LINKING DEMOGRAPHIC HISTORY AND EVOLUTION AT THE EXPANDING RANGE EDGE OF SITKA SPRUCE (*PICEA SITCHENSIS*)

by

Joane Simone Elleouet

License BOPE, Université Paul Sabatier, 2009

Master BEE, Université Montpellier II, 2012

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Forestry)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

September 2018

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Linking demographic history and evolution at the expanding range edge of Sitka spruce (*Picea sitchensis*)

submitted by Joane Elleouet in partial fulfillment of the requirements for

the degree of Doctor of Philosophy

in Forestry

Examining Committee:

Sally Aitken
Supervisor

Michael Whitlock
Supervisory Committee Member

Lori Daniels
Supervisory Committee Member

Jennifer Williams
University Examiner

Dolph Schluter
University Examiner

Additional Supervisory Committee Members:

Amy Angert
Supervisory Committee Member

Abstract

Anthropogenic climate change is shifting species ranges and exerting high selection pressures on populations of all taxa, including trees. Temperate tree species of the northern hemisphere share a history of large-scale postglacial colonization during the Quaternary, providing a natural laboratory for the study of evolutionary responses to climate fluctuations. This dissertation aims at improving our understanding of the mutual influences of demography and evolutionary patterns during range expansions in trees using *Picea sitchensis* (Sitka spruce) as a focal species.

I first focused on the most recent *P. sitchensis* expansion event in south-central Alaska to study the interplay between demography and population genetics by combining neutral genetic markers and tree ring data. This multidisciplinary approach allowed me to assess the pace of neutral evolution across five centuries of colonization. Allelic richness was efficiently recovered in the colonizing population by early, open-grown colonizers on the Kodiak Archipelago during a long phase of low population growth. However, heterozygosity remains low compared with the nearest mainland populations. These results highlight the long-term importance of early colonizing genotypes in genetics of populations and the influence of pollen dispersal in maintaining standing genetic variation during forest expansion.

Local hybridization of *P. sitchensis* colonizers with foreign pollen from white spruce (*Picea glauca*) populations occurred repeatedly during the early colonization period. However, introgression was suppressed in subsequent generations growing under a closed canopy. As the two species occupy separate climatic niches, selection against hybrids, intensified by competition, might explain this pattern. Spring precipitation tended to affect hybrid growth more negatively than pure *P. sitchensis* genotypes, but this effect was nonsignificant.

I finally assessed the extent to which demographic parameters of range expansion can be estimated from genomic data through simulations using the approximate Bayesian computation framework. Simple 3-parameter models could be successfully estimated with genetic markers developed from reduced-representation methods currently available for nonmodel species. Models of higher complexity presented challenges, especially when ongoing migration after expansion was considered, and the accuracy of results depended on the time of expansion. The demic expansion models examined here were inadequate to infer the colonization history of *P. sitchensis*.

Lay Summary

Under anthropogenic climate change, tree populations need to adapt through genetic change or colonize newly suitable territories to survive. As tree species have migrated north in response to warming after the last glaciation, their history can inform us about their future responses to current climate change. My research explores methods to estimate the demography and evolutionary changes of populations in species that have expanded their range. I studied two approaches adapted to different timescales. I first used genetic and tree ring data covering five centuries to determine the recent changes in the genetic composition of Sitka spruce at its expanded northern limit. I found that the long time it takes for trees to mature may slow down the pace of colonization but allows the establishing forest to recover genetic variation. I then examined how patterns of DNA sequence variation in natural populations can help reconstruct their history over 500 generations.

Preface

Dr. Sally Aitken and Joane Elleouet developed together the original research questions addressed in this thesis. Joane Elleouet then developed the specific hypotheses, organized data collection and performed most analyses with frequent feedback and suggestions from Dr. Sally Aitken.

Christine Chourmouzis, Bert Terhardt, Ian McLachlan, Sally Aitken, Vincent Hanlon and Jon Degner provided help with sample collection in the field for chapters 2 and 3.

Genetic data development was performed by different institutions. Probes for sequence capture in Appendix E were developed at UBC by Kay Hodgins and Sam Yeaman. Library preparation for GBS sequencing in chapters 2 and 3 was performed at Laval University's Institute for Integrative Systems Biology (IBIS). Genetic sequencing for both sequence capture data and GBS data was performed at McGill University's Genome Québec Innovation Centre.

A version of Chapter 2 has been accepted subject to revisions in The New Phytologist. Elleouet, J. S., and Aitken, S. N. The interplay between demography and neutral evolution at the expansion front of a widespread conifer, *Picea sitchensis*. It is available on bioRxiv: https://www.biorxiv.org/content/early/2018/05/22/327742

A version of Chapter 4 has been published:

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 00:1-16.

Table of Contents

List of Tables

List of Figures

Appendix Figures

List of Abbreviations

ABC	approximate Bayesian computation
cpDNA	chloroplastic DNA
CTAB	hexadecyltrimethylammonium bromide
GBS	genotyping-by-sequencing
HI	hybrid index
LDD	long-distance dispersal
P. sitchensis	*Picea sitchensis*
P. glauca	*Picea glauca*
POD	pseudo-observed dataset
PPT	precipitation
RADseq	restriction site associated DNA sequencing
RRL	reduced representation libraries
SNP	single-nucleotide polymorphism
Tave	average temperature
TMRCA	Time of most recent common ancestor

Acknowledgements

I am deeply grateful to my academic supervisor Sally Aitken, whose ideas and guidance were paramount to the initiation, development and completion of the work presented in this thesis. The subtle balance between her benevolent support and the trust and independence she entitled me to made my PhD a great learning experience and pushed me to become more self-confident. I am also grateful for the help I received from my supervisory committee. Mike Whitlock always provided amazingly quick and to-the-point feedback on my ideas and manuscripts. Lori Daniels introduced me to the field of tree ring research through personal field and laboratory teaching and access to equipment. Amy Angert's comments and suggestions during the data analysis stage improved the quality of my work.

My research financially relied on an NSERC Discovery Grant to Dr. Sally Aitken and a Strategic Recruitment Fellowship from the Faculty of Forestry of UBC. Thanks to Andrea Chan, Rosemarie Cheng, and Natasha Thompson for making the Department for Forest and Conservation Sciences a great workplace. Special thanks to Gayle Kosh, for her incredible support through the sometimes challenging times of grad school.

The invaluable contribution of labmates and friends to field work in Alaska made field trips the most delightful part of my PhD. Thank you Christine Chourmouzis, Bert Terhart, Ian Maclachlan, Sally Aitken, Jon Degner, and Vincent Hanlon. Thanks to Pia Smets for her help in pre-field preparations and administrative procedures.

Local support from experts on the Kodiak Archipelago greatly helped solve all kinds of challenges in the field. Biologists Stacey Studebaker, Bill Pyle and Ed Berg shared their knowledge and helped find appropriate sampling sites. Kodiak foresters Karl Potts and Tash Saheed (Lesnoi Inc.) provided consent, maps and equipment to safely access sampling sites in logging-active areas. I'd like to extend special thanks to Keith Coulter (Koncor Forest Products), whose generosity and logistic support on Afognak Island was beyond anything I'd hoped for.

The help of undergraduate students made tedious data-processing work much more fun and efficient: Thank you Sean King for your help and friendly company at Totem Field and in the lab, and Elissa Sweeney-Bergen for your quality work counting tree rings.

As lab work is not my primary strength, obtaining valid genetic data heavily relied on the help and expertise of Kristin Nurkowski, whose technical skills are, in my opinion, nearly magical. I thank Robin Mellway for his technical and moral support in troubleshooting in the lab, and Vincent Hanlon who made the development of a reliable DNA bark extraction protocol a fun and successful little challenge. I also acknowledge the work of Brian Boyle (IBIS, Université Laval) in GBS library preparation and Sharen Roland (McGill University and Genome Quebec Innovation Centre) in Illumina sequencing. Thanks to Raphaël Chavardès for teaching me tree-ring analyses techniques.

I am grateful to Sam Yeaman for his teaching and technical support in bioinformatics. Even after setting up his own research group, he continued helping me clarify tricky computing procedures and fix programming errors. Thanks to Kay Hodgins, Jon Degner, and Jamie Myers who also provided bioinformatics help.

Fellow graduate students are what makes grad school such a rich and enlightening experience: Ian Maclachlan, Colin Mahony, Susannah Tysor, Jon Degner, Vincent Hanlon, Rafael Cândido Ribeiro, you were amazing! Thank you Ian for your support and friendship. I wouldn't have made it past 1 year without your advice, reassurance, goofyness, technical help, britishness, coma-inducing cakes, and companionship during field trip adventures. Thank you Colin for all your wisdom, and Susannah for keeping me stoked about the fascinating world of statistics and science in general. Thank you Vincent for popping up from nowhere and turning into the perfect labmate, always in a good mood and available to exchange smart ideas and unsettle the rest of us with seemingly innocent biology questions. Thank you Jon for your to-the-point suggestions at lab meetings, and for always having an answer to my bioinformatics questions.

I thank FIP, the best radio station in the world, for keeping my spirits high on long work days.

Finally, thank you Arran, for existing.

To Kristin Nurkowski,
the brilliant alchemist who could turn blurry liquids into perfect genomic data, and helped make the world a more caring and colourful place.

Chapter 1: Introduction

Species' ranges are highly dynamic systems. All living species experience range shifts, the amplitude of which mainly depends on the type of disturbance affecting them and the timescale considered. Large climatic oscillations over geological timescales have shaped the global distribution of life during alternating glacial and interglacial eras (Davis & Shaw, 2001). On moderate timescales, range fluctuations can occur in response to events of long-distance dispersal into newly available habitats such as volcanic islands (Connor et al., 2012). Finally, the most recent worldwide factors causing range shifts are anthropogenic: they involve exotic introductions due to increasing global connectivity in anthropogenic activities and human-induced climate change. This last category of factors provides countless examples of current changes in species' distributions at various spatial scales.

1.1 Evolutionary processes at range edges

1.1.1 What determines range limits?

In the absence of a hard geographic limit (such as a coastline or an ice sheet), one can wonder what prevents species from adapting to spatial changes in the environment and expanding beyond their current range limit. Model species have shown a large capacity for adaptation during experiments of artificial selection (Rice & Estert, 1993), and numerous transplant experiments beyond a species' current range in the wild have resulted in successful growth and reproduction (Hargreaves et al., 2014). The common observation of low densities and low growth rates experienced at the edge of species distributions have led researchers to develop evolutionary models of stable range limits. Notably, Kirkpatrick and Barton (1997) created a model involving varying fitness and population densities along environmental gradients under the infinitesimal model of genetic variation, and showed that species' range limits remain at a stable equilibrium through the swamping effects of directional gene flow. However, long-term equilibrium often does not occur in nature. An awareness of the transience of species range limits

has fostered empirical studies of factors limiting the pace and extent of range shifts. Adaptation and further expansion can temporarily fail in fragmented marginal populations at range edges because of low connectivity and high inbreeding, reducing the effective population size, fitness, and response to selection of marginal populations (Pujol & Pannell, 2008). Finally, constant climatic oscillations characteristic of our planet create an ever-renewed challenge, especially for species with low generation times and low dispersal abilities, which are likely to experience a lag in tracking their suitable climate envelope (Davis & Shaw, 2001). The crucial role of demographic processes at range edges is a common denominator of all hypotheses attempting to explain the evolution of range limits. It is therefore necessary to understand these demographic processes and their effects on the genetic composition of populations at range edges.

1.1.2 The effect of demography on population genetics during range expansion

Studies involving simulations of range expansion using different dispersal models illustrate how demography influences the neutral genetic composition of populations at all timescales. Models involving a regular wave of advance over continuous space have helped identify striking changes in the frequency of mutations occurring at the wave front, a phenomenon called allele surfing (Edmonds et al., 2004). This spatial case of genetic drift is most likely to occur in small, fast-growing populations (Klopfstein et al., 2006) and creates landscape patterns that could mimic the effect of natural selection. In a bidimensional model, the same process leads to the creation of "sectors" of genetic differentiation, as evidenced by the work of Hallatscheck & Nelson (2008) on bacteria. In the context of a species invading the range of a reproductively compatible species, this phenomenon can also explain dramatic levels of introgression from the local population to the invading population at the wave front (Currat et al., 2008). Better suited to the study of organisms with low-to-medium dispersal distances, the linear serial stepping-stone colonization model with successive population bottlenecks is one of the most widely used demographic models of range expansion. With dispersal only to proximal demes, this model creates a geographic gradient of decreasing genetic diversity and increasing differentiation towards the edge of a species' range (Austerlitz et al., 1997). This outcome has been widely observed in the wild and accepted as a signature of range expansion (Kitamura et al., 2015). The number of founders and time between colonization events can modulate the extent of diversity loss along

the expansion route (Le Corre & Kremer, 1998). The wave and stepping stone models described above are suitable for the study of organisms with low dispersal abilities. If dispersal from demes further behind the expansion front is possible, as is the case for species capable of frequent long-distance dispersal (LDD), different genetic patterns arise that maintain genetic diversity (Bialozyt et al., 2006) and the level of differentiation between demes depends on the frequency of LDD. In general, recent models of expansion indicate that LDD events, even when rare, can dramatically alter the long-term genetic composition of populations in an expanding species (Austerlitz & Garnier-Géré, 2003) and disrupt patterns typically created by colonization waves (Amorim et al., 2017).

1.2 Methods to study range expansions in plants

Approaches to characterize past and current range expansions involve complementary ecological and genetic tools. They are described in sections 1.2.1 and 1.2.2, depending on the spatial and temporal scale they are best suited to, and summarized in Figure 1.1.

1.2.1 Characterizing absolute and effective dispersal patterns locally

As dispersal distance is a major determinant of the evolutionary fate of expanding populations, attempts to characterize the dispersal kernel of propagules in plants have been numerous, especially to determine levels of LDD (Nathan, 2006). Dispersal is, however, difficult to observe, either directly or indirectly. For animal-dispersed plants, direct characterization of dispersal distances relies on tedious experiments (Cramer et al., 2007; Jorge & Howe, 2009) or mechanistic models (Will & Tackenberg, 2008), whereas wind-dispersed species can be studied through mechanistic models of wind flow (Nathan et al., 2011). The limitations of such methods and the fact that demographic patterns of dispersal leave signatures in the genetic composition of populations justify the use of genetic methods to indirectly infer patterns of effective dispersal (Hamrick & Trapnell, 2011). At a small temporal and spatial scale, pedigree analyses and the neighbourhood model can help unravel immediate patterns of effective dispersal and characterize the shape of the dispersal kernel at a site (Bacles et al., 2006; Burczyk et al., 2002; Sezen et al.,

2007). This method requires sufficient amounts of genetic data and 100% sampling of the studied site. Genetic relatedness among trees at a site can also be estimated from genetic data and correlated with pairwise distance dispersal kernels, although this method has limitations due to the effects of population density (Vekemans & Hardy, 2004). The association of geographic distance with genetic distance between individuals has given birth to the vast methodologies of landscape genetics (Manel et al., 2003). These approaches are commonly used at low and medium temporal and spatial scales to determine the recent history of effective dispersal in an area (Ibrahim et al., 1996). The use of uniparentally inherited markers such as mitochondrial DNA and chloroplast DNA in landscape genetics can provide precise information on the relative importance of pollen or seed dispersal in the genetic composition of expanding species (Smouse et al., 2001). At spatial scales higher than the landscape level, genetic methods applied to the study of range expansions do not address dispersal *per se*, but rather characterize patterns of gene flow resulting from underlying effective dispersal and population size fluctuations. Such methods are described in section 1.2.2. Non-genetic methods suitable for the study of range expansion at timescales of hundreds to thousands of years involve the use of fossil pollen found in lake sediments. Such record have successfully complemented genetic inference of range shifts associated with glacial cycles, although in many cases pollen can be classified only to genus (Hu et al., 2008).

1.2.2 What can genetic data tell us about demographic history?

Simple population genetic summaries and their associated statistical tests, such as Tajima's D, a test based on the neutral theory model (Tajima, 1989), can detect past range expansion events. Some analytical tools developed more recently rely on specific population expansion models (Peter & Slatkin, 2013). Although such simple methods can help detect range expansion in populations, they do not have the power to link the observed genetic patterns to the combination of demographic parameters likely to have given rise to such patterns. An exception is the use of differentiation statistics (e.g., F_{ST}, G_{ST}) and their comparison among markers from genomes with different inheritance modes (mitochondrial, chloroplast and nuclear DNA). Used at a regional scale, these methods can provide information about the relative historic contribution of pollen versus seed dispersal during colonization (Petit et al., 2005). However, the power to

quantitatively infer parameters of species' demographic history comes from statistical phylogeography, a set of methods involving the reconstruction of gene genealogies under an assumed demographic model (Knowles & Maddison, 2002). All common phylogeographic inference methods are based on the coalescent (Kingman, 1982), a statistical framework used to build gene genealogies. Phylogeographic methods are being constantly developed and refined and their diversity is matched by the variety of demographic models and type of genetic data they can handle (Bourgeois, 2016; Excoffier & Heckel, 2006). They have successfully provided evidence for glacial refugia (Anderson et al., 2006; Petit et al., 2004), defined the number of introduction events in the case of biological invasions (Benazzo et al., 2015) and informed humans about who we are and where we come from (Schraiber & Akey, 2015). As the availability of large genomic datasets has increased, phylogenetic studies have addressed increasingly ambitious demographic models. Although recent genomic advances allowed for the development of powerful sophisticated methods (Harris & Nielsen, 2013; Li & Durbin, 2011), the amount of power gained by existing methods from larger and better datasets is not always well established, and as a result, limitations of some demographic inference methods are ill-defined. To resolve this uncertainty, analytical studies comparing methods with different dataset sizes (Terhorst et al., 2017) and simulation analyses of inference success targeting specific phylogeographic methods (Li & Jakobsson, 2012) have started to emerge. Similarly, methods that have proven useful in human demographic inference might not yet be suitable in nonmodel organisms, due to the unmatched quality of genomic datasets. The influence of genotyping uncertainty on genetic inference is also being explored (Fumagalli, 2013; Shafer et al., 2016).

1.3 Range shifts in widely distributed tree species

1.3.1 Tracking climate during glacial cycles

Along with other terrestrial organisms, temperate and boreal tree species ranges have expanded and retracted following cycles of glacial and interglacial periods during the Quaternary era. In the northern hemisphere, species repeatedly expanded from and retracted to refugia that remained ice-free (Hewitt, 2000; Petit et al., 2008). In North America, the last ice age ended

around 18,000 years ago and was followed by a global warming period called the Holocene. Tree species of North America were first thought to have expanded from refugia south of the Cordilleran and Laurentide ice sheets (Hewitt, 2000) at a high average speed of 100 to 1000 meters per year (Clark, 1998). It is now widely accepted that many taxa were also present in northern refugia during the last glaciation (Parducci et al., 2012; Shafer et al., 2010) and estimates of migration pace therefore dropped to 60-260 meters per year (Feurdean et al., 2013; McLachlan et al., 2005). Similar estimates of migration rates were found in Europe (Svenning & Skov, 2007) but higher values were estimated in Scots pine (*Pinus sylvestris*), likely due to favourable conditions and the lack of interspecific competition in this early colonizing species (Savolainen et al., 2011). Accounting for the positions of glacial refugia, Svenning and Skov (2007) determined that the distribution of most tree species in Europe was better explained by distance from refugia than by climate. It is therefore likely that many tree species were not able to track their optimal climate envelope during postglacial colonization, although their dispersal abilities have been found to match the scale of past changes in climate envelopes across landscapes (Kremer et al., 2012). These findings cannot be generalized to other continents: Hamann and Wang (2006) developed species distribution models for western North American trees and found that most species currently nearly fully occupy their climatic niche space, although some species, like *Pinus contorta* and *Pinus albicaulis*, showed unoccupied, suitable habitat north of their range. Differences between topographic features in Europe and North America (such as the orientation of mountain ranges) have previously been invoked to explain different postglacial migration outcomes in widespread tree species (Lumibao et al., 2017).

1.3.2 Effects on genetic diversity in temperate tree species

Nucleotide diversity of tree species is generally lower than in other plant taxa (Brown et al., 2004; De la Torre et al., 2017; Heuertz et al., 2006). Although this can be partly explained by low substitution rates (De la Torre et al., 2017), population bottlenecks during previous glacial periods likely played a role in this pattern and have been detected using Tajima's D- and Fay and Wu's H- statistics (Heuertz et al., 2006). The lasting genomic imprint left by postglacial migrations allows identification of the location of glacial refugia during the Pleistocene (Hu et al., 2008). In Europe, many widespread tree species display patterns of high genetic diversity in

cpDNA at intermediate latitudes and high genetic differentiation among southern latitudes. This pattern is in agreement with the scenario of expansion of species from moderately rich, isolated populations on southern peninsulas and their subsequent genetic admixture at intermediate latitudes during postglacial colonization (Petit et al., 2004). Mid-latitude populations might also have benefited from admixture with glacial refugia at similar latitudes (Magri et al., 2006). Northernmost populations have often been found to be genetically depauperate due to the erosion of diversity associated with their recent expansion. Such patterns observed in European tree species have not been as consistently detected in North-American tree species (Lumibao et al., 2017; Marsico et al., 2009), where genetic diversity is generally homogenous across species ranges. This can be explained by the lesser extent of ice during the last glaciation, allowing numerous refugial populations to survive at high latitudes (Shafer et al., 2010) and the absence of geographic barriers compared to Europe, where oceans separated peninsular refugia and East-West mountain ranges hindered northward recolonization. However, some European tree species also fail to show a phylogeographic signal of postglacial expansion, especially when looking at nuclear markers and quantitative traits (Kremer et al., 2012). This lack of geographic variation in nuclear genetic diversity on both continents can be explained by the high levels of gene flow characteristic of widespread tree species as well as selection enhancing regional differentiation (Kremer et al., 2012; Savolainen et al., 2007). Signatures of range expansion in neutral nuclear markers might also not have appeared at all due to tree-specific characteristics, including a long juvenile phase and long-distance dispersal, which have been shown to prevent the erosion of diversity along expansion axes (Austerlitz & Garnier-Géré, 2003; Le Corre et al., 1997).

1.3.3 Implications for adaptation

Forests currently cover 27% of the surface of Earth, with single species sometimes occupying contemporary geographical ranges wider than 20 degrees of latitude. For a given species, being so widely distributed requires either being differentially adapted to the diverse climatic conditions that its range encompasses or having a high degree of phenotypic plasticity (Alberto et al., 2013). Abundant evidence that boreal and temperate tree species experience fine-tuned adaptation along environmental gradients for growth and phenology in spite of high gene flow illustrate this prediction (Aitken & Bemmels, 2016; Alberto et al., 2013; Howe et al., 2003;

MacLachlan, Wang, et al., 2017; MacLachlan, Yeaman, et al., 2017). High levels of gene flow can actually increase the genetic variance of a quantitative trait even under strong locally stabilizing selection, therefore enhancing the response to selection of the recipient population. This feedback loop has been proposed as an explanation of the high level of local adaptation in forest trees (Kremer et al. 2012) and is supported by some empirical evidence (Yeaman & Jarvis, 2006). Theoretically, under the right combination of rate of environmental change, dispersal distance and migration rates, high gene flow should also enhance adaptation to a potential lag between a changing climate and population migration triggered by it (Polechová et al., 2009). Although this is a compelling explanation for the high level of local adaptation of forest trees despite recent dramatic range shifts, no empirical study has yet supported it (Kuparinen et al., 2010). Empirical studies of expanding boreal tree populations have been conducted in several widespread species. Such work typically involves characterizing molecular genetic variation in northern expanding populations in comparison with core and trailing edge populations. As part of such effort, Pyhäjärvi et al. (2007) and Wachowiak et al. (2011) used nuclear markers and several population genetic statistics to infer that the relatively high current levels of genetic diversity in northern *Pinus sylvestris* populations might result from admixture among lineages from different colonization routes. The concomitant identification of of quantitative variation through common gardens and neutral genomic variation in northern populations is an effective way to assess the mutual influence of adaptation and colonization dynamics (Chen et al., 2012; Mimura & Aitken, 2007a; Savolainen et al., 2011). This approach has been applied in a northern *Pinus sylvestris* population, where a fast and recent northward colonization history did not impair fine-scale differentiation in budset timing, suggesting rapid adaptation in the face of gene flow at a recently expanded range limit (Savolainen et al., 2011). Data extracted from common garden experiments can be used in theoretical models of species tracking their environmental optimum in space (Kremer et al., 2012; Savolainen et al., 2011) and can therefore help predict the adaptation potential of species at range limits under climate change. Transplant experiments with manipulation of temperature and precipitation regimes can also been used to assess the viability of populations in future environments at northern range limits. For instance, Rousi et al. (2017) have found a high growth and acclimation potential but unpredictable mortality in northern *Pinus sylvestris* populations under artificially-induced future climates.

1.4 *Picea sitchensis* as a study system

1.4.1 Local adaptation throughout a wide latitudinal range

With its ability to withstand salt spray and thrive in hypermaritime environments, Sitka spruce (*Picea sitchensis* (Bong.) Carr.) partially dominates Pacific coastal temperate rainforests. Its imposing stature can be observed from Fort Bragg in California up to Kodiak Island in Alaska (Figure 1.2). The northern part of its range is parapatric with a closely related species, *Picea glauca*. The two sister species are known to hybridize at several secondary contact sites on the Kenai Peninsula in Alaska (Boucher & Mead, 2006), and in British Columbia, where hybrid zones are likely maintained by ecological selection (Hamilton et al., 2013b). As a wind-pollinated, wind-dispersed tree species, *P. sitchensis* is an appealing subject for evolutionary questions such as range expansion, adaptation, and their mutual influence. Indeed, its narrow east-west range spanning 22 degrees of latitude leaves little doubt about the colonization route from its southern refugium during the last glacial period. Some northern glacial refugia have been suspected (Gapare & Aitken, 2005) but no clear evidence, genetic or otherwise, has been found. The wide array of climatic conditions that this species experiences currently allows in-depth studies of the genetic bases of adaptation in terms of growth and timing of phenological events. Mimura and Aitken (2010) set up growth chamber experiments using provenances from the core and edges of the *P. sitchensis* range and found evidence for local adaptation, including at range edges. Through monitoring of gene expression levels and genotype-phenotype association analyses on the same provenances grown in a common garden, Holliday et al. (2010) identified genes involved in growth timing and cold hardiness. Adaptation throughout the core of the range for these traits is evidenced by strong genetic clines of allelic frequencies in associated markers (Holliday, Ritland, et al., 2010; Lobo, 2011) and high Q_{ST} values (Mimura & Aitken, 2007a). However, earlier studies focusing on the northernmost populations suggest that the degree of local adaptation decreases towards the edge of the species' range. There is evidence for lower growth performances of peripheral populations. Farr & Harris (1979) showed that the growth rate based on estimates of site index of some of the northern Alaskan populations is significantly lower than predicted by latitude and growing degree days, providing potential evidence for low fitness in some of the most remote Alaskan populations. A common garden of

14 *P. sitchensis* provenances in Vancouver (BC) shows similar results from measurements on 8 year-old trees (Figure 1.3).

1.4.2 Current processes at range margins

The range of *P. sitchensis* is still expanding at its northern limit. Historical human records report a fast advance of the monospecific *P. sitchensis* forest on the Kodiak Archipelago (Griggs, 1914; Vincent, 1964). The westernmost and largest island of the group, Kodiak Island, is thought to have been colonized no more than 500 years ago (Griggs, 1937). The trees are found in dense stands on Afognak Island and the forest density tapers south-westward towards small groves of young trees (Figure 1.4). The south-west of Kodiak Island is vegetated by tundra grasses and scattered shrubby forms of alders and cottonwoods. Adaptation limitations mentioned in section 1.4.1 may have an influence on the current location of the species' range edge. Holliday et al. (2012) identified asymmetric gene flow from core Alaskan populations to the Kodiak Archipelago. Together with high inbreeding levels (Gapare & Aitken, 2005; Mimura & Aitken, 2007b), this pattern could actively be hindering adaptation. In addition, Lobo (2011) suggested that the Kodiak Island population at the northern limit of the range might lack the necessary genetic variation to adapt to the local climate and grow optimally. The recent establishment of the Kodiak population could support the alternative hypothesis that the population needs more generations to adapt to local conditions. If this is the case, the current lack of local adaptation is transient and is not actively maintain by evolutionary processes. The most recent species distribution model for *P. sitchensis* based on its realized climate niche indicates that suitable habitat extends far beyond the current range limit along the coast all the way to the Aleutian Islands (Tongli Wang, pers. com.). The current limits might therefore reflect dispersal limitations rather than adaptation failure. Tae (1997) found that *P. sitchensis* groves at the expansion front on Kodiak Island during the 20th century grew dramatically in size and number during decades after ash fall from the 1912 Novarupta volcanic eruption, which killed a large proportion of low-lying forms of plants on the island. The tephra deposited by the eruption likely reduced interspecific competition for *P. sitchensis* seedlings for about 7 years. Groves kept growing at a high pace until 1945, already established groves likely playing the role of nuclei in the establishment of new cohorts of seedlings. The ongoing range expansion on the Kodiak

Archipelago together with the current knowledge of neutral gene flow patterns and the genetics of adaptive processes makes the northern range limit of *P. sitchensis* a unique study system for examining the interplay between adaptation and range dynamics. In addition, the longevity of this species and the historical record laid down in tree rings provides opportunities for demographic and genetic studies that are not possible in herbaceous or short-lived plants.

1.5 Research outline

By studying the past movements of tree species in response to climatic fluctuations, we may be able to understand and predict future population movements in response to contemporary large-scale disturbances. Empirical and theoretical studies of postglacial migration speed as well as dispersal patterns are necessary to assess the potential of tree species to track their current climate niche. In turn, establishing the link between dispersal patterns and the resulting genetic diversity and structure in expanding populations at range edges enhances our understanding of the factors influencing the success and pace of adaptation in constantly changing environments. The work presented here tests and applies tools to understand tree population movements and their evolutionary outcome at various temporal and spatial scales through empirical analyses of population expansion in *P. sitchensis*.

Chapters 2 focuses on the northern range margin of *P. sitchensis*. In this chapter, I ask whether demographic patterns at the colonization front foster or hinder genetic diversity and population differentiation. To answer this, I reconstructed the recent colonization of the Kodiak Archipelago by *P. sitchensis* in space and time over the past 500 years, using a combination of dendrochronology and genetic analyses.

Chapter 3 examines the mutual effects of colonization and introgression at range margins by asking whether the colonization of the Kodiak Archipelago by *P. sitchensis* was accompanied by a change in admixture with the closely related parapatric species *Picea glauca*. To answer this, I characterized the spatial and temporal distribution of genetic admixture between the two

species in the studied region. In this chapter, I also attempt to determine whether the observed patterns are caused by neutral or selective mechanisms using tree ring and climatic data.

Chapter 4 explores the interplay between genetic and demographic change at larger time and spatial scales by taking a phylogeographic approach, focusing on a commonly used method, approximate Bayesian computation (ABC). In this chapter I explore the power and limitations of ABC in the inference of a recent spatial expansion using extensive simulations of a variety of scenarios. I test the effects of model complexity, sequence length, time of expansion event, and sequencing depth on the precision and accuracy of model parameter estimates. I also test the suitability of models developed in this chapter to estimate postglacial migration rates in widespread tree species through an empirical application of the method to the expansion history of *P. sitchensis* in the northern part of its range.

Finally, I conclude in Chapter 5 by summarizing findings from chapters 2 to 4, discussing how they relate to current concerns about climate-related range shifts in tree species, and addressing potential areas for research.

1.6 Figures

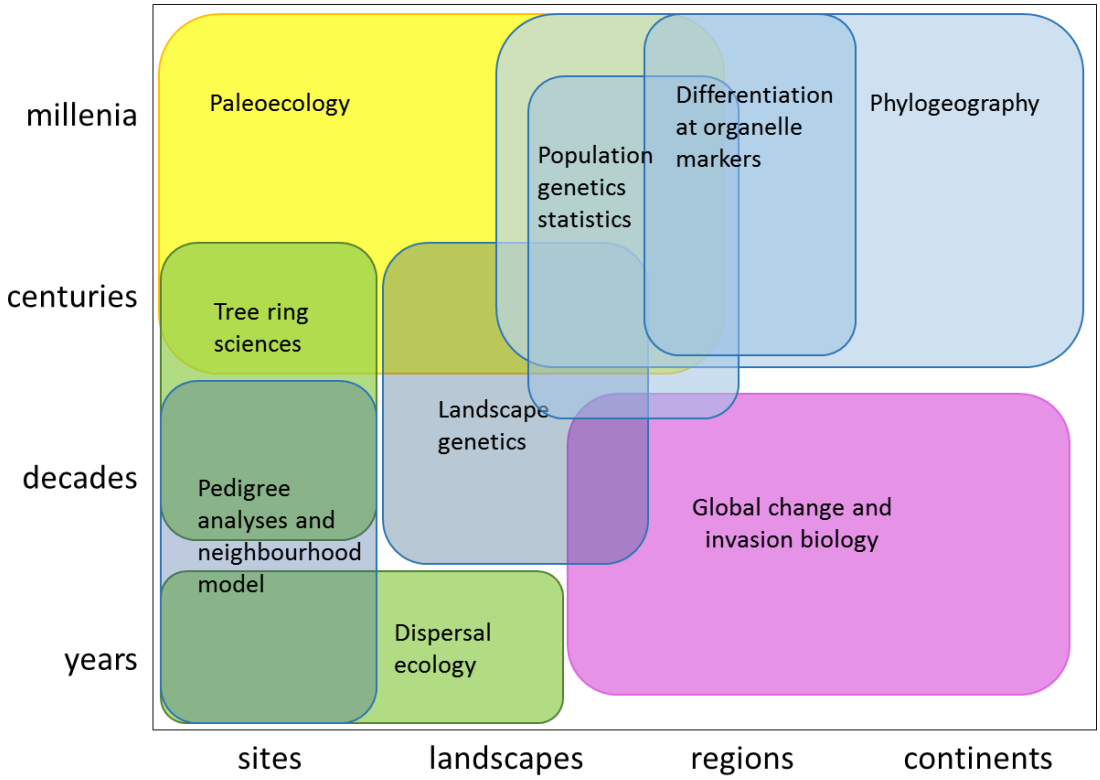


Figure 1.1 Empirical research methods contributing to the study of range expansions, modified and extended from Hampe (2011). Methods involving genetic data are displayed in blue.

Figure 1.2 Natural geographic range of *Picea sitchensis* and its sister species *Picea glauca* in western North America.

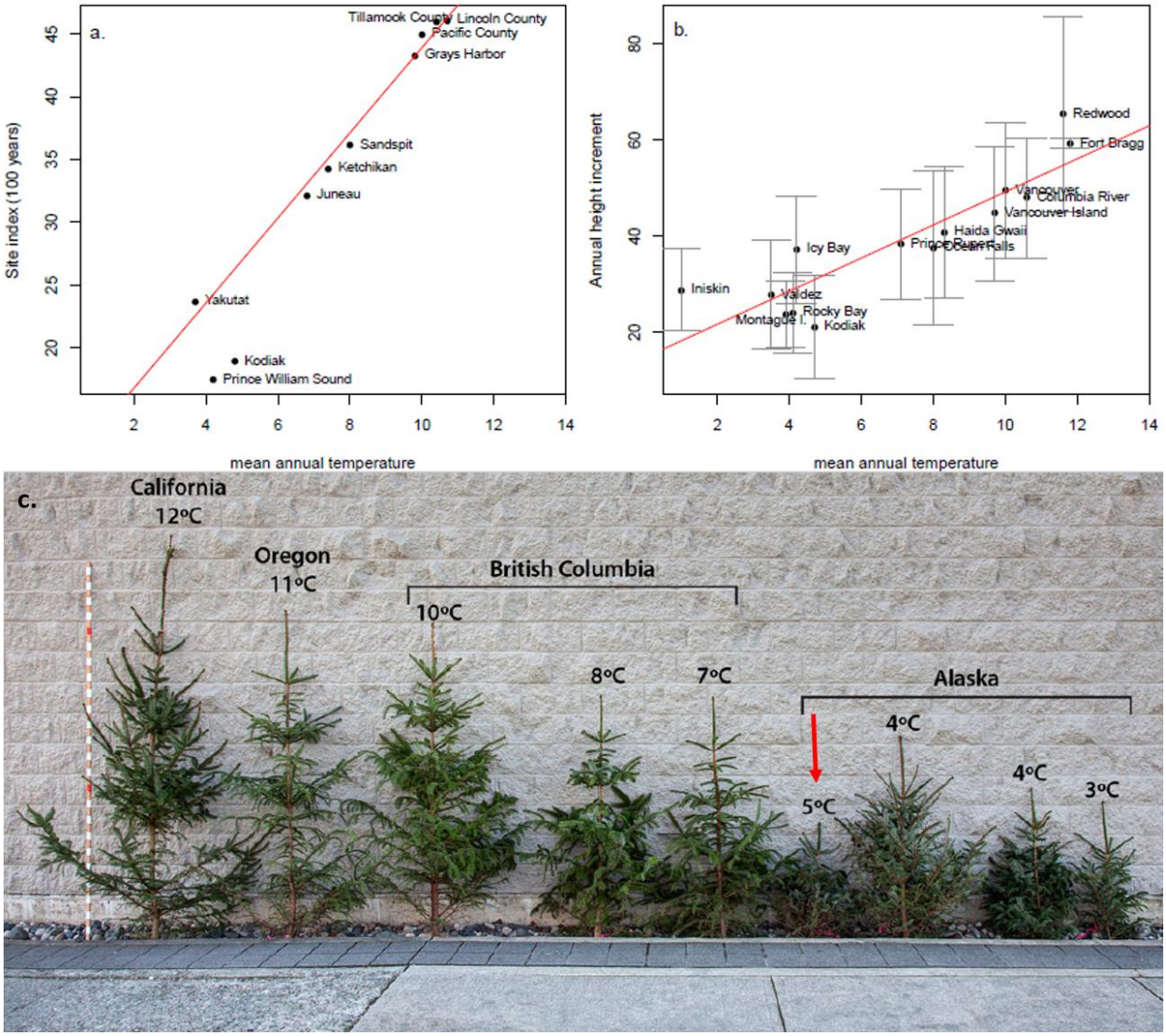


Figure 1.3 Growth in relation to mean annual temperature in two *P. sitchensis* common gardens. a. Site index, data from Farr and Harris (1979). b. average height from a UBC common garden (Joane Elleouet, unpublished). c. Photograph illustrating results from Figure 1.3b, modified from Aitken and Bemmels (2016). For each provenance, the tree with height closest to the provenance mean was selected for this photograph from the UBC common garden to illustrate growth clines across regions and climates conditions. The red arrow points to the tree representing the Kodiak provenance.

a. b.

Figure 1.4 Pictures of two forest sites at different distances of the expansion front. a. Site A5 on Afognak Island (see Figure 2.1 for exact location). b. Site K5 on Kodiak Island (southernmost site, see Figure 2.1 for exact location).

