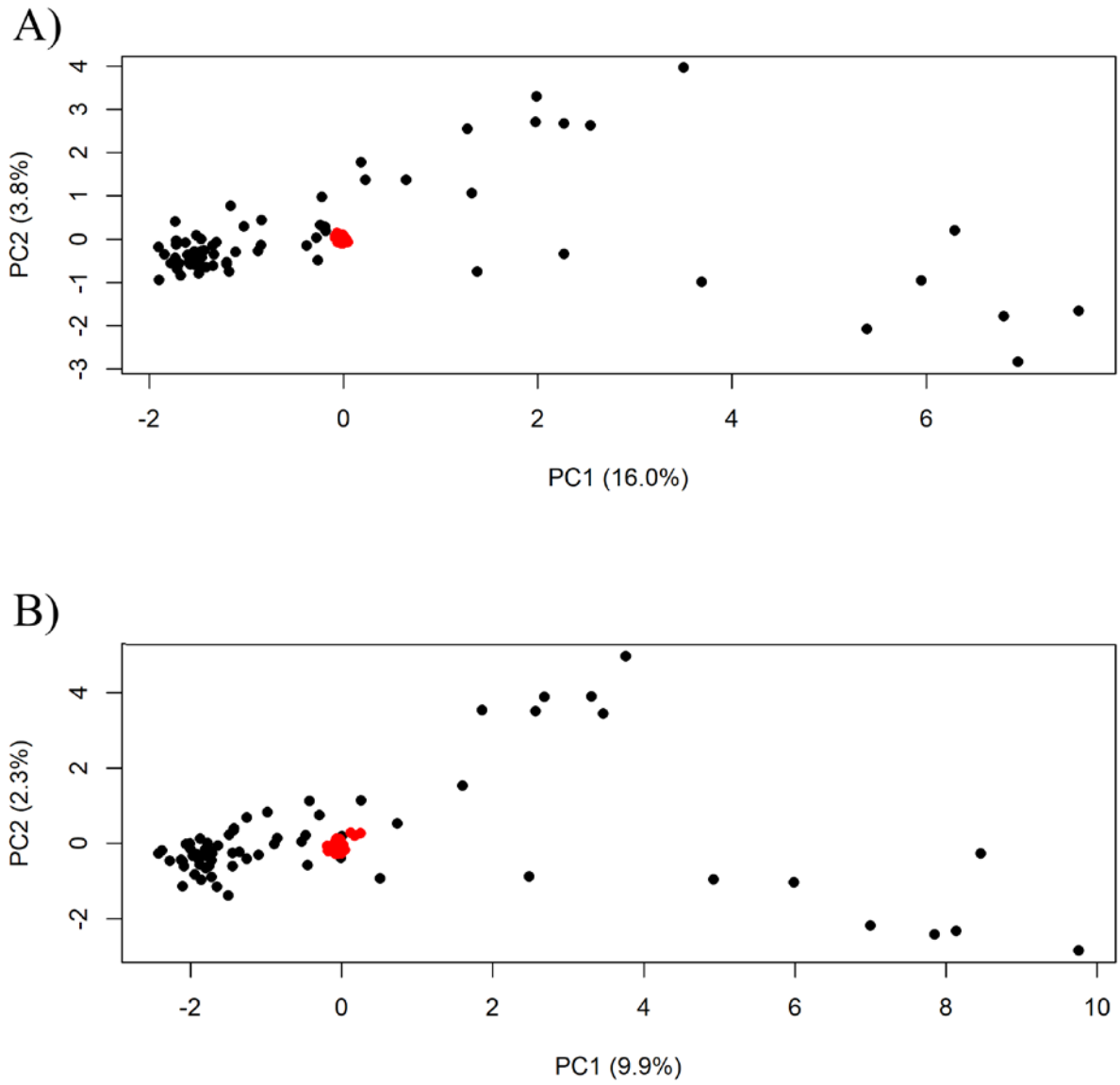

Figure 1. PCA output for GATK UnifiedGenotyper SNPs filtered using (A) relaxed and (B) stringent parameter settings. Individuals processed with Sbf1 (red) are tightly clustered relative to individuals processed with the Pst1 restriction enzyme (black)

Appendix J Cold Tolerance UnifiedGenotyper Genotypes

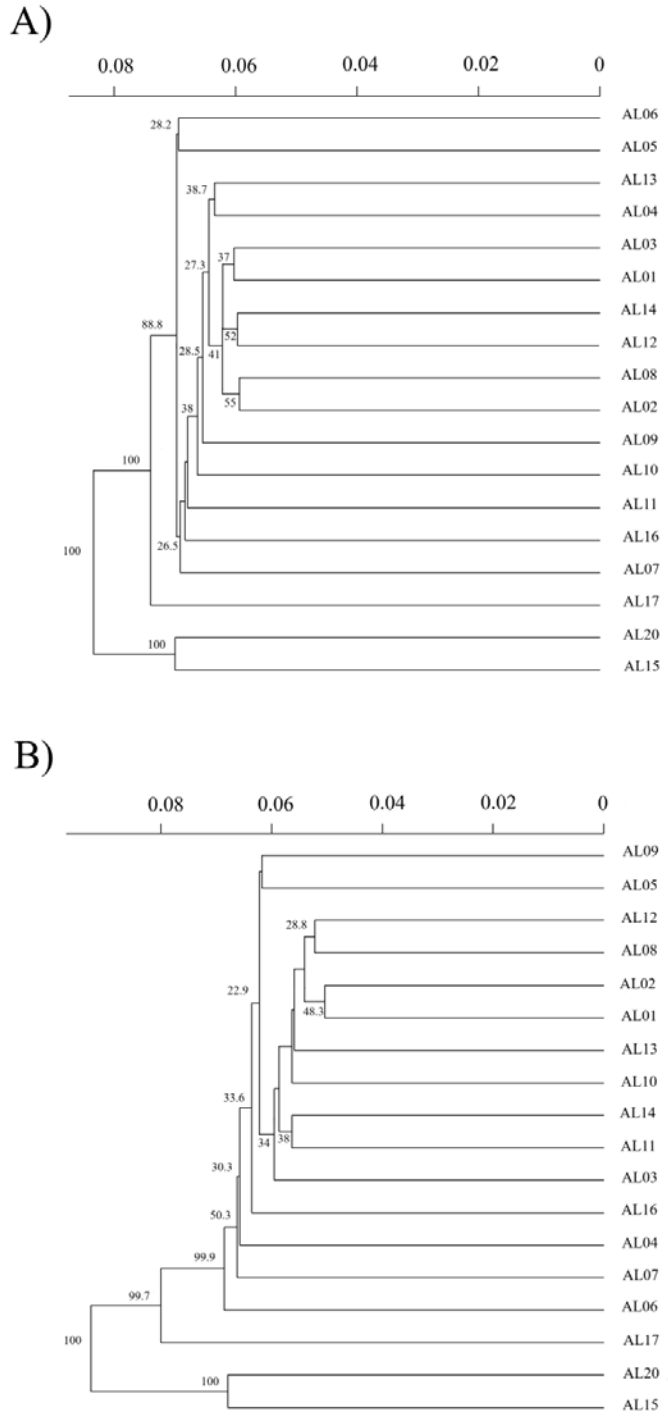

Figure 2. Dendrograms of Provesti's genetic distance for GATK UnifiedGenotyper SNPs filtered using (A) relaxed and (B) stringent settings

Appendix K: Cold Tolerance ANGSD Gentoypes

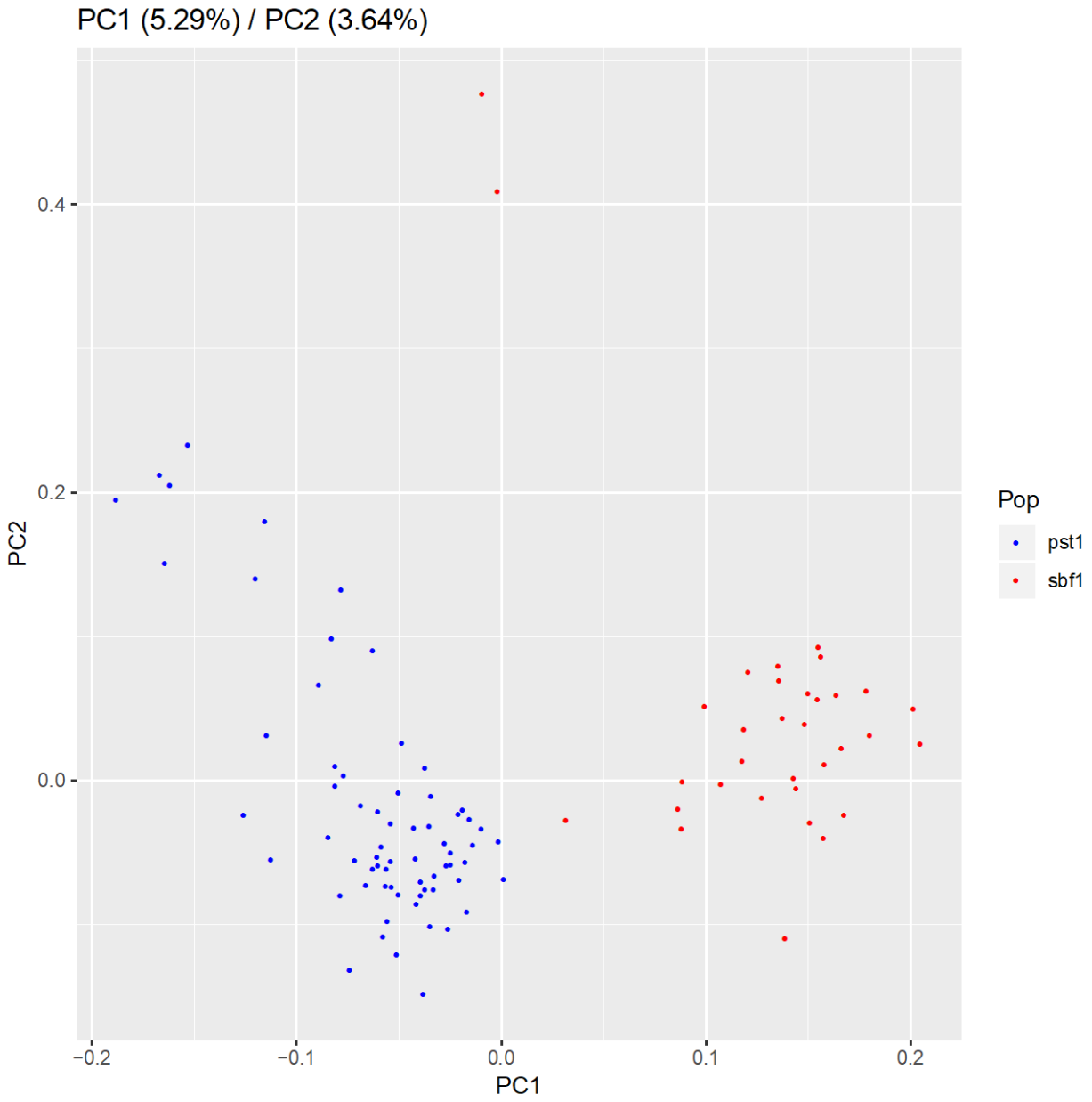

Figure 1. PCA output for SNPs genotyped using ANGSD with a minimum number of individuals of 70 out of 100 (stringent filter). There is clear separation along first two PC's according to the restriction enzyme used to process the samples (Pst1 versus Sbf1)