Chapter 2: Demographic and genetic reconstruction of the recent postglacial expansion of *Picea sitchensis*

2.1 Introduction

With an enhanced understanding of plant species migrations during past postglacial cycles and numerous observations of current climate change effects on species distributions, we are recognizing more than ever before the ubiquitous nature of range shifts. This awareness comes with a substantial literature including simulation studies of genetic changes during range expansion (Bialozyt et al., 2006; Hallatschek & Nelson, 2010; Peischl et al., 2013), empirical studies of expanding species and associated evolutionary patterns (e.g. Darling et al., 2008; Pujol & Pannell, 2008), and metaanalyses and reviews highlighting the general patterns common to — or variable among — phylogenetic and functional biological groups (Excoffier et al., 2009). Successive founder effects along colonization routes are a well-studied phenomenon causing an erosion of diversity and enhancing genetic differentiation (Excoffier et al., 2009; Hewitt, 2000; Slatkin & Excoffier, 2012). However, mechanisms related to species' dispersal and life history traits have been shown to influence the genetic outcome of range expansions and give rise to fundamentally different spatial patterns of genetic structure (Bialozyt et al., 2006; Waters et al., 2013). It is therefore not surprising that among organisms with different life history and dispersal traits, genetic effects of range expansions are not consistent (Eckert et al., 2008).

This heterogeneity is present across studies of tree species. Empirical studies at the scale of species ranges have found both higher genetic diversity (Born et al., 2008; Pluess, 2011; Shi & Chen, 2012) and lower genetic diversity (Johnson et al., 2017; Kitamura et al., 2015; Marsico et al., 2009; Mimura & Aitken, 2007b) in leading edge populations of tree species after range expansion. Temperate forest tree species are generally associated with high gene flow via wind-borne pollen across large geographic distances (Kremer et al., 2012), as well as a long lifespan and juvenile phase. These characteristics can have a strong influence on the interplay between genetic and demographic processes during range expansion. Founder effects may be buffered by high levels of gene flow (Austerlitz et al., 1997). The longevity of trees allows for founders to

persist until other propagules colonize, and their long juvenile phase forces the reliance on gene flow via foreign pollen, providing genetic diversity to the establishing population (Austerlitz et al., 2000). In addition, long-distance dispersal can prevent the erosion of genetic diversity along expansion routes (Le Corre & Kremer, 1998), promote genetic differentiation between demes (Austerlitz & Garnier-Géré, 2003) or suppress local introgression (Amorim et al., 2017).

Recent empirical studies involving exhaustive sampling and pedigree reconstruction in isolated forest stands at range edges have greatly enhanced our understanding of demographic mechanisms shaping the genetic composition of expanding populations at the local scale. Founding individuals, often arriving via long distance dispersal of seeds, play a major role at the start of establishment. Troupin et al. (2006) found that spatial genetic structure of a population shortly after range expansion strongly reflected the genetics of founding trees in a *Pinus halepensis* population. Lesser et al. (2013) identified Allee effects in an establishing *Pinus ponderosa* stand, highlighting the importance of a certain level of long-distance dispersal of founders in the establishment success of the population. A general finding in these studies is the predominance of high levels of pollen flow shortly after founding, leading to a quick recovery of genetic diversity during recruitment (Hampe et al., 2013; Lesser et al., 2013; Pluess, 2011; Sezen et al., 2007).

Here, I take a multi-scale approach to the study of expanding tree populations, combining observations of spatial and temporal patterns at the regional and local scale. In particular, I ask how demographic and spatial patterns of colonization affect genetic structure along an expansion route in space and time, and interact with potential geographic barriers to gene flow. To do so, I combine nearly five centuries worth of demographic and genetic data from several forested sites along the recent colonization route of *Picea sitchensis*. Although it is rarely the most abundant species in the southern and central part of its range, *P. sitchensis* dominates the forest cover together with mountain hemlock on the Kenai Peninsula and is the only forest tree species on the Kodiak Archipelago. Historical human records report a rapid advance of the monospecific *P. sitchensis* forest on the Kodiak Archipelago (Griggs, 1914; Vincent, 1964). The westernmost and largest island of the group, Kodiak Island, is thought to have been colonized no more than 500 years ago (Griggs, 1937). *P. sitchensis* is found in dense stands on Afognak Island and the forest density tapers south-westward towards small groves of young trees (Tae, 1997). The absence of

spruce pollen in paleoecological records on Kodiak Island (Bowman, 1934) strongly suggest that this is the first occurrence of the *P. sitchensis* forest at this site since the last glacial period. This range expansion therefore seems to be part of the long-term post-glacial colonization process of the species, with the most recent front advance having likely been facilitated by a nearby volcanic eruption in 1912, which reduced interspecific competition between *P. sitchensis* seedlings and herbaceous species (Tae, 1997). Southwestern Kodiak Island features tundra grasses and scattered shrubby forms of *Alnus viridis*, *Populus trichocarpa*, and *Betula nana*. Earlier studies focusing on *P. sitchensis* detected asymmetric gene flow from core Alaskan populations to the Kodiak Archipelago (Holliday et al., 2012), as well as a high self-fertilization rates (Gapare & Aitken, 2005; Mimura & Aitken, 2007b) and a lack of adaptive potential (Lobo, 2011) on Kodiak Island. Based on this knowledge, this chapter aims to characterize gene flow from populations on the Kenai Peninsula to the Kodiak Archipelago and identify demographic or genetic mechanisms responsible for reduced levels of genetic diversity at the expansion front.

I first conduct a demographic analysis over island and continental regions using dendrochronological methods to infer the timing and spatial structure of dispersal patterns during range expansion on the Kodiak Archipelago. To assess the current and past extent of differentiation and genetic diversity at the regional level, I then quantify genetic population structure between regions and genetic diversity within regions for different age classes at the northern range of *P. sitchensis*. This direct monitoring of genetic diversity and structure allows for the quantification of the extent and duration of potential founder effects, as well as the relative importance of early colonizers and subsequent gene flow in the accumulation of genetic diversity of the growing population. Finally, I take a closer look at sites at the expansion front, to determine the short-term genetic consequences of fine-scale dispersal patterns and demography.

2.2 Materials and Methods

2.2.1 Sampling design

I focused on the northern range of *P. sitchensis* in south-central Alaska. In 2013 and 2015, fifteen sampling sites in healthy forests with old-growth characteristics and with no evidence of

past outbreaks of spruce beetle (*Dendroctonus rufipennis*) were sampled on Kodiak Island, Afognak Island, and on the Kenai Peninsula near Seward (Figure 2.1). On the Kodiak Archipelago, sites were specifically chosen for their suitability to document initial site colonization, not post-disturbance regeneration. We therefore avoided any site with numerous canopy dominant snags or coarse woody material that would be legacies of a forest damaged by a stand replacing disturbance. Sample sizes within sites varied between 12 and 86 trees, depending on the size of the sampled site and its accessibility (Table 2.1). I classified trees into four canopy structure levels (large canopy tree, medium-sized canopy tree, sub-canopy tree, and immature sapling) and sampled sites to maximize the range of tree ages and to obtain even sample sizes from each canopy structure level. Large canopy trees were typically >70cm in diameter, and showed growth forms consistent with earlier growth in an open environment (numerous large dead branches low on the trunk and strongly tapered stems), especially on the Kodiak Archipelago. Medium-sized canopy trees were generally <70cm in diameter and showed no signs of open growth. Sub-canopy trees were mature trees that had not reached the main canopy, and immature saplings were generally no more than 2m tall. All sampled trees were separated by at least 150 meters to avoid high relatedness between individual samples. For DNA extraction, young needles were sampled whenever possible; when foliage was out of the reach of a pole pruner two 1 cm-diameter cambium disks were collected with a leather punch. Sampled materials were stored in paper envelopes in silica gel until DNA extraction. To estimate the age of individuals sampled, an increment borer was used to core each tree up to 5 times as low as possible on the trunk to obtain a wood sample that included the pith or signature thereof. A detailed description of dendrochronological methods is available in Appendix B. The age of saplings was approximated by counting major branch clusters on the stem. I used available samples of needles from 15 canopy trees in two additional sites on Shuyak Island and Port Chatham for inclusion in population genetics analyses but did not have tree ring data for these populations (Figure 2.1).

2.2.2 Dating population establishment

I used two metrics to estimate timing of population establishment at the site level: the age of canopy trees and the date of canopy closure. The age of canopy trees was obtained by averaging

age estimates of all medium and large canopy trees. This estimate has the advantage of being accurate for representing the age of live canopy trees, but does not necessarily reflect initial forest establishment if survivorship of initial colonizers was low, if tree longevity is shorter than the time since site colonization, or if trees established after a major stand-replacing disturbance (although sites were chosen on the basis of absence of cues of past disturbance). Estimating the date of canopy closure is an attempt to overcome this limitation and relies on the identification of patterns of radial growth during early life stages, based on the hypothesis that individuals growing in an open environment will experience faster growth in early stages of life compared to individuals competing under a closed canopy. Therefore I expect three scenarios: (1) that most trees establishing on forested sites with an existing closed canopy will have relatively narrow growth rings at the juvenile stage; (2) that most trees on sites in the process of initial forest colonization will have wide growth rings at the juvenile stage; (3) that formerly forest-free sites will show a temporal shift from trees with wide juvenile growth rings to trees with narrow juvenile growth rings as trees grow and compete under a increasingly closed canopy.

For each tree core, annual growth increments between years 11 and 20 were averaged to represent juvenile growth; years 1-10 were not included to avoid the effects of competition with small-statured herbaceous plants and shrubs. For individual trees, I modeled the relationship between estimated establishment date (x) and average annual juvenile growth (y) at the site level using two different regression models: a linear model $y \sim ax + b$ and a logistic model of the type $y \sim a/(1 + e^{-b(x-h)})$. Sites following scenarios (1) and (2) above are expected to show a better fit with a linear model with nonsignificant or weak slope, and sites following scenario (3) are expected to show a better fit with the logistic model. I retained the model with the lowest AIC. When sites were best fitted by a logistic model, I checked that the relationship between year of establishment and growth increment was negative (b<0) and recorded h as the estimated date of canopy closure. When the linear model was the best fit, I tested the significance of the slope (p-value for coefficient b). Site K4 was removed from this analysis due to sub-canopy trees at the site being under-represented in the sample.

2.2.3 Genotyping

To obtain putatively neutral markers for 639 trees, genotyping-by-sequencing (GBS) was used with a sbf1-msp1 double-digest protocol (Elshire et al., 2011). Libraries were sequenced with the HiSeq 2000 system, producing 100-bp single reads. I aligned the filtered reads to the *P. glauca* reference genome WS77111_V1 (Warren et al., 2015) using the bwa mem alignment algorithm. Alignment files were then input into a variant calling pipeline using functions from the program *GATK* (McKenna et al., 2010). A more detailed description of the bioinformatics processing steps is provided in 0, section A.2. After SNP calling, I removed all singletons across the 639 genotyped diploid individuals. Finally, when several SNPs were present less than 100bp apart, I retained only one of them, resulting in a final dataset with genotypes for 3244 biallelic SNPs. Unless otherwise stated, all population genetics analyses described in the following sections use a missing value cutoff of 60%. Based on preliminary tests on heterozygosity calculations, this cutoff value appears to be the best compromise between the number of usable SNPs and the completeness of the dataset. Although the GBS approach outputs datasets with considerable missing data (see Appendix A), it also provides a cost-effective genome-wide picture of genetic diversity and structure.

2.2.4 Visualizing population structure

I visualized population structure among all regions sampled using 2 methods: principal component analysis (PCA) and *Structure* clustering (Falush et al., 2003; Pritchard et al., 2000). For the PCA, I retained all genotype calls and filtered sites for missing data with a 40% cut-off value. The resulting dataset included 639 *P. sitchensis* individuals genotyped for 220 SNPs. I replaced all missing data by their mean over all individuals prior to PCA. The R packages *adegenet* (Jombart, 2008) and *ade4* (Thioulouse et al., 1997) were used to convert data files and perform a centered PCA. To characterize further population structure within the dataset, I used the program *Structure*, which detects clusters of individuals based on Hardy-Weinberg equilibrium within clusters. I first performed exploratory runs with run lengths from 10k to 100k after a 10k burn-in and 3 replicate runs for each run length. For $K \leq 4$, a run length of 50k was sufficient, whereas for $K>4$, a run length of 100k was necessary. I used the independent allelic

frequency model and the admixture ancestry model for all runs, and performed 3 runs for each value of K between 2 and 6.

2.2.5 Temporal patterns of diversity and structure

To infer the role of founder individuals and subsequent migration in the development of genetic diversity of the Kodiak-Afognak population, I selected all sites with less than 20% missing genotypic data for trees with age estimates. This dataset (120 SNPs, 412 trees) was used to estimate the year of first observation of each allele in the growing population. Using such a stringent cutoff value for missing data was necessary to adequately describe the date of appearance of each new allele in the population. A 1000-replicate randomization was applied to model the random distribution of allele accumulation curves against which to test significance of the observed results. To reconstruct the changes in gene flow patterns between continental and island populations, I assessed pairwise population differentiation between regions by computing the Weir and Cockerham (WC) F_{ST} estimator with R functions modified from the *adegenet* and *hierfstat* packages. Loci with less than 30% missing data over the whole sample were used. This cutoff value was chosen based on a preliminary exploration of F_{ST} calculations performed using different programs and subsets of the data. As the WC estimator is sensitive to unbalanced sample sizes (Bhatia et al., 2013), I randomly subsampled the larger population sample to the size of the smallest population sample. Confidence intervals around F_{ST} estimates were assessed with 1000 bootstraps.

2.2.6 Site-level summary statistics

Using SNPs that were well represented in each region (< 60% missing data across the sample), I calculated expected heterozygosity, F_{ST} and allelic richness at the site level using the R packages *adegenet*, *hierfstat* and *PopGenReport*, respectively. The latter uses the methods of El Mousadik and Petit (1996), which corrects for variable sample sizes through rarefaction. Changes in dissimilarity between sites over time on the Kodiak Archipelago were estimated using the dissimilarity calculations of Petkova et al. (2016). Estimates at the individual tree level rather than allele frequencies (such as F_{ST}) better suited the temporal analysis at the local scale due to low within-site sample sizes for each cohort. Briefly, I computed a matrix of genetic distance

between pairs of individuals using the average squared genetic difference across all well-represented SNPs (<50% missing data). I then calculated *D*, the mean genetic distance over all possible pairs of individuals from 2 distinct sites, as a measure of pairwise dissimilarity between sites. To avoid confounding effects of within-site differences and better represent genetic differences resulting from gene flow variations, I calculated between-site dissimilarity (D_b) by subtracting the average within-site dissimilarity from *D*.

2.3 Results

2.3.1 Demographic patterns from tree rings

Tree ages were successfully estimated for a total of 607 samples (N=412 on the Kodiak Archipelago, N=195 in the Seward region on the Kenai Peninsula), evenly distributed among four canopy levels. Estimated tree ages ranged from 5 to 552 years. Medium and large canopy trees were generally younger on Kodiak Island (145 years on average) than on Afognak Island and Seward (>200 years). Ages of large canopy trees differed considerably among regions with younger large canopy trees in regions closer to the range limit (Figure 2.1b). To obtain a finer resolution of the spatial demographic patterns, I calculated the age of canopy trees at sites within regions (Figure B.1). The age of canopy trees at all Seward sites was 250 to 300 years, with overlapping standard errors. Independent data indicates that Seward sites were colonized more than 1000 years ago (Jones, 2008; Mann & Hamilton, 1995). As the age of canopy trees is relatively similar across sites in Seward, I deduce that intrinsic tree mortality rather than extrinsic disturbances prevents higher ages to be reached. Therefore, assuming identical intrinsic mortality rates on the Kodiak Archipelago, the age of canopy trees will become uninformative in regard to population establishment date when reaching these values. Within Kodiak Island, two of the southernmost sites (K3 and K5) had no trees older than 200 and 135 years and a mean age of canopy trees of 140 and 67 years, respectively. At K5, the southernmost *P. sitchensis* forest located on the Southeast coast of Kodiak Island, tree ages within canopy strata were homogenous, with all medium and large canopy trees between 40 and 60 years, and large canopy trees between 55 and 135 years. At this site, trees are short and there are no suppressed trees

growing in the understory, suggesting that this site was recently colonized. In general, medium and large canopy trees are older on Afognak Island than on Kodiak Island; however, this pattern of decreasing age of canopy trees towards the species range expansion front breaks down at the local scale. Some sites in the south have an older canopy than northern sites, especially on Afognak Island (*i.e.*, A1 and A5, Figure B.1). The large variability in mean age of canopy trees among areas at a similar latitude (*i.e.*, K3 and K4, Figure B.1) suggests that colonization on the Kodiak Archipelago occurred via patchy dispersal rather than a linear advancing wave.

A signal of canopy closure was detected in juvenile growth patterns at the site level (Figure 2.2). As expected, none of the five sites on the Kenai Peninsula (S1 to S5) showed a significant relationship between growth increment and time for either the linear or the logistic model. Narrow juvenile growth rings were around 1-mm-wide, which is consistent with growth under a well-formed canopy and no evidence of stand-level disturbances over the time represented by the tree ages. Contrasting with this pattern, large canopy trees at the four considered sites on Kodiak Island (sites K1, K2, K3, and K5) show very large juvenile growth rings (3-7 mm). Large canopy trees at all five sites on Afognak Island (sites A1 to A5) show moderately large juvenile growth rings (2-4 mm). Another strikingly different pattern is that juvenile growth patterns at all eight of these nine sites was best represented by a logistic curve or a linear model with significant negative slope, suggesting that over time juvenile growth gradually decreased at all sites, with only the latest values approaching the juvenile ring width values observed in the Seward population (sites S1 to S5). K5 is the only site where juvenile ring width does not decrease, with current values sustained above 2mm. The pattern of sustained wide juvenile rings in mature trees at K5 is consistent with a relatively young stand where all trees reaching the canopy have grown in absence of intraspecific competition. For sites where a logistic curve was the best fit, estimated dates of canopy closure (h) varied little among sites. As juvenile growth rings are narrower in old trees on Afognak Island than on Kodiak Island but juvenile ring width still decreases over time, I suggest that the current mature trees on Afognak Island established under a developing canopy. The detected decrease in juvenile growth through time would correspond to a slow increase in canopy closure. An overlap between ages of medium canopy trees and suppressed sub-canopy trees at the site level supports this hypothesis.

Finally, I computed the cumulative distribution of establishment dates of canopy trees in both the Kodiak Archipelago and Seward populations (Figure 2.3a). There was a sharp increase in the cumulative number of established canopy trees on the Kodiak Archipelago around 1700. The cumulative distribution of the Seward sample suggests that this shift is not due to intrinsic mortality in trees established before 1700. Indeed, such old trees are present in the Seward population (although few trees established before 1550 were sampled). Instead, the observed shift in age distribution on the Kodiak archipelago compared to Seward might either indicate a genuine increase in establishment rate around 1700 or an elevated extrinsic mortality of trees established before that time. However, there is no known catastrophic climatic or geological event in the region from this time, nor is there any signal of it in annual rings of trees established before 1700 on the archipelago.

2.3.2 Genetic structure and diversity in space and time

Both PCA plots (Figure B.2) and *Structure* analyses (Figure B.3) suggest that population differentiation is moderate and mainly separates Seward from the other populations. In particular, the mixed ancestry of Shuyak and Port Chatham displayed in the K=2 and K=3 *Structure* bar plots as well as their position on the PCA plot suggest that the strait separating the archipelago from the mainland does not produce any marked differentiation pattern, at least not compared to similar overland distance.

To determine how the present regional pattern of population structure evolved, I analyzed the evolution of F_{ST} over 400 years between the Kodiak Archipelago and the Seward region (Figure 2.3b). Despite large confidence intervals around estimates, there is a decrease in F_{ST} from 0.15 in 1610 to 0.12 in the mid-1700s, followed by a weak, statistically nonsignificant increase to about 0.15, the current estimate. The early decrease in differentiation could indicate relatively high gene flow from the mainland to the Kodiak Archipelago during early population establishment. A shift to local recruitment likely happened in the 1700s, putting an end to the decreasing trend in genetic differentiation. This shift is coincident with the upward shift in the distribution of establishment time of canopy trees on the Kodiak Archipelago (Figure 2.3a).

Genetic diversity decreased towards the expansion front for both allelic richness and expected heterozygosity calculated over polymorphic loci (Figure 2.4). The Seward population

had the highest allelic richness, and Kodiak Island, Afognak Island, and Port Chatham had the lowest. Interestingly, Shuyak Island has a higher allelic richness than Port Chatham suggesting connectivity of the Shuyak population with other populations, possibly from *P. sitchensis* forests outside of those sampled, or from *P. glauca* populations north of the archipelago. Heterozygosity is high everywhere but on Kodiak and Afognak Islands, suggesting a local deficit of some alleles common elsewhere at the northern range edge of the species.

To determine how quickly the Kodiak-Afognak populations acquired their current allelic diversity, I built an allele-accumulation curve (Figure 2.3c) and compared it to a null model of comparable sample sizes. I found that most alleles present in the data were acquired between 1620 and the mid-1700s, a trend confirmed not to be an artefact of sampling effects.

2.3.3 Patterns of genetic structure and diversity at the colonization front

Expected heterozygosity calculated for each sampled site using all SNPs that are polymorphic ranged from 0.11 to 0.28 across the Kodiak Archipelago (Table 2.1). There is no evidence for a latitudinal decrease within the archipelago towards the edge of the range: I calculated a correlation coefficient of 0.04 between latitude and H_e (Pearson's correlation test, $p = 0.9004$). I calculated a similar correlation coefficient value between H_e and age of medium and large canopy trees (r=0.08, Pearson's correlation test, p =0.8101), and again failed to detect any erosion of genetic diversity during successive colonization of demes at the expansion front.

To test the hypothesis that colonization leads to genetic sectors on the landscape, I computed pairwise F_{ST} among sites (Figure 2.5). Larger F_{ST} values among areas on the Kodiak Archipelago than among areas on Seward would suggest the presence of such colonization-specific mechanisms at the front. However, pairwise F_{ST} values between sites on the archipelago are very low (<0.1), similar to those observed among Seward sites. This could be due to high gene flow among sites after long-distance dispersal founding events, making genetic sectors too transient to be observable in current datasets. Alternatively, it could be due to a homogeneous pool of founding individuals across the archipelago, which could occur if all propagules came from one, already depauperate source population. To test these alternative hypotheses, I took a landscape genetics approach and calculated genetic dissimilarity among sites at different distances within the Kodiak Archipelago and at four different times, and studied the change in

dissimilarity between sites over time (Figure 2.6). Dissimilarity between sites in 1710 was significantly higher than during subsequent centuries: average dissimilarity values are an order of magnitude lower in 1810, 1910 and 2010 than in 1710. This result brings support to the hypothesis that post-founding gene flow prevented initial genetic sectors from persisting.

2.4 Discussion

2.4.1 Demographic and genetic patterns of colonization

Studying the evolution of long-lived organisms such as temperate tree species is challenging because of their typically long generation time. This study shows that the concomitant use of tree ring and genetic data can turn the inconvenience of these life history characteristics into an opportunity to accurately reconstruct the demographic and genetic history of colonization over five centuries. By applying this combination of methods to several regions and sites within regions at the expansion front of *P. sitchensis*, I was able to describe demographic and neutral genetic patterns of forest establishment at the regional and local scale. I established a link between genetic diversity and spatial colonization patterns by showing that both trends of decreasing diversity and decreasing time since forest establishment towards the expansion front break down at the local scale. This provides insights into the effects of dispersal patterns on neutral evolution during range expansion. It also highlights the importance of relating the geographic scale of study to the dispersal abilities of the studied organism when testing for evolutionary trends commonly observed during range expansions. I established a second link between temporal trends in demography and genetic structure of *P. sitchensis* populations over the last five centuries at the expanding range limit: a shift from decreasing to stagnating differentiation between the establishing Kodiak population and the Seward population on the Kenai Peninsula coincides with a marked increase in successful establishment rate on the Kodiak Archipelago in the 1700s. This suggests that gene flow from continental populations was predominant in early stages of establishment, until local recruitment became the major mechanism of population growth. Most of the allelic richness on the Kodiak Archipelago was acquired during initial stages of population establishment, and the high levels of local gene flow

in later stages homogenized the genetic structure of the Kodiak population, buffering founder effects at the local scale and maintaining them at the regional scale.

2.4.2 Founder and Allee effects

Trees established as early as 1516 were sampled on the archipelago and local recruitment only appears to have become significant in the 1700s, highlighting the existence of a lag of several centuries in local recruitment. Although these results do not provide direct evidence of density-dependent population success, they echo findings of Lesser et al. (2013), who identified Allee effects in early stages of forest establishment in a *Pinus ponderosa* stand through reliance on long-distance seed and pollen dispersal for the first few centuries of population growth. Alternatively, a long period of unfavourable climate could also reduce population growth for several decades, especially in species reproducing through masting: a consistently poor climatic environment could reduce the frequency of masting years. The Little Ice Age, a general period of colder northern hemisphere climate between 1300 and 1850, is known to have been particularly pronounced at high latitudes and could therefore have forced a slow population growth and even a more pronounced founder effect than expected on the Kodiak Archipelago.

In *P. sitchensis*, wind-borne pollen is likely to be the main vector of genetic material from distant sources, although the abilities of seeds to travel by air and ocean surface currents are not well understood. The shift in differentiation trends in the mid-1700s coincides with the start of a plateau in the allele accumulation curve for the Kodiak-Afognak region. This result can also be related to Roques et al. (2012), who found through models of colonization waves that Allee effects prevent the erosion of genetic diversity along colonization routes: populations at the front accumulate genetic diversity during establishment through reliance on populations behind the expansion front. If a lag in local recruitment through Allee effects is common in populations at the edge of forest tree species, mechanisms described in Roques et al. (2012) could partly explain why many studied forest expansions do not show a classic decrease in diversity at the expansion front. In the case of *P. sitchensis*, when populations had a low density of mature trees they produced a relatively small pollen cloud compared to external sources, until large mature trees became the main producers of pollen locally present. However, it is worth noting that expected heterozygosity over polymorphic sites remained low on Kodiak and Afognak Islands,

although the expansion process on the Kodiak Archipelago did not result in local erosion of allelic richness. This suggests that although allelic diversity was recovered largely during colonization, many alleles remain at low frequencies in the newly established population. The genotype of early colonizers might therefore have a long-lasting influence in the establishing population.

2.4.3 Genetic structure at the expansion front

P. sitchensis started establishing at most sites on the Kodiak Archipelago before 1700. Evidence for continental sources dominating gene flow prior to the mid-1700s indicates that long-distance dispersal from the mainland played an important role in initial recruitment. Models of colonization through patches from long-distance dispersal show that genetic sectors are expected to arise (Hallatschek & Nelson, 2010). However, I found no evidence for spatial genetic sectors in the current sample. I propose two explanations for this lack of spatial genetic structure. First, it seems that most founders colonized Kodiak and Afognak Islands from just a few source populations that were already somewhat depleted in alleles. Indeed, it seems that the low allelic richness observed on a regional scale on the Kodiak Archipelago could be due to the low levels of allelic richness of source populations at the tip of the Kenai Peninsula: levels of allelic richness on Kodiak and Afognak Islands are similar to those in Port Chatham, the closest continental population sampled. Low levels of allelic richness in Port Chatham may be explained by lower population sizes due to a fragmented landscape, as the region is highly mountainous and suitable habitat is restricted to narrow valleys between ice-capped mountains and numerous sinuous fjords. Also, as allelic richness is higher on Shuyak Island than in Port Chatham, multiple source populations might contribute to the maintenance of genetic diversity in regions at the expansion front. The second factor likely to explain the absence of genetic sectors is high gene flow within the Kodiak Archipelago during the last 250 years. I found a decrease in mean pairwise dissimilarity among sites at the expansion front between 1700 and subsequent centuries, suggesting that initial spatial patterns of genetic structure from founding individuals did not persist or develop further due to high levels of subsequent gene flow across the archipelago.

2.4.4 Demographic estimates of establishment times: power and limitations

Estimates of establishment time, age of canopy trees and time of canopy closure all showed a clear demographic signal of population establishment on Kodiak Island, and a less clear signal for Afognak Island. As these estimates were also calculated for the Seward population, known to have established several thousand years ago, the power and limitations of these estimates are delineated. The age of canopy trees is only informative for about 300 years, and the estimated time of canopy closure confirmed the recent nature of forests on Kodiak Island and — to a lesser extent — Afognak Island. The complementarity of these two estimates is best illustrated with the Afognak Island site A5: this site is similar to Seward sites in age of canopy trees (Figure B.1) but juvenile growth rings show a signal of increasing canopy density throughout the 19th century, a pattern not observed in Seward (Figure 2.2). Although time of canopy closure is useful for inferring the absence of a closed canopy when the first trees established, the temporal change in juvenile radial growth couldn't be modeled by a logistic curve, or was better fitted by a linear model. This can be due to the initiation of intraspecific competition suppressing growth being too recent, or to spatial heterogeneity within stands. In addition, I found little variability between sites in estimates of time of canopy closure. The parameter *h* from the logistic curve modelling may not be the most informative measure of establishment time. Visually inspecting juvenile ring width profiles over time (Figure 2.2) might provide more information about stand establishment than extracting a single value from these profiles.

2.4.5 Implications for long-lived wind-pollinated species

Results suggest that the evolutionary potential of wind-pollinated tree species is more likely to be limited by slow demographic growth than by slow accumulation of genetic diversity. In spite of a slow initial population growth, allelic richness recovered during this period up to levels comparable to nearby source populations. In addition, geographic barriers to gene flow are weak despite the studied population being isolated from the continent by a 70 km-wide ocean strait. The demographic lag observed in this and other studies of tree populations suggests an Allee effect, whereby the reproductive ability of establishing trees is limited firstly by a long phase of juvenile growth, and secondly by a higher dependence on foreign pollen fertilizing local mature trees when population densities are low. This potential Allee effect could keep a colonizing tree

population in a vulnerable state and contribute to the migration lag of tree species tracking their suitable niche space, especially in the context of rapid anthropogenic climate change. However, such a lag is also likely to be at the origin of an efficient recovery of genetic diversity after a founding event, as several studies have shown a predominance of pollen of foreign origin during early population establishment (Hampe et al., 2013; Lesser et al., 2013; Pluess, 2011). In both cases, management through planting trees from diverse, carefully selected provenances could accelerate successful establishment and adaptation in populations of conservation concern (Aitken & Whitlock, 2013). In general, understanding demographic processes during range shifts and their effect on evolutionary potential is necessary as climate change is shifting species' suitable niches nearly everywhere on the planet. This chapter illustrates that dispersal potential and temporal patterns of population growth are important factors influencing population expansion and adaptation. The effects of other mechanisms such as hybridization, competition, and natural selection also need to be assessed in order to predict or help species movements in response to a rapidly changing world.

2.5 Tables and Figures

Table 2.1 Nomenclature, description, and diversity estimates of sampled sites. Size=sample size; H_e=population-level heterozygosity averaged across loci; mean canopy age=average tree age across large canopy and medium canopy trees, in years.

site	region	latitude	longitude	size	H_e	mean canopy age
S1	Seward	60.253	-149.357	37	0.283	294
S2	Seward	60.213	-149.369	36	0.29	288
S3	Seward	60.19	-149.558	48	0.268	285
S4	Seward	60.107	-149.354	48	0.251	319
S5	Seward	60.06	-149.444	26	0.202	266
PC	Afognak	59.224	-151.703	15	0.123	NA
Sh	Afognak	58.563	-152.555	15	0.183	NA
A1	Afognak	58.298	-152.329	24	0.119	206
A2	Afognak	58.231	-152.475	37	0.155	223
A3	Afognak	58.209	-152.585	12	0.12	315
A4	Afognak	58.145	-152.433	25	0.121	239
A5	Afognak	58.118	-152.541	86	0.15	310
K1	Kodiak	57.845	-152.422	52	0.134	203
K2	Kodiak	57.638	-152.437	49	0.134	147
K3	Kodiak	57.619	-152.334	50	0.123	207
K4	Kodiak	57.617	-152.219	31	0.136	201
K5	Kodiak	57.428	-152.344	48	0.13	66

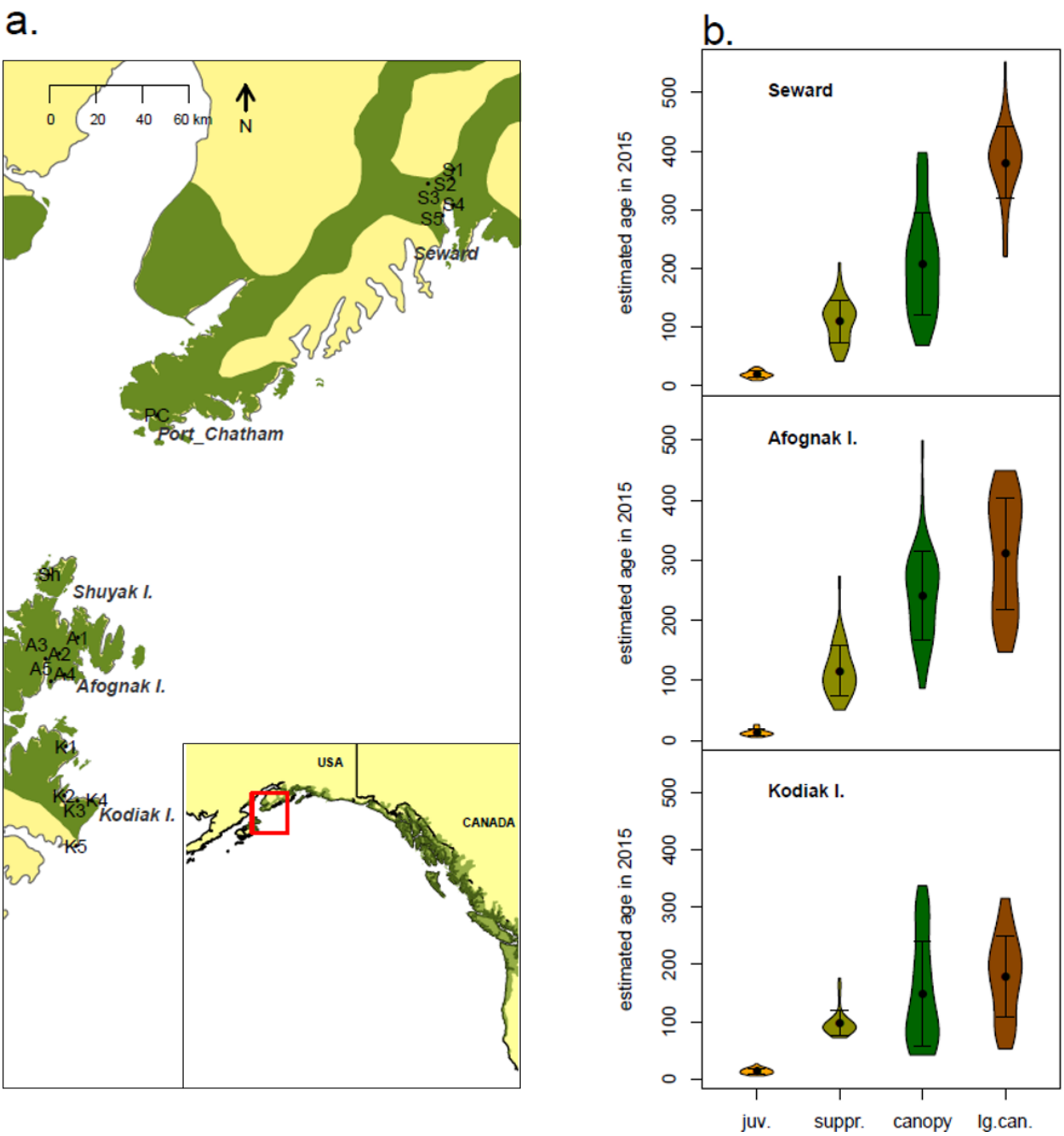


Figure 2.1 Sampled sites and tree ages in south-central Alaska. a. Map of sites with the whole range of *P. sitchensis* (green area) in inset. b. Violin plots of individual tree ages for each canopy structure level within each region sampled, with mean and standard deviation displayed in black. juv.=juvenile tree; suppr.=sub-canopy tree; canopy=medium canopy tree; lg.can.=large canopy tree

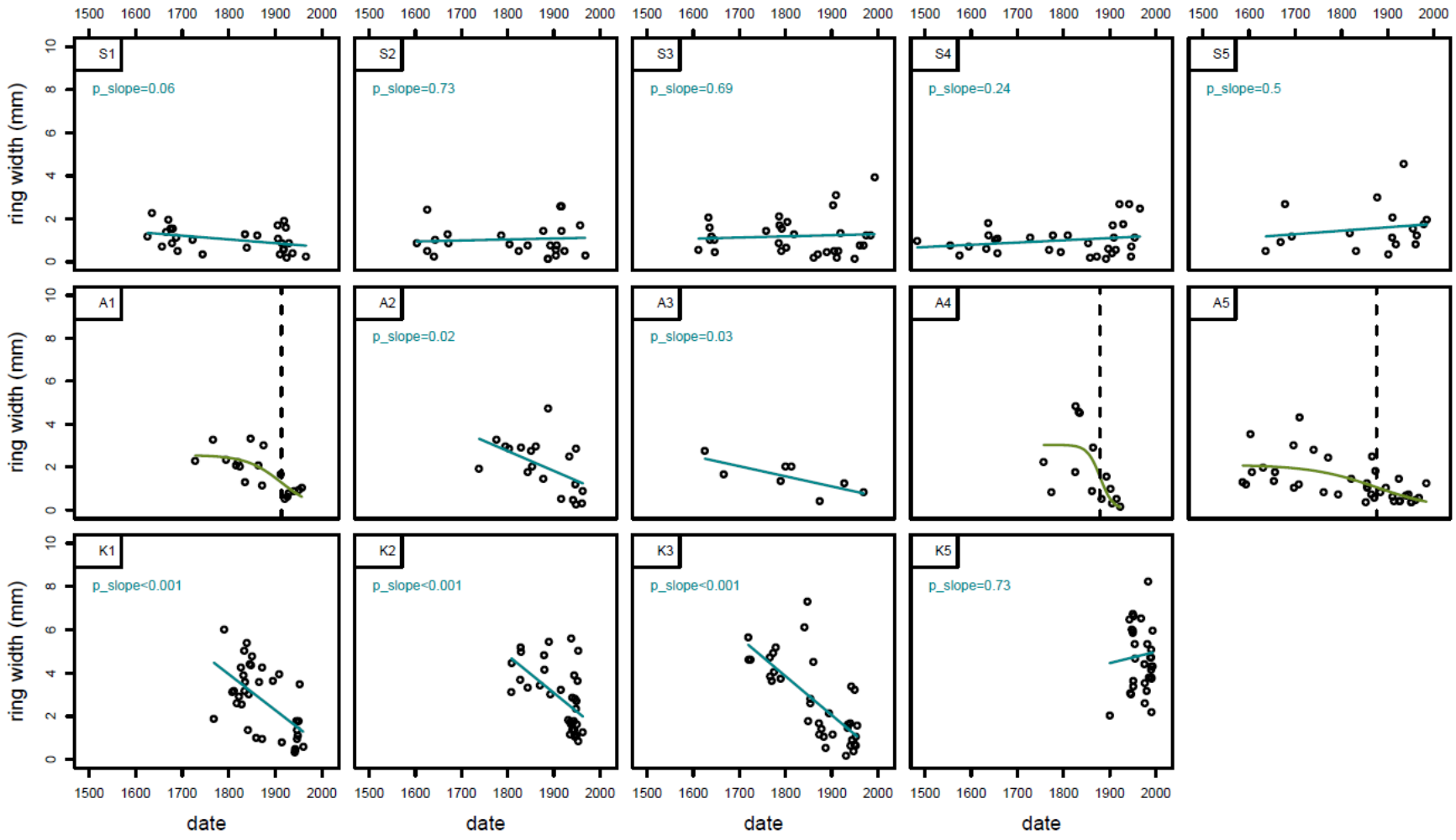


Figure 2.2 Individual annual radial growth between years 10 and 20 vs. date of establishment. The best model among linear and logistic models is displayed in blue with the p-value of the slope (linear model) or in green (logistic model). In the latter case the estimated time of canopy closure (h) is represented by a dashed line.

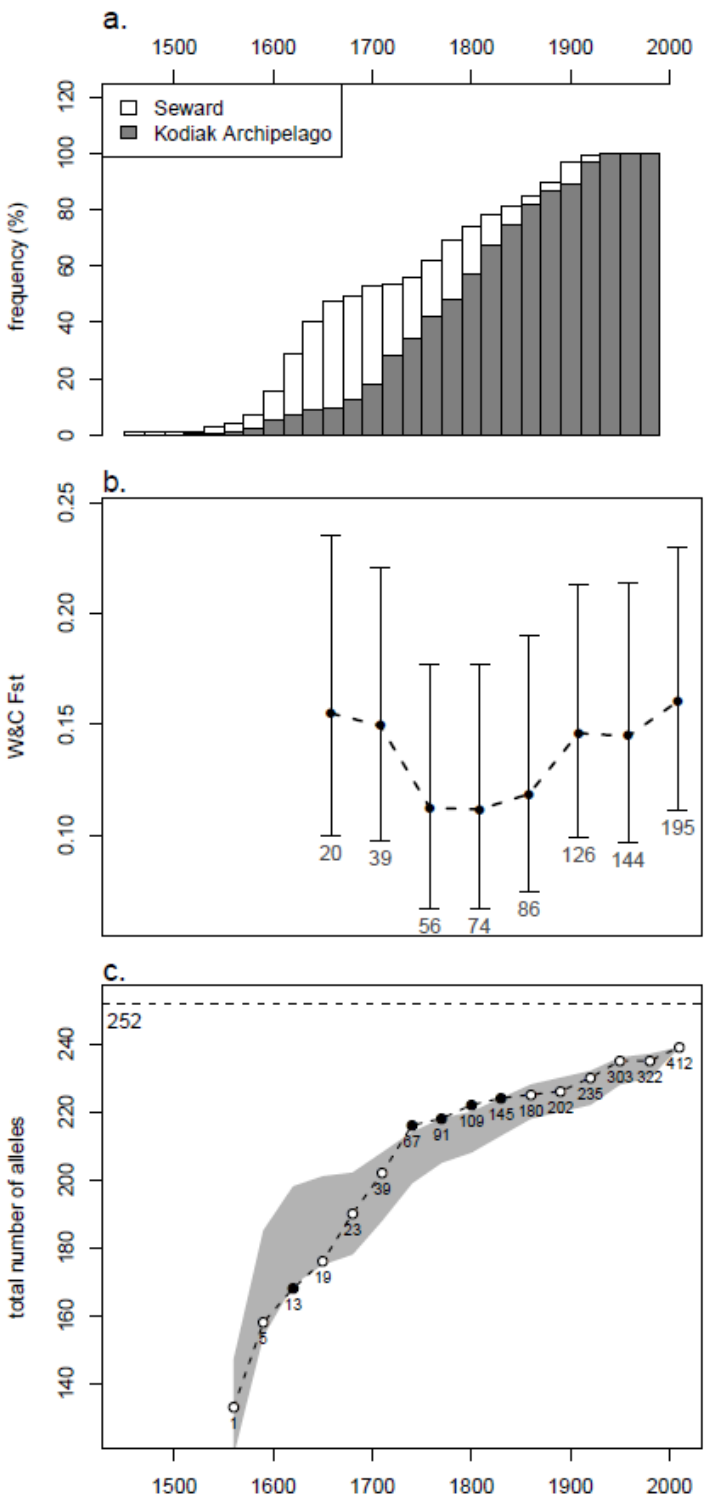


Figure 2.3 Demographic and genetic changes over time on the Kodiak Archipelago in relation to the Kenai Peninsula. a. Cumulative distribution of establishment dates for canopy trees for the Kodiak Archipelago and the Seward region. b. Temporal change in F_{ST} between the Kodiak Archipelago and Seward. F_{ST} values are calculated for the cumulative sample (each sample associated with a date is made of all individuals alive at this date). Error bars represent confidence intervals from 1000 bootstraps. The number of individuals per population is indicated below error bars. c. Allele accumulation curve for the Kodiak Archipelago with 95% interval band from 1000 random permutations (grey). Numbers under data points correspond to cumulative sample sizes. Data points with values outside the 95% confidence band are represented with filled circles. The top dotted line correspond to the maximum number of alleles in the whole sample of 639 individuals.

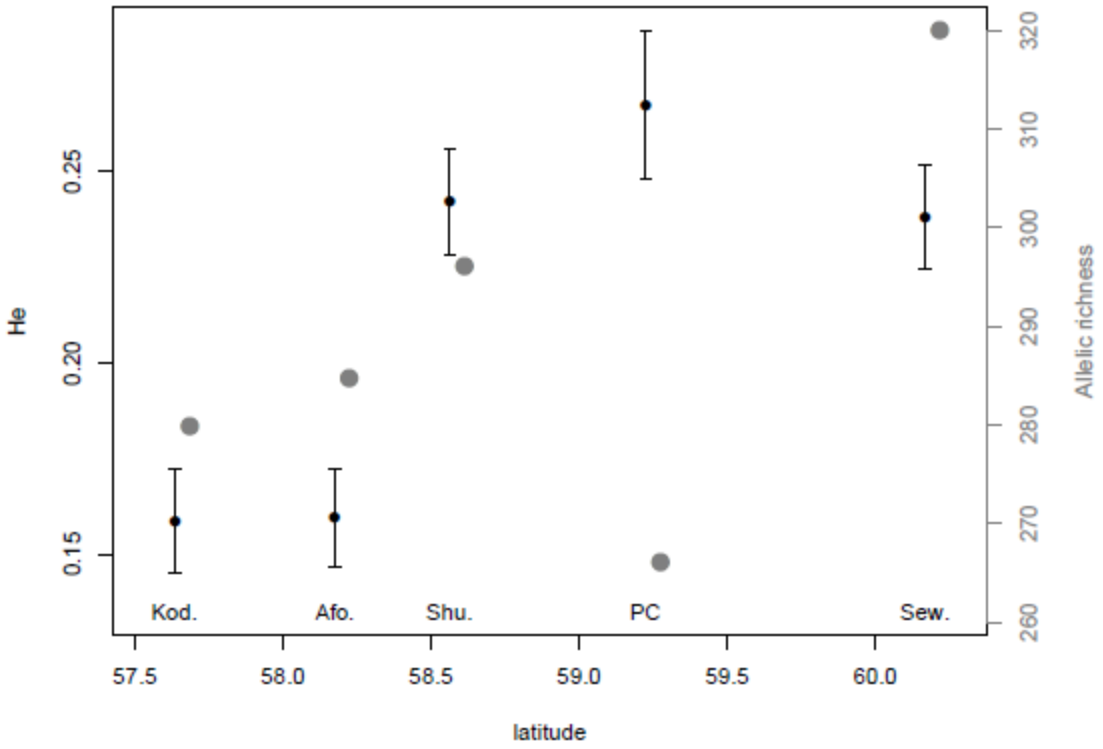


Figure 2.4 Regional expected heterozygosity over variable SNPs (black) and estimated allelic richness (grey) vs. latitude or regions. Error bars on He are standard errors of the mean. Kod.=Kodiak Island, Afo.=Afognak Island, Shu.=Shuyak Island, PC=Port Chatham, Sew.=Seward.

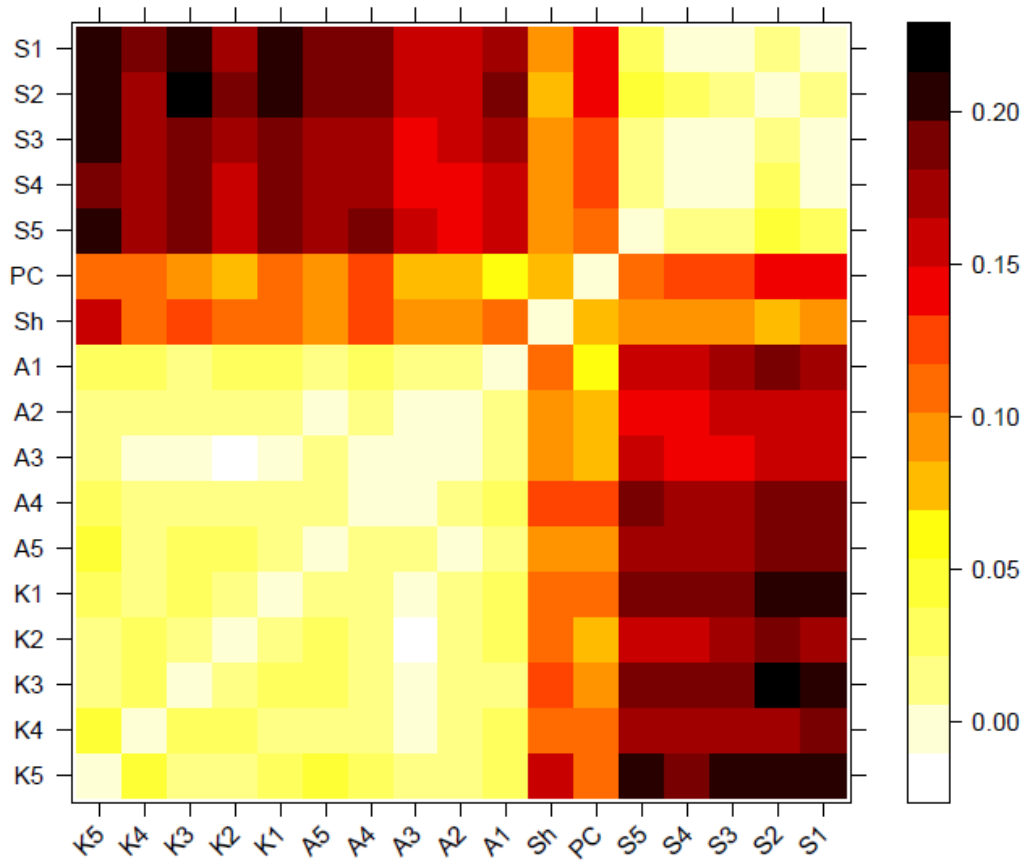


Figure 2.5 Heat map of pairwise F_{ST} values between sampled sites.

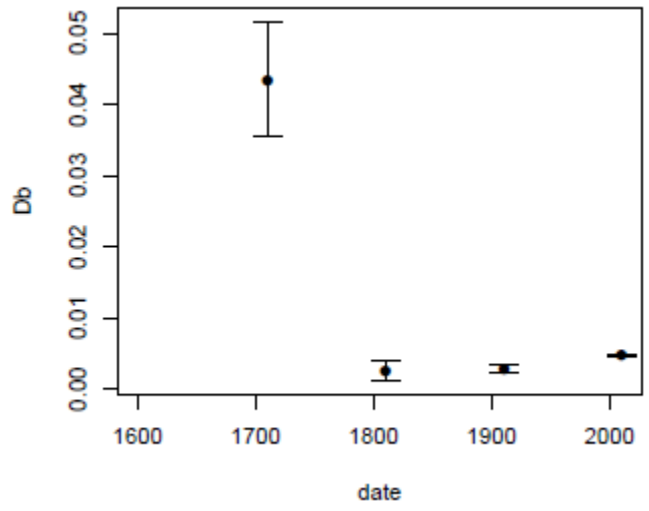


Figure 2.6 Temporal change in mean pairwise dissimilarity between sites on the Kodiak Archipelago. Error bars represent standard error of the mean.

Chapter 3: Patterns and effects of *Picea sitchensis* admixture with a closely related species during range expansion

3.1 Introduction

Chapter 2 linked dispersal and demographic patterns to evolutionary outcomes in a spatially expanding tree population. Extrinsic forces can also play a major role in the demographic and evolutionary trajectory of an expanding species at its range limit. Different climatic conditions as well as biotic interactions (competition, parasitism, and predation) shape the spatial distribution of a colonizing population and its persistence in the new environment. If hybridization is possible between a local species and the invading species, it can fundamentally alter the genetic makeup of the colonizing population and its adaptive capacity. Examples of enhanced hybridization during range expansion are numerous (reviewed in Currat et al., 2008), with evidence for adaptive introgression and speciation during colonization (Rieseberg et al., 2007). Although natural selection is often put forward as an obvious explanation to why hybridization during range expansion is so common, Currat et al. (2008) have also shown that neutral processes can be sufficient to explain introgression of a local species into the genome of an invading population. Indeed, demographic mechanisms of colonization can drive introgression of the local species' genes into the genome of the introduced species, at rates up to 100% if the interbreeding rate exceeds 10%. Although this dramatic result can be observed under a variety of demographic conditions including a wide range of interbreeding rates and population densities (Currat & Excoffier, 2011), Amorim et al. (2017) have shown that frequent long-distance dispersal events can limit introgression and even completely prevent it if the expansion wave front is targeted. As introgression between *Picea glauca* and *Picea sitchensis* is common in the central and northern range of *P. sitchensis*, the expanding *P. sitchensis* population on the Kodiak Archipelago provides an opportunity to develop empirical knowledge about the interplay between range expansion and introgression. Although no species able to interbreed with *P. sitchensis* was present on the archipelago during its colonization, the expanding range edge is within the reach of pollen dispersal from nearby *P. glauca* and hybrid populations north and east

of it (Figure 3.1). The goals of this chapter are to characterize the spatial and temporal patterns of introgression between an expanding *P. sitchensis* population and neighbouring *P. glauca* populations, to determine whether it originates from recent hybridization events, and to assess the potential role of natural selection in the formation of the observed patterns.

There is a rich history of introgression studies in the context of range expansion in tree species, especially conifers (Du Fang et al., 2009). There are several reasons for this. First, members of the Pinaceae are organisms of choice to study the interplay between gene flow levels and introgression within the genome. This is because different genetic material can be inherited either biparentally via nuclear DNA, or from one parent (maternally for mitochondrial DNA, paternally for chloroplast DNA). These different modes of inheritance and their associated levels of gene flow are informative to understand determinants of introgression and the demographic history and phylogeny of species complexes (Bouillé et al., 2010; Du Fang et al., 2009). Second, postglacial colonization often involves secondary contact between sister species re-expanding towards the pole after a glacial period, resulting in repeated introgression events and the maintenance of weak reproductive barriers (Jaramillo-Correa et al., 2009). At its northern expanding range edge, *P. sitchensis* coexists and hybridizes with its sister species *P. glauca* at multiple contact zones on the Kenai Peninsula (Boucher & Mead, 2006) and in Iniskin Bay area. Although there is no known intrinsic reproductive barrier between the two species, they occupy distinctive environmental niches. This explains the relative stability of hybrid zones and the preservation of species on either side of them in the hybrid zones of British Columbia (Hamilton & Aitken, 2013). Climatic niche differences are most strongly associated with differences in the annual amount of precipitation as rain and low winter temperatures (Hamilton et al., 2013a; Hamilton & Aitken, 2013). The fact that *P. glauca* individuals and *P. glauca* - *P. sitchensis* hybrids can be found outside of their usual climatic niche due to demographic mechanisms specific to range expansion prompts questions about their viability in the new environment. Therefore, after identifying patterns and causes of introgression at the expansion front of *P. sitchensis*, this chapter addresses the effect of climate on individuals with different levels of ancestry in the recently established *P. sitchensis* population, in comparison to populations further from the migration front.

Chapter 2 illustrated how past forest density and canopy closure could be tracked using changes in individual ring-width patterns over time. Variations in annual tree ring width are also a powerful tool in determining the effect of climate on tree growth. When paired with temporal series of annual variation in climatic variables, time series of tree ring widths can help identify the set of conditions that limit radial growth for a particular tree species in a forest stand. The main determinants of climate sensitivity for a forest tree population are usually measured with dendroclimatic methods and mostly depend on species and geographic location (Babst et al., 2013; Büntgen et al., 2007; Martin-Benito & Pederson, 2015). However, a recent study of an *Abies alba* population showed that Carpathian populations originating from different glacial refugia responded to different climatic drivers of radial growth, suggesting that radial growth patterns could differ among more subtle phylogenetic resolutions than the species level (Bosela et al., 2016). As studies coupling dendrochronological and genetic information have started to appear in recent years, it still remains unclear how genetic characteristics of trees can influence their radial growth responses to climate. Here, I assess the influence of interspecific admixture on radial growth patterns and their relationship with seasonal climatic variables.

Most of the tree ring and genetic data analysed in this chapter comes from the same material as in Chapter 2. I first complemented the *P. sitchensis* genetic data obtain by GBS with genotypes from a *P. glauca* outgroup to quantify the extent of genetic admixture of *P. sitchensis* populations with *P. glauca* at the northern range limit of the species. I especially ask whether introgression was enhanced or suppressed during colonization of the Kodiak Archipelago. I then develop hypotheses to explain the observed patterns of admixture. I attempt to determine whether admixed genotypes originated from distinct hybridization events on the Kodiak Archipelago or from more ancient mechanisms (introgression from past secondary contact in the continental *P. sitchensis* source populations or incomplete lineage sorting). Finally, I use annual variations in radial tree growth to test whether admixed and pure *P. sitchensis* trees in the establishing population show different growth patterns that translate into different responses to climate.

3.2 Materials and methods

3.2.1 Geographic location, sampling and genotyping

I used the same *P. sitchensis* samples as in Chapter 2, including 639 individuals sampled across the northern species' range on the Kodiak Archipelago and the Kenai Peninsula. In addition to these *P. sitchensis* samples, needles from 30 mature *P. glauca* mature trees were collected at Denali Park Village, Alaska (63.719136 N ; 148.812888 W) for genotyping. This sample was selected to serve as an outgroup to identify *P. glauca* genetic ancestry in the *P. sitchensis* sample, and is believed to be far enough from the *P. sitchensis* range to be devoid of hybrids. It is also part of the same Alaskan phylogenetic group as *P. glauca* populations parapatric with my *P. sitchensis* sample from the Kenai Peninsula (Anderson et al., 2006). The 669 collected individuals were genotyped using genotyping-by-sequencing (GBS). The procedure is described in Chapter 2 and Appendix A. The 6,644 high-quality SNPs obtained after sequencing, genotype calling and quality filtering were further filtered to remove singletons. I also removed loci that had less than 40% representation in any of the three main populations in the sample (Kodiak Archipelago, Seward, and Denali). When several SNPs were present on the same read (*i.e.*, less than 100bp apart), I retained only the first one. The resulting 338 polymorphic SNPs were used in subsequent genetic structure analyses.

3.2.2 Defining a hybrid index and reference groups

I performed Bayesian clustering using *Structure* (Falush et al., 2003) with the number of clusters *K* varying from 2 to 6. Run lengths were set to 10^5, with an initial 10^3-burnin period. I implemented 3 replicate runs for each value of *K* to ensure the consistency of results. I used the independent allelic frequency model and the admixture ancestry model for all runs. These analyses were set to run unsupervised, *i.e.*, without prior population information. The dominant cluster in the Denali sample was identified and hybrid index (HI) was defined as the proportion of this cluster in any given individual tree. HI therefore represents the proportion of *P. glauca* ancestry in a given individual. I then defined species reference groups by selecting all individuals with HI<0.01 (*P. sitchensis* reference) and HI>0.99 (*P. glauca* reference). These groups were used to identify diagnostic markers and calculate interspecific heterozygosity.

3.2.3 Interspecific heterozygosity and hybridization patterns

To infer hybridization patterns that gave rise to the currently observed admixture levels on the Kodiak Archipelago, I examined the relationship between interspecific heterozygosity (int.h) and hybrid index (HI) for each individual. For an individual, interspecific heterozygosity is defined as the heterozygosity over diagnostic sites (sites with one different, fixed allele in each parental species). When represented jointly with HI values, int.h values are useful to assess the genetic structure of hybrid populations (Fitzpatrick, 2012). With a large enough set of ancestry informative markers, interspecific heterozygosity makes it possible to assess whether hybrids with intermediate HI values originate from a recent hybridization event (relatively high int.h) or are the result of more ancient introgression or ancestral polymorphism at the genetic loci considered (low int.h). Indeed, linkage between ancestry-informative alleles in the first generation of hybridization would translate into high int.h values in first- and second-generation hybrids. This linkage would be broken down by recombination in subsequent generations. Shared polymorphisms inherited from ancient secondary contact would therefore be characterized by much lower int.h values. Similarly, if alleles predominant in one species have been kept at low frequencies in the other species through incomplete lineage sorting, it is unlikely that a set of individuals would be heterozygous for all or most of them; the average int.h should therefore also be low if polymorphisms originate from incomplete lineage sorting.

Here, as there are no fully diagnostic SNPs in the sample, I selected all SNPs with $F_{ST}>0.7$ between the two reference groups (see section 3.2.3) as ancestry informative markers. Interspecific heterozygosity was calculated as the average heterozygosity over these selected SNPs, for each individual with a HI value between 0.01 and 0.99. The joint distribution of int.h and HI was then represented in a triplot (Fitzpatrick, 2012). As the markers used here are not fully diagnostic, it is not easy to discriminate between the two sets of hypotheses of recent hybridization (within the last 2 generations) and ancestral polymorphism/ancient introgression. Therefore, to help interpret the observed genetic structure of hybrids, I simulated F1, F2 and F1 backcrosses with each parental group (N=200 for each progeny) using the *adegenet* R package (Jombart, 2008), and selecting the two reference groups as parental genotypes. I represented the joint distribution of int.h and HI for these simulated progenies on the triplot together with results

from observed hybrids. HI for simulated data was calculated in a similar fashion to HI for the observed data: for each simulated cross, the 200 simulated genotypes and the 2 reference groups were input in *Structure* with the same run parameters as for the observed data (see section 3.2.3), but with prior population information for the two reference groups. Interspecific heterozygosity was calculated for both observed and simulated data using the R package *introgress* (Gompert & Buerkle, 2010).

To characterize temporal changes in hybrid genotypes I also assessed interspecific heterozygosity in relation to individual dates of establishment (inferred in Chapter 2).

3.2.4 Dendroclimatic analysis

To assess the climatic determinants of radial growth in pure Sitka and admixed genotypes, correlations between tree ring chronologies and local climate variables need to be established. To create chronologies, I used the *dplr* R package (Bunn, 2008). Tree ring processing methods prior to detrending are described in Appendix B. I individually detrended crossdated ring-width series from all medium and large canopy trees by fitting a cubic smoothing spline with a 50% frequency response cut-off and a smoothing wavelength of 50 years. This process removes long-term growth trends. This step is necessary to filter out age-related and stand dynamics effects on radial growth. I then applied an autoregressive time series model to remove all but the high-frequency signal. The output can be correlated with interannual climatic variations (Cook & Kairiukstis, 1990). Four standard chronologies were created by averaging detrended ring-width series for trees stratified by region (Seward and the Kenai Peninsula) and genetic ancestry (Sitka parental group and hybrids with $int.h \geq 0.5$), using the *dplr* R package (Bunn, 2008). I obtained annual seasonal values of temperature and precipitation for the Kenai Peninsula and the Seward region using *ClimateWNA* (Wang et al., 2012). The chronologies and climate records were trimmed to a period between 1940 (earliest records from local weather stations) and 2010 (last date of high precision measurement for tree rings). Correlation analyses were performed using custom functions modified from the *dcc* function in the R package *bootRes* (Zang & Biondi, 2013). This function calculates correlation coefficients between chronologies and climate variables with associated confidence intervals calculated from a built-in bootstrap method involving resampling among years. To establish statistical significance among admixture levels

within regions, I developed an alternative bootstrap procedure involving the creation of 1000 chronologies by randomly selecting subsamples of tree ring series of each admixture class with replacement.

3.2.5 Non-parametric growth patterns analysis

I complemented the dendroclimatic analysis with a cluster analysis of individual ring-width series following single detrending. As the level of *P. glauca* - *P. sitchensis* admixture of an individual tree might influence aspects of radial growth that are independent of climatic influences, I studied differences in tree ring sequence patterns without prior assumptions on the factors driving growth variation. To do this, I performed a single detrending of individual ring-width series using a cubic smoothing spline with a 50% frequency response cut-off and a smoothing wavelength of 70 years. Although this detrending process dampens some medium- and long-term growth trends potentially affected by the level of admixture, it is a more conservative representation of individual radial growth variations than double detrending using the 50-year smoothing spline and autocorrelative method described in section 3.2.4. Applying this standardization is especially important in the case of the Kodiak Archipelago sample, where substantial tree density and canopy changes have occurred at different sites and times during the studied period (Figure 2.2). I computed a dissimilarity matrix of all detrended ring-width series truncated to the period 1950-2010 using a dissimilarity index developed by Chouakria and Nagabhushan (2007). This index takes into account both absolute values and direction of change in values from one year to the next by combining a conventional distance measure (here the Euclidean distance D) with a temporal correlation index *CORT*. If $S_1=\{u_1,u_2,\ldots u_p\}$ and $S_2=\{v_1,v_2,\ldots v_p\}$ are two tree ring index series, then

$$CORT(S_1,S_2)=\frac{\sum_{i=1}^{p-1}(u_{i+1}-u_i)(v_{i+1}-v_i)}{\sqrt{\sum_{i=1}^{p-1}(u_{i+1}-u_i)^2}\sqrt{\sum_{i=1}^{p-1}(v_{i+1}-v_i)^2}}$$

In the calculation of the dissimilarity index D, the relative importance of CORT over the Euclidean distance d_e is modulated by an inverse exponential function with a single tuning parameter, s.

$$D(S_1, S_2) = d_e . \frac{2}{1 + e^{s.CORT}}$$

After a few exploratory analyses, I set *s*=2, which results in CORT contributing 76.2% of the final dissimilarity index *D*. Hierarchical clustering was performed on the resulting dissimilarity matrix using the Ward agglomeration method in the *hclust* R package.

3.3 Results

3.3.1 Population structure and hybrid index

Posterior probability values of *Structure* models with different numbers of clusters form a plateau starting at K=3 (Figure C.1). The three clusters generally correspond to Denali, Seward, and the Kodiak Archipelago (Figure C.3). This is consistent with *Structure* analyses results in Chapter 2, which used the same sample excluding the Denali population and supported two clusters that primarily separated Seward and the Kodiak Archipelago. Runs with K=4 generally showed the same results as K=3 runs, with a low-frequency fourth cluster that is absent in most individuals (Figure C.4). Using the percentage of the major cluster in Denali in each individual as a measure of their *P. glauca* ancestry seems like a legitimate choice. However, a striking result is the apparent mixed ancestry of the Denali sample: half of its individuals harboured *P. sitchensis* ancestry, up to 50%, for all values of K tested in sampled regions (Figure 3.2, top panel) and sites (Figure 3.2, bottom panel). This could be due to incomplete lineage sorting and/or introgression after secondary contact. The latter process is most likely involved, as individuals with 50% of *P. sitchensis* alleles would be unlikely to exist solely under the incomplete lineage sorting hypothesis.

Focusing on the major cluster in the Denali population, K=2 runs generally show the same structure as K=3 runs for the Kodiak Archipelago: on Afognak and Kodiak Islands, a minority of individuals (7.5% and 1.8% respectively) harbour more than 10% of *P. glauca* genetic ancestry, with wide-ranging proportions in both regions (up to 67%). Conversely, K=2 and K=3 analyses showed completely different results for the Seward population. K=2 analyses

characterize the Seward sample as a group of individuals that nearly uniformly display about 30% *P. glauca* ancestry (Figure C.2), whereas K=3 analyses result in Seward being represented by its own cluster with a slightly higher proportion of trees with *P. glauca* ancestry than on the Kodiak Archipelago (Figure C.3). From here on, individual hybrid index is defined as the percentage of ancestry from the major cluster in Denali in the K=3 *Structure* analysis. Figure 3.2 displays the geographic distribution of hybrid index and suggests a spatial decrease in number of admixed individuals towards the *P. sitchensis* expansion front.

3.3.2 Hybridization patterns

The *P. glauca* ancestry observed in *P. sitchensis* trees on the Kodiak Archipelago could have originated from several non-exclusive mechanisms: hybridization events with pollen from nearby *P. glauca* sources during colonization, more ancient hybridization involving ancestors on the Kenai Peninsula, or incomplete lineage sorting. None of the 338 SNPs in the datasets were truly diagnostic between the two species groups, with most loci showing low F_{ST} values (Figure 3.3a). This likely indicates the preservation of ancestral polymorphism between the two species, either through incomplete lineage sorting or extensive secondary contact. To determine whether recent hybridization occurred, I examined the relationship between hybrid index and interspecific heterozygosity in a triplot. All hybrids with high hybrid index on the Kodiak Archipelago showed high interspecific heterozygosity values (>0.5), illustrating very little to no recombination between ancestry informative loci since hybridization occurred (Figure 3.3b). The position of simulated hybrids is informative to interpret the distribution of observed hybrid genotypes in the bidimensional space of the triplot. Five of the hybrids sampled on the Kodiak Archipelago overlap with the distribution of simulated F1 and one overlaps with the distribution of simulated F1 backcrossed to *P. sitchensis*. Six hybrids are situated between the F1 and F1 backcrossed to *P. sitchensis* distributions. One hybrid is situated between the F1 and F1 backcrossed to *P. glauca* distributions. None overlaps with simulated F2s or would overlap with more advanced generation hybrids, as such groups would be distributed in the lower part of the triplot (Fitzpatrick, 2012). All hybrids are located close to the “legs” of the triangle, which correspond to the theoretical maximum interspecific heterozygosity for the corresponding hybrid index value, assuming diagnostic genetic markers. This suggests that high-level hybrids on the

archipelago originated from interbreeding between two nearly pure parental genotypes or between a F1 hybrid genotype and a pure or nearly pure parental genotype. A symmetric pattern is observed in the Denali sample, for which hybrids are also close to the maximal interspecific heterozygosity given their hybrid index value. This confirms that the high hybrid index values observed in the Denali sample are mostly due to hybridization in the last few generations. The fact that loci used for the calculation of interspecific heterozygosity are not fully diagnostic did not allow for complete differentiation between different hybrid-class equivalents and explains the presence of data points outside of the triplot. However, as the vast majority of trees on the Kodiak Archipelago have no *P. glauca* ancestry and as pollen disperses typically much further than seeds, the most likely scenario leading to the observed patterns is that hybridization occurred between local mother trees (of hybrid or pure *P. sitchensis* genotype) and pollen from a nearby *P. glauca* population.

3.3.3 Distribution of hybrid index in time and forest structure

To determine the temporal evolution of admixture during colonization of the Kodiak Archipelago, I examined the distribution of hybrid index and interspecific heterozygosity in time over the last four centuries of colonization history captured by the data (Figure 3.4). None of the trees established on Afognak Island before 1670 harbour any *P. glauca* ancestry. Hybridization with *P. glauca* pollen on the archipelago occurred exclusively between the late 1600s and early 1800s. All trees established later had hybrid index estimates below 0.06. On Kodiak Island, hybrids that are amongst the earliest samples in the early 1700s are mostly *P. sitchensis*. Hybridization events during the 1800s and early 1900s resulted in a few individuals with high hybrid indices. For both areas, one would expect to observe intermediate-level recombinant hybrids in cohorts subsequent to hybridization events. However, such individuals are mostly absent from the sample. This result is striking when considering the distribution of hybrids across canopy levels on the archipelago compared to the mainland. At all sites, four canopy levels (large canopy tree, medium canopy tree, understory mature tree, juvenile tree) were sampled as evenly as possible (Chapter 2). Whereas hybrids with $0.1 < HI < 0.9$ are homogenously distributed among canopy levels in the Seward population, they are only present as canopy trees on the Kodiak Archipelago, and are absent from sub-canopy trees and saplings

(Figure 3.5). This observation is unlikely to be due to low sample size limitations according to a Chi-squared test performed on the distribution of hybrids among cohorts on the archipelago (Figure 3.5b, $\chi^2 = 19.4$, df = 3, $p = 2.2\times10^{-4}$). It might alternatively be due to selection disfavouring hybrids under a closed canopy with high levels of competition among trees.

3.3.4 Growth sensitivity to seasonal climate variables

Climatic conditions within the *P. sitchensis* range differ slightly between the Kodiak Archipelago and the Kenai Peninsula (Figure 3.6); however, within the Kenai Peninsula they differ dramatically between *P. sitchensis* areas (e.g., Seward) and *P. glauca* areas (e.g., Nikiski). (Figure 3.1b). Nikiski receives 2 to 4 times less rain in all seasons than Seward, with much weaker fluctuations in precipitation between seasons and between years. Continentality (difference between mean warmest month and mean coldest month temperatures) is also slightly higher in Nikiski. These differences correspond to expected climate niche differences between *P. glauca* and *P. sitchensis*. As hybrids on the Kodiak Archipelago are most likely first- or second-generation hybrids, it is possible that these individuals are maladapted to conditions found on the Kodiak Archipelago, making them somewhat less fit and altering their growth patterns. To test the hypothesis that hybrids and pure Sitka genotypes have different radial growth responses to local climate variables, I correlated standard chronologies of each group to seasonal values of temperature and precipitation over a period of 70 years, which corresponds to the longest period of availability of high-quality monthly climate data (Figure 3.7). In general, both regions showed similar climate-growth responses, with a major, positive influence of summer temperature on radial growth. However, trees on the archipelago showed a negative correlation with spring and summer precipitation that was not as strong in the Seward population. There was no significant difference in correlation coefficients with temperature or precipitation between hybrids and pure *P. sitchensis* genotypes in each of the two regions. Although confidence intervals overlap, there were consistent differences between hybrids and pure *P. sitchensis* across regions: growth consistently correlated more negatively with the amount of spring precipitation in hybrid than in pure *P. sitchensis* trees. Another noticeable difference is the lack of sensitivity to spring temperature in hybrids on the Kodiak Archipelago, contrasting with the positive correlation with this variable observed in pure *P. sitchensis* growth.

3.3.5 General differences in growth patterns

To test whether differences could be observed in general growth patterns between hybrids and pure genotypes, I performed hierarchical clustering on a pairwise tree-ring index dissimilarity matrix. While all trees within regions clustered together, hybrids did not, even within regions (Figure 3.8). I further explored the clustering results by setting the number of clusters to 5 based on visual assessment of the dendrogram and by mapping the distribution of clusters within sites (Figure 3.9). Cluster 1 was exclusively present in the Seward population. The Kodiak Archipelago was composed of clusters 2 and 3. In this region, we found that younger trees were most often assigned to cluster 3 whereas older trees were assigned to cluster 2 (Figure 3.10). Clusters 4 and 5 were only represented by 3 and 4 individuals, respectively. Chronologies of ring-width indices grouped by clusters are represented on Figure 3.9c for the three most common clusters. Clusters 2 and 3 seem to differentiate from cluster 1 through wider fluctuations of radial growth. This might be due to the lack of buffering of disturbances and climatic fluctuations in a forest with a younger canopy.

3.4 Discussion

3.4.1 Main results

All sampled populations including the *P. glauca* population from Denali were admixed, and no diagnostic marker could be found in the dataset of 336 polymorphic, genome-wide SNPs used for this analysis. Together with a relatively low interspecific differentiation level ($F_{ST} = 0.2$), this confirms that introgression has been extensive between *P. glauca* and *P. sitchensis*, probably due to secondary contacts along the borders of the species' range (Boucher & Mead, 2006; Hamilton & Aitken, 2013; Jones, 2008). However, geographic patterns of admixture could still be identified: I observed a general decline in *P. glauca* ancestry levels from Seward to regions closer to the *P. sitchensis* expansion front. The large majority of trees sampled on the Kodiak Archipelago harboured pure Sitka genotypes, and *P. glauca* ancestry was carried by individuals with high hybrid indices. Although rare, admixed individuals on the Kodiak Archipelago

harboured high-heterozygosity levels characteristics of first- or second-generation hybrids (F1, or *P. glauca*-backcrossed F1), suggesting that hybridization was most likely occurring during colonization between establishing *P. sitchensis* or hybrid trees and *P. glauca* pollen. The hybrids established on Afognak Island mostly between 1700 and 1800, and somewhat later on Kodiak Island, in the 1800s and as recently as 1930 in the southernmost and most recently colonized *P. sitchensis* site (K5 in Chapter 2). This suggests repeated pollen flow from *P. glauca* populations to the archipelago.

Surprisingly, recombinant hybrid genotypes seem to be absent from the samples. In addition, all hybrids were canopy trees, and no *P. glauca* ancestry was detected in lower canopy levels or in juvenile trees. This raises questions about the viability of hybrids in the colonizing environment. The observed patterns could result from a reduced fitness of first or second-generation hybrids. This could occur if the reproductive success of hybrids was reduced through lower seed production or asynchronous cone phenology relative to pure Sitka genotypes. It could also be due to lower fitness at early life stages leading to weaker competitive abilities. In this case, the lack of interspecific competition during early stages of colonization may have allowed hybrids to establish and contribute to the developing canopy, while interspecific competition under the canopy after canopy closure could have prevented hybrids from establishing. Lower fitness of hybrid juvenile trees could originate from susceptibility to insects, disease or climatic events, or to less efficient growth, although no evidence for the latter has been found (Hamilton et al., 2013b). An alternative hypothesis is the temporary superiority of *P. glauca* or hybrid genotypes over pure *P. sitchensis* during the Little Ice Age, as the observed high proportion of hybrid establishment coincides with the last Little Ice Age maximum, in the first half of the 18th century.

As fitness is most often approximated by growth measurements in wild tree populations, I compared annual radial growth patterns of non-recombinant hybrid and pure *P. sitchensis* genotypes. If hybrids were less competitive than pure *P. sitchensis* trees through different growth responses to environmental conditions, such differences should be evident in radial growth patterns of mature canopy trees, especially if the underlying cause is climatic maladaptation. Hamilton et al. (2013b) found that the geographic occurrence of *P. sitchensis* vs. *P. glauca*

correlates with the amount of annual precipitation, with *P. glauca* habitat being drier. Although I found that radial growth of hybrids tended to be more negatively affected by high amounts of precipitation than radial growth of pure *P. sitchensis* trees, the difference was not significant. One would also expect to *P. glauca* phenotypes to be more cold-hardy in the winter and adapted to avoid spring frost injury through later bud break timing. A reduced response to winter and spring temperatures from hybrids compared to pure *P. sitchensis* on the Kodiak Archipelago was noticeable, but not significant. In general, if growth response differences are real, they are too weak to be significant with the sample size available in this analysis. Although probably influential, genotypic effects might be overshadowed by other factors affecting radial growth: based on hierarchical structuring analyses, ring width was primarily affected by geographic location and tree age.