Appendix K Cold Tolerance ANGSD Gentoypes

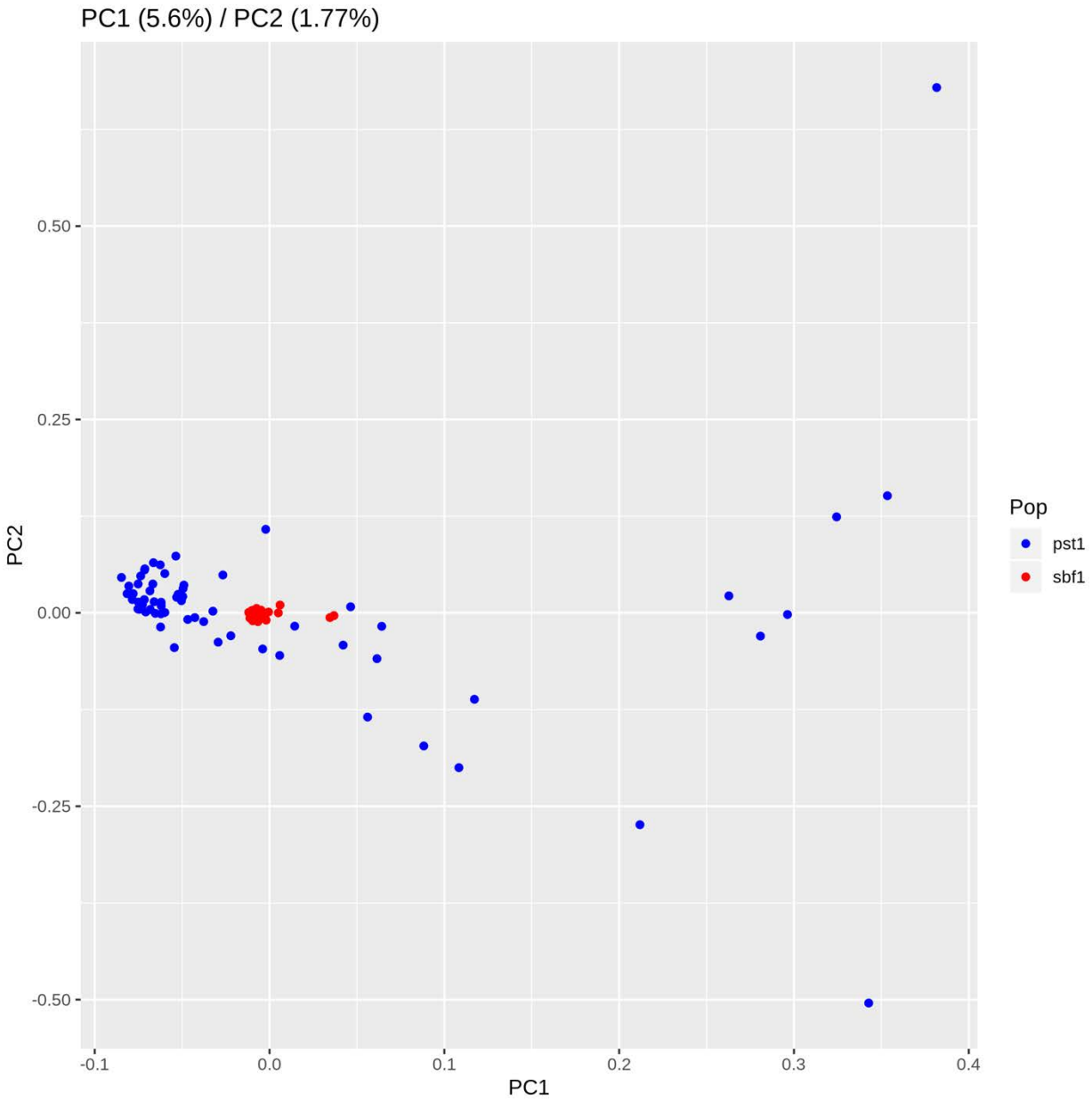

Figure 2. PCA output for SNPs genotyped using ANGSD with a minimum number of individuals of 50 out of 100 (standard filter). Individuals processed with Sbf1 (red) are tightly clustered relative to individuals processed with the Pst1 restriction enzyme (blue)

Appendix K Cold Tolerance ANGSD Gentoypes

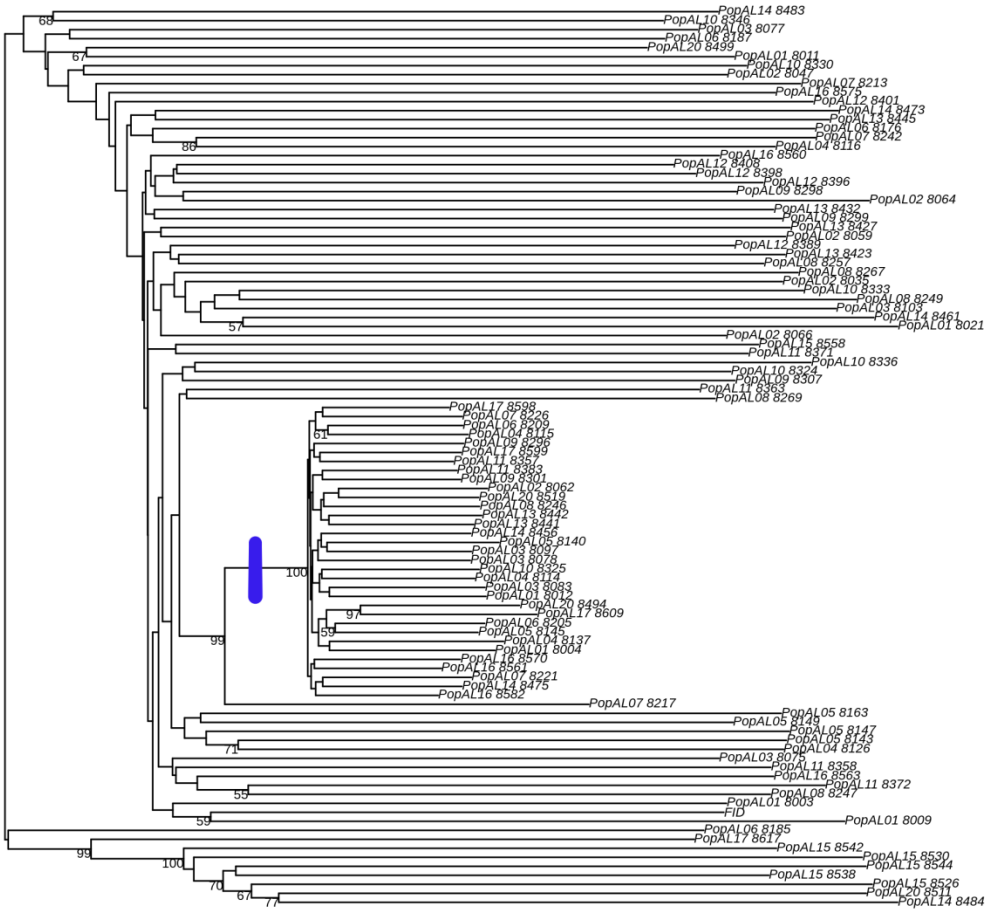

Figure 3. Dendrogram for SNPs genotyped with ANGSD using the standard filter (min_ind = 50). Branching by the restriction enzyme used to process samples is visible with Sbf1 samples denoted in blue.

Appendix K Cold Tolerance ANGSD Gentoypes

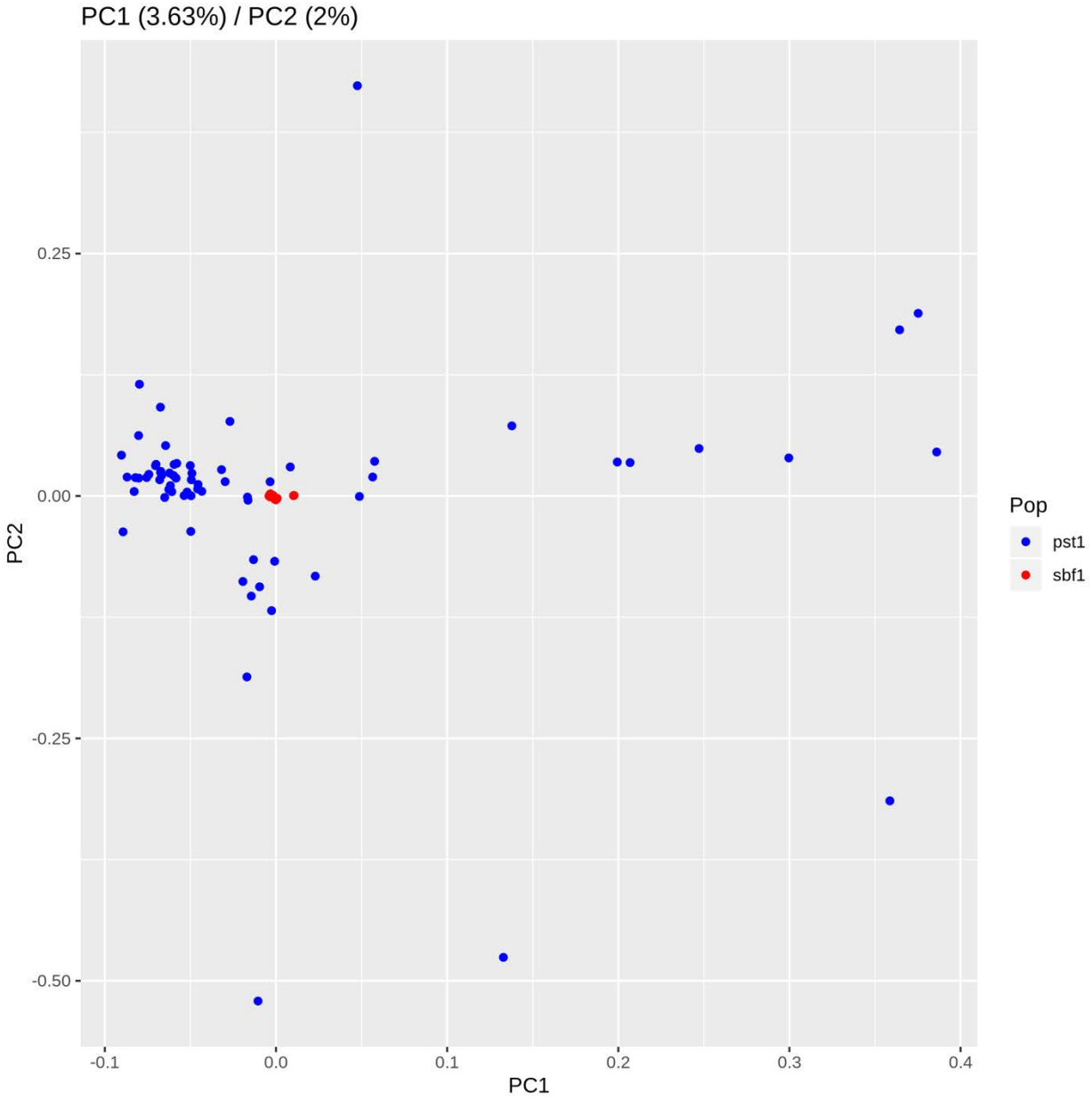

Figure 4. PCA output for SNPs genotyped using ANGSD with a minimum number of individuals of 30 out of 100 (relaxed filter). Individuals processed with Sbf1 (red) are tightly clustered relative to individuals processed with the Pst1 restriction enzyme (blue)

Appendix K Cold Tolerance ANGSD Gentoypes

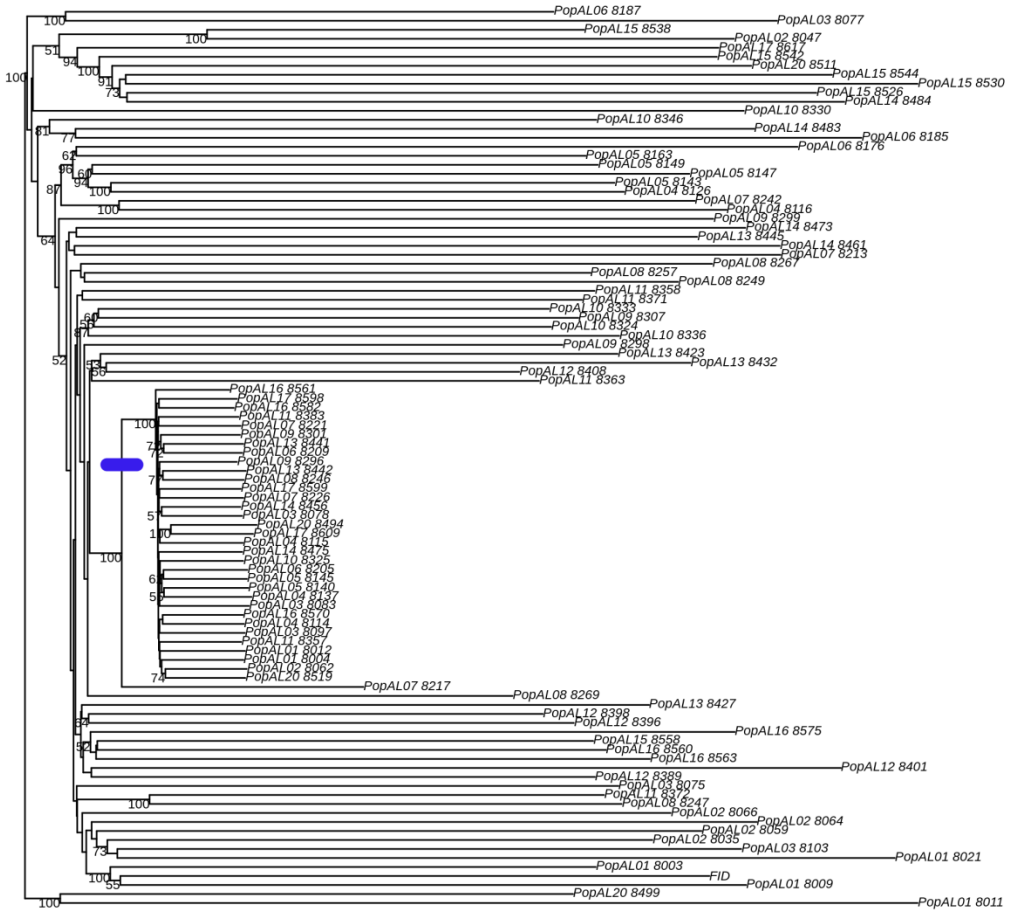

Figure 5. Dendrogram for SNPs genotyped with ANGSD using the relaxed filter (min_ind = 30). Branching by the restriction enzyme used to process samples is visible with Sbf1 samples denoted in blue.

Appendix K Cold Tolerance ANGSD Gentoypes

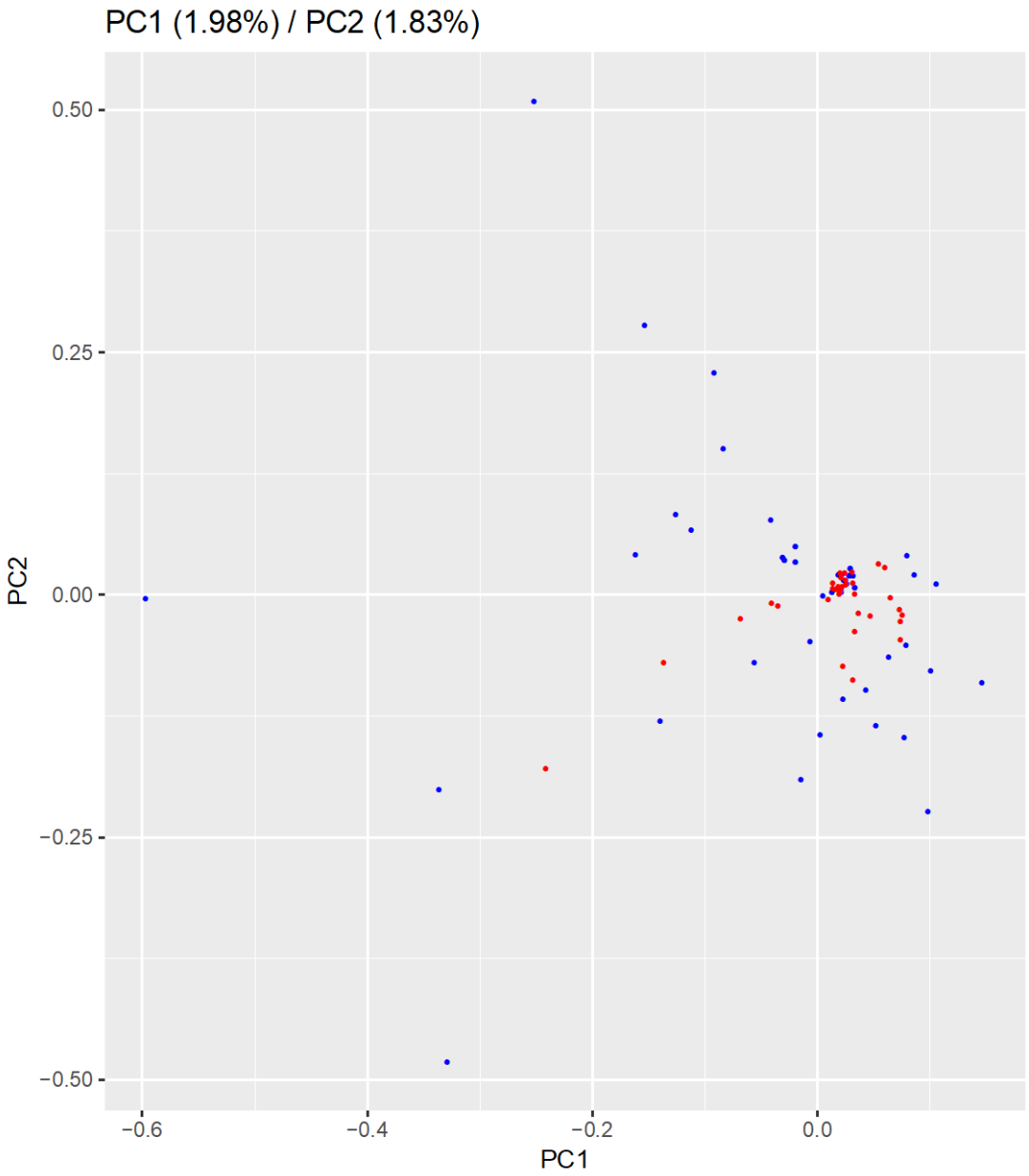

Figure 6. PCA output for SNPs genotyped using ANGSD with a minimum number of individuals of 30 out of 100 (relaxed filter). In this figure, SNPs within 125 nucleotides were not filtered. Individuals processed with Sbf1 (red) are tightly clustered relative to individuals processed with the Pst1 restriction enzyme (blue)