3.4.2 Limitations of radial growth analyses

Several factors limit the power of the tree ring analyses presented here. First, the low proportion of hybrid genotypes found in *P. sitchensis* populations prevented the subdivision of hybrids into more precisely defined hybrid classes and limited the statistical power to identify phenotypic differences between genotype classes. Obtaining a sufficient sample of hybrid genotypes would require a larger sampling and genotyping effort. The second phenotypic analysis, involving hierarchical clustering of ring-width index series, had the advantage of organizing individual time series of radial growth by similarity without making assumptions about pre-defined factors such as regional or genetic groups. However, this analysis presented some limitations too. Detrending ring-width series prior to calculations is necessary to account for variation due to tree age, tree core origin (compression or tension wood) and microsite or stand-level variations. As only high-frequency residual variation is retained after detrending, potential effects of low frequency events are dampened. Periods of reduced growth during a particular life stage or climatic period as well as differential long-term responses to an episodic climatic event (such as drought) would produce such undetectable trends. Detrending methods that only remove age-related trends such as the negative exponential model or the linear model with a negative slope are usually preferable, but this is most suitable for trees that established in open conditions and are not affected by subsequent stand-level disturbances. Sites on the Kodiak Archipelago

underwent large and independent canopy changes due to colonization (Figure 2.2) and therefore required the removal of low and medium-frequency variation in individual tree ring series to allow for the comparison of growth trends between genotypes across sites. The use of tree ring data in association with genetic data has recently appeared in the literature, addressing diverse applications such as genome-wide association studies (Heer et al., 2018; Housset et al., 2016) and low-resolution phylogeography (Bosela et al., 2016). This chapter, together with these few examples, illustrate the potential for further studies coupling genetic and tree ring data to foster our understanding of forest adaptation to changing environments.

3.4.3 Conclusions

Through genetic analyses of interspecific hybridization between an expanding species (*P. sitchensis*) and nearby populations of a sister species (*P. glauca*), this chapter investigates the extent of hybridization during colonization. Using tree ring data, I reconstructed the succession of observed patterns in the last 400 years and showed that pollen flow from *P. glauca* populations has occurred but has only led to successful genetic admixture in very early stages of colonization at any given site. Genetic processes such as population expansion with a growing proportion of local-to-foreign pollen could explain the population-wide reduction in *P. glauca* ancestry over time and space along the range expansion axis. This chapter indirectly addressed the potential role of selection against first-generation hybrids or backcrossed genotypes by studying their radial growth patterns. Although weak but predictable differences in sensitivity to precipitation were detected between hybrids and pure Sitka genotypes, further analyses would be needed to understand the phenotypic traits and life stages involved in potential selective forces preventing further introgression at the expansion front. As many theoretical studies have allowed accurate and fascinating predictions about short- and long-term evolutionary outcomes of interspecific gene flow at range margins, additional empirical studies are needed to determine how common such outcomes are in the wild.

3.5 Figures

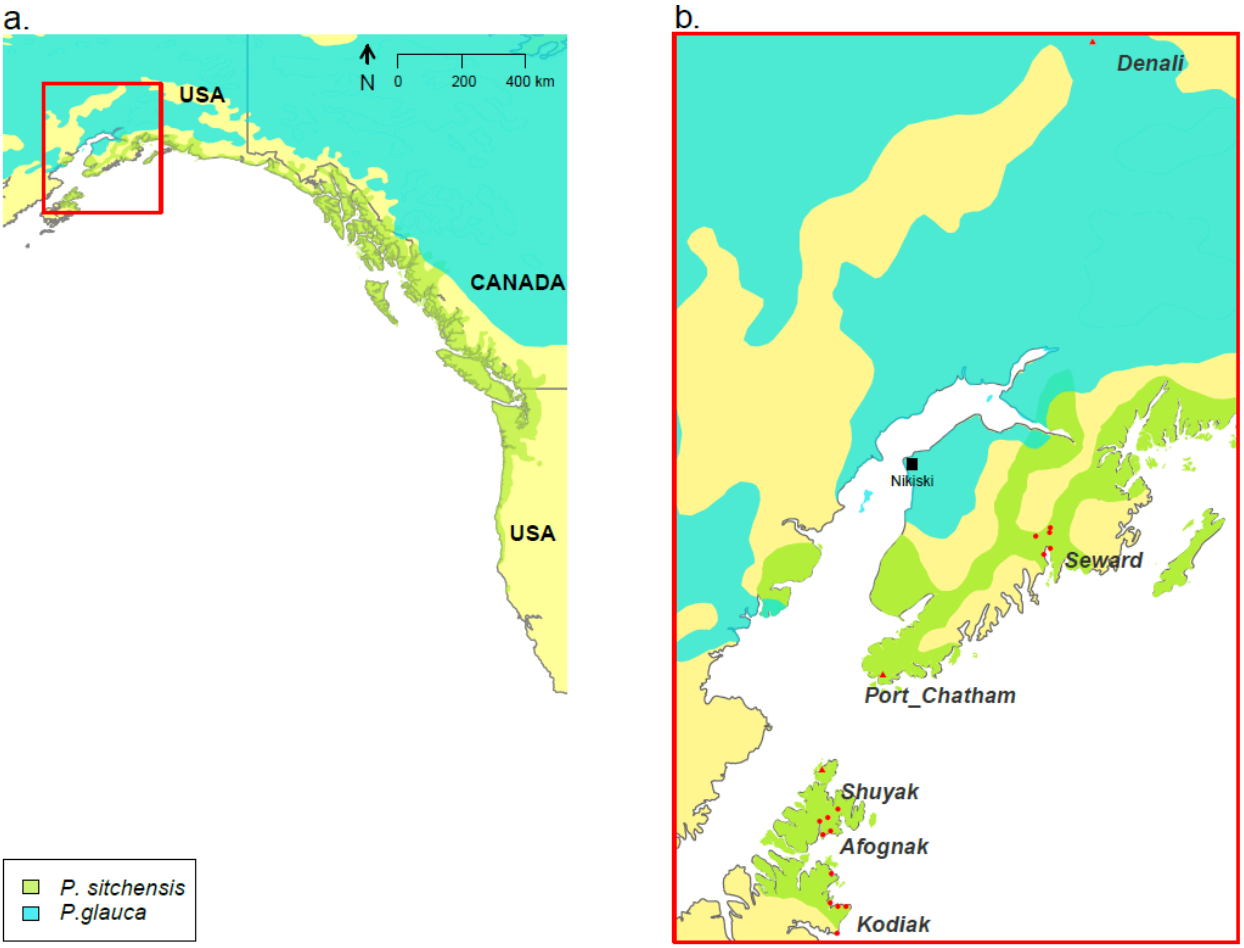


Figure 3.1 Geographic range of *Picea sitchensis* and *Picea glauca.* a. In western North-America b. In the studied region.

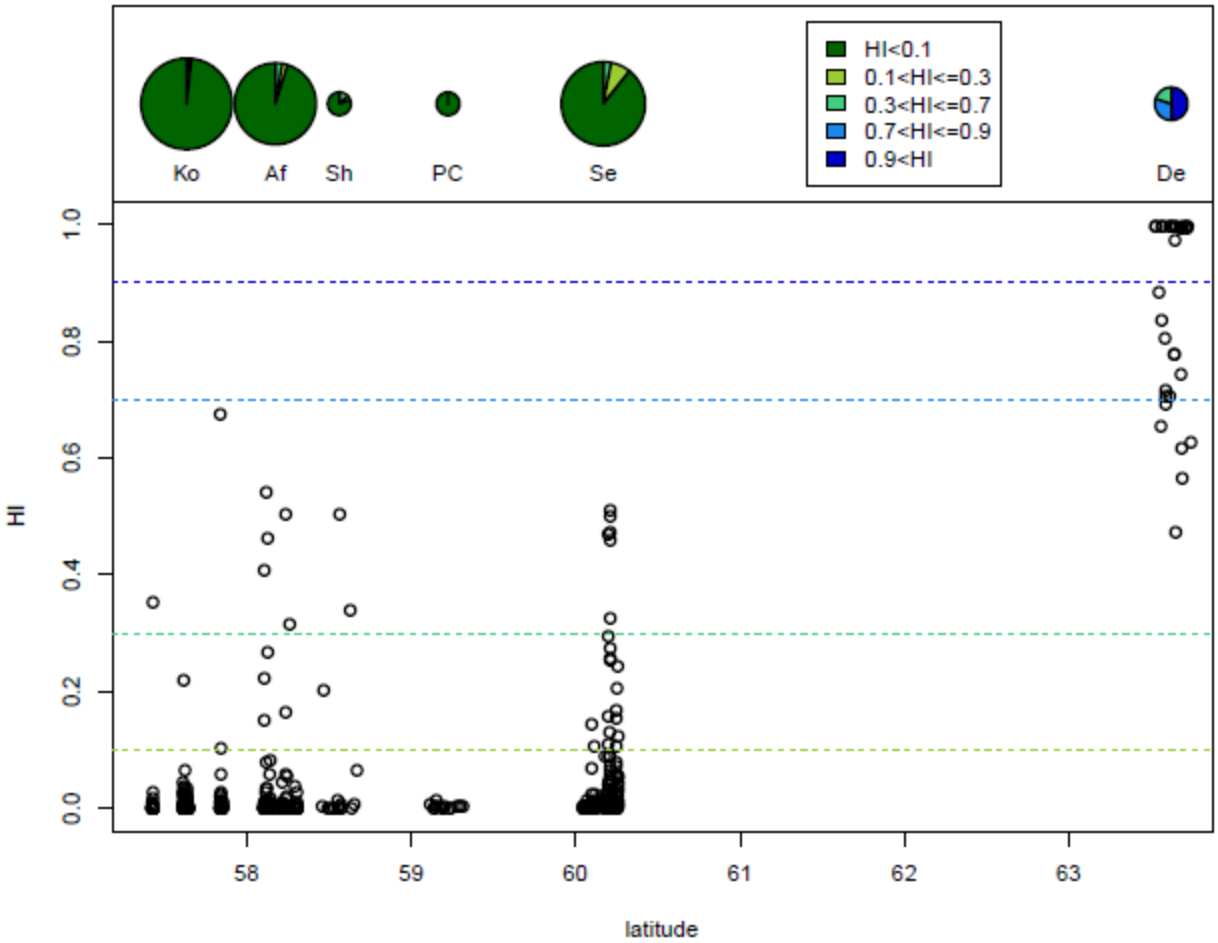


Figure 3.2 Geographic distribution of hybrid index in sampled regions (top panel) and sites (bottom panel). A slight horizontal jitter (factor 0.1) was added to the bottom panel for visibility purposes.

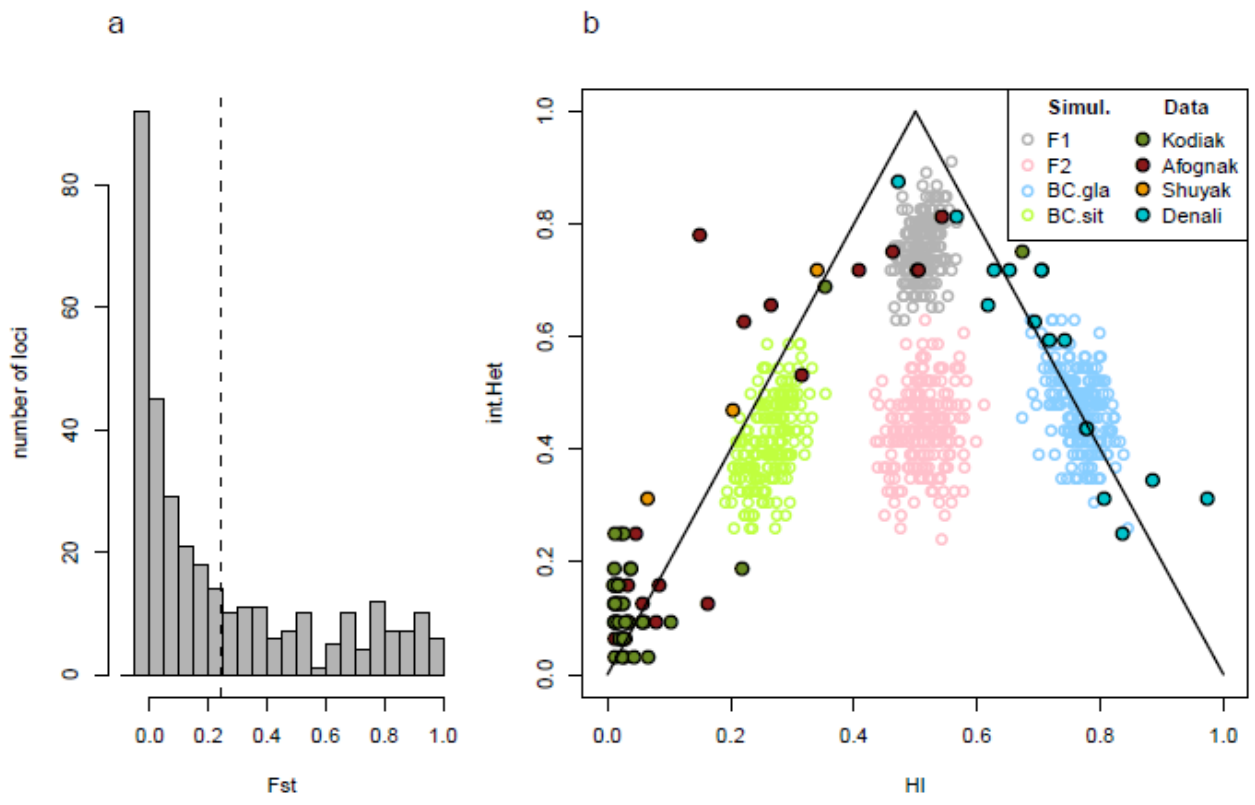


Figure 3.3 Locus-specific F_{ST} and interspecific heterozygosity a. Distribution (bars) and average (dotted line) of locus-specific F_{ST} between reference groups. b. Triangle plot of interspecific heterozygosity vs. hybrid index for Kodiak Archipelago and Denali hybrids, in comparison with hybrid class simulations. Observed data: only individuals with hybrid index between 0.01 and 0.99 are represented. Hybrid class simulations: N=200 for each cross. BC.gla = F1 backcross to *P. glauca*; BC.sit = F1 backcross to *P. sitchensis*.

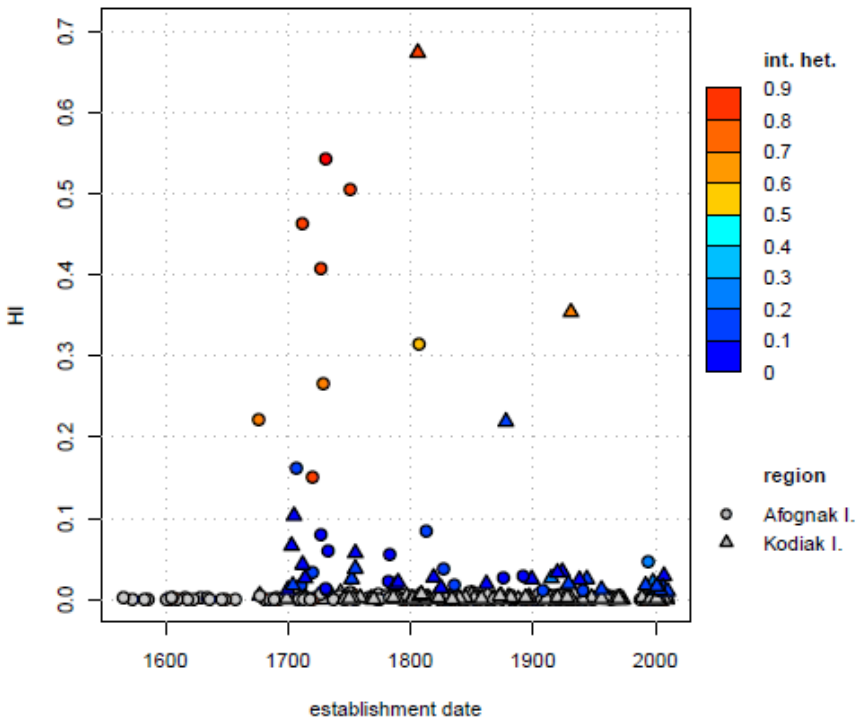


Figure 3.4 Temporal change in hybrid index and interspecific heterozygosity on Afognak and Kodiak Islands. Grey points correspond to the parental *P. sitchensis* genotype group.

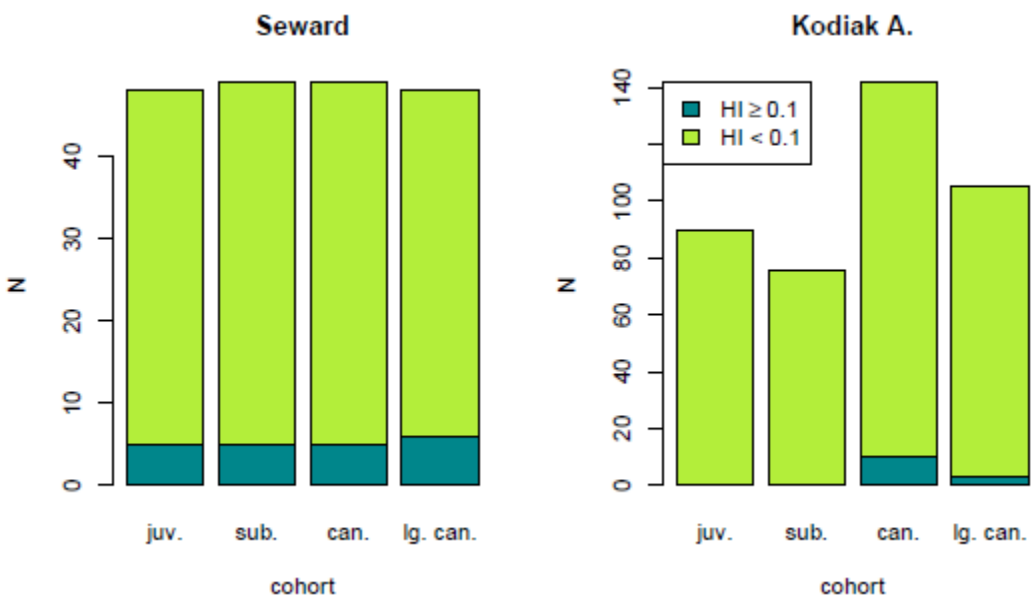


Figure 3.5 Distribution of hybrids among forest canopy structure levels.

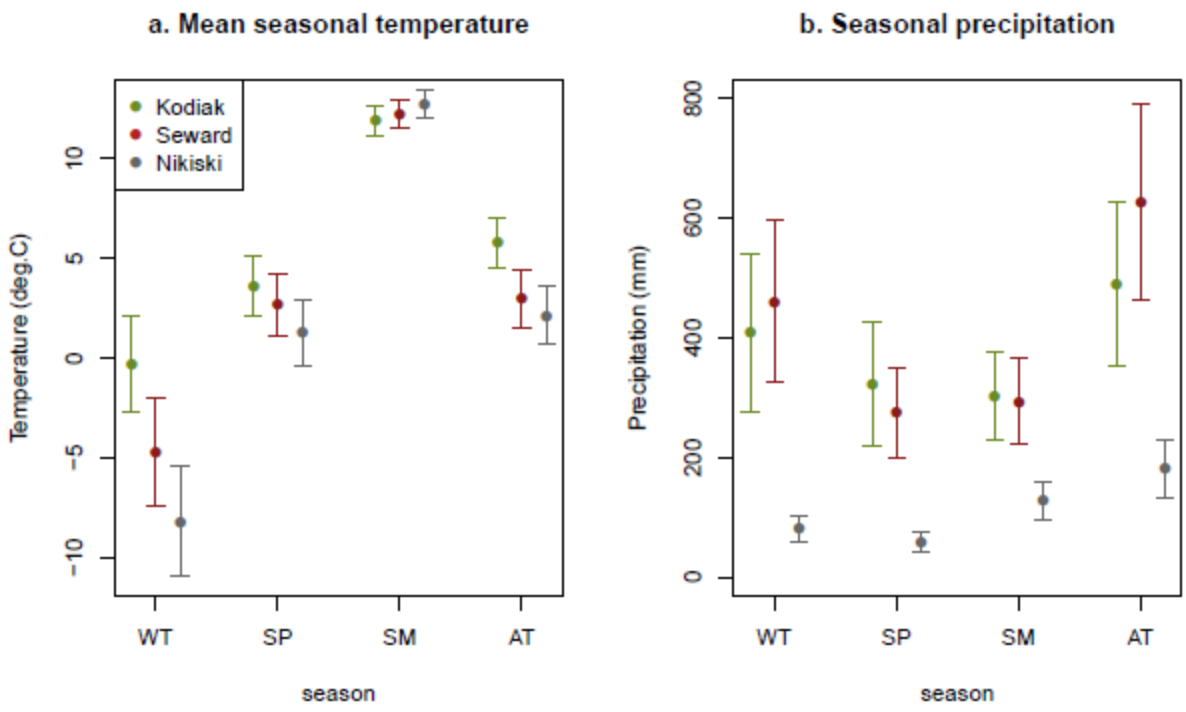


Figure 3.6 Mean seasonal temperature and precipitation at three locations in the studied regions and Nikiski (northern Kenai Peninsula, see Figure 3.1b). Error bars represent standard deviation (N=70 years).

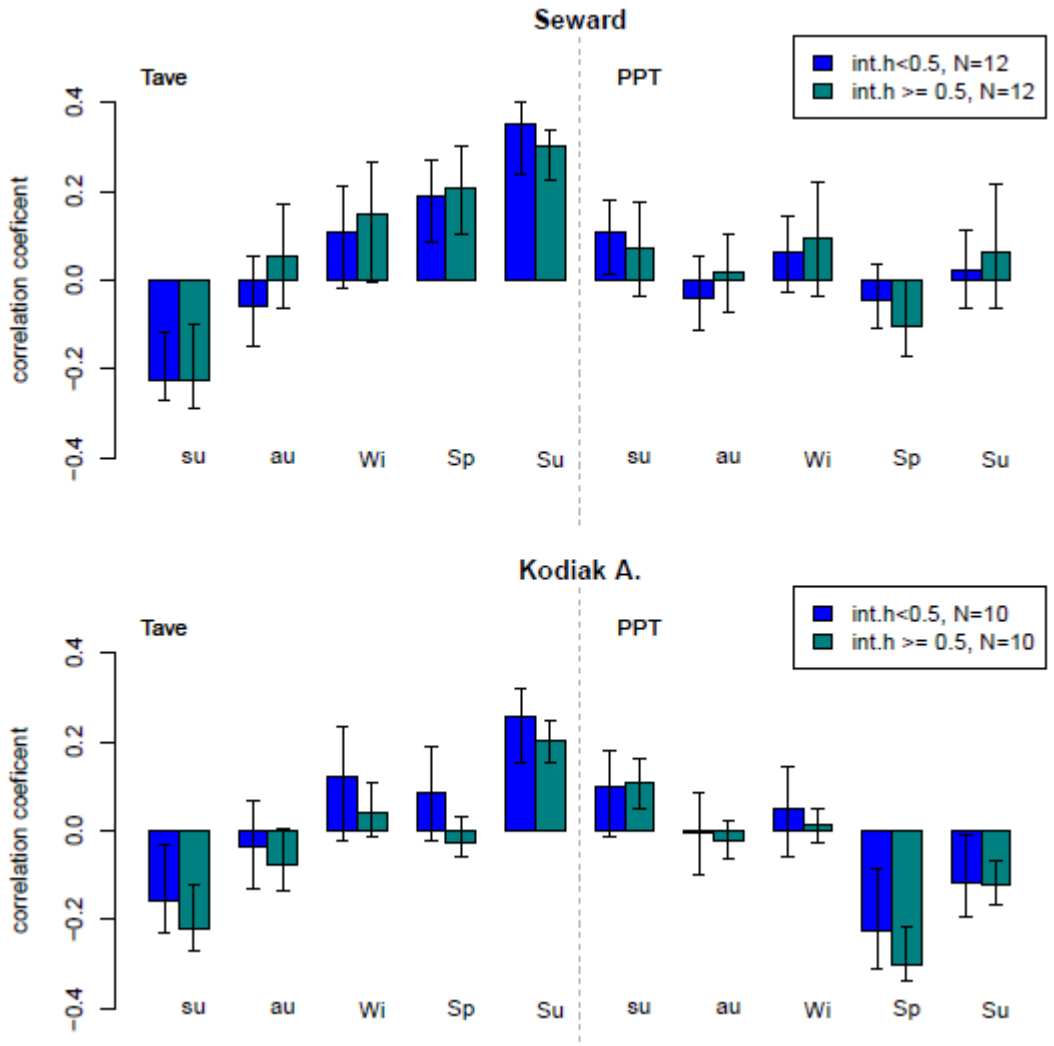


Figure 3.7 Correlation between tree ring index and average temperature (Tave) and total amount of precipitation (PPT) by season (su=previous summer, au=previous autumn, Wi=winter, Sp=spring, Su=current summer) for pure Sitka and for hybrid genotypes. Results for the period between previous summer (su) and current summer (Su) are displayed for the Kenai Peninsula (top graph) and the Kodiak Archipelago (bottom graph).

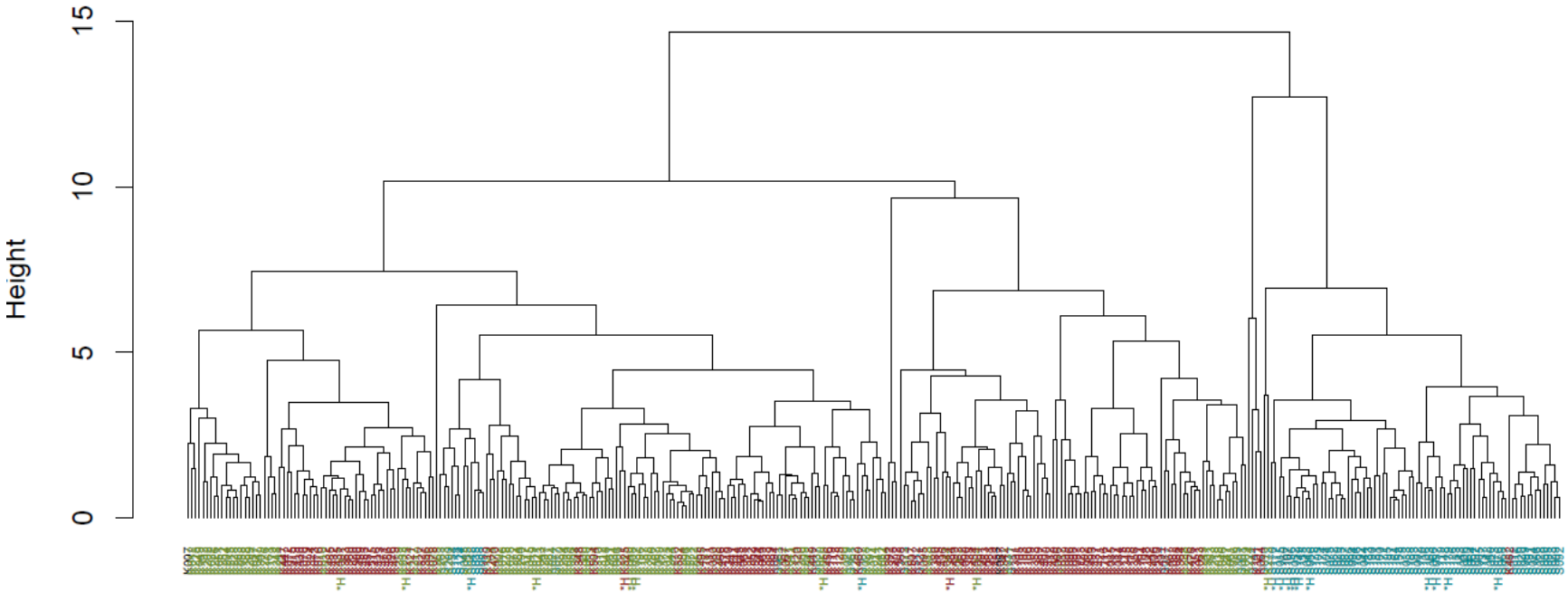


Figure 3.8 Dendrogram of hierarchical clustering of tree ring index series. Colours represent regions (blue=Seward, red=Afognak, green=Kodiak). Hybrids with int.h ≥ 0.5 are represented with the *H symbol.

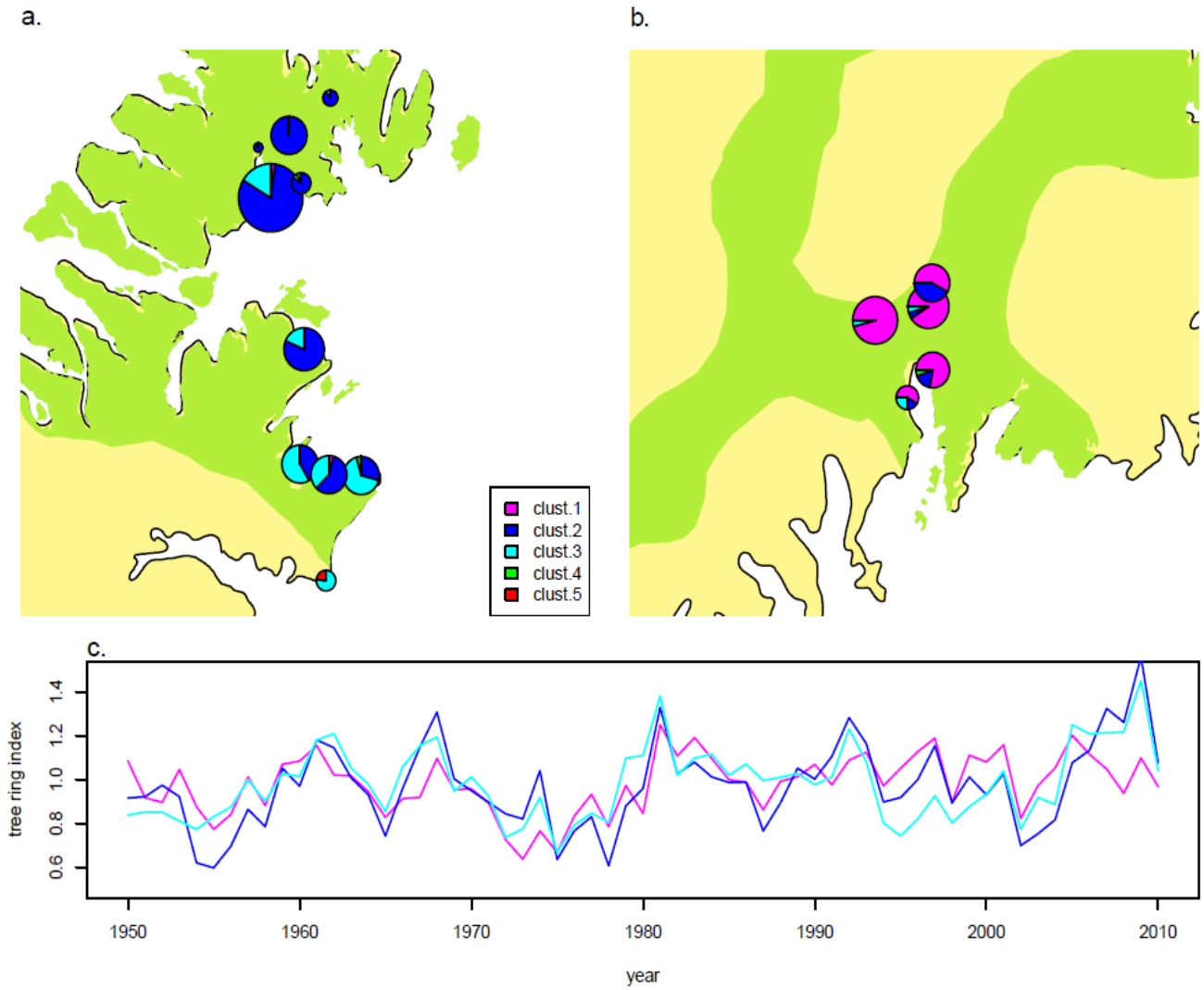


Figure 3.9 Geographic distribution of hierarchical clusters and representation of the average time series profile for the three common clusters. a. Distribution on the Kodiak Archipelago, b. Distribution in the Seward population, c. Average tree ring index profiles for clusters 1, 2 and 3.

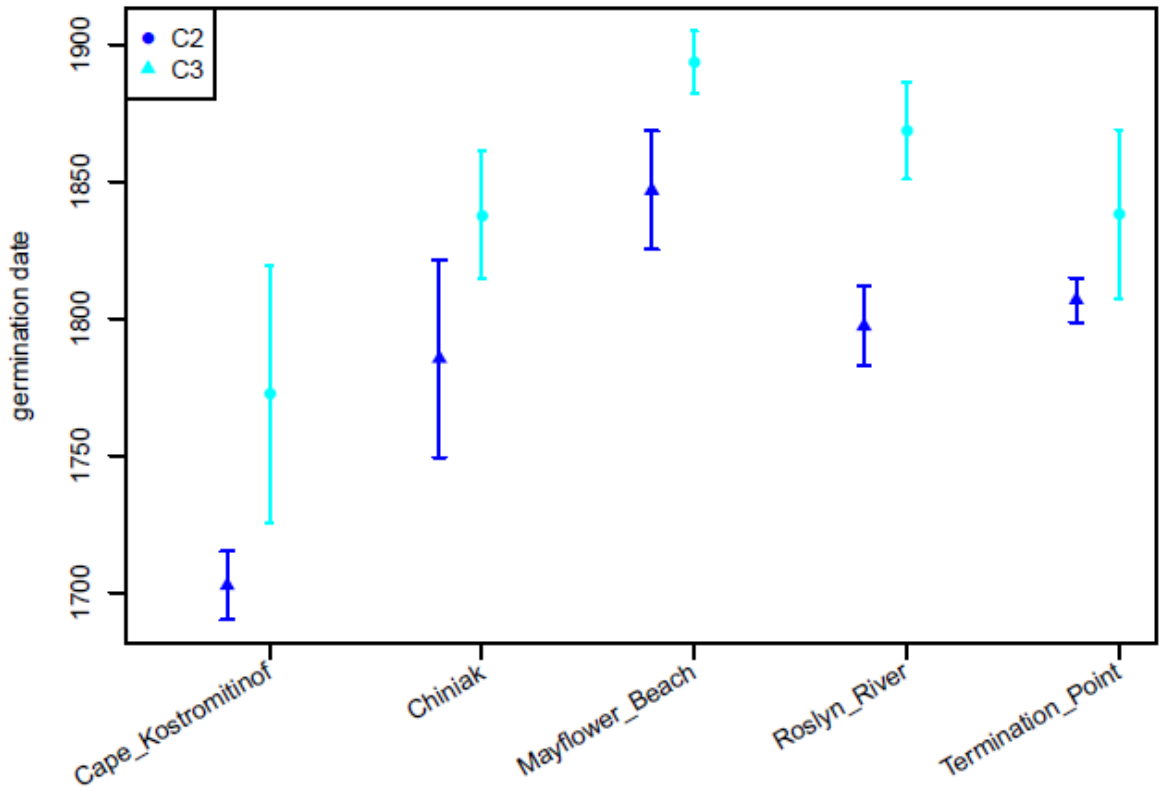


Figure 3.10 Establishment dates of trees belonging to tree ring index clusters 2 and 3 (from Figure 3.8) within areas where both clusters are represented on the Kodiak Archipelago.

Chapter 4: Power and limitations of approximate Bayesian computation in demographic inference of spatial expansion

4.1 Introduction

Chapters 2 and 3 used genetic data to evaluate evolutionary and demographic processes of range expansion over only the few most recent generations of population establishment. Patterns of DNA variation among individuals can however be used to unravel more ancient events in the history of populations. The aim of this chapter is to assess a method for detecting historical demographic events using genetic data from a few individuals sampled in contemporary populations. I use a population simulation approach and subsequently apply the assessed methods to the postglacial migration of *Picea sitchensis* in British Columbia and Alaska. Rapid progress in sequencing technologies at the start of the 21st century has allowed the inference of increasingly complex demographic models, by using increasingly complete genomic datasets. However, this increase in amount of data and complexity of demographic scenarios necessitates updated statistical methods for analysis and inference. Tackling large genetic datasets with inherent errors and uncertainties requires sophisticated techniques for marker development. In parallel, inferring complex historic demographic scenarios with several populations and numerous demographic parameters necessitates efficient algorithms to provide accurate parameter estimates and model validation measures. Reviews and improvements of methods have recently emerged (Schraiber & Akey, 2015), illustrating the fast pace of change in the field of statistical genetics. However, the efficiency of inference methods for different types of demographic models as well as effects of completeness of genomic datasets need to be understood to ensure quality and accuracy of inferences.

4.1.1 Demographic inference in natural populations of nonmodel organisms

In less than 30 years, human demographic inference has taken a leap, evolving from the evidence for a single African origin of all humans using a few non-recombining mitochondrial markers (Cann et al., 1987), to the inference of highly complex demographic scenarios using whole

genomes (Harris & Nielsen, 2013). Although there is still room for improvement in demographic inference of human populations (Schraiber & Akey, 2015), human genomics is at the leading edge of inference from DNA data. Unfortunately, the state-of-the-art statistical inference techniques applied to human data are currently out of reach for studies of natural populations of nonmodel organisms. Knowledge from demographic inference of these species is, however, crucial: it is often the most efficient way to determine how to manage invasive species (Benazzo et al., 2015; Guillemaud et al., 2010), to conserve endangered species or ecosystems (Chan et al., 2014; Dussex et al., 2014; Lopez et al., 2006; Quéméré et al., 2012), and to predict the future distribution and abundance of widespread species that are of economical or ecological importance (Holliday, Yuen, et al., 2010; Zinck & Rajora, 2016). The good news is the genomic revolution has reached nonmodel organisms, creating a spectrum of levels of genetic knowledge across a broad range of taxa. Using a few microsatellites or moderate-sized panels of resequenced SNPs is still common practice (Li et al., 2010; Zinck & Rajora, 2016), but most current studies of nonmodel species now use genomic methods to extract markers for inference. In recent years, sequencing whole genomes of nonmodel species has become feasible in some organisms with small genomes (Boitard et al., 2016; Liu et al., 2014) and has allowed the inference of detailed demographic models using Approximate Bayesian Computation (ABC) or Pairwise Sequential Markovian Coalescent (PSMC) (Nadachowska-Brzyska et al., 2013). For organisms with larger genomes or for studies with lower data requirements, reduced-representation library (RRL) sequencing, through either targeted capture or restriction enzymes, is widely applied (Davey et al., 2011). RRL techniques involving restriction enzymes (commonly referred to as restriction site-associated DNA sequencing (RADseq) or genotyping-by-sequencing (GBS)) output a large number of short sequences (100bp, or longer with paired-end sequencing) from across the genome and have proven useful in population genetics studies and inference involving maximum likelihood methods based on the site frequency spectrum (SFS) or ABC methods (Narum et al., 2013). Most recently, the number of published drafts of whole genomes for nonmodel species has increased dramatically, granting access to longer sequences through the second category of genomic markers: targeted enrichment. This approach allows the use of linkage information for population genetics inference (Li & Jakobsson, 2012).