Appendix K Cold Tolerance ANGSD Gentoypes

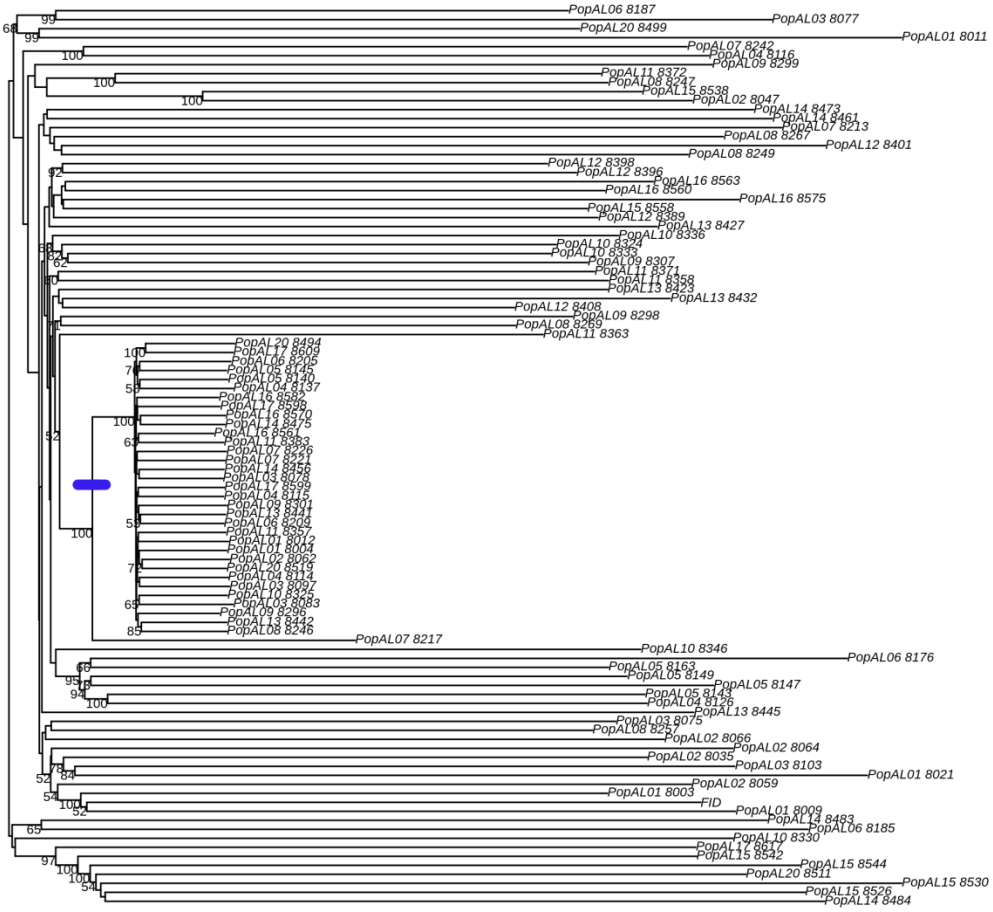

Figure 7. Dendrogram for SNPs genotyped with ANGSD using the relaxed filter (min_ind = 30) without filtering SNPs within 125 nucleotides. Branching by the restriction enzyme used to process samples is visible with Sbf1 samples denoted in blue.

Appendix L: Proportion Variation Explained in randomForest

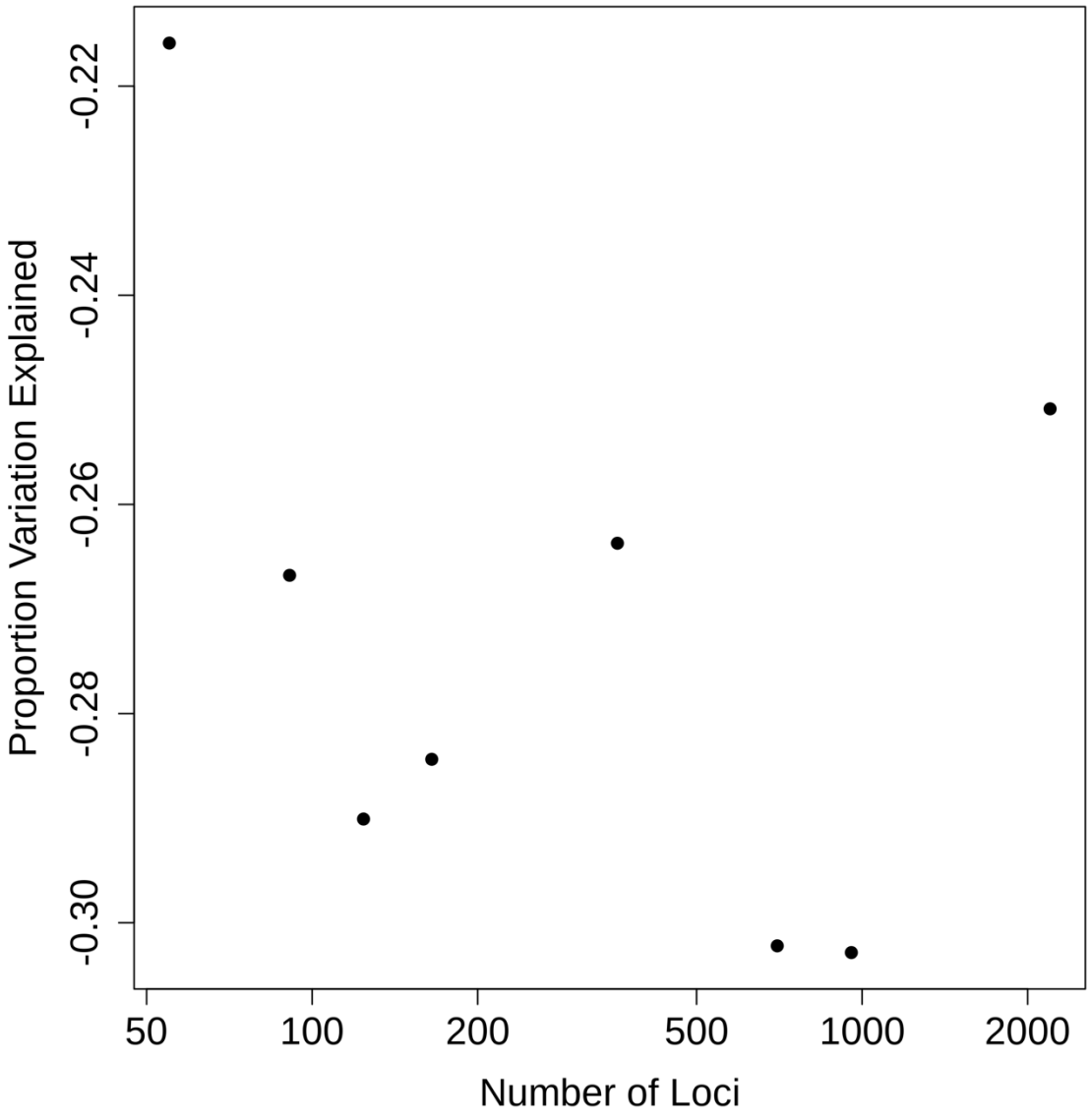

Figure 1. Spring cold injury is very poorly predicted by genotypic variation in top-ranked subsets (top 2%, 3%, 4%, 5%, 10%, 20%, 30%) of 2,197 SNPs generated using ANGSD. Rankings reflect importance values approximated during random forest analysis with all loci considered at each node.

Appendix L: Proportion Variation Explained in randomForest

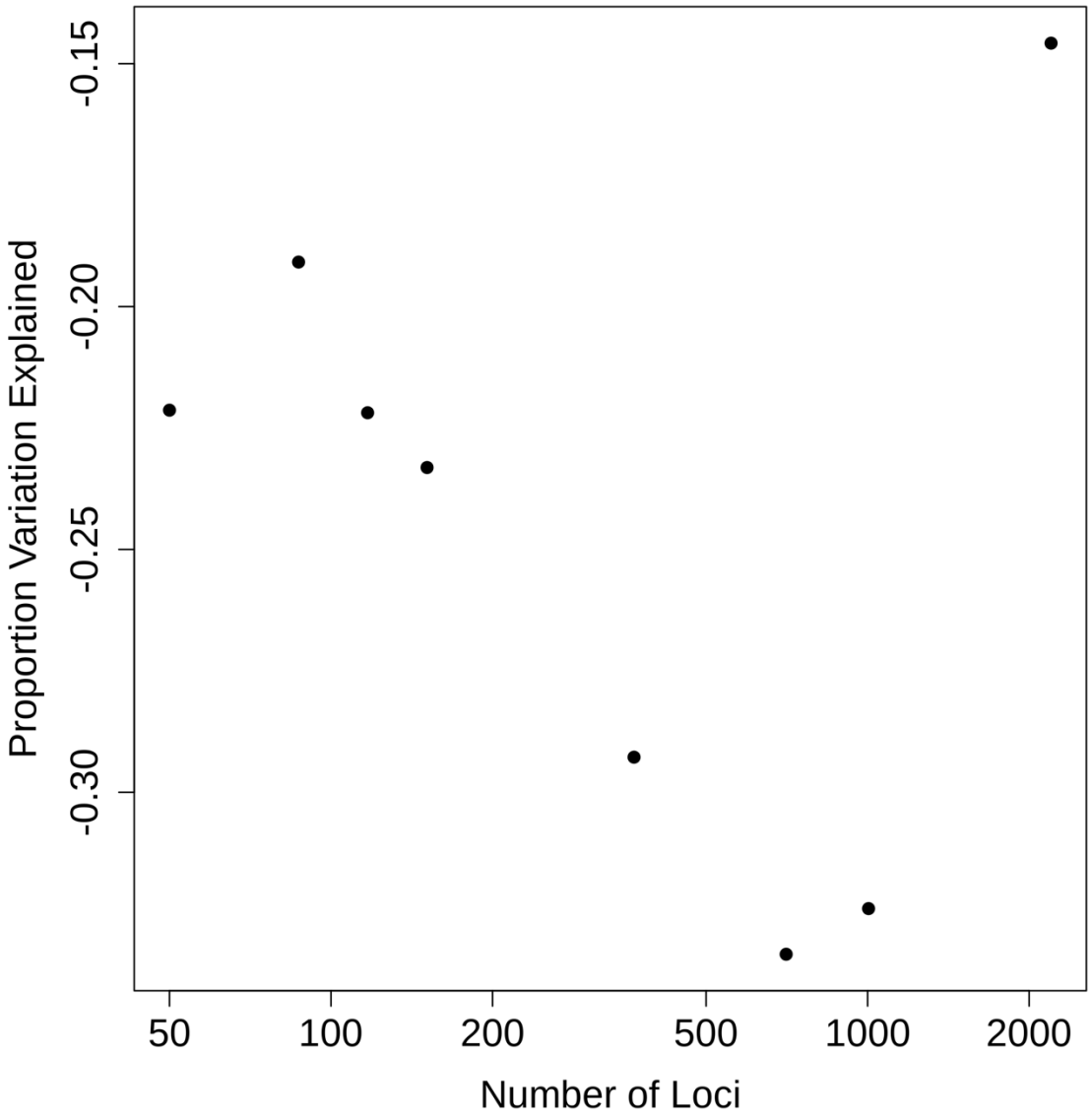

Figure 2. Autumn cold injury is very poorly predicted by genotypic variation in top-ranked subsets (top 2%, 3%, 4%, 5%, 10%, 20%, 30%) of 2,197 SNPs generated using ANGSD. Rankings reflect importance values approximated during random forest analysis with all loci considered at each node.

Appendix M: randomForest Cross-Validation Output

Random forest cross-validation output:

```
> fm <-randomForest(x=dat2, y = dat$Resid, importance=TRUE ,proximity=TRUE,
mtry=length(dat1$loc), ntree=200000)

> rf.crossValidation(fm, dat2, n=100, bootstrap=TRUE)
Bootstrap sampling is being applied, p=0.1 argument is ignored
running: regression cross-validation with 100 iterations
Fit MSE = 0.002481332
Fit percent variance explained = -10.43
Median permuted MSE = 0.001117015
Median permuted percent variance explained = 49.64
Median cross-validation RMSE = 0.04850987
Median cross-validation MBE = 0.001675272
Median cross-validation MAE = 0.0388751
RMSE cross-validation error variance = 2.019821e-05
MBE cross-validation error variance = 0.0001132603
MAE cross-validation error variance = 1.539646e-05
```

Appendix N: Reference for SNP Predictors of Winter Cold Tolerance

Table 1. Reference sequence from Siberian larch draft genome v. 1.0 for SNPs (red) identified as predictors of winter cold injury via random forest analysis for 100 subalpine larch trees representing 18 populations from the northern portion of the range.

SNP	Sequence	Annotation
Pseudo_151_6087320	GAAAACCTTTTAGAAAAGATGTCAGAACCTGCCCCTGCAGGTTTAT ATCCGGGCTTTAACTTAGCCTGGCCGTCCATGTATTATTAGGATTTA ACACAGCC	*L. sibirica* annotation mRNA
Pseudo_225_5038052	CGGTCCGCCGCTAAGGACGCTTCTACAGACTACAATTCAGGCAGCG CTGCCACCCGATTTTCATGCTGGGCTCTTCCCGGTTCGCTCGCCGTT ACTAGGGG	BLASTn 5.8S ribosomal RNA
Pseudo_247_15753552	AAGAATACAAACTATGACCTCGAAAAATAATGGCAGGCACAACAA CAGGGTTGAGGGAATAATATAATACAAAGTATCTAACATCCCTTCA CATAAGCTTA	
Pseudo_271_24459634	GTATCATCTTATCTATGTAATAAATTTTGATGATATTAAATGAGAA AACTAGAATTTTGACTTCCATGTTATTATTTCTATATATACAATCTA TAAACATA	
Pseudo_349_4095395	TTATGGAGAGCACAGATTGAGCGAAAAGCACCAGTTGCTGTTAGG AATTCAGCAGAAACCTCAAATAGAGCAGCATGACAAGGTTCAAGA GATTGACAACA	*L. sibirica* annotation mRNA
Pseudo_405_917624	TATATTCCCTGCAGGGAAGGGGCTAATAAATGGGAAATCATTGACT TCAGTTGCATACTTTGGGAAAGCACGAACAAGCCAACTACACACA AACTCTGTCA	*L. sibirica* annotation mRNA
Pseudo_709_6262212	TAGAGTGATTGATGAGTCGCGAATCTATAGGATGGTGTCTATTGCC AATGGACGTGTCCGTTTTCAAGATTGATTCTACTACTTGAAAGATA ACATTCCCT	
Pseudo_734_3455909	AGTAACCACAATTATTATTCCAGGGCACACACACACAGTTTCCTAA CAAAATTCTAACCTACATGATTCCCAAATCGGGTGTTCCATAGTTA AGTAACCAC	

Appendix O: Reference for Overlapping BayeScan F_{ST} Outlier SNPs

Table 1. Reference sequence from Siberian larch draft genome v. 1.0 for SNPs (red) identified as BayeScan F_{ST} outliers in both the range-wide Pst1 dataset genotyped with ANGSD and the range-wide Sbf1 dataset genotyped with GATK UnifiedGenotyper

SNP	Sequence	Annotation
Pseudo_222_20364504	TCTTGTTCTAAGGTTTTAATTTGCCAATCCTTCCAAGCGAATTGGATCAGCTCAC ATTTTTGGATATTGTTGGATTCATCCTGCAGGATATTGTTGGATTC	
Pseudo_248_20818947	ATCCATGGCATTCACGTTCATGATTGTCAGCCTGCAGGAAATGTTACGGTGAGG GACTCTCCTACGCATTACGGGTGGAGACCTATATGCGATGGAGATGG	BLASTn mRNA from *Picea sitchensis* and *Picea glauca*
Pseudo_272_3301043	GTAGAATTGGCAGGTGGCTGCCCCTGCAGGCCAAAGATTATCTCTGTCATTGCC TTGACAATCGTTGGATCCTCGCTGTAGCTCTCGCAGCAGCGCCCCCA	BLASTn mRNA from *Picea sitchensis* and *Picea glauca*
Pseudo_334_391595	AAATAACAAAACCATACTTTCAAAAATATGATGTGAGAAATCAACTAAAACGTT ATCCACCAAAACCATGCCTCTAAAATAAAAAATTTACAAAACAATGT	
Pseudo_446_532154	GTCTTGCTGCGTTTGACGTTCGATAATGTCTGTATGATTGCCTTCGGCGTGGATC CCGGCTGTCTCAGACTTGGACTCCCAGAGATACCCTTCGCCAAGGC	*L. sibirica* annotation mRNA
Pseudo_638_8142505	GTATAGTTTATCACGTCTAATATTCCTGCAGGATAATCCCTGTAAAGAGATACTC CTTCCTGTAAAATCACGACGGAAATTTCAATCAAATTTATGTTGTT	
Pseudo_705_2951648	CCAGCATCGACATCGACATCATCAGGAGGCATGATCTTGGTGAAGCCCGGCTAG GGTTCTCGGCAGAAGCCTGCGGAGAAGAAGAAGATGAACCGGCATCA	BLASTn hypothetical proteins in *Pinus taeda* and *Pinus lambertiana*
Pseudo_820_4506024	CACAATATGCAGCTTAATGACTGCAGCTATATACATCCTGCAGGTTTTTACTGTG CAGTGGCTTTGCAACTGTAGTTGAACTATATTCACAGCCTCCAGGC	

Appendix P: bayenv2 Neutral Covariance Matrices

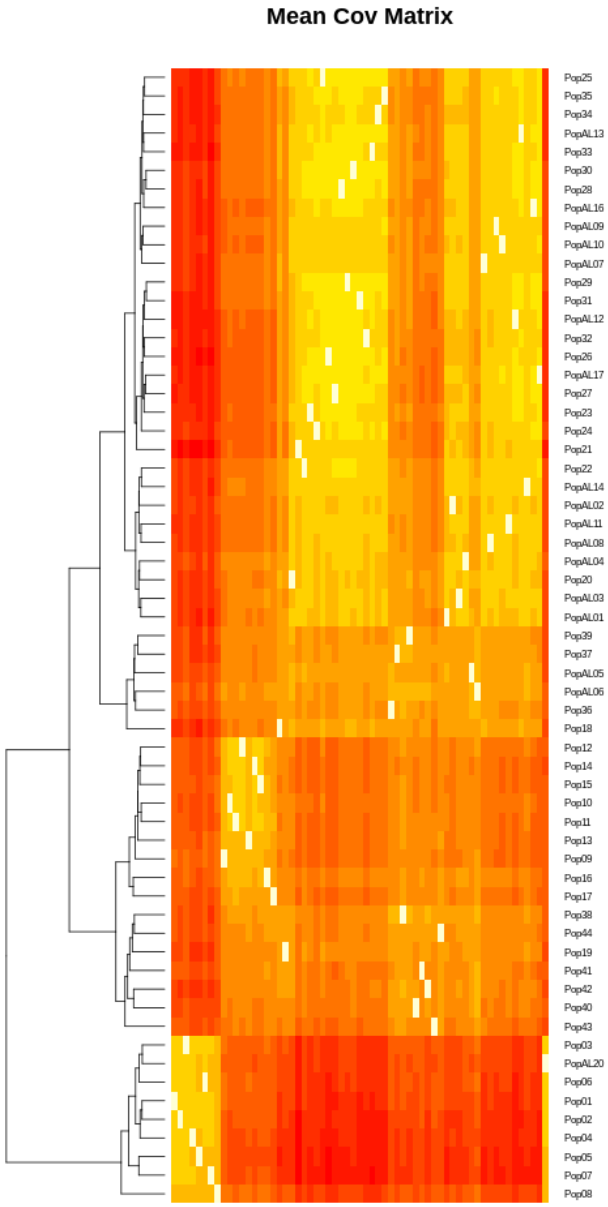

Figure 1. The mean covariance matrix for the range-wide Sbf1 GATK UnifiedGenotyper dataset represents the neutral genetic structure of subalpine larch. The first split separates the Cascade Range from the Rocky Mountains. The second split separates the southern and central Rockies from the Northern Rockies, as in Chapter 2, where this data was previously analyzed. Within the northern Rockies, an additional “southern” cluster is detectable.