4.1.2 Approximate Bayesian Computation and other approaches

In this chapter, my aim is to explore ABC for datasets obtained from reduced-representation library sequencing in nonmodel organisms, especially tree species, and to compare the results obtained with those from a SFS approach based on approximation of the composite likelihood (Excoffier & Foll, 2011). I also apply the ABC procedure examined in this chapter to the northward postglacial expansion of *Picea sitchensis* (Appendix E). I chose to explore ABC because of its versatility: It accommodates a wide spectrum of demographic models and dataset types. ABC has been reviewed in a number of publications and its algorithms and techniques are being refined constantly (Bertorelle et al., 2010; Csilléry et al., 2010; Lintusaari et al., 2016; Marin et al., 2012; Sunnaker et al., 2013). For applications in demographic inference using genetic data, the general ABC method involves the following steps. First, a large number of datasets are simulated under a specific demographic model using the coalescent (Kingman, 1982). Parameters used for simulations are drawn from prior distributions that are pre-defined by the user. The simulated datasets are then compared to the observed dataset through calculation of summary statistics. Finally, simulated datasets with the closest vector of statistics to the vector of observed summary statistics are selected. A regression adjustment based on the local relationship between statistics and parameters is then usually performed to approximate the posterior distribution of each model parameter from the parameter values of selected simulations. ABC is suitable when inferring models for which the likelihood function is intractable, as it relies on approximating the likelihood function using a large number of simulations. However, each one of the numerous steps in the implementation of ABC requires users to make empirical decisions. There is particularly a need to improve our understanding of the relationship between the type of markers obtained to build genetic datasets and the way genetic data is subsequently summarized and its power to tease apart demographic models and produce accurate parameter estimates.

4.1.3 Previous work exploring ABC

The need to test the inference power of datasets for demographic models of interest has been recognized in recent years, both in terms of model selection and parameter estimation. Robert et al. (2011) warned against the use of insufficient summary statistics in ABC model choice, opening the door to improved methods for model testing and the associated choice of summary

statistics (Marin et al., 2014; Prangle et al., 2013). Among theoretical results and general guidelines, Marin et al. (2014) suggested the use of different sets of summary statistics for estimation and model selection. Several studies show the use of preliminary simulations testing parameter estimation and model choice with different number and length of markers and number of individuals (Sousa et al., 2012; Stocks et al., 2014), type of molecular markers (Cabrera & Palsbøll, 2017), and choice of summary statistics and models considered (Benazzo et al., 2015; Guillemaud et al., 2010; Li & Jakobsson, 2012; Sousa et al., 2012; Stocks et al., 2014). As most scientists have switched to using genome-wide data, there is a need to expand this set of simulation studies to test and understand the power of different types of genomic data. As part of such an effort, Li and Jakobsson (2012) simulated large, phased genomic datasets comparable to human genomic datasets at the time. Under 2-population split models, they found that ABC produces accurate estimates for most but not all parameters and concluded ABC is well suited to large genomic datasets summarized with LD-based statistics. Robinson et al. (2014) tested the effect of the number and length of unphased genomic sequences and compared them to the effect of the number of individuals sequenced for the inference of three-population admixture models. They found that increasing the number and length of sequences was more beneficial than increasing sample size. Shafer et al. (2015) investigated the power of ABC on short diploid sequences obtained by GBS. They focused on a wide range of simple 1-population and 2-population models with bottleneck, growth, migration and a combination of these parameters. They found that population size changes such as ancient temporary bottlenecks would not be inferred correctly regardless of the number of markers available. This set of studies provides valuable information about the use of genomic data in ABC. Our aim is to extend this knowledge by directly comparing ABC results from molecular markers obtained with different types of RRL sequencing techniques, different sequencing effort allocations, and different levels of genomic knowledge. This will hopefully help future ABC users who do not have access to complete genomic data to select methods and develop genomic datasets that are best suited to answer the demographic questions they are addressing.

4.1.4 General model and datasets

I focused on estimating parameters for a set of 2-population models of demic expansion that are applicable to studies of species invasion, reintroduction, or natural colonization. I tested the power of ABC on these models using a range of marker sets obtainable by RRL methods: datasets with a large number of short genomic reads would correspond to single-end GBS sequencing, whereas fewer but longer diploid sequences correspond to a targeted enrichment approach. For each type of dataset, I quantified the potential benefits of knowing the gametic phase of sequence markers by including or excluding linkage-related statistics at the data-summarizing step. I expect to observe an improvement in the inference for datasets with long sequences. For each model assessed, I also tested the effect of time since colonization. I hypothesize that recent events might be inferred more accurately with datasets containing linkage information, due to the generally higher rate of recombination compared to mutation, and to the potential information contained in long haplotypes. This part of the analysis is also motivated by the fact that overestimates of divergence times are a common result of demographic inference in empirical studies (Holliday, Yuen, et al., 2010) and this upward bias has been found for some demographic scenarios in simulation studies (Benazzo et al., 2015). I therefore aim to explore this potential bias by testing increasingly old events within the same models. As NGS techniques require a trade-off between sample size and individual sequencing depth and are characterized by high genotyping errors, I explore the effect of different trade-offs at different sequencing error rates. Fumagalli (2013) found that increasing sample size at the cost of decreasing depth was beneficial in the inference of diversity measures and population structure. Here, I extend this hypothesis to ABC inference. I also compare ABC results with those obtained from an approximate likelihood method using the site frequency spectrum from simulated reduced-representation libraries. As they provide millions of genome-wide SNPs without ascertainment bias, restriction enzyme-based genomic sequencing techniques seem to be particularly well suited to SFS-based inference methods. Comparing SFS results with ABC results on a range of models and datasets will inform future work on demographic inference in nonmodel organisms. Finally, I apply ABC to populations in northern range of *P. sitchensis*, using one of the demographic models assessed in this chapter and genetic markers obtained by sequence capture (Appendix E).

4.2 Methods

4.2.1 Demographic models

I focused on a basic 2-population model of demic expansion (Figure 4.1). A pre-existing population, population 1, is of constant size N_1. At time T_{EXP} before present, the spatial population expansion begins: population 2 is created by 2 migrants from population 1. Population 2 then grows exponentially between times $t=T_{EXP}$ and t=0 (the present) to size N_2 at t=0. The rate of population growth *r* is defined by the other parameter values through the formula $r=log(\frac{N_{02}}{N_2})/T_{EXP}$. Model 1 therefore has just 3 independent unknown parameters: N_1, N_2, and T_{EXP}. I created additional models of increasing complexity by adding parameters. In models 2 and 4, the number of founders of population 2, N_{02}, is unknown (Figure 4.1a and Figure 4.1d).. In models 3 and 4, migration is allowed from population 1 to population 2, with the parameter m_{21} describing a per-generation migration rate (Figure 4.1c and Figure 4.1d). Model 3 is used in Appendix E to infer population history in *P. sitchensis*. In all four models described above, the mutation rate and the recombination rate are fixed. I chose wide and uniform parameter priors for population sizes to accommodate a wide range of types of organisms, and a log-uniform prior for the timing of the expansion event, as this study intends to focus on recent rather than ancient expansion events (Table 4.1).

4.2.2 Generating sets of coalescent simulations

For each of the four models, I created a set of 1 million simulations with each of the five types of datasets described below, with a fixed number of 10 diploid individuals sampled per population. For datasets corresponding to single-end RADseq sequencing techniques, I simulated 10,000 independent DNA sequences of 100bp each. For datasets corresponding to sequence capture methods, I created 100 independent DNA sequences of 10kb each. Additionally, I explored a range of possible configurations between these two types of datasets (Table 4.2). With 4 models and 5 types of datasets, I obtained a total of 20 combinations of models and datasets, each with a million simulations. I used the program *scrm* (Staab et al., 2015), which simulates datasets by creating the ancestral recombination graph following the Wiuf and Hein method (1999). I used

custom R scripts (R Core Team, 2016) inspired by scripts from Shafer et al. (2015) to compute the simulations.

4.2.3 Summary statistics

For each simulation I computed all summary statistics available in the program *msABC* (Pavlidis et al., 2010). The available statistics include diversity statistics (number of segregating sites and θ estimates) and summaries of the SFS (Tajima's D and Fay and Wu's H). These statistics were calculated on each sequence marker for each population and for the whole sample. The available statistics also include summaries of the 2d-SFS: differentiation measures such as the pairwise FST and the number of private and shared polymorphisms. Finally the Thomson estimator of TMRCA and its variance were calculated for each population and for the whole sample. To test the effect that knowing haplotype information has on inference, the ABC analysis was performed twice on each combination of model and dataset type. The first time, I summarized data using only the statistics mentioned above, which are calculated at the SNP level and therefore are available when the gametic phase of the diploid sequences is unknown. The second inference was performed on the same dataset, but additional statistics based on linkage information were added to the previously described set of statistics to summarize the data: Zns for populations 1, 2, and the whole sample was calculated (Kelly, 1997), as well as dvk and dvh (Depaulis & Veuille, 1998). These additional statistics are calculated at the haplotype level and so are only available in cases where the gametic phase of the diploid sequences is known. For each set of simulations, I computed the mean and variance of every statistic over all polymorphic sequence markers in the dataset. As a result, 60 statistics were computed for datasets with known gametic phases (hereafter referred to as "phased"), and 50 statistics were computed for datasets with unknown gametic phases (hereafter referred to as "unphased"). As the haplotype phase and ancestral allelic states are unknown in the *P. sitchensis* datasets, only the statistics that do not rely on such information are used in the empirical analysis (Appendix E).

Using a high number of statistics to summarize genetic data has harmful effects on the quality of the ABC inference, a problem commonly referred to as the "curse of dimensionality" (Blum et al., 2013). I used the partial least squares (PLS) method implemented in *ABCtoolbox* to reduce the number of statistics to 5-7 PLS components (see Appendix A1 for details).

4.2.4 Pseudo-observed datasets

For each set of 1M simulations, I created a corresponding set of 100 pseudo observed datasets (PODs), with parameters randomly chosen from the same priors as for the set of 1M simulations. By doing so I assumed that priors are reliable and reflect the true, unknown distribution of the PODs. These were then summarized with the same summary statistics as their corresponding set of 1M simulations.

4.2.5 ABC estimation

I performed the ABC estimation using each POD as the observed dataset to obtain parameter estimates. The standard *Estimate* algorithm from the program *ABCtoolbox* was used for all ABC computations to create posterior probabilities from the corresponding set of 1M simulations, with a post-sampling regression adjustment through ABC-GLM (Leuenberger & Wegmann, 2010). I fixed the tolerance parameter to 10^{-3}, a compromise between having a tolerance threshold value as low as possible (Li & Jakobsson, 2012) and keeping an appropriate number of simulations for posterior estimation.

4.2.6 Validation

For each combination of model and type of dataset, I computed a measure of precision and accuracy called the relative prediction error (RPE), the ratio of the mean squared error over the variance of the prior, which follows equation (2):

$$(2)\ \varepsilon = \frac{\sum_{j=1}^{j=i}(\widehat{\theta_J} - \theta_j{}^*)^2}{Var(\theta)} \times \frac{1}{i}$$

where $Var(\theta)$ is the variance of the prior distribution and *i* is the number of observations. The RPE was computed on 1,000 PODs. The advantage of using the RPE as a validation statistic is that it directly indicates the contribution of the genetic dataset to the estimation of the posterior. Another attractive feature of the RPE is that it allows comparisons between parameters, as it scales from 0 (precise estimate) to 1 and beyond (in the case of a consistent bias in estimation).

As an additional measure of precision, the 95% highest posterior density interval (HDI) was calculated on a set of 100 PODs for each combination of model and dataset type. This measure is defined as the shortest continuous interval with an integrated posterior density of a certain value (Wegmann et al., 2010). For each combination of model and dataset type I reported the 95% HDI coverage, i.e. the number of times (out of 100) the true parameter value fell within the 95% HDI, expecting values close to 95.

4.2.7 Testing the effect of T_{EXP} on parameter estimation

To test the effect of the time of expansion on the precision of the ABC estimation, I created 100 PODs for each set of 1M simulations and 12 fixed values of $\log T_{EXP}$ spanning the prior range. RPE and 95%HDI were calculated from the results of each set of 100 PODs.

4.2.8 Effect of sequencing effort allocation and sequencing error

The main challenge when developing genomic markers is managing sequencing and variant calling errors. Sequencing a large number of individuals might increase the precision of population genetics inference, but with a fixed sequencing budget, this comes at the cost of reduced individual sequencing depth, which in turn can affect variant calling and estimation of allelic frequencies (Fumagalli, 2013). I explored this challenge focusing on model 2 (4 parameters with number of founders and no migration) and dataset type 2 (5,000 sequences of 200bp). I chose a realistic fixed sequencing effort and derived 3 fixed sampling strategies from it: 250 sampled individuals at a mean individual depth of 4, 100 individuals with depth 10, and 20 individuals with depth 50. I then incorporated three per-nucleotide sequencing error rates (0, 10^{-2}, 10^{-3}), and applied them to each category described above. The resulting 9 categories of PODs, as well as "perfect" datasets (no depth sampling and no error) were all simulated using the same 10 parameter combinations. Further details about the creation of "imperfect" PODs can be found in Appendix A1. Once these imperfect PODs were created and summarized, ABC was performed to estimate their true parameter values. Two additional sets of 1M simulations needed to be created to match the number of individuals sampled per population: one with 100 diploids per population, and the second with 250. It the latter case, I only created 610,000 simulations

because of computation time limitations. The same tolerance (10^{-3}) as all other runs was used for the estimation.

4.2.9 Comparing ABC and SFS estimation

I simulated 10,000 independent DNA sequences of 100bp each for the 4 demographic models 10 times. The resulting 40 datasets were input into both *ABCtoolbox* and *fastsimcoal2*, which uses the SFS to approximate a composite likelihood from a large number of simulations through a conditional maximization algorithm (Appendix A1). I compared the results from the two methods using RPE, 95% credible intervals and 95% confidence intervals.

4.3 Results

The 20 combinations of models and datasets used as input for ABC simulations (Table 4.1, Table 4.2) resulted in a total of 2×10^7 simulated datasets available for analysis, 2×10^5 training simulation sets and 2×10^3 PODs. Each set of 1M simulations was used in two runs of estimation: one including all summary statistics available in *msABC*, the other one excluding statistics based on linkage information, for a total of 40 ABC estimations.

4.3.1 Effect of model complexity on the precision of parameter estimates

In general, the ability to infer demographic history declined rapidly as model complexity increased. The simplest model (1), estimating only population sizes N_1 and N_2 and the log-transformed time of expansion T_{EXP}, allowed the expansion event to be dated accurately. Models 2 and 3 each had 4 parameters: model 2 included the number of founders N_{02} and model 3 allowed migration from population 1 to population 2 (m_{21}). For both model 2 and 3, $\log T_{EXP}$ was inferred with slightly lower precision than for model 1. Finally, scenarios corresponding to model 4, which had all 5 parameters, failed to be correctly inferred.

Not all parameter estimates were sensitive to the addition of parameters in the models: the precision of contemporary population size estimates N_1 and N_2 were independent of model complexity. RPE values for N_1, which was constant over generations, were mostly below 0.05

for the four models assessed (Figure 4.2). The widths of 95% highest posterior density intervals varied between 3,000 and 60,000. For N_2, the contemporary population 2 size after exponential growth, 95% HDI intervals were about as wide as the prior range, indicating a failure to estimate this parameter in all four models (Figure 4.3).

The expansion time T_{EXP} was generally well estimated in model 1, which is the simplest 3-parameter model (Figure 4.2) with no migration between demes and the number of founders set to 2. For this model, the RPE was mostly below 0.1. The precision of $\log T_{EXP}$ estimation was almost as high for the two 4-parameter models, where the number of founders N_{02} (model 2) is unknown and needs to be estimated, or where migration from population 1 to population 2 is likely (model 3). For these two models, the RPE is below 0.2. The ABC analysis of the 5-parameter model (model 4) was unable to recover the true T_{EXP} value.

Estimates of the number of founders of population 2 (N_{02}) and migration rate from population 1 to 2 (m_{21}) were surprisingly imprecise in models of low complexity (model 2 and 3) and could not be recovered at all in model 4 (Figures 4.2 and 4.3).

Models 1 to 4 all rely on population 2 growing exponentially from T_{EXP} to the present time. I tested whether demographic parameters could be estimated more successfully in a model where population 2 goes through a single sudden population change instead of exponential growth. I created a new set of 1M simulations based on model 2 (where N_{02} is a varying parameter) and dataset type 1 (many short sequences) and a smaller prior range for T_{EXP} (2-500 generations). In the new model the size of population 2 changes from N_{02} to N_2 at $T_{EXP}/10$ and remains constant before and after $T_{EXP}/10$. These modifications brought no improvements to any of the parameter estimates (Table D.1).

4.3.2 Do sequence length and linkage-related statistics improve the estimation?

The addition of linkage statistics available in *msABC* brought no notable improvement in the RPE and 95% HDI of parameter estimates for all models (Figures 4.2 and 4.3). It even seems to make the estimation of N_1 less precise in some cases for model 1, 2 and 4, although this pattern is inconsistent across dataset types. ABC performance on models 3 and 4 seemed to be slightly

more dependent on sequence length, with the inference on large sequences marginally benefiting from haplotype information.

4.3.3 Quality of parameter estimates across prior ranges

For each parameter, I visualized estimated values and 95% HDI of ABC results in relation to true parameter values to assess performance over the prior range. Results for the 3-parameter model (model 1) and dataset types 1 and 5 are shown in Figure 4.4. Consistently across models, estimates of N_2, N_{02}, and m_{21} are largely inaccurate regardless of the true value, with HDI ranges as wide as the prior range. Conversely, N_1 estimates are accurate in all models regardless of the true N_1 value. Unlike N_1, the values of T_{EXP} have an impact on the precision of their respective estimates. Accuracy and precision of T_{EXP} estimates for models 1 and 3 decrease with increasing true value. Interestingly, the opposite pattern is observed for model 2: more recent events are less precisely inferred than ancient ones. Results for model 4 show a "cross" pattern where most PODs' $\log T_{EXP}$ values are correctly estimated but some PODs with extreme $\log T_{EXP}$ values show estimates at the opposite extreme. This pattern suggests a complex multivariate relationship between model parameters and statistics.

4.3.4 Effect of the time of the expansion event on the estimation

I tested whether older expansion events are generally more difficult to characterize than recent ones within the time range specified by the prior. To do this, I studied the effect of the true T_{EXP} value on the precision of parameter estimates. I find different trends among the 4 models (Figures 4.5 and D.4-D.8). The precision of inference on model 1 is higher at low T_{EXP} values and decreases at $\log T_{EXP}>4$. Conversely, for model 2, older events are generally better inferred: estimates of T_{EXP} and N_{02} increase in precision as T_{EXP} increases, as shown by the RPE (Figure D.2) and the 95% HDI (Figures D.6-D.8). Model 3 shows the best results for moderately recent expansion events ($3 < \log T_{EXP} < 4$), as shown by RPE and 95% HDI of T_{EXP} and m_{21} (Figures D.3 and D.7). However, this was not verified by the empirical application, where supposedly more recent expansion events in *P. sitchensis* were not more successfully infered than ancient ones (Appendix E). Finally, results for model 4 show high values of RPE and 95% HDI for all parameters, with RPE values mostly above 0.5 (Figures D.4-D.8).

4.3.5 Effect of sequencing effort allocation and sequencing error

Focusing on model 2 and datasets of 5,000 × 200bp sequences, I simulated sequencing and variant calling for three different sample size and depth combinations. The RPE of parameter estimates for 13 tested PODs is represented in Figure 4.6. Depth of sequencing (dp) has very little effect on the precision of estimates: only N_1 and $\log T_{EXP}$ have a marginally higher RPE when sequencing depth is simulated. Error rates affect N_{02} estimates at low depth (N=250, dp=250), as well as $\log T_{EXP}$ estimates at low sample size (N=20, dp=50). The estimation is otherwise robust to introduced errors. For a given set of PODs (e.g. N=250, dp=4), the precision lost in a parameter estimate because of an error rate of 0.01 (N_{02}) is gained on another parameter (N_1), reflecting the limitations of the model estimation process rather than the effect of sequencing error. However, the results suggest that choosing a larger sample size with a shallower individual sequencing depth improves estimation over other strategies, especially for the estimation of $\log T_{EXP}$.

4.3.6 Comparing ABC with SFS estimation using an approximate composite likelihood

Figures 4.7 and 4.8 illustrate the performance of ABC and approximate composite likelihood from the SFS for all models performed with datasets of 10^3 100-kb sequences. Both methods gave similar results in terms of precision of parameter estimates. The SFS-based method performed slightly better than ABC in the model with migration (model 3), but the precision of ABC estimates was superior for model 2 (Figure 4.7). The approximate composite likelihood method generally provided narrower 95% confidence intervals (Figure 4.8).

4.4 Discussion

I explored the ability of approximate Bayesian computation to characterize a recent event of spatial expansion from one population of constant size to a new and growing population, a model which can be broadly applied to studies of species range expansion, invasion biology, or reintroduction of endangered species. I found that regardless of model complexity, estimates of the size of the growing, newly founded population (N_2) are poor. However, this did not prevent

successful estimation of other parameters (N_1, $\log T_{EXP}$, and in restricted cases N_{02}). Failure to estimate N_2 does not come as a surprise: estimates of past changes in effective population size from one punctual sampling event commonly rely on linkage information between markers, a calculation not readily available in ABC packages (Beaumont, 2003). My result that models of higher complexity are harder to estimate was expected, but in the case of the expansion models I considered, this trend leads surprisingly quickly to a complete failure to estimate any parameter, as soon as 5 parameters are involved. While expansion timing was precisely estimated in the 3- and 4-parameter models, it could not be recovered in the 5-parameter model. ABC on model 2 (the 4-parameter model including the number of founders but no subsequent migration) successfully estimated all parameters (except N_2) for old expansion events. In contrast, for model 3, the 4-parameter model including migration between demes, estimations were more successful for recent events. These results highlight the potential importance of taking into account the timing of an expansion event when predicting estimation success for a given demographic model. The difficulty of estimating the time of a founding event with subsequent migration was also reported by Robinson et al. (2014); however, I show here that for a moderately recent event (10 to 100 generations), it is possible. An application of ABC to the colonization history of *P. sitchensis* using model 3 is described in Appendix E.

4.4.1 Implications of including haplotype information

Analyses based on unphased sequences exploring similar models to those used here have shown encouraging results (Robinson et al., 2014). However, no study to date has explicitly compared datasets of phased and unphased sequences using the same models and same amount of data. Here, I quantified the benefits of using phased haplotype sequences over single SNPs by including or leaving out LD-based and haplotype-level statistics at the data summarization step of the ABC inference. Surprisingly, haplotype information did not substantially improve the precision of parameter estimates, even when 10-kb sequences were used as markers. Li and Jakobsson (2012) explored ABC with similar 2-population split models and a similar fixed population-wise per-generation recombination rate as in our study. When they tested different combinations of summary statistics, their results did not demonstrate any obvious superiority of LD-based statistics over SNP-based statistics. They concluded that the selected summary

statistics should capture as many different aspects of the data as possible, with as little redundancy as possible. Potentially, phasing the data may not have improved inferences because the extent of linkage that the chosen statistics are sensitive to differs from the linkage actually present in the simulated data. Future work when dealing with phased data would require developing expectations of LD levels and creating or choosing statistics that cover the extent of LD likely to be present in the data.

One needs to be aware of the difficulties associated with the use of LD information. Firstly, ABC on phased data requires reasonable knowledge of recombination rates and variability across the genome. The recombination rate needs to be included as a parameter along with demographic parameters, or as a nuisance parameter with a hyperprior. Secondly, simulating the coalescent with recombination is a complicated process and comes at high computational costs (McVean & Cardin, 2005). With high recombination rates or very long sequences, coalescent simulations might take so long to run that one would instead use a more efficient inference method than ABC. Moreover, translating genome-wide observed data into a set of summary statistic values that are readily useable by ABC programs and comparable to simulated datasets can be a challenge. File input formats in most programs are currently not compatible with sequence information, and many summary statistics programs do not offer haplotype-level calculations. Thirdly, when aligning reads to a fragmented and incomplete reference genome, as is often the case for nonmodel organisms, defining haplotypes can be tricky. One also needs to address problems of sequencing errors, paralogous sequences and imperfect mapping. Inevitable sequencing uncertainties will affect haplotype statistics more strongly than single-SNP diversity measures. Data processing errors and filters can severely bias inferences, to the extent of supporting the wrong demographic model, as revealed by Shafer et al. (2016). Finally, targeted sequence capture will result in thousands of markers of various lengths. Setting up simulations that correspond closely to an observed dataset requires approximating the distribution of sequence lengths, and this may also affect inferences, especially if variances of summary statistics are included at the data summarization step. Considering the difficulty of obtaining reliable haplotype information in nonmodel organisms, the potential difficulties of adapting the use of long sequences to currently available ABC programs, and computational

time, our results tend to suggest that using SNP-level information from GBS-type data is preferable over targeted sequence capture.

4.4.2 Choosing summary statistics

It is important to note that all the results presented here are only valid in the context of the choice of summary statistics. In the present study, I decided to use the first and second moment of all statistics available in *msABC*, and to reduce the dimensionality with a PLS transformation. Several previous publications have performed simulations either using the two first moments of summary statistics (Li & Jakobsson, 2012) or only using the mean (Shafer et al., 2015). To our knowledge, only Robinson et al. (2014) tested the use of 4 moments for summary statistics for models of divergence with admixture. They compared their results with those obtained using only the mean and found that the mean alone was sufficient. Although the two first moments may not be the most representative summaries for some statistics, adding higher-level moments will come at a computational cost.

It is widely recognized that choosing a set of summary statistics is probably the most challenging step for ABC users. For instance, the optimal set of statistics for parameter estimation in a given model might differ from the optimal set of statistics to discriminate between demographic models. As insufficient summary statistics have detrimental effects on model selection (Robert et al., 2011), Fernhead and Prangle (2012) introduced "semi-automatic ABC", which relies on an ABC pilot run and a subsequent linear regression to choose the most appropriate set of summary statistics. Similarly, *ABCtoolbox 2.0* implements a statistical selection step based on the incremental assessment of inference power with the addition of summary statistics. However, documentation is lacking for this new feature of the program. These improvements constitute a promising step towards a more rigorous statistical framework for the automatic selection of ABC summary statistics.

4.4.3 Sequencing effort: go large and shallow!

I found that "imperfect" datasets created with a high number of individuals sequenced at a low individual depth seemed to perform consistently better for most parameters than datasets with fewer individuals and higher depth. This is consistent with Fumagalli (2013), who studied the

same trade-offs on diversity statistics under various demographic settings. This result seems to hold even with simulations with moderate or high sequencing error rates, although this is difficult to conclude with confidence considering the large bootstrapped confidence intervals (Figure 4.6). It is worth noting that if the error rate is not properly estimated during the genotype calling process, more errors will be present in the final dataset and it is likely that ABC results will be impacted for all sequencing strategies, especially those with low depth. As ABC summary statistics rely on the SFS and not on individual genotypes, I suggest that future ABC users sequence large sample sizes at low depth. In this case, estimating the SFS or derived statistics following methods such as described in Nielsen et al. (2012) and Fumagalli et al. (2014) has proven more successful than genotype calling in inferring the SFS. There is unfortunately no straightforward program or pipeline of compatible programs incorporating these methods into an ABC framework. One possibility is to summarize the SFS into quantiles and to use the latter as summary statistics in a classic ABC run. Such a process would need to be further tested.

4.4.4 Comparing ABC to other methods

I did not find large differences in the precision of parameter estimates between ABC and the SFS-based likelihood method implemented in *fastsimcoal2*. Shafer et al. (2015) found a similar result while comparing the performance of ABC with a SFS-based inference implemented in *δaδi* (Gutenkunst et al., 2009). They found that *δaδi* tends to overestimate the time of population split and bottleneck events, a trend not supported by my findings with *fastsimcoal2*. In addition to parameter estimation, Shafer et al. (2015) tested the performance of both methods for model selection and found ABC more accurate, especially in the case of bottleneck scenarios. The advantage of ABC lies in its versatility: the general statistical method poses no constraint on the type of demographic models and the nature of genetic datasets. However, the practical application of ABC to complex genomic datasets currently involves the development of custom bioinformatics pipelines to link together programs with different data formatting requirements (Appendix E). There is a need for user-friendly ABC programs adapted to the type of genomic datasets currently available.

ABC has proven moderately useful for demographic inference with long, genome-wide haplotypes but comparisons with alternative approaches are scarce. Notable examples include Nadachowska-Brzyska et al. (2013), who used ABC and PSMC in a complementary way. Robinson et al. (2014) compared their ABC results with an exact likelihood method developed by Lohse et al. (2011) and found that ABC resulted in more uncertainty, especially in model comparisons. As ABC performance with linkage information needs to be further explored, comparisons to emerging analytical methods based on whole genomes or long sequences such as MSMC (Schiffels & Durbin, 2014) or identity-by-descent haplotype sharing (Harris & Nielsen, 2013) will greatly help refine methods for demographic inference using data at a genomic scale.

Theoretical improvements of ABC methods are emerging rapidly. Although the results presented here do not show that ABC benefits greatly from the use of more complete genomic datasets, the versatility of ABC might be key to its useful applications in a wide variety of fields, even those progressing rapidly such as population genetics. Constant methodological improvement, however, requires regular updates to available ABC programs.

4.5 Tables

Table 4.1 Model parameters with their associated prior ranges. U=uniform distribution; logU=log-uniform distribution

estimated in models	Parameter	Symbol	Prior range	Unit
-	Mutation rate	μ	9×10^{-9}	-
-	recombination rate	R	10^{-8}	-
1,2,3,4	population size 1	N_1	$U(10^4:10^5)$	ind.
1,2,3,4	population size 2	N_2	$U(10^4:10^5)$	ind.
1,2,3,4	time of expansion	T_{EXP}	$logU(2:10^4)$	gen
2,4	initial population size 2	N_{02}	$U(2:10^3)$	ind.
3,4	migration rate from 1 to 2	m_{21}	$U(10^{-3}:10^{-2})$	-

Table 4.2 Description of the five types of simulated datasets

	number of sequences	sequence length (bp)	number of diploid individuals
1	10,000	100	20
2	5,000	200	20
3	1,000	1,000	20
4	500	2,000	20
5	100	10,000	20

4.6 Figures

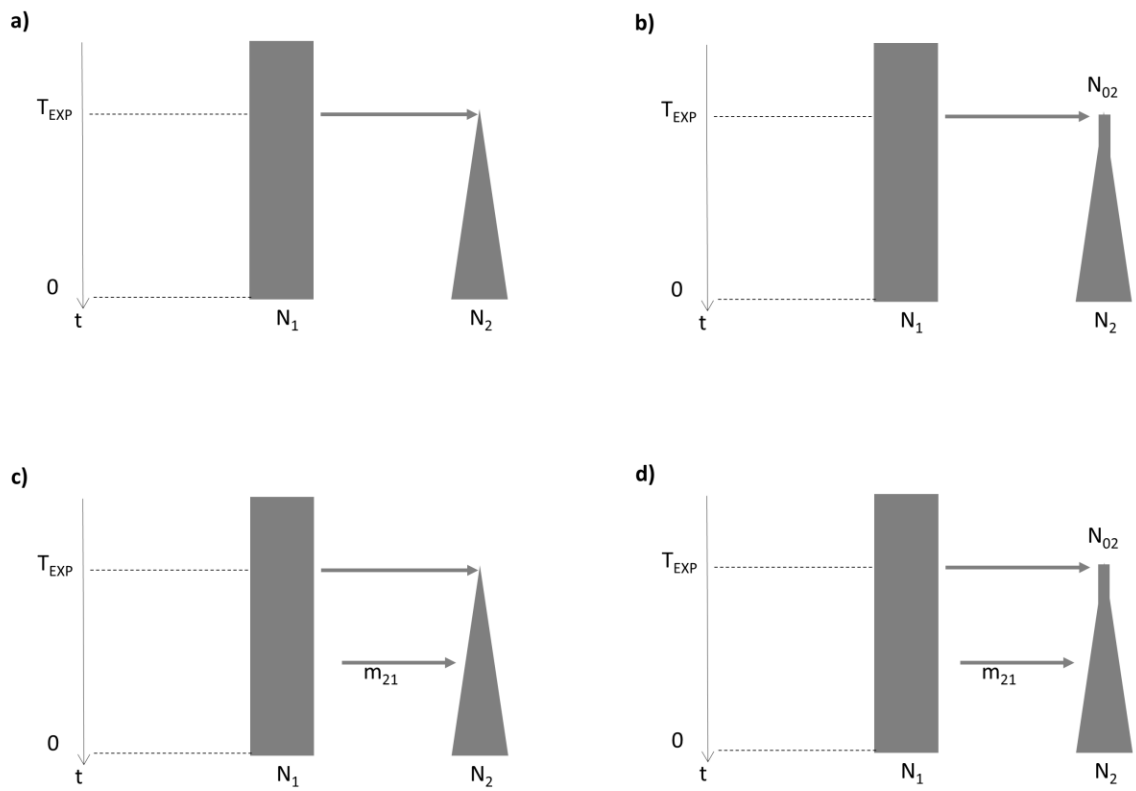


Figure 4.1 Demographic models for ABC analysis. a. Model 1: A three-parameter model of expansion featuring colonization of new population 2 by 2 diploid individuals from population 1 at time T_{EXP}. Population 1 is of constant size N_1, whereas population 2 grows exponentially to size N_2, its size at present. b. Model 2: the number of founders of population 2 is a variable parameter. c. Model 3: a per-generation migration rate from population 1 to population 2 is added as a parameter. d. Model 4 includes all 5 parameters: N_1, N_2, T_{EXP}, N_{02}, and m_{21}.

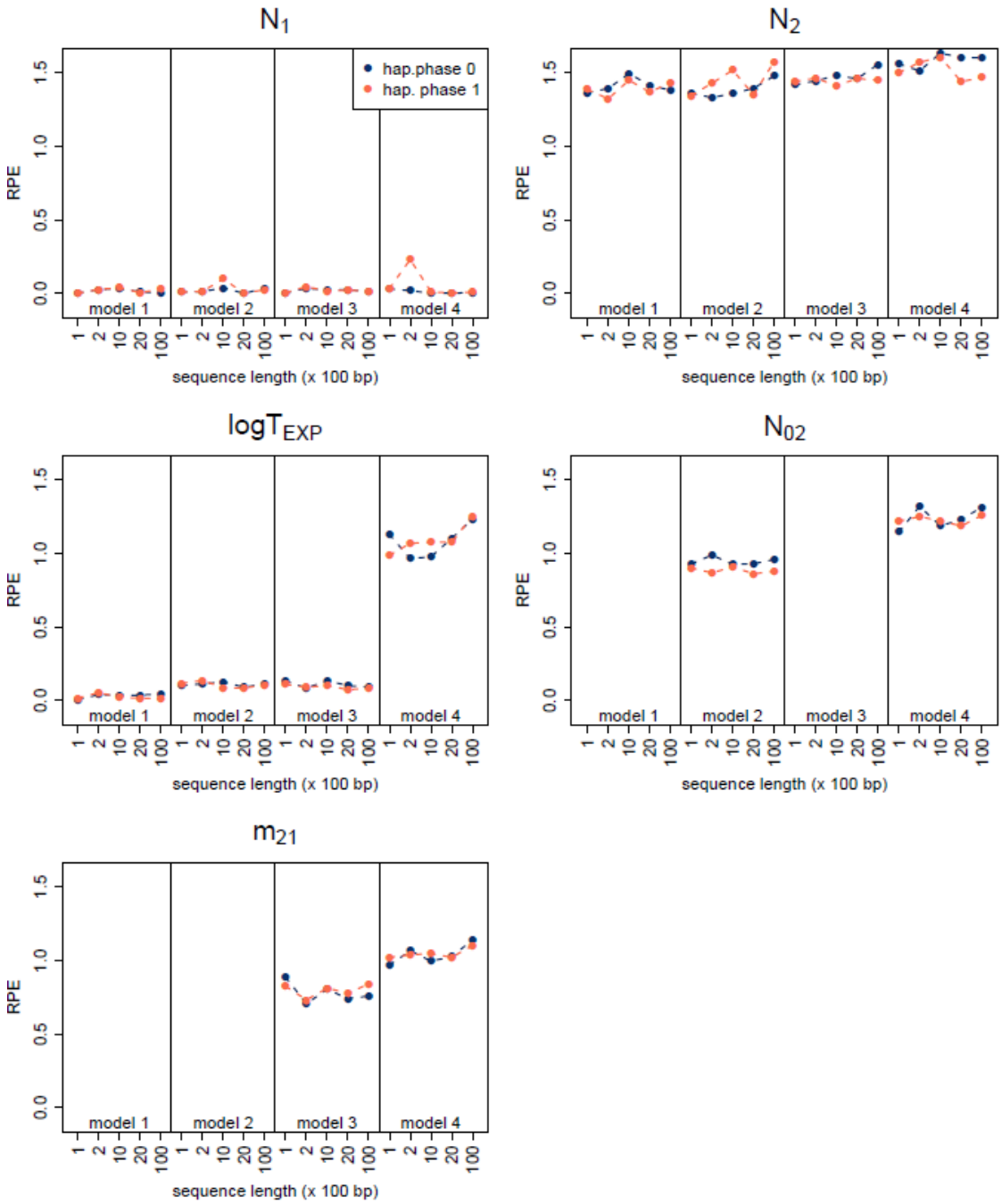


Figure 4.2 Relative prediction error (RPE) calculated from the results of ABC analyses of 20 different combinations of demographic models and sampling designs (x-axis). For each combination, ABC was performed on simulated datasets summarized with statistics including linkage-based measures (hap. phase 1) and on the same set of simulations summarized with only SNP-based statistics (hap. phase 0). RPE values were calculated from the ABC estimation results of 1000 datasets with parameter values randomly drawn from their prior distributions.

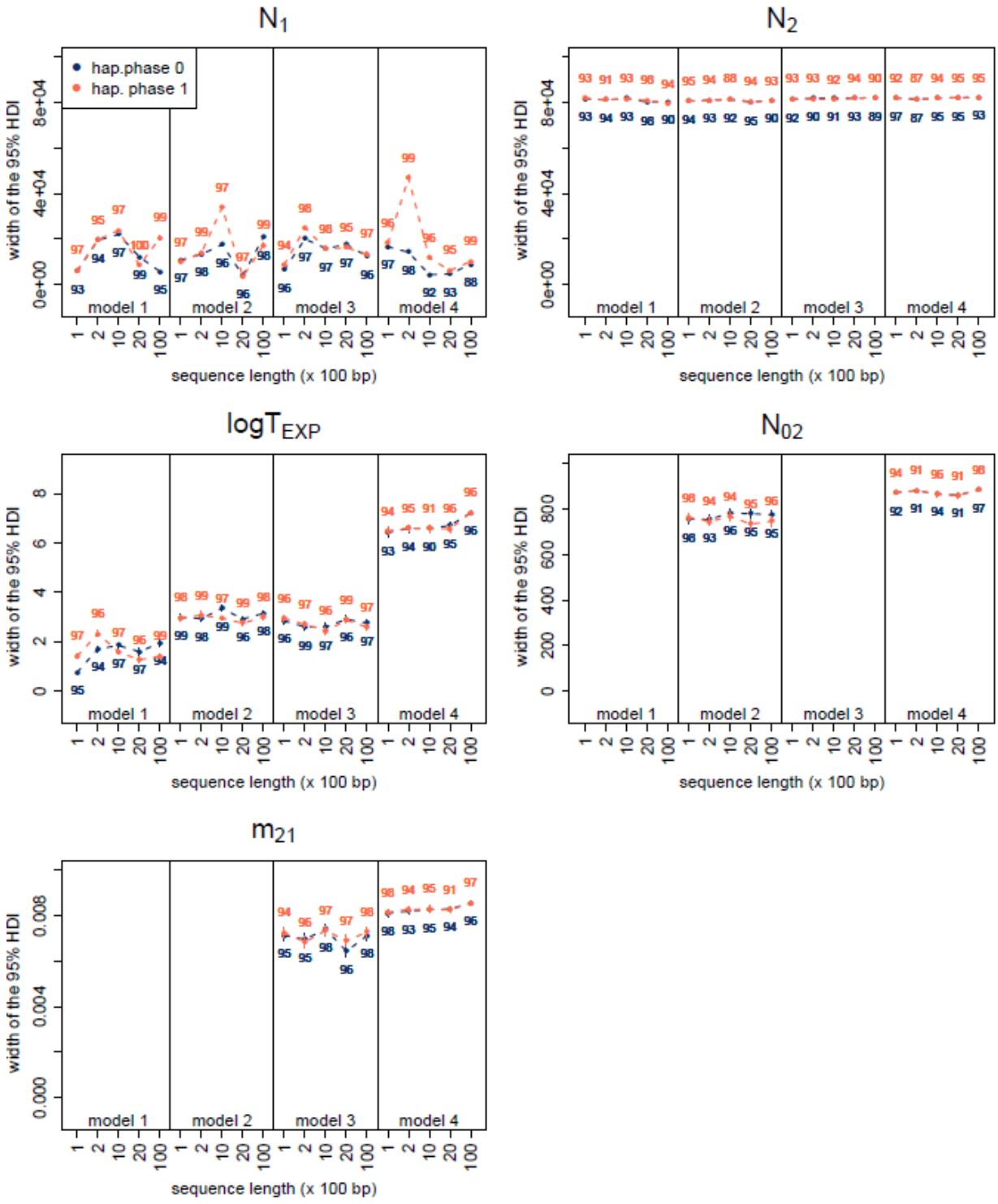


Figure 4.3 Width of the 95% highest posterior density intervals calculated from the results of ABC analyses of 20 different combinations of demographic models and sampling designs. Error bars represent standard errors (N=100 PODs). See caption of Figure 4.2 for more details.

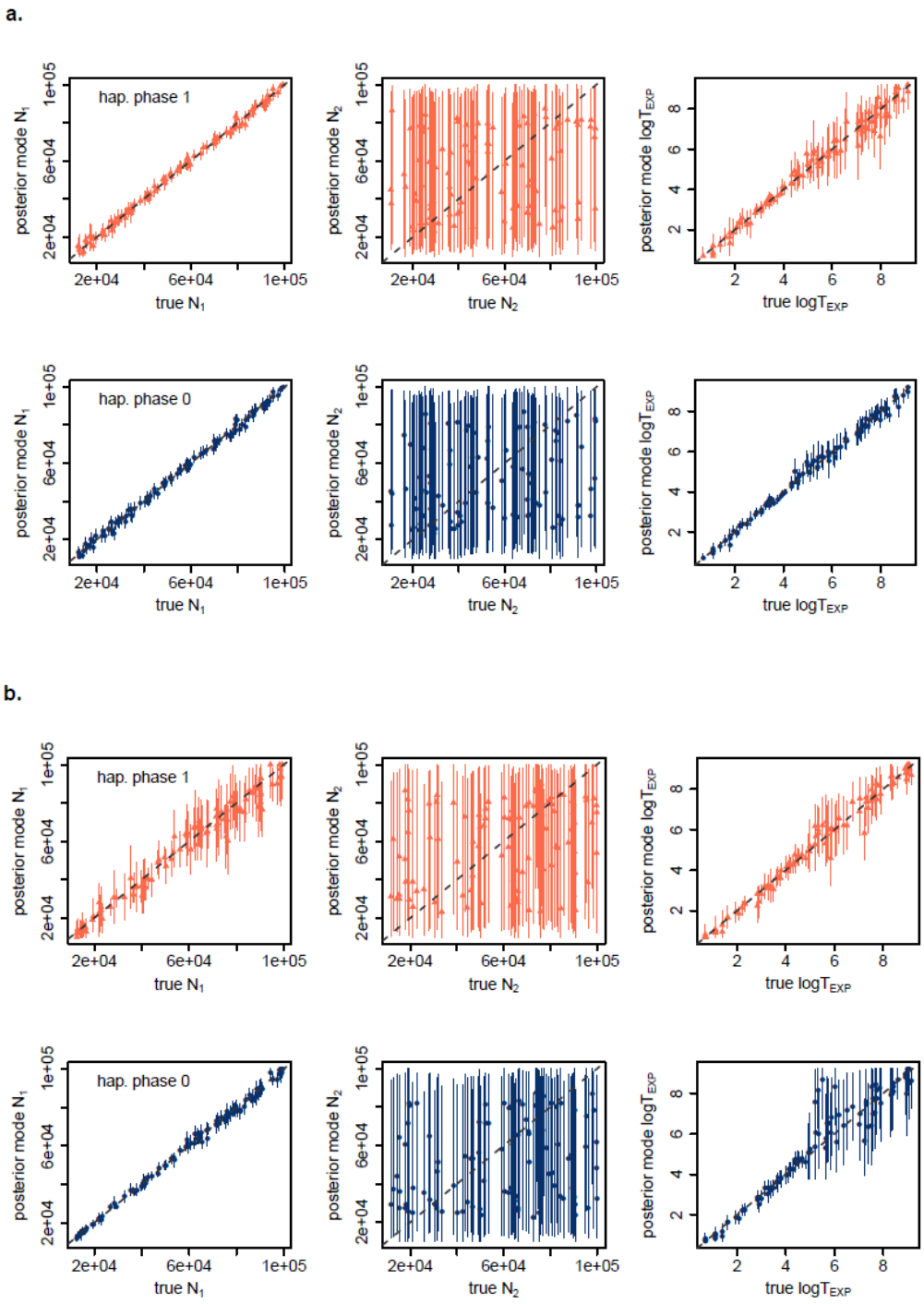


Figure 4.4 Accuracy of parameter estimates for model 1. Each data point corresponds to the estimated value of the parameter vs. the true parameter value for one POD, for a total of 100 PODs. Error bars correspond to the 95% HDI around the estimate. a) Results with datasets of type 1. b) Results with datasets of type 5. Top panels shows results on unphased datasets, bottom panels shows results for phased datasets.

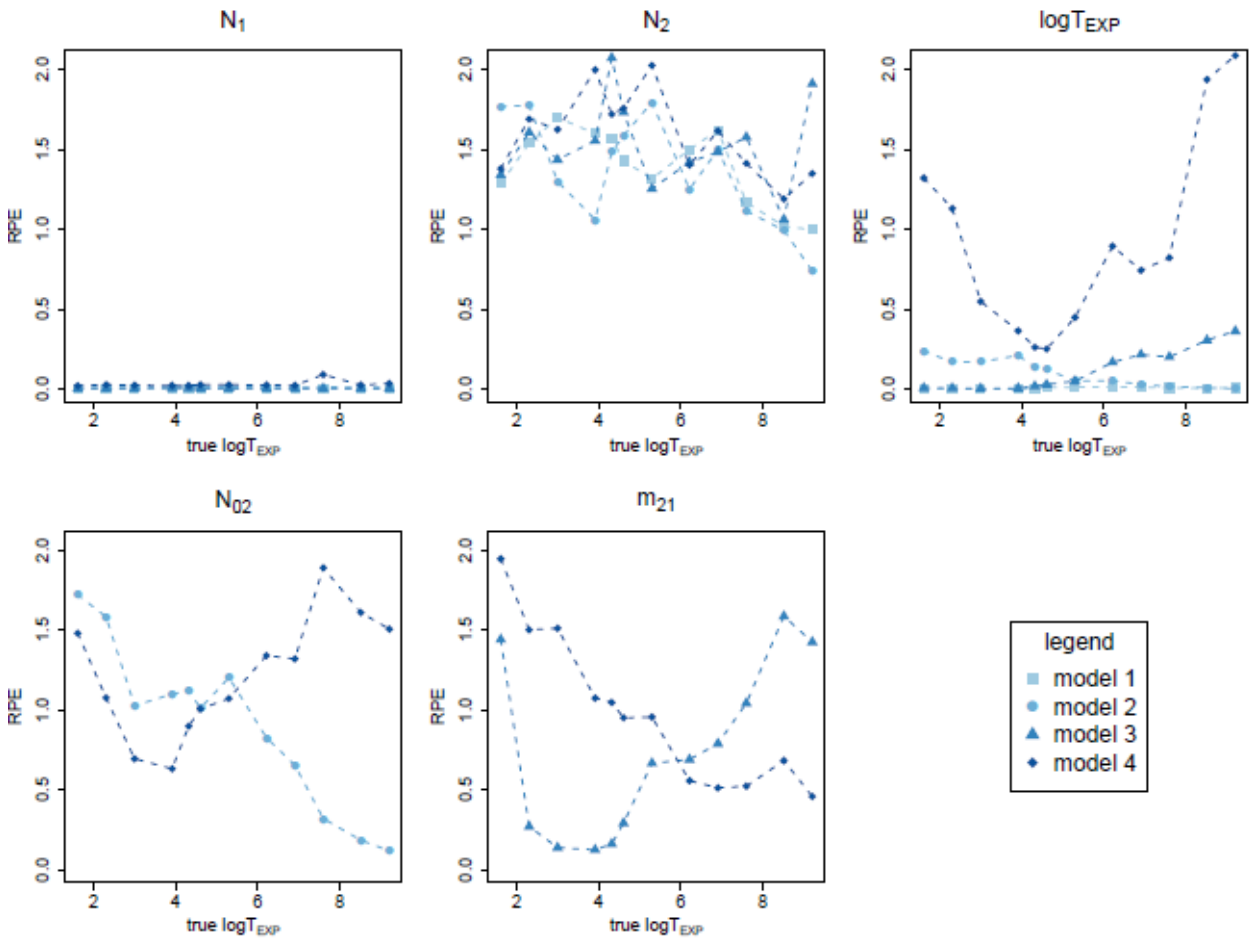


Figure 4.5 RPE of model parameters for different fixed values of T_{EXP}. Results are shown for ABC runs with datasets of type 1 (10k sequences, 100-bp long). For a given parameter, results from different models are shown in the same plot window with different characters and colours. To see results for other model-dataset combinations as well as 95% HDI results, see Appendix D, Figures D1-D8.

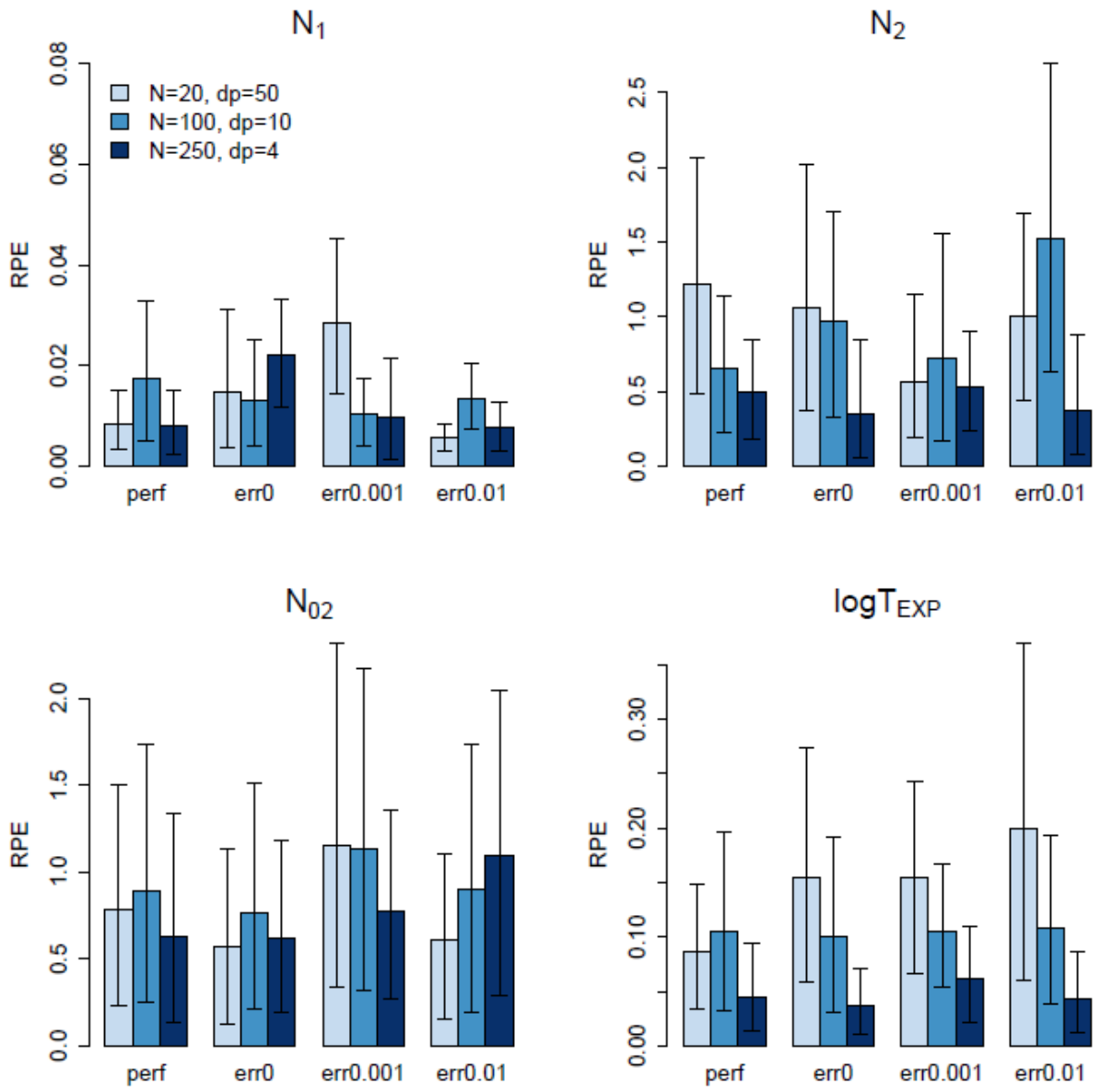


Figure 4.6 RPE and bootstrapped confidence intervals of model 2 parameters under different sequencing strategies and per-nucleotide error rates. N corresponds to the number of diploid individuals sequenced, dp to the mean individual sequencing depth. "perf" corresponds to perfect datasets whereas "err0", "err0.001" and "err0.01" correspond to datasets where the sequencing process was simulated, with depth sampling and errors introduced at rates 0, 0.001, and 0.01 substitutions per nucleotide respectively. 13 PODs were used for each treatment.

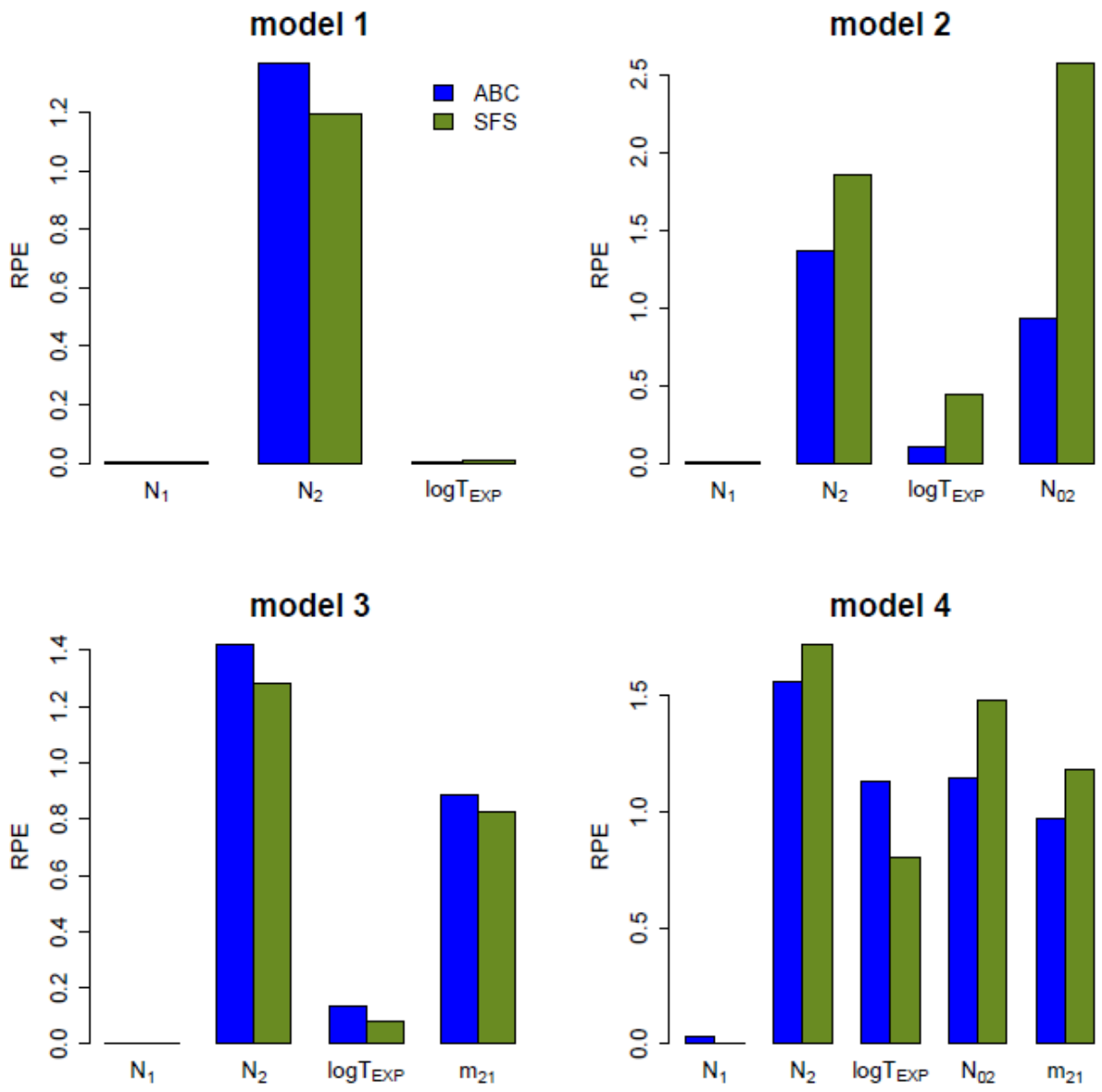


Figure 4.7 RPE calculated from 100 datasets for models 1 to 4 using two different inference methods: ABC, computed on SNP-level summary statistics, and approximate composite likelihood, computed from the SFS. In both cases, datasets had 10,000 sequences of 100bp genotyped in 20 diploid individuals.

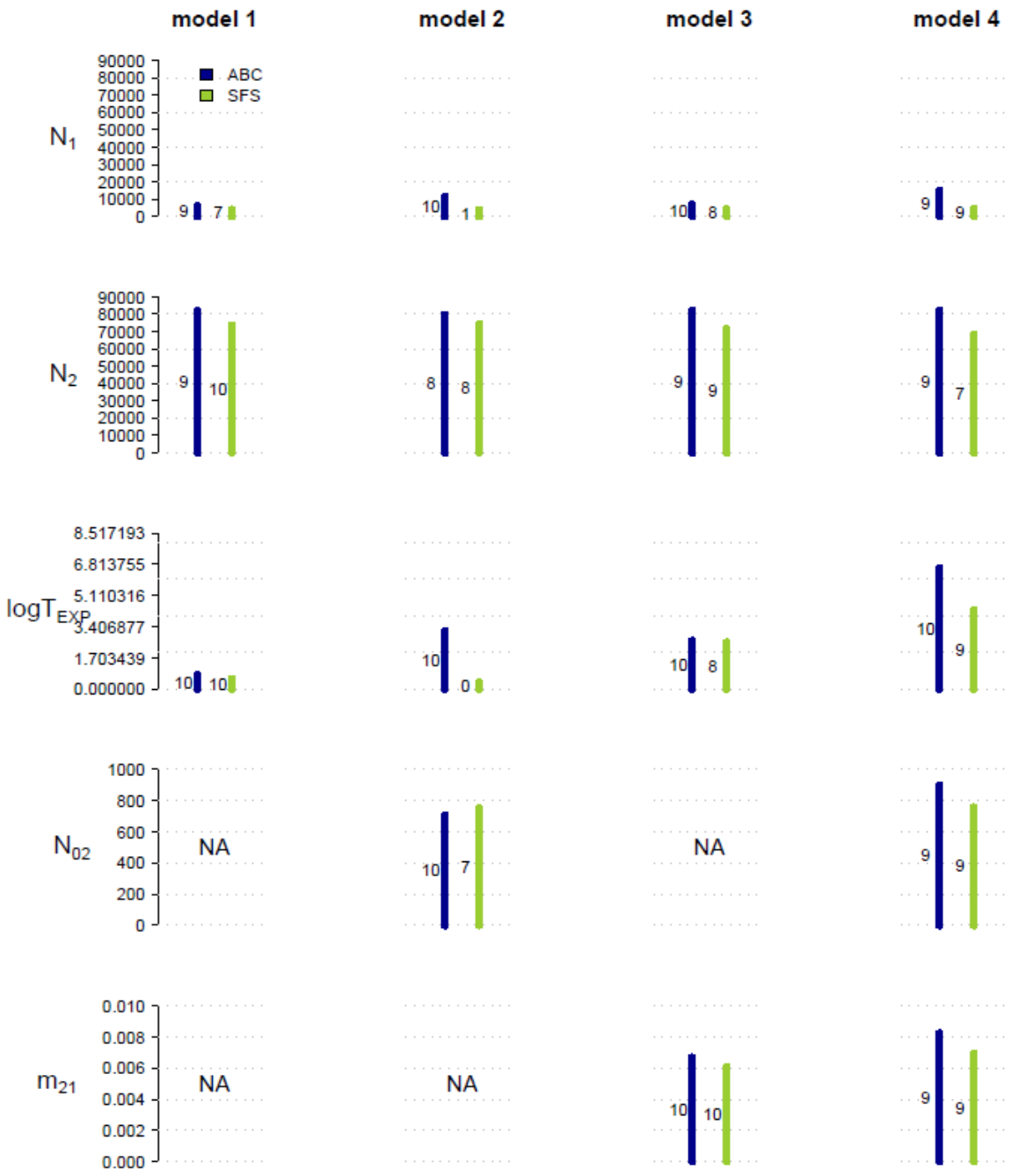


Figure 4.8 Width of the 95% HDI from ABC results, compared to 95% CI from the SFS inference method. For each of the four demographic models, the same 10 simulated datasets were used as pseudo-observed datasets for both the ABC and the SFS runs. HDI and CI widths were calculated from 100 bootstraps. Numbers correspond to the coverage of 95% CI (out of 10 PODs). PODs had 10,000 sequences of 100bp genotyped in 20 diploid individuals.

Chapter 5: Conclusion

5.1 The past: a natural laboratory to predict the future

The current pace of anthropogenic climate change raises great concerns about the health and persistence of communities, species and populations all over the planet. As for many other taxa, persistence of tree populations will either necessitate adaptation in already colonized areas or migration into novel territory (Aitken et al., 2008). The history of temperate tree species on all continents of the northern hemisphere involved similar climate-induced changes during the Quaternary, with large-scale post-glacial northward migration into new habitat at the colonization front and continuous local adaptation in established populations behind the front (Davis & Shaw, 2001). Examining the effects of these past climatic changes such as population migration rates and current local adaptation is therefore the best way to predict the future of currently established tree populations (Petit et al., 2008). In particular, characterizing the potential pace of colonization and adaptation of tree populations is of paramount importance to determine whether tree species will be able to track their climatic niches or persist in the southernmost areas of their ranges. As the genetic composition of individuals and populations is one of the most persisting types of evidence of past demographic changes, many tools and methods involving population genetics have been developed to interpret population changes in time and space, each suited to a particular spatial and temporal scale (Figure 1.1). The research presented in this thesis aims at improving our understanding of the interplay between demographic and genetic patterns. It addresses several aspects of their mutual influences in natural populations using Sitka spruce (*Picea sitchensis*) as a study organism. As several approaches were used in this research, it is critical to consider their strengths and limitations at the spatial and temporal scales they can address.

5.2 Conservation genomics and phylogeography

Next-generation sequencing (NGS) techniques have provided exciting new insights into management prospects for species of conservation or economic concern, through the identification of conservation units, assessments of connectivity among fragmented populations, and the detection of inbreeding or local adaptation (Garner et al., 2016). Phylogeographic methods are a strong component of the conservation genetics toolkit through their use in unravelling population history and therefore finding the cause of population decline, characterizing the effects of landscape fragmentation, and determining routes of introduction of invasive species. The field of phylogeography has been positively impacted by the shift from genetic to genomic datasets, as the complex demographic history of wild populations cannot be precisely estimated without a set of markers that are representative of the whole genome. Many algorithms and associated programs have been developed to perform demographic inference of population history (Bourgeois, 2016; Excoffier & Heckel, 2006; Schraiber & Akey, 2015). The need to test the power of such methods and assess their limitation has started being addressed recently (Li & Jakobsson, 2012; Robinson et al., 2014; Shafer, Gattepaille, et al., 2015). Chapter 4 aims to contribute to this body of literature by extending our knowledge of the possibilities associated with one of the most popular and flexible methods, approximate Bayesian computation (ABC). By focusing on demographic inference using the types of genomic datasets obtained from the two most common reduced-representation sequencing methods, sequence capture and genotyping-by-sequencing (GBS), this research specifically addressed the application of ABC to nonmodel species, for which whole-genome sequencing methods are not yet fully available. GBS methods particularly hold great promise in term of trade-offs between cost and potential applications to conservation (Narum et al., 2013). Results from Chapter 4 highlight the effect of model complexity on the success of demographic inference, focusing on a widely applicable 2-population spatial expansion model.

Recommendations for the practical implementation of ABC on this type of models were also provided to help orient potential future ABC users in their choice of genomic technique for marker development and distribution of sequencing effort. In particular, the simulation study performed in Chapter 4 showed a discrepancy in inference success among datasets generated

from recent vs. more ancient expansion events. This highlights the limitations associated with large prior distributions for model parameters. Because different summary statistics might be sensitive to different parameter ranges, it is important to select summary statistics most sensitive to the expected order of magnitude of the expansion event, and choose prior distributions and prior ranges accordingly. By tailoring the ABC process to the specific demographic models and timing considered to have shaped the history of populations, it will theoretically be possible to infer events that occurred at different time scales, for instance, ancestral secondary contact and recent expansion. One could do so by performing ABC independently for each event using differentially optimized steps of ABC and feeding the result of one analysis into the other in a dynamic manner.

Chapter 4 also discussed the lack of comprehensive user-friendly software for the application of demographic inference in nonmodel organisms with genomic data, a concern that has been raised in the past (Shafer, Wolf, et al., 2015). In my opinion, conservation biologists should not need advanced programming skills to perform demographic inference on species of conservation concern. Currently, many steps of the ABC procedure are still challenging to put in practice with genomic data. The extraction of useful data from raw sequencing reads without the loss of precious information about diversity characteristics requires extensive research into poorly-documented software features. In Appendix E, I developed elaborate scripts to be able to use *P. sitchensis* sequences, including monomorphic ones, as genetic markers in demographic inference of postglacial expansion. The specificity of summary statistic computation programs in terms of available calculations and input format also presented great challenges. The availability of user-friendly computational methods in phylogeography could make research more efficient and help reduce the gap between fundamental research in ecology and evolution and its applications to management and species conservation.

5.3 Adaptation and genetic diversity: the case of temperate tree species

Climate change is causing range shifts in nearly all terrestrial taxa (Parmesan, 2006). Temperate tree species are no exception to this, with evidence for upward and northward range shifts

observed in many northern hemisphere tree species (Iverson & McKenzie, 2013). For already established populations, a change in current climatic conditions may not only create a mismatch between local phenotypes and new conditions (Allen et al., 2010; Cleland et al., 2007), lowering overall population fitness, but also create an imbalance in biological interactions with commensals and parasites, leading to high mortality in affected populations (Bentz et al., 2010). The recovery of populations from climate change-induced disturbances relies on the existence locally of phenotypes allowing individuals to adapt to new conditions. Climatic adaptation in temperate and boreal tree species involves fine-tuned phenological responses to seasonal changes in climatic conditions (Howe et al., 2003). The genetic architecture of quantitative traits involved local adaptation is typically characterized by many genes of small effect (Alberto et al., 2013). The extent of adaptation of tree populations under climate change therefore relies on the genetic variance of quantitative traits, which in turn depends on genetic diversity in populations. In tree species and especially conifers, nucleotide diversity is low, but adaptive diversity among populations is high (Savolainen & Pyhäjärvi, 2007). This suggests that adaptive capacity and resilience of tree populations depends on the maintenance of standing genetic variation through gene flow. Understanding patterns of gene flow within and among tree populations and identifying the factors that influence them can helps predict the outcome of climate change. As populations at the front of expanding ranges are likely to experience new environmental conditions, I focused this research on populations at range margins. The research presented in Chapter 2 and 3 is part of such effort. The concomitant analysis of tree rings and genome-wide neutral markers provided an empirical illustration of the pace of recovery of genetic diversity and gene flow patterns in a recently established population at the range limit of a widespread conifer species, *Picea sitchensis*.

5.4 Genetic patterns of spatial expansion: empirical evidence for theoretical predictions

5.4.1 Intraspecific patterns

Model- and simulation-based studies of the evolutionary fate of populations expanding at range limits have greatly helped creating a framework of potential outcomes of colonization in tree

species. We know from past theoretical work that successive colonization events create a gradient of decreasing heterozygosity that is stronger when the time between colonizing events is short and long-distance seed dispersal is rare (Austerlitz & Garnier-Géré, 2003; Le Corre & Kremer, 1998). Similarly, Allee effects at the expansion front allow the recovery of genetic diversity (Roques et al., 2012) and high levels of pollen dispersal attenuate the loss of genetic diversity in diffusion models of spatial expansion (Austerlitz & Garnier-Géré, 2003). The findings of Chapter 2 contribute to the emerging body of literature providing empirical support to the abundant theoretical work describing genetic processes at expanding range limits (Excoffier et al., 2009). The main results from Chapter 2 are in line with several previous empirical studies. First, low population growth dominated the Kodiak Archipelago for several hundred years before the 18th century, and in spite of low population growth, this is the period during which new alleles accumulated most rapidly in the population, suggesting that allelic richness is quickly recovered during early stages of population establishment before local recruitment becomes substantial. Similar findings are described in Lesser & Jackson (2013) and could probably be observed in most conifer species for which gene flow occurs predominantly via pollen dispersal. The following scenario provides an interpretation of this result. In the case of isolated habitat, such as the Kodiak Archipelago for *P. sitchensis*, occasional long-distance seed dispersal is indispensable to initiate colonization. If some seeds disperse and establish successfully, the founders originating from such events need to grow to maturity, produce female cones, and intercept pollen to finally produce offspring. It takes additional decades for offspring to, in turn, become reproductively mature. As seed dispersal events are rare and have a low chance to lead to successful establishment without facilitation from previously established trees, founders by long distance seed dispersal are likely sparse, and probably rely entirely on foreign pollen for reproduction, assuming self-pollination is not common. While limiting the pace of spatial expansion, this unusually long period of reliance on foreign pollen from diverse sites in nearby populations can allow for an efficient recovery of genetic diversity after the population bottleneck triggered by founding events. In Chapter 2 I also showed that a period of fast population growth on the Kodiak Archipelago after initial establishment was accompanied by a weakening of the initially higher genetic differentiation among different sites at the front, and an end to the erosion of differentiation between the archipelago and continental populations. These

patterns are consistent with the hypothesis of a local pollen cloud becoming dominant and homogenizing the population at the regional level. Indeed, after several generations have built up population density in an establishing population, local pollen should become much more abundant than foreign pollen. Evidence for these shifts in temporal patterns at multiple spatial scales is one of the strength of analyses in Chapter 2 bringing novelty to the empirical literature of range expansions studies in tree species. Finally, the findings that allelic frequencies in the population remained relatively stable during demographic growth and F_{ST} with continental population only insignificantly increased suggests that the genetic composition of the recently established *P. sitchensis* population largely reflects the genotypes of its oldest trees, stressing the fundamental role of early colonizers in the long-term genetic composition of establishing populations. This hypothesis is supported by Troupin et al. (2006) at a more limited spatial and temporal scale.

5.4.2 Range expansion and species boundaries

With the ongoing redistribution of climatic envelopes everywhere on earth, range shifts are likely to create contact between previously geographically separate species. Species boundaries are permeable between phylogenetically close lineages, and some range shifts will result in the creation of new hybrid zones. In addition, existing hybrid zones maintained by climatic gradients will also shift. The genomes of widespread, wind-dispersed tree species are especially permeable: examples of hybrid zones between closely related tree species are numerous (De la Torre, 2015), and this feature has been related to other characteristics such as long generation times, large pollen dispersal distances, and a predominantly outcrossing mating system. Spruce hybrid complexes have been intensively studied in North America, with no evidence for intrinsic reproductive barriers between sister species (De la Torre et al., 2013; Hamilton et al., 2013a). These spruce hybrid zones are maintained by climatic gradients (Hamilton et al., 2015). In the light of this knowledge, it is not surprising that historical secondary contacts during the Holocene resulted in significant introgression between *Picea* species in North America. The recent colonization of *P. sitchensis* on the Kodiak Archipelago, surrounded by *P. glauca* and *P. sitchensis* populations, provided an unprecedented opportunity to examine extremely recent patterns of hybridization between the two species and reconstruct the evolution of these patterns

over several generations. Although hybridization can result in the formation of adaptive genotypes in novel environments, theoretical work also suggests that neutral introgression at range margins with closely related species can be enhanced in a context of spatial expansion (Currat & Excoffier, 2011). In Chapter 3 I took advantage of a unique natural research setup involving both an accurate description of range expansion (P. *sitchensis* colonization of the Kodiak Archipelago, Chapter 2) and potential hybridization with a parapatric species (*P. glauca*). The results of this chapter are both interesting and insufficient to draw firm conclusions. Indeed, they describe compelling evidence that continuous pollen flow from nearby *P. glauca* populations produced early-generation hybrids on the Kodiak Archipelago in early colonizers across sampled sites. They also suggest an unexpected absence of recombinant hybrids in subsequent generations, and a complete lack of *P. glauca* ancestry in the genetic composition of trees grown under a fully formed canopy. These patterns suggest that selection is acting against hybrids in late stages of colonization. The most obvious hypothesis explaining the nature of selection against hybrids is that pollen flow from *P. glauca* population is maladaptive, introducing to the mild, wet maritime archipelago genotypes that are adapted to more continental climates with higher continentality and lower amounts of precipitation. My analysis of radial growth on the Kodiak Archipelago suggested hybrids and pure *P. sitchensis* differ in their sensitivity to precipitation, but these differences were not significant. Although the results of Chapter 3 are not definitive, they spark questions related to topics of importance with respect to adaptation concerns, such as the influence of direction of gene flow. The typically high levels of gene flow in tree species has been invoked to explain the adaptive potential of populations, but if gene flow to a marginal population is dominated by populations adapted to different conditions, then they may prevent local adaptation. Although there is little empirical evidence supporting this gene-swamping hypothesis in tree species, this is likely to become common at southern range limits of northern hemisphere species. Indeed, the colonization of preadapted genotypes is likely at leading range limits, but trailing range limits will experience climatic conditions present nowhere else in the species range, with the only possible gene flow coming from core populations adapted to different climates. Studies examining conditions allowing adaptation to local climates under maladaptive gene flow would therefore be useful for predicting the future distribution of tree species.

5.4.3 Strengths and limitations of dendrogenetic approaches

Pairing dendrochronology and genetics in the study of population history takes full advantage of the exceptional longevitiy of conifers and provides a great opportunity to understand fine-scale processes of tree population expansion. The longevity of temperate trees and their growth variability is captured in their annual rings, and can allow a detailed reconstruction of past demographic changes over several centuries. When associated with genetic data from individual trees, demographic changes observed in natural populations can be directly associated with their evolutionary effect, as shown in Chapter 2. Applying the monitoring of recent changes in allelic frequencies to the study of adaptive loci is one promising future application of the combination of tree ring and genetic methods, and would contribute to identifying recent adaptive responses to climate change in natural populations. The identification of age cohorts within a tree population sample can also help analyse the contribution of distinct populations or species to the population gene pool at different periods, as shown in Chapter 3. With adequate tree core sampling, it is even possible to compare absolute radial growth amounts between genetically distinct individuals in a population and identify more or less adaptive genotypes in regard to certain local conditions (Housset et al., 2016).

However, dendrogenetic approaches come with intrinsic limitations, the most obvious one being the absence in the sample of trees that have died prior to sampling. Because the cumulative mortality of trees increases inherently as one goes back in time, the inference of evolutionary states and processes that happened further in the past will be less reliable. Indeed, establishment dates calculated from the tree ring data will be less likely to reflect population establishment in the more distant than in the more recent past. In addition to this bias due to tree mortality, the high sensitivity of tree rings to a myriad of different environmental and intrinsic factors challenges the interpretation of variation in tree ring width patterns. Statistical detrending of tree ring series is necessary to isolate the frequency of variation associated with the phenomenon of interest, but leads to an inherent loss of information. Averaging variation over individual trees is also necessary before performing correlation analyses with genotype classes or processes at the stand level. This leads to a loss in statistical power to detect the effect of genotypes or environmental factors on tree radial growth.

5.5 Implications for potential management practices

Whether temperate tree species will be able to track their environmental niche and adapt to new conditions arising from rapid climate change is a concerning question, and active management solutions proposed recently, such as assisted migration for species of economic or conservation concern, are ready to be put to practice (Aitken & Bemmels, 2016). Research presented in this dissertation supports the long-standing idea that pollen flow across large geographic areas contributes to the maintenance of high levels of genetic diversity in tree species, and emphasizes its fundamental role during early stages of colonization. The main concern about populations in regions under adaptive pressure might therefore not be related to low effective population sizes and genetic isolation, even in newly established populations at range margins. Instead, the direction of gene flow connecting populations across regions might be of crucial importance: in this light, the southernmost populations of temperate and boreal tree species might struggle to adapt under gene flow from source populations further north. Chapter 2 also shows that the genetic composition of a newly established population will largely reflect the genotype of founding trees for several subsequent generations. This result echoes simulation work conducted by Kuparinen et al. (2010), which suggests that high mortality and periodic large-scale disturbances such as wildfires increase the speed of local adaptation to climate by removing old trees and boosting selection during the subsequent enhanced recruitment period. These results should be considered in the framework of assisted gene flow, especially in widespread species where large-scale harvesting and planting is common practice: anticipating the climatic conditions that trees selected for planting will be adapted to when they reach high fertility, decades or centuries later, will probably contribute to keep these partially managed forests adapted to a changing climate.