Appendix P: bayenv2 Neutral Covariance Matrices

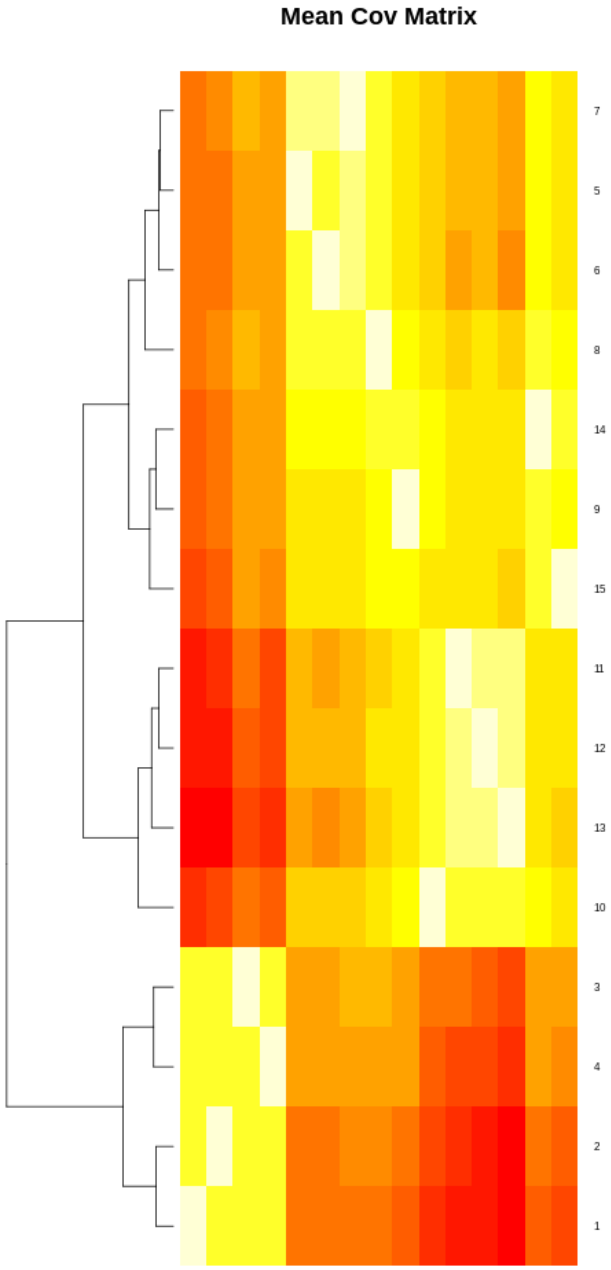

Figure 2. The mean covariance matrix for the range-wide Pst1 ANGSD dataset accurately identifies the neutral structure of subalpine larch. The first split separates the Cascade Range from the Rocky Mountains. The second split separates the southern Rockies from the central and southern Rockies from the northern Rockies.

Appendix Q: Reference for Overlapping bayenv2 SNPs

Table 1. Reference sequence from Siberian larch draft genome v. 1.0 for SNPs (red) correlated with environmental gradients as per bayenv2 analyses for both the range-wide Pst1 ANGSD dataset and Sbf1 GATK UnifiedGenotyper dataset

SNP	Sequence	Annotation
Pseudo_47_1721469	CTTTGGCTACGCTGTATTATGGCGAGTATATGAACTCGGGACCT GGGGCAGCAACTGCTGAGCGAGTTCAATGGCCTGGGTATCACGC CATGACCATTTCC	
Pseudo_112_5519243	TTGCCAGTGCGCTGTTTCTGCCTCTCTTCATTGGTGAACATCCTC GCGGTGGGATCAATATACGACACAGCGAAGTTGAGCTATCGAA GGGACTCCAATCT	
Pseudo_198_22285064	GGCCTGCAGGTAAAATTAAGACATTGTCTTTGACCAATTTTGAT ACAAAGAGGTTACAAATAATTTTGGAGAATGGTATTCCAATTGT CAGTAATCAAGTG	BLASTn nearby match to *Picea sitchensis* and *Picea glauca* mRNA
Pseudo_236_2642487	GTTAAGCAACGCAGTCATGACCTGGCAAGGTAGTTGAGAAGTCA GTTTGGCGCATATGGAAAATCTGTTACTTCATTGAGAATGTTTCC CTGCAGGTTTTC	
Pseudo_252_21191933	ACCGCGACTTGCAGATGGCTGCATTGTATCCCGCGAATCTCAGA GTGTTCATTACAAATCGGCGGGAACCACTCTTTTGCTCCGCATTG TCGGGAGCATTA	*L. sibirica* annotation mRNA
Pseudo_254_22353006	AGAACCTGCAGGAATAAATTTATAGCAAAATTCAAAATCAGATG TTTGTACGCTCATTAGTAAGCAACGAAACTGTAAAACAGCTCAT ATCACATGCAATA	*L. sibirica* annotation mRNA
Pseudo_259_19308369	AACCCAAAACGAAGTAGTCTCGACTTATTCCCAAGATTAAATCA ACATTAATAGAAACCCAATTTATACAAGACACTGCACCGAATCC CTAAAAAATTCAC	
Pseudo_297_6014296	TCTGTAGCCATGTTTAATCTTGTCCTTTGTTGTTTAAGACCCTCCT GAATGGCAGCTGCACTACCTGTGCTATGCTCCCTGTTTACGAAA GGGGTTGTCTC	*L. sibirica* annotation mRNA

SNP	Sequence	Annotation
Pseudo_330_14254778	AAATACAAGAAAAGCACGTAAAGAAAATAGGAAAACAGGTAAA ATAATAACAGACAGGAGCGTGATTTTGAAACAAATCAACCGGCC AAGGCCGTTCTGTA	
Pseudo_351_6175164	TTACTATTGCTACAGGCTACTGTGATCGACATGGTTTGGTTGGTG TCAGCCGGTGTGACTTGCTACAAGCCACAGAGCTATGGTACTAT CCCTAATCCAAA	
Pseudo_369_4132364	TAAAGAATTTGGGAACTTGGGAGTTCTACCCTTGAGAGGGTAAT GGAAGCTTGATGGAGCACCTGCAGGGATTTCAATCCTTTTGAGG AGAGAAATAACCA	
Pseudo_377_15881506	CCCACGACACCGAAAAGCAGGCCAAAACAAATGATCAGAACAT CTGCAATCAGCTGAGCCAAAGAACAAGTATCCCTGCAGGCGGTG AAGTGAAATAGAGC	BLASTn match to mRNA from *Picea glauca*
Pseudo_404_2501874	GTTCAACATTCCTGCAGGCATGTGATAGACTATGCAGTTCCTTCT ACTGTCACAAAAGATAACGTCTATTTGATTGGAAGCGCTCCCAG TTCCTCGCTAAC	BLASTn nearby match to mRNA from *Picea glauca*
Pseudo_408_7240070	CAGTGGCACGTTAGCTCGTTTAACTTGTTTTTCCATAAAATTGTC TATGTGGCTTCTCTATCTTAGAGTTCTTTCTTAATGAAATGCCCA TATTTAGTGAA	
Pseudo_508_9553361	ACAGAGCATAATTACTTGCATAAAAAAAAAGGCGAAAATACCT CAGAAACGGAGAACAAGTTTTCCCGGCGAGTAAGAAAGACGGT GTTGAAGAAAAACAA	*L. sibirica* annotation mRNA
Pseudo_530_3964246	CTTTACGGACATTAGTGCCCGTAAAGCCCAAAGTAAGTTTTTTG GTAACAGTTTTCATCTTTCCCTTGTTGAAATCTTGTTCCTGTGCCT GCAGGTTTGTG	*L. sibirica* annotation mRNA
Pseudo_561_10943560	TTGCAATGAGACAGGACCTATGCGAGCAAGTATATTCCTATTAT TGCAAGCGTTAGTGAACATCAGCCACCTACCTGGCCTTCATATTT TATGGATATAAA	
Pseudo_596_1954353	CTAGCTTATTGACGAGTGACGCAGGTACGACTTCATTAACTCTA AATTCCATGAATCCCAACTTAAACTTGGGCTTGCCGTCACTTGTT	