References

Aitken, S. N., & Bemmels, J. B. (2016). Time to get moving: assisted gene flow of forest trees. *Evolutionary Applications*, *9*, 271–290. https://doi.org/10.1111/eva.12293

Aitken, S. N., & Whitlock, M. C. (2013). Assisted gene flow to facilitate local adaptation to climate change. *Annual Review of Ecology, Evolution, and Systematics*, *44*, 367–388. https://doi.org/10.1146/annurev-ecolsys-110512-135747

Aitken, S. N., Yeaman, S., Holliday, J. A., Wang, T., & Curtis-McLane, S. (2008). Adaptation, migration or extirpation: climate change outcomes for tree populations. *Evolutionary Applications*, *1*, 95–111. https://doi.org/10.1111/j.1752-4571.2007.00013.x

Alberto, F. J., Aitken, S. N., Alía, R., González-Martínez, S. C., Hänninen, H., Kremer, A., … Savolainen, O. (2013). Potential for evolutionary responses to climate change - evidence from tree populations. *Global Change Biology*, *19*, 1645–1661. https://doi.org/10.1111/gcb.12181

Allen, C. D., Macalady, A. K., Chenchouni, H., Bachelet, D., McDowell, N., Vennetier, M., … Cobb, N. (2010). A global overview of drought and heat-induced tree mortality reveals emerging climate change risks for forests. *Forest Ecology and Management*, *259*, 660–684. https://doi.org/10.1016/j.foreco.2009.09.001

Amorim, C. E. G., Hofer, T., Ray, N., Foll, M., Ruiz-Linares, A., & Excoffier, L. (2017). Long-distance dispersal suppresses introgression of local alleles during range expansions. *Heredity*, *118*, 135. https://doi.org/10.1038/hdy.2016.68

Anderson, L. L., Hu, F. S., Nelson, D. M., Petit, R. J., & Paige, K. N. (2006). Ice-age endurance: DNA evidence of a white spruce refugium in Alaska. *Proceedings of the National Academy of Sciences of the United States of America*, *103*, 12447–50. https://doi.org/10.1073/pnas.0605310103

Austerlitz, F., & Garnier-Géré, P. H. (2003). Modelling the impact of colonisation on genetic diversity and differentiation of forest trees: interaction of life cycle, pollen flow and seed long-distance dispersal. *Heredity*, *90*, 282–90. https://doi.org/10.1038/sj.hdy.6800243

Austerlitz, F., Jung-Muller, B., Godelle, B., & Gouyon, P.-H. (1997). Evolution of coalescence times, genetic diversity and structure during colonization. *Theoretical Population Biology*, *51*, 148–164. https://doi.org/10.1006/tpbi.1997.1302

Austerlitz, F., Mariette, S., Machon, N., Gouyon, P., & Godelle, B. (2000). Effects of colonization processes on genetic diversity : differences between annual plants and tree species. *Genetics*, *154*, 1309–1321.

Babst, F., Poulter, B., Trouet, V., Tan, K., Neuwirth, B., Wilson, R., … Frank, D. (2013). Site- and species-specific responses of forest growth to climate across the European continent: Climate sensitivity of forest growth across Europe. *Global Ecology and Biogeography*, *22*, 706–717. https://doi.org/10.1111/geb.12023

Bacles, C. F. E., Lowe, A. J., & Ennos, R. A. (2006). Effective seed dispersal across a fragmented landscape. *Science*, *311*, 628–628. https://doi.org/10.1126/science.1121543

Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics*, *164*, 1139–1160.

Benazzo, A., Ghirotto, S., Vilaça, S. T., & Hoban, S. (2015). Using ABC and microsatellite data to detect multiple introductions of invasive species from a single source. *Heredity*, *115*, 262–272. https://doi.org/10.1038/hdy.2015.38

Bentz, B. J., Régnière, J., Fettig, C. J., Hansen, E. M., Hayes, J. L., Hicke, J. A., … Seybold, S. J. (2010). Climate change and bark beetles of the western United States and Canada: direct and indirect effects. *BioScience*, *60*, 602–613. https://doi.org/10.1525/bio.2010.60.8.6

Bertorelle, G., Benazzo, A., & Mona, S. (2010). ABC as a flexible framework to estimate demography over space and time: Some cons, many pros. *Molecular Ecology*, *19*, 2609–2625. https://doi.org/10.1111/j.1365-294X.2010.04690.x

Bhatia, G., Patterson, N., Sankararaman, S., & Price, A. L. (2013). Estimating and interpreting F_{ST}: The impact of rare variants. *Genome Research*, *23*, 1514–1521. https://doi.org/10.1101/gr.154831.113

Bialozyt, R., Ziegenhagen, B., & Petit, R. J. (2006). Contrasting effects of long distance seed dispersal on genetic diversity during range expansion. *Journal of Evolutionary Biology*, *19*, 12–20. https://doi.org/10.1111/j.1420-9101.2005.00995.x

Birol, I., Raymond, A., Jackman, S. D., Pleasance, S., Coope, R., Taylor, G. A., … Jones, S. J. M. (2013). Assembling the 20 Gb white spruce (Picea glauca) genome from whole-genome shotgun sequencing data. *Bioinformatics (Oxford, England)*, *29*, 1492–7. https://doi.org/10.1093/bioinformatics/btt178

Blankenberg, D., Gordon, A., Von Kuster, G., Coraor, N., Taylor, J., & Nekrutenko, A. (2010). Manipulation of FASTQ data with Galaxy. *Bioinformatics*, *26*, 1783–1785. https://doi.org/10.1093/bioinformatics/btq281

Blum, M. G. B., Nunes, M. A., Prangle, D., & Sisson, S. A. (2013). A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science*, *28*, 189–208. https://doi.org/10.1214/12-STS406

Boitard, S., Rodríguez, W., Jay, F., Mona, S., & Austerlitz, F. (2016). Inferring population size history from large samples of genome-wide molecular data: an approximate Bayesian computation approach. *PLOS Genetics*, *12*, e1005877–e1005877. https://doi.org/10.1371/journal.pgen.1005877

Born, C., Kjellberg, F., Chevallier, M.-H., Vignes, H., Dikangadissi, J.-T., Sanguié, J., … Hossaert-McKey, M. (2008). Colonization processes and the maintenance of genetic diversity: insights from a pioneer rainforest tree, *Aucoumea klaineana*. *Proceedings. Biological Sciences / The Royal Society*, *275*, 2171–9. https://doi.org/10.1098/rspb.2008.0446

Bosela, M., Popa, I., Gömöry, D., Longauer, R., Tobin, B., Kyncl, J., … Büntgen, U. (2016). Effects of post-glacial phylogeny and genetic diversity on the growth variability and climate sensitivity of European silver fir. *Journal of Ecology*, *104*, 716–724. https://doi.org/10.1111/1365-2745.12561

Boucher, T. V., & Mead, B. R. (2006). Vegetation change and forest regeneration on the Kenai Peninsula, Alaska, following a spruce beetle outbreak, 1987–2000. *Forest Ecology and Management*, *227*, 233–246. https://doi.org/10.1016/j.foreco.2006.02.051

Bouillé, M., Senneville, S., & Bousquet, J. (2010). Discordant mtDNA and cpDNA phylogenies indicate geographic speciation and reticulation as driving factors for the diversification of the genus *Picea*. *Tree Genetics & Genomes*, *7*, 469–484. https://doi.org/10.1007/s11295-010-0349-z

Boulesteix, A.-L., & Strimmer, K. (2007). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, *8*, 32–44. https://doi.org/10.1093/bib/bbl016

Bourgeois, Y. (2016). Going down the rabbit hole: a review on how to link genome-wide data with ecology and evolution in natural populations. *bioRxiv*, 052761. https://doi.org/10.1101/052761

Bowman, P. W. (1934). Pollen analysis of Kodiak bogs. *Ecology*, *15*, 97–100.

Box, G. E. P. (1979). Robustness in the Strategy of Scientific Model Building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in Statistics* (pp. 201–236). Academic Press. https://doi.org/10.1016/B978-0-12-438150-6.50018-2

Brown, G. R., Gill, G. P., Kuntz, R. J., Langley, C. H., & Neale, D. B. (2004). Nucleotide diversity and linkage disequilibrium in loblolly pine. *Proceedings of the National Academy of Sciences*, *101*, 15255–15260. https://doi.org/10.1073/pnas.0404231101

Bunn, A. G. (2008). A dendrochronology program library in R (dplR). *Dendrochronologia*, *26*, 115–124. https://doi.org/10.1016/j.dendro.2008.01.002

Büntgen, U., Frank, D. C., Kaczka, R. J., Verstege, A., Zwijacz-Kozica, T., & Esper, J. (2007). Growth responses to climate in a multi-species tree-ring network in the Western Carpathian Tatra Mountains, Poland and Slovakia. *Tree Physiology*, *27*, 689–702.

Burczyk, J., Adams, W. T., Moran, G. F., & Griffin, A. R. (2002). Complex patterns of mating revealed in a *Eucalyptus* regnans seed orchard using allozyme markers and the neighbourhood model. *Molecular Ecology*, *11*, 2379–2391. https://doi.org/10.1046/j.1365-294X.2002.01603.x

Cabrera, A. A., & Palsbøll, P. J. (2017). Inferring past demographic changes from contemporary genetic data: A simulation-based evaluation of the ABC methods implemented in DIYABC. *Molecular Ecology Resources*. https://doi.org/10.1111/1755-0998.12696

Cann, R. L., Stoneking, M., & Wilson, A. C. (1987). Mitochondrial DNA and human evolution. *Nature*, *325*, 31–36. https://doi.org/10.1038/325031a0

Carstens, B. C., Brennan, R. S., Chua, V., Duffie, C. V., Harvey, M. G., Koch, R. A., … Sullivan, J. (2013). Model selection as a tool for phylogeographic inference: an example from the willow *Salix melanopsis*. *Molecular Ecology*, *22*, 4014–4028. https://doi.org/10.1111/mec.12347

Chan, Y. L., Schanzenbach, D., & Hickerson, M. J. (2014). Detecting concerted demographic response across community assemblages using hierarchical approximate Bayesian

computation. *Molecular Biology and Evolution*, *31*, 2501–15. https://doi.org/10.1093/molbev/msu187

Chen, Jun. (2012). *Conifer evolution, from demography and local adaptation to evolutionary rates. Examples from the Picea genus*. Uppsala University.

Chen, Jun, Källman, T., Gyllenstrand, N., & Lascoux, M. (2010). New insights on the speciation history and nucleotide diversity of three boreal spruce species and a Tertiary relict. *Heredity*, *104*, 3–14. https://doi.org/10.1038/hdy.2009.88

Chen, Jun, Källman, T., Ma, X., Gyllenstrand, N., Zaina, G., Morgante, M., … Lascoux, M. (2012). Disentangling the Roles of History and Local Selection in Shaping Clinal Variation of Allele Frequencies and Gene Expression in Norway Spruce (Picea abies). *Genetics*, genetics.112.140749. https://doi.org/10.1534/genetics.112.140749

Chouakria, A. D., & Nagabhushan, P. N. (2007). Adaptive dissimilarity index for measuring time series proximity. *Advances in Data Analysis and Classification*, *1*, 5–21. https://doi.org/10.1007/s11634-006-0004-6

Clark, J. S. (1998). Why trees migrate so fast: confronting theory with dispersal biology and the paleorecord. *The American Naturalist*, *152*, 204–24. https://doi.org/10.1086/286162

Cleland, E. E., Chuine, I., Menzel, A., Mooney, H. A., & Schwartz, M. D. (2007). Shifting plant phenology in response to global change. *Trends in Ecology & Evolution*, *22*, 357–365. https://doi.org/10.1016/j.tree.2007.04.003

Connor, S. E., Leeuwen, J. F. N. van, Rittenour, T. M., Knaap, W. O. van der, Ammann, B., & Björck, S. (2012). The ecological impact of oceanic island colonization – a palaeoecological perspective from the Azores. *Journal of Biogeography*, *39*, 1007–1023. https://doi.org/10.1111/j.1365-2699.2011.02671.x

Cook, E. R., & Kairiukstis, L. A. (Eds.). (1990). *Methods of Dendrochronology*. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-015-7879-0

Coombe, L., Warren, R. L., Jackman, S. D., Yang, C., Vandervalk, B. P., Moore, R. A., … Birol, I. (2016). Assembly of the complete Sitka spruce chloroplast genome using 10X Genomics' GemCode sequencing data. *PLOS ONE*, *11*, e0163059. https://doi.org/10.1371/journal.pone.0163059

Cornille, A., Giraud, T., Bellard, C., Tellier, A., Le Cam, B., Smulders, M. J. M., … Gladieux, P. (2013). Postglacial recolonization history of the European crabapple (*Malus sylvestris Mill.*), a wild contributor to the domesticated apple. *Molecular Ecology*, *22*, 2249–2263. https://doi.org/10.1111/mec.12231

Cottrell, J. E., Krystufek, V., Tabbener, H. E., Milner, A. D., Connolly, T., Sing, L., … Dam, B. C. van. (2005). Postglacial migration of *Populus nigra* L.: lessons learnt from chloroplast DNA. *Forest Ecology and Management*, *206*, 71–90. https://doi.org/10.1016/j.foreco.2004.10.052

Cramer, J. M., Mesquita, R. C. G., & Bruce Williamson, G. (2007). Forest fragmentation differentially affects seed dispersal of large and small-seeded tropical trees. *Biological Conservation*, *137*, 415–423. https://doi.org/10.1016/j.biocon.2007.02.019

Csilléry, K., Blum, M. G. B., Gaggiotti, O. E., & François, O. (2010). Approximate Bayesian computation (ABC) in practice. *Trends in Ecology & Evolution*, *25*, 410–8. https://doi.org/10.1016/j.tree.2010.04.001

Currat, M., & Excoffier, L. (2011). Strong reproductive isolation between humans and Neanderthals inferred from observed patterns of introgression. *Proceedings of the National Academy of Sciences*, *108*, 15129–15134. https://doi.org/10.1073/pnas.1107450108

Currat, M., Ruedi, M., Petit, R. J., & Excoffier, L. (2008). The hidden side of invasions: massive introgression by local genes. *Evolution; International Journal of Organic Evolution*, *62*, 1908–20. https://doi.org/10.1111/j.1558-5646.2008.00413.x

Darling, E., Samis, K. E., & Eckert, C. G. (2008). Increased seed dispersal potential towards geographic range limits in a Pacific coast dune plant. *The New Phytologist*, *178*, 424–35. https://doi.org/10.1111/j.1469-8137.2007.02349.x

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, *12*, 499–510. https://doi.org/10.1038/nrg3012

Davis, M. B., & Shaw, R. G. (2001). Range shifts and adaptive responses to Quaternary climate change. *Science (New York, N.Y.)*, *292*, 673–9. https://doi.org/10.1126/science.292.5517.673

De la Torre, A. R. (2015). Genomic admixture and species delimitation in forest trees. In *Evolutionary Biology: Biodiversification from Genotype to Phenotype* (pp. 287–303). Cham: Springer.

De la Torre, A. R., Li, Z., Van de Peer, Y., & Ingvarsson, P. K. (2017). Contrasting Rates of Molecular Evolution and Patterns of Selection among Gymnosperms and Flowering Plants. *Molecular Biology and Evolution*, *34*, 1363–1377. https://doi.org/10.1093/molbev/msx069

De la Torre, A. R., Wang, T., Jaquish, B., & Aitken, S. N. (2013). Adaptation and exogenous selection in a *Picea glauca* × *Picea engelmannii* hybrid zone: implications for forest management under climate change. *New Phytologist*, *201*, 687–699. https://doi.org/10.1111/nph.12540

Depaulis, F., & Veuille, M. (1998). Neutrality tests based on the distribution of haplotypes under an infinite-site model. *Molecular Biology and Evolution*, *15*, 1788–1790. https://doi.org/10.1093/oxfordjournals.molbev.a025905

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*, 491. https://doi.org/10.1038/ng.806

Du Fang, K., Petit, R. J., & Liu Jian Quan. (2009). More introgression with less gene flow: chloroplast vs. mitochondrial DNA in the *Picea asperata* complex in China, and comparison with other conifers. *Molecular Ecology*, *18*, 1396–1407. https://doi.org/10.1111/j.1365-294X.2009.04107.x

Duncan, R. P. (1989). An evaluation of errors in tree age estimates based on increment cores in kahikatea (*Dacrycarpus dacrydioides*). *New Zealand Natural Sciences*, *16*, 1–37.

Dussex, N., Wegmann, D., & Robertson, B. C. (2014). Postglacial expansion and not human influence best explains the population structure in the endangered kea (*Nestor notabilis*). *Molecular Ecology*, *23*, 2193–2209. https://doi.org/10.1111/mec.12729

Eckert, C. G., Samis, K. E., & Lougheed, S. C. (2008). Genetic variation across species' geographical ranges: the central-marginal hypothesis and beyond. *Molecular Ecology*, *17*, 1170–88. https://doi.org/10.1111/j.1365-294X.2007.03659.x

Edmonds, C. a, Lillie, A. S., & Cavalli-Sforza, L. L. (2004). Mutations arising in the wave front of an expanding population. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 975–9. https://doi.org/10.1073/pnas.0308064100

El Mousadik, A., & Petit, R. J. (1996). High level of genetic differentiation for allelic richness among populations of the argan tree [*Argania spinosa* (L.) Skeels] endemic to Morocco. *Theoretical and Applied Genetics*, *92*, 832–839.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, *6*, e19379. https://doi.org/10.1371/journal.pone.0019379

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., & Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genetics*, *9*, e1003905. https://doi.org/10.1371/journal.pgen.1003905

Excoffier, L., & Foll, M. (2011). *fastsimcoal*: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics*, *27*, 1332–1334. https://doi.org/10.1093/bioinformatics/btr124

Excoffier, L., Foll, M., & Petit, R. J. (2009). Genetic consequences of range expansions. *Annual Review of Ecology, Evolution, and Systematics*, *40*, 481–501. https://doi.org/10.1146/annurev.ecolsys.39.110707.173414

Excoffier, L., & Heckel, G. (2006). Computer programs for population genetics data analysis: a survival guide. *Nature Reviews Genetics*, *7*, 745–758. https://doi.org/10.1038/nrg1904

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

Farr, W. A., & Harris, A. (1979). Site Index of Sitka spruce along the Pacific coast related to latitude and temperatures. *Forest Sciences*, *25*, 145–153.

Fearnhead, P., & Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *74*, 419–474. https://doi.org/10.1111/j.1467-9868.2011.01010.x

Feurdean, A., Bhagwat, S. A., Willis, K. J., Birks, H. J. B., Lischke, H., & Hickler, T. (2013). Tree migration rates: narrowing the gap between inferred postglacial rates and projected rates. *PLOS ONE*, *8*, e71797. https://doi.org/10.1371/journal.pone.0071797

Fitzpatrick, B. M. (2012). Estimating ancestry and heterozygosity of hybrids using molecular markers. *BMC Evolutionary Biology*, *12*, 131.

Fumagalli, M. (2013). Assessing the effect of sequencing depth and sample size in population genetics inferences. *PLoS One*, *8*, e79667.

Fumagalli, M., Vieira, F. G., Linderoth, T., & Nielsen, R. (2014). *ngsTools*: methods for population genetics analyses from next-generation sequencing data. *Bioinformatics*, *30*, 1486–1487. https://doi.org/10.1093/bioinformatics/btu041

Gao, J., Wang, B., Mao, J.-F., Ingvarsson, P., Zeng, Q.-Y., & Wang, X.-R. (2012). Demography and speciation history of the homoploid hybrid pine *Pinus densata* on the Tibetan Plateau. *Molecular Ecology*, *21*, 4811–4827. https://doi.org/10.1111/j.1365-294X.2012.05712.x

Gapare, W. J., & Aitken, S. N. (2005). Strong spatial genetic structure in peripheral but not core populations of Sitka spruce [*Picea sitchensis* (Bong.) Carr.]. *Molecular Ecology*, *14*, 2659–67. https://doi.org/10.1111/j.1365-294X.2005.02633.x

Garner, B. A., Hand, B. K., Amish, S. J., Bernatchez, L., Foster, J. T., Miller, K. M., … Luikart, G. (2016). Genomics in conservation: case studies and bridging the gap between data and application. *Trends in Ecology & Evolution*, *31*, 81–83. https://doi.org/10.1016/j.tree.2015.10.009

Garrick, R. C., Bonatelli, I. A. S., Hyseni, C., Morales, A., Pelletier, T. A., Perez, M. F., … Carstens, B. C. (2015). The evolution of phylogeographic data sets. *Molecular Ecology*, *24*, 1164–1171. https://doi.org/10.1111/mec.13108

Gavin, D. G., Fitzpatrick, M. C., Gugger, P. F., Heath, K. D., Rodríguez-Sánchez, F., Dobrowski, S. Z., … Williams John W. (2014). Climate refugia: joint inference from fossil records, species distribution models and phylogeography. *New Phytologist*, *204*, 37–54. https://doi.org/10.1111/nph.12929

Gompert, Z., & Buerkle, C. A. (2010). introgress: a software package for mapping components of isolation in hybrids. *Molecular Ecology Resources*, *10*, 378–384. https://doi.org/10.1111/j.1755-0998.2009.02733.x

Griggs, R. F. (1914). Observations on the edge of the forest in the Kodiak region of Alaska. *Bulletin of Torrey Botanical Club*, *41*, 381–385.

Griggs, R. F. (1937). Timberlines as indicators of climatic trends. *Science*, *85*, 251–255.

Gugger, P. F., Sugita, S., & 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. https://doi.org/10.1111/j.1365-294X.2010.04622.x

Guillemaud, T., Beaumont, M. A., Ciosi, M., Cornuet, J.-M., & Estoup, A. (2010). Inferring introduction routes of invasive species using approximate Bayesian computation on microsatellite data. *Heredity*, *104*, 88–99. https://doi.org/10.1038/hdy.2009.92

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP

frequency data. *PLOS Genetics*, *5*, e1000695. https://doi.org/10.1371/journal.pgen.1000695

Hallatschek, O., & Nelson, D. R. (2008). Gene surfing in expanding populations. *Theoretical Population Biology*, *73*, 158–70. https://doi.org/10.1016/j.tpb.2007.08.008

Hallatschek, O., & Nelson, D. R. (2010). Life at the front of an expanding population. *Evolution; International Journal of Organic Evolution*, *64*, 193–206. https://doi.org/10.1111/j.1558-5646.2009.00809.x

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

Hamilton, G., Currat, M., Ray, N., Heckel, G., Beaumont, M., & Excoffier, L. (2005). Bayesian estimation of recent migration rates after a spatial expansion. *Genetics*, *170*, 409–17. https://doi.org/10.1534/genetics.104.034199

Hamilton, J. A., & Aitken, S. N. (2013). Genetic and morphological structure of a spruce hybrid (*Picea sitchensis x P. glauca*) zone along a climatic gradient. *American Journal of Botany*, *100*, 1651–1662. https://doi.org/10.3732/ajb.1200654

Hamilton, J. A., Lexer, C., & Aitken, S. N. (2013a). Differential introgression reveals candidate genes for selection across a spruce (*Picea sitchensis x P . glauca*) hybrid zone. *The New Phytologist*, *197*, 927–38. https://doi.org/10.1111/nph.12055

Hamilton, J. A., Lexer, C., & Aitken, S. N. (2013b). Genomic and phenotypic architecture of a spruce hybrid zone (*Picea sitchensis* × *P. glauca*). *Molecular Ecology*, *22*, 827–41. https://doi.org/10.1111/mec.12007

Hamilton, J. A., Torre, A. R. D. la, & Aitken, S. N. (2015). Fine-scale environmental variation contributes to introgression in a three-species spruce hybrid complex. *Tree Genetics & Genomes*, *11*, 817. https://doi.org/10.1007/s11295-014-0817-y

Hampe, A. (2011). Plants on the move: The role of seed dispersal and initial population establishment for climate-driven range expansions. *Acta Oecologica*, *37*, 666–673. https://doi.org/10.1016/j.actao.2011.05.001

Hampe, A., Pemonge, M., & Petit, R. J. (2013). Efficient mitigation of founder effects during the establishment of a leading-edge oak population. *Proceedings of the Royal Society B*.

Hamrick, J. L., & Trapnell, D. W. (2011). Using population genetic analyses to understand seed dispersal patterns. *Acta Oecologica*, *37*, 641–649. https://doi.org/10.1016/j.actao.2011.05.008

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

Hargreaves, A. L., Samis, K. E., & Eckert, C. G. (2014). Are species' range limits simply niche limits writ large? A review of transplant experiments beyond the range. *The American Naturalist*, *183*, 157–73. https://doi.org/10.1086/674525

Harris, K., & Nielsen, R. (2013). Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genetics*, *9*, e1003521–e1003521. https://doi.org/10.1371/journal.pgen.1003521

Harvey, M. G., Smith, B. T., Glenn, T. C., Faircloth, B. C., & Brumfield, R. T. (2016). Sequence capture versus restriction site associated DNA sequencing for shallow systematics. *Systematic Biology*, *65*, 910–924. https://doi.org/10.1093/sysbio/syw036

Heer, K., Behringer, D., Piermattei, A., Bässler, C., Brandl, R., Fady, B., … Opgenoorth, L. (2018). Linking dendroecology and association genetics in natural populations: Stress responses archived in tree rings associate with SNP genotypes in silver fir (*Abies alba* Mill.). *Molecular Ecology*, *27*, 1428–1438. https://doi.org/10.1111/mec.14538

Heuertz, M., Paoli, E. D., Källman, T., Larsson, H., Jurman, I., Morgante, M., … Gyllenstrand, N. (2006). Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce [*Picea abies* (L.) Karst]. *Genetics*, *174*, 2095–2105. https://doi.org/10.1534/genetics.106.065102

Hewitt, G. (2000). The genetic legacy of the Quaternary ice ages. *Nature*, *405*, 907–13. https://doi.org/10.1038/35016000

Holliday, J. A., Ritland, K., & Aitken, S. N. (2010). Widespread, ecologically relevant genetic markers developed from association mapping of climate-related traits in Sitka spruce (*Picea sitchensis*). *The New Phytologist*, *188*, 501–14. https://doi.org/10.1111/j.1469-8137.2010.03380.x

Holliday, J. A., Suren, H., & Aitken, S. N. (2012). Divergent selection and heterogeneous migration rates across the range of Sitka spruce (*Picea sitchensis*). *Proceedings. Biological Sciences / The Royal Society*, *279*, 1675–83. https://doi.org/10.1098/rspb.2011.1805

Holliday, J. A., Yuen, M., Ritland, K., & Aitken, S. N. (2010). Postglacial history of a widespread conifer produces inverse clines in selective neutrality tests. *Molecular Ecology*, *19*, 3857–64. https://doi.org/10.1111/j.1365-294X.2010.04767.x

Holmes, R. L. (2000). Cofecha. *Laboratory of Tree-Ring Research*.

Housset, J. M., Carcaillet, C., Girardin, M. P., Xu, H., Tremblay, F., & Bergeron, Y. (2016). In situ comparison of tree-ring responses to climate and population genetics: the need to control for local climate and site variables. *Frontiers in Ecology and Evolution*, *4*. https://doi.org/10.3389/fevo.2016.00123

Howe, G. T., Aitken, S. N., Neale, D. B., Jermstad, K. D., Wheeler, N. C., & Chen, T. H. H. (2003). From genotype to phenotype : unraveling the complexities of cold adaptation in forest trees. *Canadian Journal of Botany*, *1266*, 1247–1266. https://doi.org/10.1139/B03-141

Hu, F. S., Hampe, A., & Petit, R. J. (2008). Paleoecology meets genetics: deciphering past vegetational dynamics. *Frontiers in Ecology and the Environment*, *7*, 371–379. https://doi.org/10.1890/070160

Ibrahim, K. M., Nichols, R. A., & Hewitt, G. M. (1996). Spatial patterns of genetic variation generated by different forms of dispersal during range expansion, *77*, 282–291.

Iverson, L. R., & McKenzie, D. (2013). Tree-species range shifts in a changing climate: detecting, modeling, assisting. *Landscape Ecology*, *28*, 879–889. https://doi.org/10.1007/s10980-013-9885-x

Jaramillo-Correa, J. P., Beaulieu, J., Khasa, D. P., & 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. https://doi.org/10.1139/X08-181

Johnson, J. S., Gaddis, K. D., Cairns, D. M., Konganti, K., & Krutovsky, K. V. (2017). Landscape genomic insights into the historic migration of mountain hemlock in response to Holocene climate change. *American Journal of Botany*, *104*, 439–450. https://doi.org/10.3732/ajb.1600262

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

Jones, M. C. (2008). *Climate and vegetation history from late-glacial and Holocene peat from the Kenai Peninsula, Alaska: a record of pollen, macrofossils, stable isotopes, and carbon storage.* Columbia University.

Jorge, M. L. S. P., & Howe, H. F. (2009). Can forest fragmentation disrupt a conditional mutualism? A case from central Amazon. *Oecologia*, *161*, 709–718. https://doi.org/10.1007/s00442-009-1417-7

Kelly, J. K. (1997). A test of neutrality based on interlocus associations. *Genetics*, *146*, 1197–1206.

Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and Their Applications*, *13*, 235–248. https://doi.org/10.1016/0304-4149(82)90011-4

Kirkpatrick, M., & Barton, N. H. (1997). Evolution of a species' range. *The University of Chicago Press for The American Society of Naturalists*, *150*, 1–23.

Kitamura, K., Matsui, T., Kobayashi, M., Saitou, H., Namikawa, K., & Tsuda, Y. (2015). Decline in gene diversity and strong genetic drift in the northward-expanding marginal populations of *Fagus crenata*. *Tree Genetics & Genomes*, *11*. https://doi.org/10.1007/s11295-015-0857-y

Klopfstein, S., Currat, M., & Excoffier, L. (2006). The fate of mutations surfing on the wave of a range expansion. *Molecular Biology and Evolution*, *23*, 482–90. https://doi.org/10.1093/molbev/msj057

Knowles, L. L., & Maddison, W. P. (2002). Statistical phylogeography. *Molecular Ecology*, *11*, 2623–2635.

Kremer, A., Ronce, O., Robledo-Arnuncio, J. J., Guillaume, F., Bohrer, G., Nathan, R., … Schueler, S. (2012). Long-distance gene flow and adaptation of forest trees to rapid climate change. *Ecology Letters*, 378–392. https://doi.org/10.1111/j.1461-0248.2012.01746.x

Kuparinen, A., Savolainen, O., & Schurr, F. M. (2010). Increased mortality can promote evolutionary adaptation of forest trees to climate change. *Forest Ecology and Management*, *259*, 1003–1008. https://doi.org/10.1016/j.foreco.2009.12.006

Lauterjung, M. B., Bernardi, A. P., Montagna, T., Candido-Ribeiro, R., da Costa, N. C. F., Mantovani, A., & dos Reis, M. S. (2018). Phylogeography of Brazilian pine (*Araucaria angustifolia*): integrative evidence for pre-Columbian anthropogenic dispersal. *Tree Genetics & Genomes*, *14*. https://doi.org/10.1007/s11295-018-1250-4

Le Corre, V., & Kremer, A. (1998). Cumulative effects of founding events during colonisation on genetic diversity and differentiation in an island and stepping-stone model. *Journal of Evolutionary Biology*, *11*, 495–495. https://doi.org/10.1007/s000360050102

Le Corre, V., Machon, N., Petit, R. J., & Kremer, A. (1997). Colonization with long-distance seed dispersal and genetic structure of maternally inherited genes in forest trees: a simulation study. *Genetical Research*, *69*, 117–125.

Lesser, M. R., & Jackson, S. T. (2013). Contributions of long-distance dispersal to population growth in colonising *Pinus ponderosa* populations. *Ecology Letters*, *16*, 380–9. https://doi.org/10.1111/ele.12053

Lesser, M. R., Parchman, T. L., & Jackson, S. T. (2013). Development of genetic diversity, differentiation and structure over 500 years in four ponderosa pine populations. *Molecular Ecology*, *22*, 2640–52. https://doi.org/10.1111/mec.12280

Leuenberger, C., & Wegmann, D. (2010). Bayesian computation and model selection without likelihoods. *Genetics*, *184*, 243–252. https://doi.org/10.1534/genetics.109.109058

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, *475*, 493–496. https://doi.org/10.1038/nature10231

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, S., & Jakobsson, M. (2012). Estimating demographic parameters from large-scale population genomic data using approximate Bayesian computation. *BMC Genetics*, *13*, 22–22. https://doi.org/10.1186/1471-2156-13-22

Li, Y., Stocks, M., Hemmila, S., Kallman, T., Zhu, H., Zhou, Y., … Lascoux, M. (2010). Demographic histories of four spruce (*Picea*) species of the Qinghai-Tibetan Plateau and neighboring areas inferred from multiple nuclear loci. *Molecular Biology and Evolution*, *27*, 1001–1014. https://doi.org/10.1093/molbev/msp301

Lintusaari, J., Gutmann, M. U., Dutta, R., Kaski, S., & Corander, J. (2016). Fundamentals and recent developments in approximate Bayesian computation. *Systematic Biology*, syw077-syw077. https://doi.org/10.1093/sysbio/syw077

Liu, S., Lorenzen, E. D., Fumagalli, M., Li, B., Harris, K., Xiong, Z., … Wang, J. (2014). Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. *Cell*, *157*, 785–794. https://doi.org/10.1016/j.cell.2014.03.054

Lobo, N. (2011). *Clinal variation at putatively adaptive polymorphisms in mature populations of Sitka spruce*. University of British Columbia.

Lohse, K., Harrison, R. J., & Barton, N. H. (2011). A general method for calculating likelihoods under the coalescent process. *Genetics*, *189*, 977–987. https://doi.org/10.1534/genetics.111.129569

Lopez, A. D., Mathers, C. D., Ezzati, M., Jamison, D. T., & Murray, C. J. (2006). Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. *Lancet*, *367*, 1747–1757.

Lumibao, C. Y., Hoban, S. M., & McLachlan, J. (2017). Ice ages leave genetic diversity "hotspots" in Europe but not in Eastern North America. *Ecology Letters*, *20*, 1459–1468. https://doi.org/10.1111/ele.12853

MacLachlan, I. R., Wang, T., Hamann, A., Smets, P., & Aitken, S. N. (2017). Selective breeding of lodgepole pine increases growth and maintains climatic adaptation. *Forest Ecology and Management*, *391*, 404–416. https://doi.org/10.1016/j.foreco.2017.02.008

MacLachlan, I. R., Yeaman, S., & Aitken, S. N. (2017). Growth gains from selective breeding in a spruce hybrid zone do not compromise local adaptation to climate. *Evolutionary Applications*, *11*, 166–181. https://doi.org/10.1111/eva.12525

Magri, D., Vendramin, G. G., Comps, B., Dupanloup, I., Geburek, T., Gömöry, D., … de Beaulieu, J. (2006). A new scenario for the Quaternary history of European beech populations: palaeobotanical evidence and genetic consequences. *New Phytologist*, *171*, 199–221. https://doi.org/10.1111/j.1469-8137.2006.01740.x

Mahony, C. R., Cannon, A. J., Wang, T., & Aitken, S. N. (2017). A closer look at novel climates: new methods and insights at continental to landscape scales. *Global Change Biology*, *23*, 3934–3955. https://doi.org/10.1111/gcb.13645

Manel, S., Schwartz, M. K., Luikart, G., & Taberlet, P. (2003). Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology & Evolution*, *18*, 189–197. https://doi.org/10.1016/S0169-5347(03)00008-9

Mann, D. H., & Hamilton, T. D. (1995). Late Pleistocene and Holocene paleoenvironments of the North Pacific coast. *Quaternary Science Reviews*, *14*, 449–471. https://doi.org/10.1016/0277-3791(95)00016-I

Marin, J.-M., Pillai, N. S., Robert, C. P., & Rousseau, J. (2014). Relevant statistics for Bayesian model choice. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*, 833–859.

Marin, J.-M., Pudlo, P., Robert, C. P., & Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing*, *22*, 1167–1180. https://doi.org/10.1007/s11222-011-9288-2

Marsico, T. D., Hellmann, J. J., & Romero-Severson, J. (2009). Patterns of seed dispersal and pollen flow in *Quercus garryana* (Fagaceae) following post-glacial climatic changes. *Journal of Biogeography*, *36*, 929–941. https://doi.org/10.1111/j.1365-2699.2008.02049.x

Martin-Benito, D., & Pederson, N. (2015). Convergence in drought stress, but a divergence of climatic drivers across a latitudinal gradient in a temperate broadleaf forest. *Journal of Biogeography*, *42*, 925–937. https://doi.org/10.1111/jbi.12462

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*, 1297–1303. https://doi.org/10.1101/gr.107524.110

McLachlan, J. S., Clark, J. S., & Manos, P. S. (2005). Molecular indicators of tree migration capacity under rapid climate change. *Ecology*, *86*, 2088–2098.

McVean, G. A. T., & Cardin, N. J. (2005). Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*, 1387–1393. https://doi.org/10.1098/rstb.2005.1673

Mimura, M., & Aitken, S. N. (2007a). Adaptive gradients and isolation-by-distance with postglacial migration in *Picea sitchensis*. *Heredity*, *99*, 224–32. https://doi.org/10.1038/sj.hdy.6800987

Mimura, M., & Aitken, S. N. (2007b). Increased selfing and decreased effective pollen donor number in peripheral relative to central populations in *Picea sitchensis*. *American Journal of Botany*, *94*, 991–998.

Mimura, M., & Aitken, S. N. (2010). Local adaptation at the range peripheries of Sitka spruce. *Journal of Evolutionary Biology*, *23*, 249–58. https://doi.org/10.1111/j.1420-9101.2009.01910.x

Nadachowska-Brzyska, K., Burri, R., Olason, P. I., Kawakami, T., Smeds, L., & Ellegren, H. (2013). Demographic divergence history of pied flycatcher and collared flycatcher inferred from whole-genome re-sequencing data. *PLoS Genetics*, *9*, e1003942. https://doi.org/10.1371/journal.pgen.1003942

Narum, S. R., Buerkle, C. A., Davey, J. W., Miller, M. R., & Hohenlohe, P. A. (2013). Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology*, *22*, 2841–2847. https://doi.org/10.1111/mec.12350

Nathan, R. (2006). Long-distance dispersal of plants. *Science*, *313*, 786–788.

Nathan, R., Katul, G. G., Bohrer, G., Kuparinen, A., Soons, M. B., Thompson, S. E., … Horn, H. S. (2011). Mechanistic models of seed dispersal by wind. *Theoretical Ecology*, *4*, 113–132. https://doi.org/10.1007/s12080-011-0115-3

Neale, D. B., Langley, C. H., Salzberg, S. L., & Wegrzyn, J. L. (2013). Open access to tree genomes: the path to a better forest. *Genome Biology*, *14*, 120.