SNP	Sequence	Annotation
	GAGAATATCTCT	
Pseudo_602_8643462	AATCAAATTATCCTGCAGGTTTACCTTTTTCAAAGACCTTTCTCA TTTATTTGTGGACTGTTCTAATCCAGGATATAAAATTTTGTTCCA CTATTGAGTAA	
Pseudo_603_10102834	TGCTAGTTAAATTGAGCGCTAAAGATTAACCTGTATAGCAATAA GCATATCCAATTGCGTACCCTTATGCACCTTGGGTTAAAGTGTTC CTTTCAGTGCAG	
Pseudo_655_5097907	GTTCTAAGCTACTTTATACTGAAACAAAAGCAGCATCTATTATG ATCCCTCAGAAGTACCCCCCTAATCATCCTAGCCTGCAGGTTTCC AGGTGCCATAGA	*L. sibirica* annotation mRNA
Pseudo_678_759073	TGCCCCCCCCTCTCCCTGGAGTTTCATACATCGTAGGACACAGGC TAAAAGCACCTGTAATAGTAAAGGCATGAATCTAAAGACTGGAC CCTGCAGGACAT	
Pseudo_680_3876454	CAGGGACCAATTCGTTTTAAGCAAACATATGGCATCAAATGATG GGGCTGACCTGCAGGATATACTTTTGAATGTTCTCCCTTAGGCTA CACTATTGCAGT	
Pseudo_696_4766119	CAGGTGGTTCGATGATCGATTTTCAGTGCTACTGCGTGGAATGA GCAATCGCGTGGAACGAACAATGAACAACGGCGGCACGGGGAG TCAGGACTTTGTTC	
Pseudo_707_5269576	CTGCAGGCTTTTGGAACTATGCGCATCCGTCAGAACCCAATACG AAAGTGGCCATACATAACGATATGGATGGATATTGACCAAGGCA GTGAGCATTAATG	
Pseudo_741_262116	AAGCCCTGAGGGAGCAAACAATTTGATGTTTTCAGTTGCGGCTT TGAAATAGCCCATTGCTCACAGTGTCACACGTTCCAGCTCATGA TCCAATCCATTGC	
Pseudo_817_2066018	TCAGAATGCAGGTGTACTCTCCTGCAGGAAAGGAGCTCACTGCT AACCTCAACTTCTAACCCGCCTGCGGACCGACCAAGTTCGAGTT TAGCTGACTCGCG	

SNP	Sequence	Annotation
Pseudo_819_1710072	GTGACAAAAATCACAATGTAGCAAAATGCTGTAACATACCCCAG AGGCATATTGAAAATCATGGTTGGCAGTCCATTGTTGAGAGTGA TAGTCAATTAGAT	
Pseudo_820_4506024	CACAATATGCAGCTTAATGACTGCAGCTATATACATCCTGCAGG TTTTTACTGTGCAGTGGCTTTGCAACTGTAGTTGAACTATATTCA CAGCCTCCAGGC	
Pseudo_940_3109473	ACCTGCTTTAAGAAATATATGATCGTGAAAATATATCACAACTA AAGGTTTAAAAACAACCCTGCAGGAAATATGTCAAAAGTGCCA GAAAGTTCTTTCCA	*L. sibirica* annotation mRNA

Appendix R: Overlapping Climatic Relationships for Sbf1 SNPs

Table 1. Overlap between range-wide Sbf1 GATK Unified Genotyper SNPs with bayenv2 Bayes factors >= 3.0 associated with different climate variables

	AHM	bFFP	CMD	DD_0	DD5	eFFP	EMT	Eref	EXT	FFP	MAP	MAT	MCMT	MSP	MWMT	NFFD	PAS	SHM	TD	Tmin_at
AHM	63																			
bFFP	9	59																		
CMD	13	3	40																	
DD_0	9	14	7	56																
DD5	10	19	8	39	68															
eFFP	4	27	3	31	26	71														
EMT	10	12	8	47	33	34	57													
Eref	11	4	11	10	18	4	6	54												
EXT	11	8	11	26	36	11	22	34	61											
FFP	6	39	3	24	26	48	24	6	11	64										
MAP	37	10	10	0	2	2	0	5	3	5	57									
MAT	9	17	8	47	51	29	39	16	35	26	1	59								
MCMT	9	10	8	41	30	27	45	8	21	20	1	35	49							
MSP	15	4	12	15	12	9	15	8	9	5	10	13	16	34						
MWMT	10	18	8	36	62	24	30	20	40	24	2	47	26	11	73					
NFFD	5	33	5	32	40	45	29	4	18	47	4	36	25	9	36	74				
PAS	33	11	4	0	3	3	0	8	3	8	44	1	0	6	4	3	61			
SHM	18	4	20	17	15	10	18	10	14	8	11	15	18	22	13	10	6	44		
TD	7	3	4	8	6	8	11	4	2	8	4	7	11	8	7	6	5	5	42	
Tmin_at	7	24	9	44	37	52	45	4	18	40	3	41	37	13	35	51	2	15	9	73

Appendix S: bayenv2 Output for SNPs Predictors of Winter Cold Tolerance

Table 1. Bayes factors and Spearman correlation coefficients for a subset of the eight SNPs identified by random forest analysis as predictors of winter cold injury in 100 subalpine larch trees representing 18 populations from the northern portion of the range

SNP	Range-wide dataset used	Climate Variable	Bayes Factor	Spearman correlation
Pseudo_151_6087320	.	.	.	.
Pseudo_225_5038052	.	.	.	.
Pseudo_247_15753552	Sbf1	Tmin_at	5.382	-0.229
		MAT	3.993	-0.229
		MWMT	3.420	-0.219
		MCMT	3.428	-0.219
		MSP	16.65	0.286
		DD_0	3.813	0.213
		DD5	3.600	-0.225
		NFFD	5.893	-0.226
		eFFP	3.120	-0.214
		EMT	4.888	-0.224
Pseudo_271_24459634	Pst1	MSP	8.410	0.427
Pseudo_349_4095395	.	.	.	.
Pseudo_405_917624	Sbf1	TD	5.336	-0.270
		MSP	3.779	-0.208
Pseudo_709_6262212	.	.	.	.
Pseudo_734_3455909`	Sbf1	AHM	3.968	-0.165

Appendix T: bayenv2 Output for SNPs Identified as F_{ST} Outliers

Table 1. Output from Pst1 bayenv2 analysis (Bayes Factors and Spearman correlation coefficients) for SNPs identified as BayeScan F_{ST} outliers in both the range-wide Pst1 and Sbf1 datasets

SNP	Climate Variable	Bayes Factor	Spearman Correlation
Scaffold_446_532154	Tmin_at	3.048343	0.142
Scaffold_446_532154	MAT	3.309500	0.396
Scaffold_446_532154	MAP	5.120333	0.385
Scaffold_446_532154	AHM	4.266	-0.323
Scaffold_446_532154	PAS	6.253233	0.357
Scaffold_638_8142505	MCMT	9.675	-0.138
Scaffold_638_8142505	TD	7304.5	-0.115
Scaffold_638_8142505	MSP	203.32	-0.147
Scaffold_638_8142505	SHM	1939.1	0.243
Scaffold_638_8142505	bFFP	12.841	0.402
Scaffold_638_8142505	CMD	128.69	0.219
Scaffold_820_4506024	Eref	15.425	-0.459
Scaffold_820_4506024	CMD	8.091	-0.383

Table 2. Output from Sbf1 bayenv2 analysis (Bayes Factors and Spearman correlation coefficients) for SNPs identified as BayeScan F_{ST} outliers in both the range-wide Pst1 and Sbf1 datasets

SNP	Climate Variable	Bayes Factor	Spearman Correlation
Scaffold_820_4506024	Tmin_at	14.832	-0.237
Scaffold_820_4506024	MAT	13.380	-0.231
Scaffold_820_4506024	MWMT	48.187	-0.301
Scaffold_820_4506024	MCMT	5.169	-0.197
Scaffold_820_4506024	DD_0	8.042	0.203
Scaffold_820_4506024	DD5	32.227	-0.284
Scaffold_820_4506024	NFFD	36.707	-0.255
Scaffold_820_4506024	bFFP	6.411	0.185
Scaffold_820_4506024	eFFP	9.946	-0.236
Scaffold_820_4506024	FFP	17.890	-0.233
Scaffold_820_4506024	EMT	5.133	-0.179
Scaffold_820_4506024	EXT	3.258	-0.134

Appendix U: bayenv2 Output for Overlapping bayenv2 SNPs

Table 1. Output from bayenv2 analysis (Bayes Factors and Spearman correlation coefficients) for range-wide Pst1 dataset for SNPs identified as overlapping between Pst1 and Sbf1 datasets

SNP	Climate Variable	Bayes Factor	Spearman Correlation
Scaffold_112_5519243	MCMT	5.576	0.196
Scaffold_198_22285064	MWMT	4.698	-0.349
Scaffold_198_22285064	AHM	4.468	-0.274
Scaffold_236_2642487	TD	3.824	0.380
Scaffold_252_21191933	TD	19.885	0.012
Scaffold_252_21191933	SHM	3.285	0.011
Scaffold_254_22353006	MCMT	6.561	-0.514
Scaffold_259_19308369	MCMT	4.069	0.042
Scaffold_259_19308369	eFFP	3.719	0.177
Scaffold_297_6014296	MSP	4.096	0.365
Scaffold_297_6014296	SHM	4.196	-0.275
Scaffold_297_6014296	NFFD	3.039	0.373
Scaffold_297_6014296	bFFP	12.538	-0.449
Scaffold_297_6014296	FFP	8.616	0.314
Scaffold_297_6014296	CMD	5.015	-0.305
Scaffold_330_14254778	Eref	5.969	-0.369
Scaffold_351_6175164	SHM	124.137	0.249
Scaffold_351_6175164	CMD	9.619	0.168
Scaffold_369_4132364	MWMT	7.105	-0.359
Scaffold_369_4132364	EXT	3.666	-0.325
Scaffold_369_4132364	Eref	4.192	-0.303
Scaffold_377_15881506	Tmin_at	3.229	0.169
Scaffold_377_15881506	MCMT	4.390	0.143
Scaffold_377_15881506	SHM	3.410	0.145
Scaffold_377_15881506	EMT	6.131	0.199
Scaffold_404_2501874	Eref	6.597	0.316
Scaffold_408_7240070	bFFP	3.645	0.431
Scaffold_47_1721469	bFFP	5.598	-0.406
Scaffold_508_9553361	MCMT	14.277	-0.236
Scaffold_508_9553361	TD	3.353	0.002
Scaffold_508_9553361	MSP	110.212	0.389
Scaffold_508_9553361	SHM	9.397	-0.415
Scaffold_508_9553361	bFFP	14.591	-0.318
Scaffold_508_9553361	Eref	4.002	-0.388

SNP	Climate Variable	Bayes Factor	Spearman Correlation
Scaffold_508_9553361	CMD	45.710	-0.475
Scaffold_530_3964246	DD5	39.125	-0.240
Scaffold_530_3964246	EXT	16.264	-0.307
Scaffold_561_10943560	MWMT	3.061	0.304
Scaffold_561_10943560	Eref	4.393	0.333
Scaffold_596_1954353	TD	4.383	0.461
Scaffold_602_8643462	MAP	8.428	0.490
Scaffold_602_8643462	AHM	19.481	-0.515
Scaffold_602_8643462	PAS	3.653	0.365
Scaffold_603_10102834	bFFP	3.888	-0.441
Scaffold_655_5097907	TD	8.225	-0.083
Scaffold_678_759073	Tmin_at	683.263	-0.362
Scaffold_678_759073	MAT	1432.803	-0.265
Scaffold_678_759073	MWMT	1206.133	-0.387
Scaffold_678_759073	MCMT	3830.743	-0.468
Scaffold_678_759073	MSP	33.896	0.259
Scaffold_678_759073	DD_0	2660.290	0.370
Scaffold_678_759073	DD5	517.160	-0.179
Scaffold_678_759073	NFFD	15.731	-0.270
Scaffold_678_759073	eFFP	44.165	-0.441
Scaffold_678_759073	FFP	9.368	-0.348
Scaffold_678_759073	EMT	8905.333	-0.443
Scaffold_678_759073	EXT	186.750	-0.464
Scaffold_678_759073	Eref	15.125	-0.383
Scaffold_680_3876454	bFFP	7.923	-0.387
Scaffold_696_4766119	MCMT	4.985	-0.063
Scaffold_696_4766119	TD	41.015	0.034
Scaffold_696_4766119	MSP	3.736	-0.142
Scaffold_696_4766119	SHM	18.859	0.350
Scaffold_696_4766119	bFFP	8.934	0.257
Scaffold_696_4766119	CMD	41.945	0.388
Scaffold_707_5269576	MAT	6.296	0.123
Scaffold_707_5269576	MWMT	5.304	0.289
Scaffold_707_5269576	DD_0	3.522	-0.107
Scaffold_707_5269576	DD5	5.661	0.180
Scaffold_707_5269576	EXT	4.561	0.139
Scaffold_741_262116	MAP	4.589	0.568
Scaffold_741_262116	AHM	3.289	-0.586
Scaffold_817_2066018	SHM	31.582	0.262

SNP	Climate Variable	Bayes Factor	Spearman Correlation
Scaffold_819_1710072	TD	3.632	-0.165
Scaffold_820_4506024	Eref	15.425	-0.459
Scaffold_820_4506024	CMD	8.091	-0.383
Scaffold_940_3109473	NFFD	3.366	0.300
Scaffold_940_3109473	bFFP	4.622	-0.291
Scaffold_940_3109473	FFP	3.768	0.300

Table 2. Output from bayenv2 analysis for range-wide Sbf1 dataset for SNPs identified as overlapping between Pst1 and Sbf1 datasets

SNP	Climate Variable	Bayes Factor	Spearman Correlation
Scaffold_112_5519243	Tmin_at	6.220	0.256
Scaffold_112_5519243	MAT	3.142	0.224
Scaffold_112_5519243	DD5	3.548	0.204
Scaffold_112_5519243	NFFD	8.153	0.239
Scaffold_112_5519243	bFFP	3.405	-0.166
Scaffold_112_5519243	eFFP	3.477	0.177
Scaffold_112_5519243	FFP	4.451	0.187
Scaffold_198_22285064	MAP	6.827	-0.186
Scaffold_198_22285064	MSP	3.245	-0.160
Scaffold_198_22285064	AHM	13.617	0.192
Scaffold_198_22285064	PAS	3.226	-0.162
Scaffold_236_2642487	AHM	4.316	-0.169
Scaffold_252_21191933	bFFP	4.315	-0.155
Scaffold_252_21191933	eFFP	3.171	0.114
Scaffold_252_21191933	FFP	4.907	0.135
Scaffold_254_22353006	MAP	13.893	-0.257
Scaffold_254_22353006	AHM	12.036	0.242
Scaffold_254_22353006	PAS	12.831	-0.265
Scaffold_259_19308369	Eref	4.473	-0.199
Scaffold_259_19308369	CMD	4.098	-0.128
Scaffold_297_6014296	Tmin_at	3.033	0.200
Scaffold_297_6014296	MSP	4.926	-0.244
Scaffold_297_6014296	SHM	3.305	0.233
Scaffold_297_6014296	eFFP	3.656	0.166
Scaffold_297_6014296	EMT	3.213	0.253
Scaffold_330_14254778	MAP	3.375	0.199
Scaffold_330_14254778	AHM	7.756	-0.170
Scaffold_330_14254778	eFFP	3.101	-0.239

SNP	Climate Variable	Bayes Factor	Spearman Correlation
Scaffold_351_6175164	MWMT	3.438	-0.196
Scaffold_351_6175164	DD5	3.306	-0.204
Scaffold_351_6175164	EXT	4.511	-0.226
Scaffold_369_4132364	Tmin_at	4.180	0.195
Scaffold_369_4132364	MCMT	4.034	0.182
Scaffold_369_4132364	eFFP	5.022	0.221
Scaffold_369_4132364	EMT	4.910	0.199
Scaffold_377_15881506	bFFP	4.577	0.178
Scaffold_377_15881506	CMD	4.430	0.204
Scaffold_404_2501874	Tmin_at	5.337	-0.204
Scaffold_404_2501874	NFFD	14.198	-0.242
Scaffold_404_2501874	bFFP	30.243	0.228
Scaffold_404_2501874	eFFP	3.859	-0.144
Scaffold_404_2501874	FFP	17.523	-0.209
Scaffold_408_7240070	SHM	4.108	0.147
Scaffold_47_1721469	bFFP	3.007	0.128
Scaffold_508_9553361	SHM	5.173	0.156
Scaffold_508_9553361	CMD	6.579	0.205
Scaffold_530_3964246	CMD	4.290	0.221
Scaffold_561_10943560	eFFP	4.032	-0.098
Scaffold_561_10943560	FFP	4.228	-0.119
Scaffold_596_1954353	TD	3.146	-0.156
Scaffold_596_1954353	bFFP	3.219	0.141
Scaffold_602_8643462	MAP	5.471	-0.165
Scaffold_602_8643462	PAS	9.844	-0.189
Scaffold_603_10102834	EXT	5.023	0.205
Scaffold_603_10102834	Eref	9.861	0.250
Scaffold_655_5097907	MAP	4.579	-0.186
Scaffold_655_5097907	PAS	4.069	-0.152
Scaffold_678_759073	SHM	5.443	-0.162
Scaffold_678_759073	CMD	4.781	-0.166
Scaffold_680_3876454	Tmin_at	3.345	0.169
Scaffold_680_3876454	MAT	3.276	0.186
Scaffold_680_3876454	MWMT	3.162	0.151
Scaffold_680_3876454	DD5	4.999	0.181
Scaffold_680_3876454	CMD	3.891	0.160
Scaffold_696_4766119	Tmin_at	3.435	0.199
Scaffold_696_4766119	MAT	3.891	0.187
Scaffold_696_4766119	MCMT	5.449	0.258

SNP	Climate Variable	Bayes Factor	Spearman Correlation
Scaffold_696_4766119	DD_0	4.921	-0.210
Scaffold_696_4766119	EMT	5.237	0.251
Scaffold_707_5269576	Tmin_at	43.838	0.298
Scaffold_707_5269576	MAT	5.005	0.184
Scaffold_707_5269576	MWMT	3.164	0.134
Scaffold_707_5269576	MCMT	3.628	0.196
Scaffold_707_5269576	DD_0	4.441	-0.191
Scaffold_707_5269576	DD5	4.880	0.168
Scaffold_707_5269576	NFFD	88.452	0.288
Scaffold_707_5269576	bFFP	35.825	-0.217
Scaffold_707_5269576	eFFP	41.874	0.264
Scaffold_707_5269576	FFP	109.012	0.277
Scaffold_707_5269576	EMT	7.460	0.236
Scaffold_741_262116	NFFD	3.059	-0.165
Scaffold_817_2066018	NFFD	3.193	-0.190
Scaffold_817_2066018	bFFP	4.139	0.196
Scaffold_819_1710072	MAT	4.260	-0.255
Scaffold_819_1710072	MWMT	5.442	-0.260
Scaffold_819_1710072	DD5	6.748	-0.282
Scaffold_819_1710072	NFFD	3.304	-0.228
Scaffold_819_1710072	bFFP	3.402	0.223
Scaffold_819_1710072	EXT	7.133	-0.281
Scaffold_820_4506024	Tmin_at	14.832	-0.237
Scaffold_820_4506024	MAT	13.380	-0.231
Scaffold_820_4506024	MWMT	48.187	-0.301
Scaffold_820_4506024	MCMT	5.169	-0.197
Scaffold_820_4506024	DD_0	8.042	0.203
Scaffold_820_4506024	DD5	32.227	-0.284
Scaffold_820_4506024	NFFD	36.707	-0.255
Scaffold_820_4506024	bFFP	6.411	0.185
Scaffold_820_4506024	eFFP	9.946	-0.236
Scaffold_820_4506024	FFP	17.890	-0.233
Scaffold_820_4506024	EMT	5.133	-0.179
Scaffold_820_4506024	EXT	3.258	-0.134
Scaffold_940_3109473	Tmin_at	3.627	0.191
Scaffold_940_3109473	MAT	7.090	0.208
Scaffold_940_3109473	MWMT	4.581	0.191
Scaffold_940_3109473	MCMT	4.686	0.190
Scaffold_940_3109473	DD_0	7.025	-0.220

SNP	Climate Variable	Bayes Factor	Spearman Correlation
Scaffold_940_3109473	DD5	3.504	0.155
Scaffold_940_3109473	EMT	4.886	0.217
Scaffold_940_3109473	EXT	4.322	0.157
Scaffold_940_3109473	Eref	3.599	0.142
Scaffold_940_3109473	CMD	3.462	0.171