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., & Wang, J. (2012). SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing Data. *PLoS ONE*, *7*, e37558. https://doi.org/10.1371/journal.pone.0037558

Pan, J., Wang, B., Pei, Z.-Y., Zhao, W., Gao, J., Mao, J.-F., & Wang, X.-R. (2015). Optimization of the genotyping-by-sequencing strategy for population genomic analysis in conifers. *Molecular Ecology Resources*, *15*, 711–722. https://doi.org/10.1111/1755-0998.12342

Parducci, L., Jørgensen, T., Tollefsrud, M. M., Elverland, E., Alm, T., Fontana, S. L., … Willerslev, E. (2012). Glacial survival of boreal trees in northern Scandinavia. *Science (New York, N.Y.)*, *335*, 1083–6. https://doi.org/10.1126/science.1216043

Parmesan, C. (2006). Ecological and evolutionary responses to recent climate change. *Annual Review of Ecology, Evolution, and Systematics*, *37*, 637–669. https://doi.org/10.1146/annurev.ecolsys.37.091305.110100

Pavlidis, P., Laurent, S., & Stephan, W. (2010). msABC: a modification of Hudson's ms to facilitate multi-locus ABC analysis. *Molecular Ecology Resources*, *10*, 723–727. https://doi.org/10.1111/j.1755-0998.2010.02832.x

Peischl, S., Dupanloup, I., Kirkpatrick, M., & Excoffier, L. (2013). On the accumulation of deleterious mutations during range expansions. *Molecular Ecology*, *22*, 5972–82. https://doi.org/10.1111/mec.12524

Peteet, D. M. (1986). Modern pollen rain and vegetational history of the Malaspina Glacier district, Alaska. *Quaternary Research*, *25*, 100–120. https://doi.org/10.1016/0033-5894(86)90047-5

Peter, B. M., & Slatkin, M. (2013). Detecting range expansions from genetic data. *Evolution; International Journal of Organic Evolution*, *67*, 3274–89. https://doi.org/10.1111/evo.12202

Petit, R. J., Bialozyt, R., Garnier-Géré, P., & Hampe, A. (2004). Ecology and genetics of tree invasions: from recent introductions to Quaternary migrations. *Forest Ecology and Management*, *197*, 117–137. https://doi.org/10.1016/j.foreco.2004.05.009

Petit, R. J., Duminil, J., Fineschi, S., Hampe, A., Salvini, D., & Vendramin, G. G. (2005). Comparative organization of chloroplast, mitochondrial and nuclear diversity in plant populations. *Molecular Ecology*, *14*, 689–701. https://doi.org/10.1111/j.1365-294X.2004.02410.x

Petit, R. J., Hu, F. S., & Dick, C. W. (2008). Forests of the past: a window to future changes. *Science*, *320*, 1450–2. https://doi.org/10.1126/science.1155457

Petkova, D., Novembre, J., & Stephens, M. (2016). Visualizing spatial population structure with estimated effective migration surfaces. *Nature Genetics*, *48*, 94–100. https://doi.org/10.1038/ng.3464

Pluess, A. R. (2011). Pursuing glacier retreat: genetic structure of a rapidly expanding *Larix decidua* population. *Molecular Ecology*, *20*, 473–85. https://doi.org/10.1111/j.1365-294X.2010.04972.x

Polechová, J., Barton, N., & Marion, G. (2009). Species' range: adaptation in space and time. *The American Naturalist*, *174*, E186-204. https://doi.org/10.1086/605958

Prangle, D., Fearnhead, P., Cox, M. P., Biggs, P. J., & French, N. P. (2013). Semi-automatic selection of summary statistics for ABC model choice. *Statistical Applications in Genetics and Molecular Biology*, *13*, 67–82. https://doi.org/10.1515/sagmb-2013-0012

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

Pujol, B., & Pannell, J. R. (2008). Reduced responses to selection after species range expansion. *Science*, *321*, 96–96. https://doi.org/10.1126/science.1157570

Pyhäjärvi, T., García-Gil, M. R., Knürr, T., Mikkonen, M., Wachowiak, W., & Savolainen, O. (2007). Demographic History Has Influenced Nucleotide Diversity in European Pinus sylvestris Populations. *Genetics*, *177*, 1713–1724. https://doi.org/10.1534/genetics.107.077099

Quéméré, E., Amelot, X., Pierson, J., Crouau-Roy, B., & Chikhi, L. (2012). Genetic data suggest a natural prehuman origin of open habitats in northern Madagascar and question the deforestation narrative in this region. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 13028–33. https://doi.org/10.1073/pnas.1200153109

R Core Team. (2016). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.* Retrieved from https://www.R-project.org/

Rice, W., & Estert, E. (1993). Laboratory experiments on speciation: What have we learned in 40 years? *Evolution*, *47*, 1637–1652.

Rieseberg, L. H., Kim, S.-C., Randell, R. A., Whitney, K. D., Gross, B. L., Lexer, C., & Clay, K. (2007). Hybridization and the colonization of novel habitats by annual sunflowers. *Genetica*, *129*, 149–165. https://doi.org/10.1007/s10709-006-9011-y

Robert, C. P., Cornuet, J. M., Marin, J. M., & Pillai, N. S. (2011). Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 15112–15117. https://doi.org/10.1073/Pnas.1102900108

Robinson, J. D., Bunnefeld, L., Hearn, J., Stone, G. N., & Hickerson, M. J. (2014). ABC inference of multi-population divergence with admixture from unphased population genomic data. *Molecular Ecology*, *23*, 4458–4471. https://doi.org/10.1111/mec.12881

Roques, L., Garnier, J., Hamel, F., & Klein, E. K. (2012). Allee effect promotes diversity in traveling waves of colonization. *Proceedings of the National Academy of Sciences*, *109*, 8828–8833. https://doi.org/10.1073/pnas.1201695109

Rousi, M., Possen, B. J. M. H., Ruotsalainen, S., Silfver, T., & Mikola, J. (2017). Temperature and soil fertility as regulators of tree line Scots pine growth and survival—implications for the acclimation capacity of northern populations. *Global Change Biology*, *24*, e545–e559. https://doi.org/10.1111/gcb.13956

Savolainen, O., Kujala, S. T., Sokol, C., Pyhäjärvi, T., Avia, K., Knürr, T., … Hicks, S. (2011). Adaptive potential of northernmost tree populations to climate change, with emphasis on

Scots pine (Pinus sylvestris L.). *The Journal of Heredity*, *102*, 526–36. https://doi.org/10.1093/jhered/esr056

Savolainen, O., & Pyhäjärvi, T. (2007). Genomic diversity in forest trees. *Current Opinion in Plant Biology*, *10*, 162–167. https://doi.org/10.1016/j.pbi.2007.01.011

Savolainen, O., Pyhäjärvi, T., & Knürr, T. (2007). Gene flow and local adaptation in trees. *Annual Review of Ecology, Evolution, and Systematics*, *38*, 595–619.

Schiffels, S., & Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, *46*, 919–925. https://doi.org/10.1038/ng.3015

Schraiber, J. G., & Akey, J. M. (2015). Methods and models for unravelling human evolutionary history. *Nature Reviews Genetics*, *16*, 727–740. https://doi.org/10.1038/nrg4005

Sezen, U. U., Chazdon, R. L., & Holsinger, K. E. (2007). Multigenerational genetic analysis of tropical secondary regeneration in a canopy palm. *Ecology*, *88*, 3065–3075.

Shafer, A. B. A., Cullingham, C. I., Côté, S. D., & Coltman, D. W. (2010). Of glaciers and refugia: a decade of study sheds new light on the phylogeography of northwestern North America. *Molecular Ecology*, *19*, 4589–4621. https://doi.org/10.1111/j.1365-294X.2010.04828.x

Shafer, A. B. A., Gattepaille, L. M., Stewart, R. E. A., & Wolf, J. B. W. (2015). Demographic inferences using short-read genomic data in an approximate Bayesian computation framework: In silico evaluation of power, biases and proof of concept in Atlantic walrus. *Molecular Ecology*, *24*, 328–345. https://doi.org/10.1111/mec.13034

Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., & Wolf, J. B. W. (2016). Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods in Ecology and Evolution*, n/a-n/a. https://doi.org/10.1111/2041-210X.12700

Shafer, A. B. A., Wolf, J. B. W., Alves, P. C., Bergström, L., Bruford, M. W., Brännström, I., … Zieliński, P. (2015). Genomics and the challenging translation into conservation practice. *Trends in Ecology & Evolution*, *30*, 78–87. https://doi.org/10.1016/j.tree.2014.11.009

Shi, M., & Chen, X. (2012). Leading-edge populations do not show low genetic diversity or high differentiation in a wind-pollinated tree. *Population Ecology*, *54*, 591–600. https://doi.org/10.1007/s10144-012-0332-7

Slatkin, M., & Excoffier, L. (2012). Serial founder effects during range expansion: a spatial analog of genetic drift. *Genetics*, *191*, 171–81. https://doi.org/10.1534/genetics.112.139022

Smouse, P. E., Dyer, R. J., Westfall, R. D., & Sork, V. L. (2001). Two-generation analysis of pollen flow across a landscape. I. Male gamete heterogeneity among females. *Evolution*, *55*, 260. https://doi.org/10.1554/0014-3820(2001)055[0260:TGAOPF]2.0.CO;2

Sousa, V. C., Beaumont, M. A., Fernandes, P., Coelho, M. M., & Chikhi, L. (2012). Population divergence with or without admixture: selecting models using an ABC approach. *Heredity*, *108*, 521–530. https://doi.org/10.1038/hdy.2011.116

Staab, P. R., Zhu, S., Metzler, D., & Lunter, G. (2015). *scrm*: efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics (Oxford, England)*, *31*, 1680–2. https://doi.org/10.1093/bioinformatics/btu861

Stocks, M., Siol, M., Lascoux, M., & De Mita, S. (2014). Amount of information needed for model choice in approximate Bayesian computation. *PLoS ONE*, *9*, 1–13. https://doi.org/10.1371/journal.pone.0099581

Sunnaker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., & Dessimoz, C. (2013). Approximate Bayesian computation. *PLoS Computational Biology*, *9*. https://doi.org/10.1371/journal.pcbi.1002803

Suren, H., Hodgins, K. A., Yeaman, S., Nurkowski, K. A., Smets, P., Rieseberg, L. H., … Holliday, J. A. (2016). Exome capture from the spruce and pine giga-genomes. *Molecular Ecology Resources*, *16*, 1136–1146. https://doi.org/10.1111/1755-0998.12570

Svenning, J.-C., & Skov, F. (2007). Could the tree diversity pattern in Europe be generated by postglacial dispersal limitation? *Ecology Letters*, *10*, 453–460. https://doi.org/10.1111/j.1461-0248.2007.01038.x

Tae, K. (1997). *Processes Controlling Range Expansion of Sitka Spruce on Kodiak Island*. University of Alaska Fairbanks.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*, 585–595.

Terhorst, J., Kamm, J. A., & Song, Y. S. (2017). Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics*, *49*, 303–309. https://doi.org/10.1038/ng.3748

Thioulouse, J., Chessel, D., Dole´dec, S., & Olivier, J.-M. (1997). ADE-4: a multivariate analysis and graphical display software. *Statistics and Computing*, *7*, 75–83. https://doi.org/10.1023/A:1018513530268

Troupin, D., Nathan, R., & Vendramin, G. G. (2006). Analysis of spatial genetic structure in an expanding *Pinus halepensis* population reveals development of fine-scale genetic clustering over time. *Molecular Ecology*, *15*, 3617–30. https://doi.org/10.1111/j.1365-294X.2006.03047.x

Vekemans, X., & Hardy, O. J. (2004). New insights from fine-scale spatial genetic structure analyses in plant populations. *Molecular Ecology*, *13*, 921–935. https://doi.org/10.1046/j.1365-294X.2004.02076.x

Vincent, R. E. (1964). The origin and affinity of the biota of the Kodiak Island group, Alaska. *Pacific Sci.*, *18*, 119–125.

Wachowiak, W., Salmela, M. J., Ennos, R. A., Iason, G., & Cavers, S. (2011). High genetic diversity at the extreme range edge: nucleotide variation at nuclear loci in Scots pine (Pinus sylvestris L.) in Scotland. *Heredity*, *106*, 775–787. https://doi.org/10.1038/hdy.2010.118

Wang, T., Hamann, A., Spittlehouse, D. L., & Murdock, T. Q. (2012). ClimateWNA—High-resolution spatial climate data for Western North America. *Journal of Applied Meteorology and Climatology*, *51*, 16–29. https://doi.org/10.1175/JAMC-D-11-043.1

Warren, R. L., Keeling, C. I., Yuen, M. M. S., Raymond, A., Taylor, G. A., Vandervalk, B. P., … Bohlmann, J. (2015). Improved white spruce (*Picea glauca*) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism. *The Plant Journal: For Cell and Molecular Biology*, *83*, 189–212. https://doi.org/10.1111/tpj.12886

Waters, J. M., Fraser, C. I., & Hewitt, G. M. (2013). Founder takes all: density-dependent processes structure biodiversity. *Trends in Ecology & Evolution*, *28*, 78–85. https://doi.org/10.1016/j.tree.2012.08.024

Wegmann, D., Leuenberger, C., Neuenschwander, S., & Excoffier, L. (2010). *ABCtoolbox*: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics*, *11*, 116. https://doi.org/10.1186/1471-2105-11-116

Will, H., & Tackenberg, O. (2008). A mechanistic simulation model of seed dispersal by animals: Simulation model of seed dispersal by animals. *Journal of Ecology*, *96*, 1011–1022. https://doi.org/10.1111/j.1365-2745.2007.01341.x

Wiuf, C., & Hein, J. (1999). Recombination as a point process along sequences. *Theoretical Population Biology*, *55*, 248–259.

Yeaman, S., & Jarvis, A. (2006). Regional heterogeneity and gene flow maintain variance in a quantitative trait within populations of lodgepole pine. *Proceedings. Biological Sciences / The Royal Society*, *273*, 1587–93. https://doi.org/10.1098/rspb.2006.3498

Zang, C., & Biondi, F. (2013). Dendroclimatic calibration in R: The bootRes package for response and correlation function analysis. *Dendrochronologia*, *31*, 68–74. https://doi.org/10.1016/j.dendro.2012.08.001

Zinck, J. W. R., & Rajora, O. P. (2016). Post-glacial phylogeography and evolution of a wide-ranging highly-exploited keystone forest tree, eastern white pine (*Pinus strobus*) in North America: single refugium, multiple routes. *BMC Evolutionary Biology*, *16*, 56–56. https://doi.org/10.1186/s12862-016-0624-1

Appendices

Appendix A Methods and challenges applying GBS to *P. sitchensis*, a nonmodel organism with a large genome

A.1 Introduction

There is no widely accepted definition of a genomic dataset, nor is there a list of diagnostic features that differentiate it from a classic genetic dataset. These features would probably describe genome-wide polymorphisms, such as a dense sampling of markers across the genome with no ascertainment bias, as well as, maybe, genome-wide information about their linkage such as precise chromosomal position or the availability of long sequences. The number of available markers to represent genome-wide variation is in all cases one of the key distinctive features of genomic datasets for population genetics research. Since 2010, the median number of SNPs in datasets for phylogeographic studies published in Molecular Ecology has sharply increased from a few hundred to tens of thousands (Garrick et al., 2015). This shift reflects the period when genomic sequencing studies, mainly RADseq methods such as GBS, became widely available and applied to population genetics studies on nonmodel organisms.

The success of RADseq methods in producing high-quality, large datasets, depends on each of the main steps of the procedure: the genome reduction method, the sequencing technique and sequencing effort, the availability of a closely related genome for alignment, and the availability of external genomic resources for quality control. The application of GBS to *P. sitchensis* for chapters 2 and 3 presented challenges. One of them is intrinsic: a large, highly repetitive genome of about 20 Gbp makes it difficult to obtain a representative sample of the genome-wide variation in populations of the species. In addition, the only reference genomes currently available to map sequenced reads are highly fragmented draft genomes (~ 3.35M contigs) of interior spruce (*P. glauca x engelmanii*) or white spruce (*P. glauca*). Other genetic resources from the species were developed prior to the work presented in this thesis (Holliday et al., 2010), and genomic resources have been under recent development (Coombe et al., 2016).

Here, I report the development of the genomic dataset used in analyses in chapters 2 and 3. After describing in details the followed procedure, I present the output with intermediate results from each main step of the process. I finally discuss the steps of the protocol that led to an unexpected loss of information, or that were not optimal and could be improved to better suit the genetic characteristics of *P. sitchensis* in relation to current technology.

A.2 Methods

DNA was extracted from needle samples and bark samples using a protocol modified from Macherey-Nagel Nucleospin Plant II extraction kit and custom CTAB protocols, respectively. Individual samples were sent to Laval University's Institute for Integrative Systems Biology for library preparation. Each individual sample received a molecular barcode. Individuals were then pooled into 6 batches of 96 and 2 batches of 48. A double-digest method involving the restriction enzymes sbf1 (marker-anchoring enzyme) and msp1 (to reduce fragment size) was applied. The resulting libraries were sent to Génome Québec Innovation Centre (McGill University) for sequencing. Each library was duplicated and sequenced on two lanes using the HiSeq 2000 system and producing 100-bp single reads. We demultiplexed the sequenced libraries and filtered for quality with the program *Stacks*. We applied a threshold of 10 for the raw PHRED quality score for a read to be kept. The *fastx* command *fastq_quality_trimmer* was then applied to the sequences to remove sequences shorter than 20 nucleotides. Then, the *fastx* command *fastq_quality_filter* was applied to the selected fastq sequences to remove sequences with less than 90% of nucleotides with a PHRED quality score higher than 10. After sequence filtering, we aligned the filtered reads to the reference genome WS77111_V1 (Warren et al., 2015) using the *bwa mem* alignment algorithm. The resulting bam files were sorted and indexed using the program *samtools* (Li et al., 2009). I then created a reduced genome only containing mapped contigs. Contigs were grouped into 1000 master-scaffolds for further data processing. I re-aligned the fastq files to this reduced genome, and sorted and indexed resulting bam files using the same tools as previously stated. Bam files from the same individuals sequenced on different lanes were merged using the *MergeSamFiles* command from *Picard* (http://broadinstitute.github.io/picard/). Read groups were added to individual bam files using the

AddOrReplaceReadGroups command of the *Picard* program. The resulting alignment files were then input into a variant calling pipeline implemented using the program *GATK* (McKenna et al., 2010). The first step uses the *HaplotypeCaller* function to output individual gvcf files with a minimum base quality score of 10 and a maximum of one alternate allele. Individual gvcf files were then combined into 6 batches using *CombineGVCF*, a necessary step when processing several hundred individuals. *GenotypeGVCF* was then run on the entire set of cohorts, with a minimum PHRED confidence threshold of 20 for calling a variant site. Finally, I applied hard filters on the resulting callset following *GATK* recommendations as of June 2017, using the *SelectVariants* and *VariantFiltration* functions on all SNPs.

A.3 Results

Sequencing

Each lane of sequencing produced between 2.10^{-8} and $2.6.10^{-8}$ reads (Figure A.1a). Demultiplexing (1 lane=48 individuals) and filtering yielded between 1.10^{-6} and 4.10^{-6} reads per individual (Figure A.1b). To determine whether the sequencing effort was insufficient in respect to library complexity (*i.e.* mean number of restriction fragments from different cutsite positions), I selected the sequencing output from 2 libraries and cumulatively subsampled reads to mimic the accumulation of reads using the program *seqtk* (https://github.com/lh3/seqtk), and performed de novo alignments using *Stacks*. The results of this simulation show a plateau of tags (unique restriction fragments) with increasing number of reads (Figure A.2), suggesting that the sequencing effort was sufficient. I performed the same analysis on reads from 3 individual *P. sitchensis* samples from library P2_l1 and found the same result at the individual level (Figure A.3). However, an important result is the difference in number of tags typically identified in the P2_l1 library (~15M) versus its individuals (10k to 50k), a 30 to 150-fold difference. This suggests that individuals did, to a large extent, not contain the same tags.

Alignment

Albeit lower than for *P. glauca*, mapping of *P. sitchensis* reads to the *P. glauca* reference genome gave good results: most individuals had a mapping success above 60% (Figure A.4). This reflects findings from Suren et al. (2016), who applied sequence capture developed for *P. glauca* to other *Picea* and *Pinus* species and found a similar capture success of congeneric species to this of the focal species. 1,979,875 contigs from the *P. glauca* reference genome were mapped to. This represents 50% of the contigs, and a much higher number than I had expected. Indeed, both simulations using the R package *simRAD* and simple *grep* command tests performed on the reference genome estimated the number of cutsites for the *sbf1* enzyme to 110k. The number of mapped contigs should therefore have been lower than 110k.

Genotype calling

After filtering, a total of 630,777 polymorphic SNPs were identified. However, a vast majority of them were genotyped in too few individuals to be of any use in population genetics analyses (Figure A.5). Indeed, 98% of SNPs had a genotype count of less than 20 (out of 669 individuals).

A total of 6344 SNPs polymorphic over 639 *P. sitchensis* and 30 *P. glauca* individuals and with less than 95% missing data were the basis of all population genetics analyses of chapters 2 and 3.

A.4 Discussion

Depending on the subset of individuals used for analysis and the chosen threshold for missing data, the number of SNPs typically used in analyses in chapters 2 and 3 was in the order of hundreds, a somewhat disappointing output for a next-generation sequencing procedure. Sequencing and alignment to the *P. glauca* genome were relatively successful and yielded expected results. Additional analyses showed that the number of de novo alignment tags is one to two orders of magnitude higher in libraries reads than in reads grouped by individual samples. Also, the number of reference contigs mapped to was much higher than expected. These two results suggest that unwanted issues occurred during library preparation. The double-digest procedure to cut DNA at *sbf1* cutsites and reduce fragment lengths with *msp1* seem to have

partly failed, leading to the amplification and sequencing of many unique sheared fragments probably not linked to any *sbf1* cutsite. This would have decreased the amplification rate and sequencing rate of fragments linked to actual *sbf1* cutsites and originating from all individual DNA samples. The fact that most reads amplified and sequenced were not specific to any cutsite can explain the output of hundreds of thousands of SNPs only present in one or a few individuals. Whether the partial failure of library preparation is due to the specific choice of enzyme combination or other technical issues such as deteriorated DNA or amplification problems is hard to determine. In all cases, the choice of restriction enzyme and shearing method are critical in organisms with large repetitive genomes like *P. sitchensis*. Especially, choosing protocols favoring a reduction of library complexity such as restriction enzymes with rare cutsites should increase the chance of obtaining good data. This is the case of *sbf1*. Being a 10-base cutter, this enzyme cuts at a frequency 10 to 15 times lower than the popular Pst1 or Nsi1 (both 6-base cutters). Another important consideration is the avoidance highly repetitive sequences. This can be achieved by selecting a methylation-sensitive enzyme, which sbf1 is not. In a useful attempt to improve techniques applying RADseq methods to conifers, Pan et al. (2015) tested the effect of several combinations of restriction enzymes, multiplexing levels and variant calling pipelines on the success of SNP dataset development for pine species. They obtained good results (7,000-14,000 SNPs) with pst1, and related the success of this method to the low complexity of obtained libraries thanks to the methylation sensitivity of the enzyme, and to a fragment size selection by polyacrylamide gel electrophoresis (PAGE). Although these and my results do not provide a gold standard for GBS techniques in conifer species, they provide complementary knowledge necessary to improve genomic techniques for marker development in a group of ecologically important species.

A.5 Figures

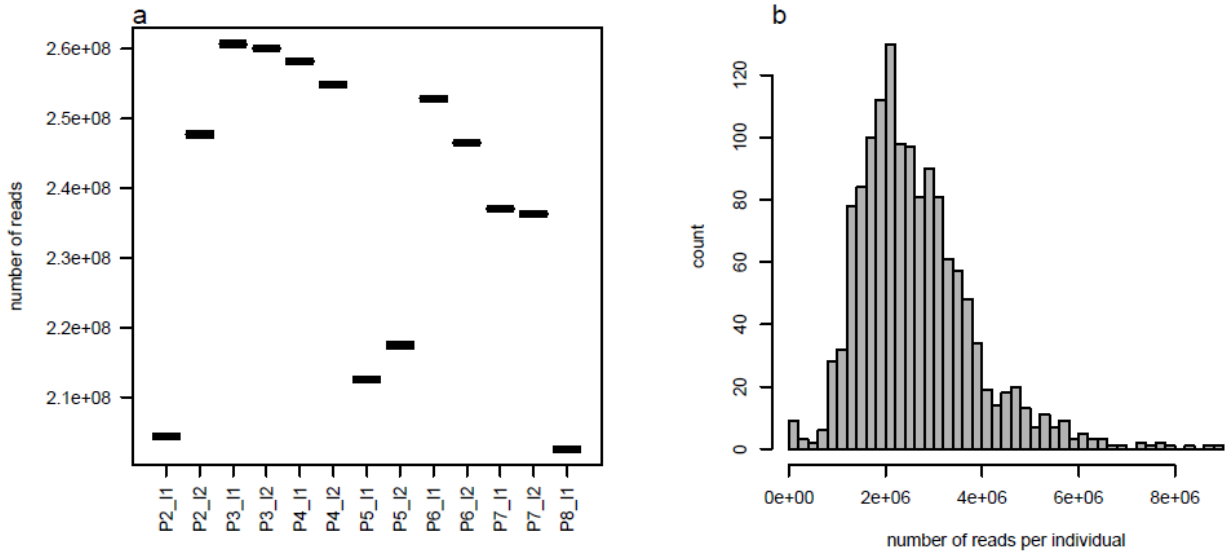


Figure A.1 Number of GBS reads. a. Number of reads in each sequencing lane; b. Distribution of the number of GBS reads per lane per sampled tree.

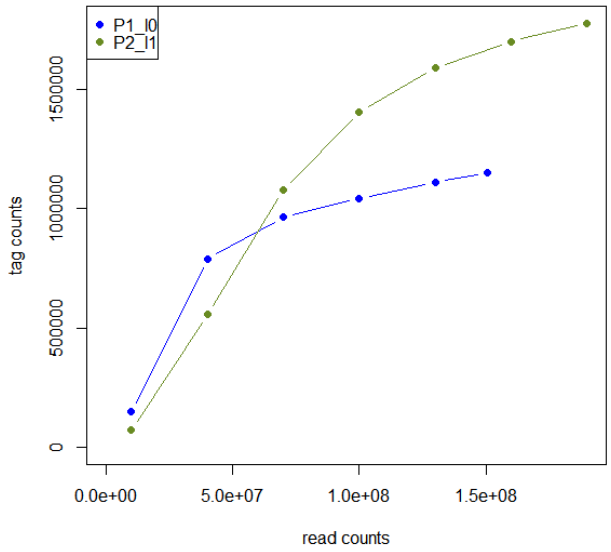


Figure A.2 Simulated accumulation of tags for 2 libraries from de novo alignments with increasing number of reads.

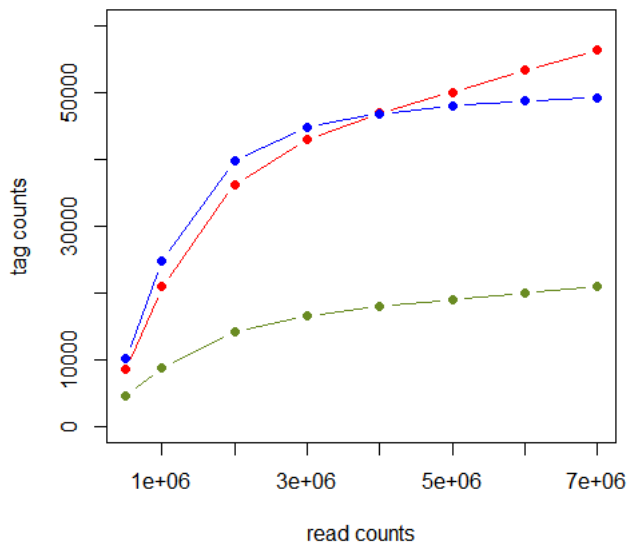


Figure A.3 Simulated accumulation of tags for 3 individuals from de novo alignments with increasing number of reads.

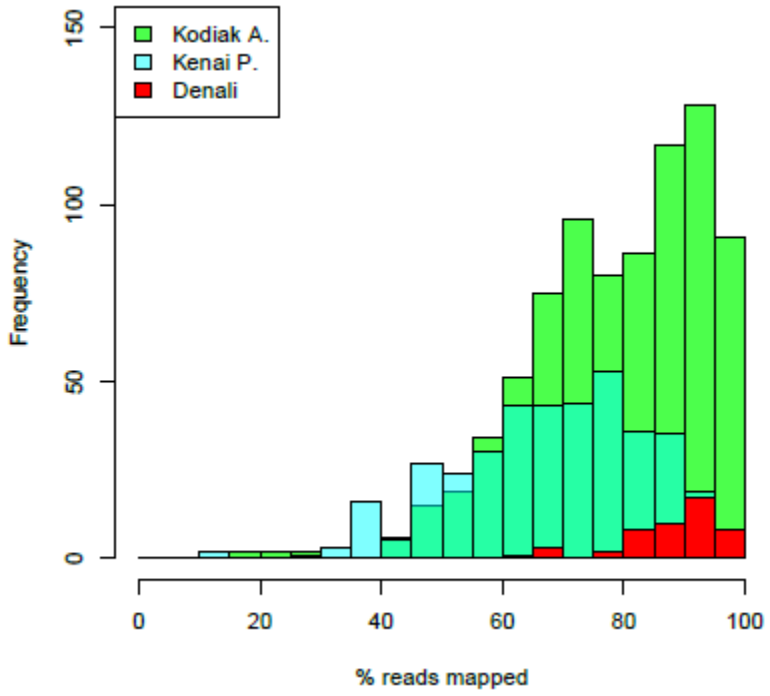


Figure A.4 Distribution of alignment mapping success for 3 regions.

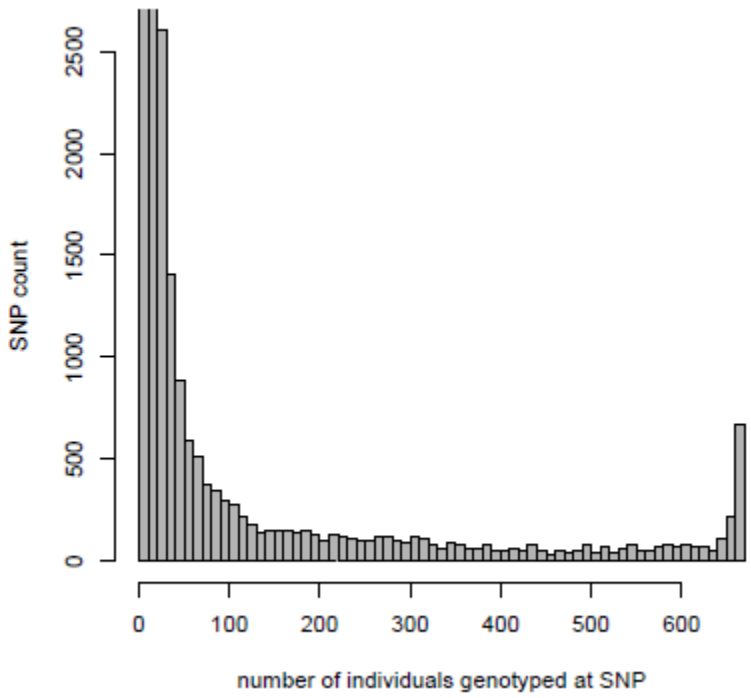


Figure A.5 distribution of genotype counts after genotype calling and filtering. The two first classes reach 609,976 and 7757 respectively, and have been cut on the y axis for visibility purposes.

Appendix B Supplemental materials and figures for Chapter 2

B.1 Methods

Extracting age and ring width information from tree cores

Tree cores were kept in the freezer until processing. After drying and mounting, cores were sanded using a belt sander and grain from 120 up to 600. Each core was scanned with a resolution of 1200 dpi. Tree rings were counted and measured using the program *Coorecorder (Cybis)*. To ensure there were no dating errors or false rings and to account for missing rings, ring-width series were crossdated visually using the program *Cdendro* (*Cybis*), and statistically using the program *Cofecha* (Holmes, 2000). Briefly, I first built site-specific collections with *Cdendro* by choosing a few cores that crossdated well as an initial collection, and crossdating new cores to the updated chronology from the growing collection; I then compared chronologies among site-specific collections to ensure consistent successful crossdating across the whole sample. I then submitted individual tree ring series to *Cofecha* and iteratively corrected tree ring measurements on *Cdendro* and *Cofecha* until no more problems were detected in the *Cofecha* output. To estimate germination dates from cores, two corrections were applied. For cores that did not intercept the pith, a geometric correction was applied to estimate the number of missing rings (Duncan 1989). To take into account the number of years for trees to grow to coring height, linear models were fitted to predict age from the height of sampled juvenile trees. I fitted linear models to explain the number of nodes (e.g. age) with two fixed effects: height and growth conditions (“open” or “closed”). I predicted the age at coring height for Kodiak and Afognak Island canopy trees that had wide rings closest to the pith, indicating initial growth with little competition in open conditions, using coefficients of the Kodiak-Afognak linear model for “open” conditions. Coefficients for the “closed” conditions were used to correct for coring height in sub-canopy trees with narrow initial rings. I predicted the age at coring height for all trees in the Seward region using coefficients of the Kodiak-Afognak linear model for “closed” conditions, given their narrow initial rings.

B.2 Results

Population structure: PCA and *Structure* results

I performed a PCA on all samples (Figure B.2). PC1 carries 14.72% of the variance and mostly represents a continuous isolation-by-distance pattern among the sampled *P. sitchensis* populations. PC2 and subsequent axes carry very little of the variance and do not show any clear geographic pattern. *Structure* results with K=2, K=3, and K=4 are presented in Figure B.2a-c, and posterior probabilities are shown in Figure B.2d. Analysis of posterior probability values for each set of runs supports the existence of two clusters, mainly partitioning the sample into a Kodiak-Afognak group and Seward group. Under this model, I observe a mixed (but Kodiak-dominated) ancestry for Shuyak Island and Port Chatham between the two *P. sitchensis* clusters. K=3 further partitions the Kodiak-Afognak group in two equal-sized groups.

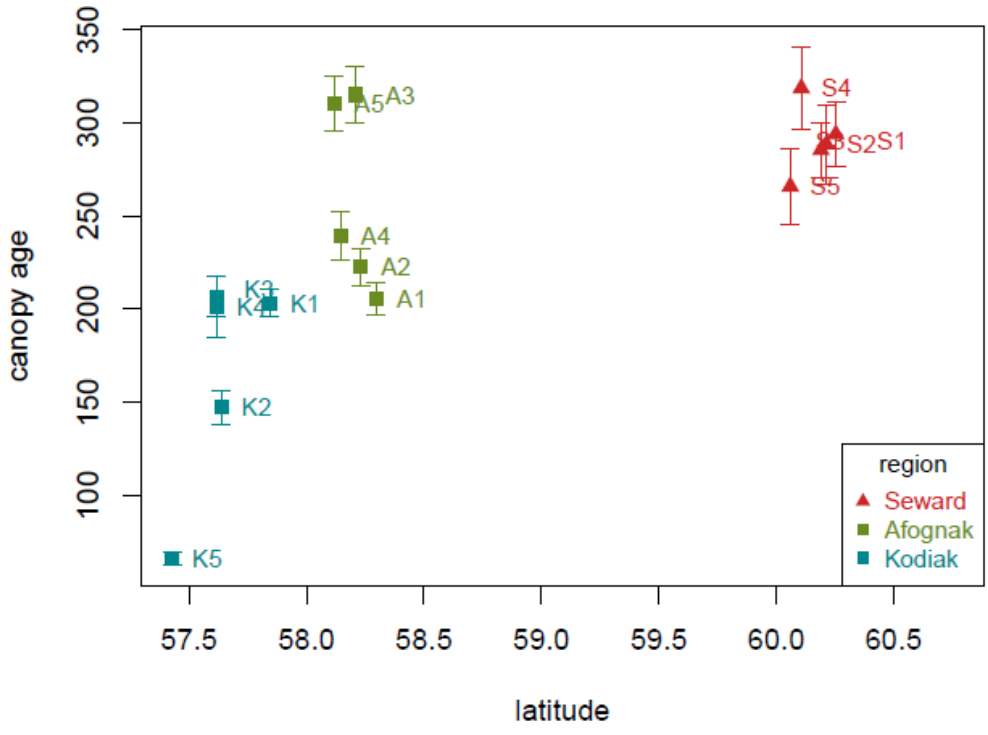


Figure B.1 Mean age of canopy trees vs. latitude of sampled sites. Error bars represent the standard error of the mean.

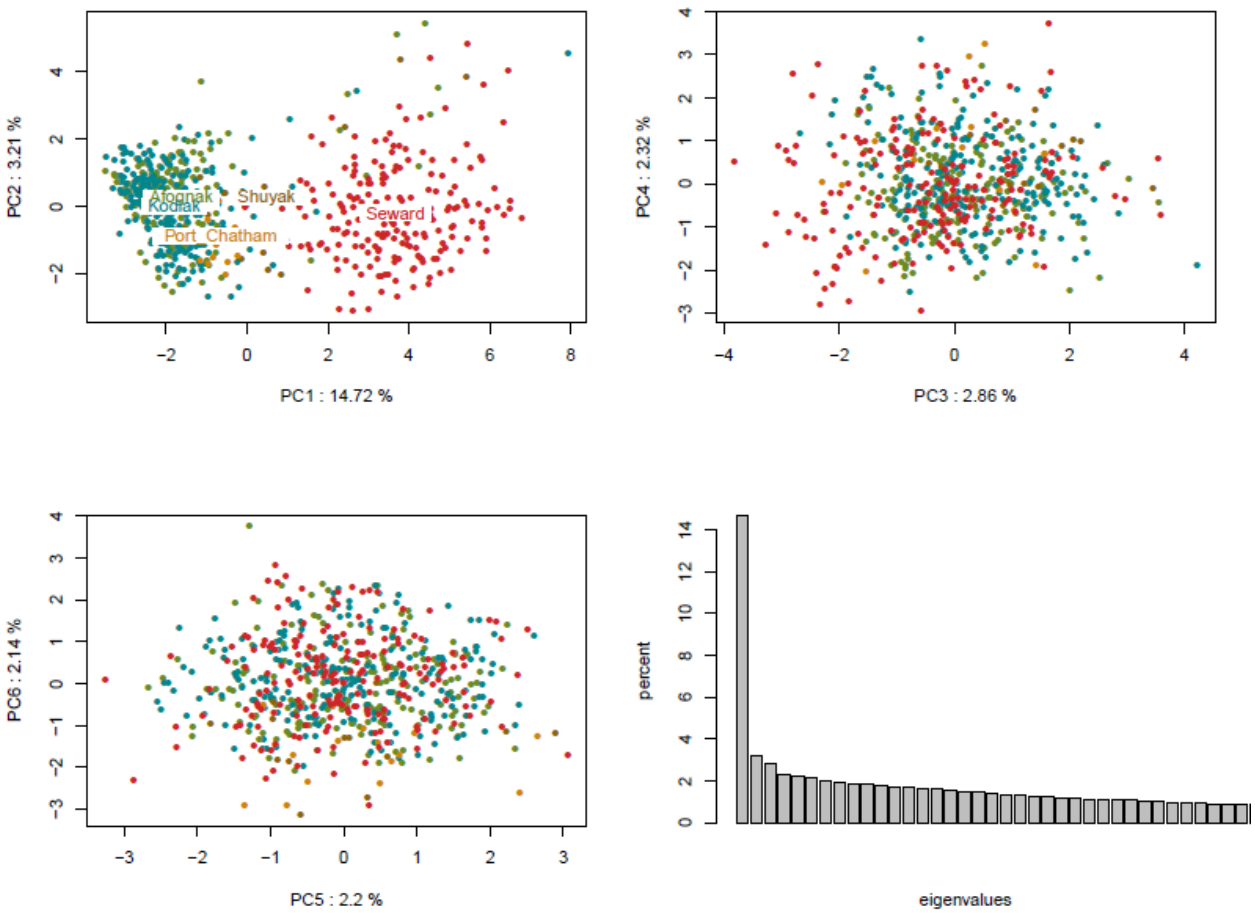
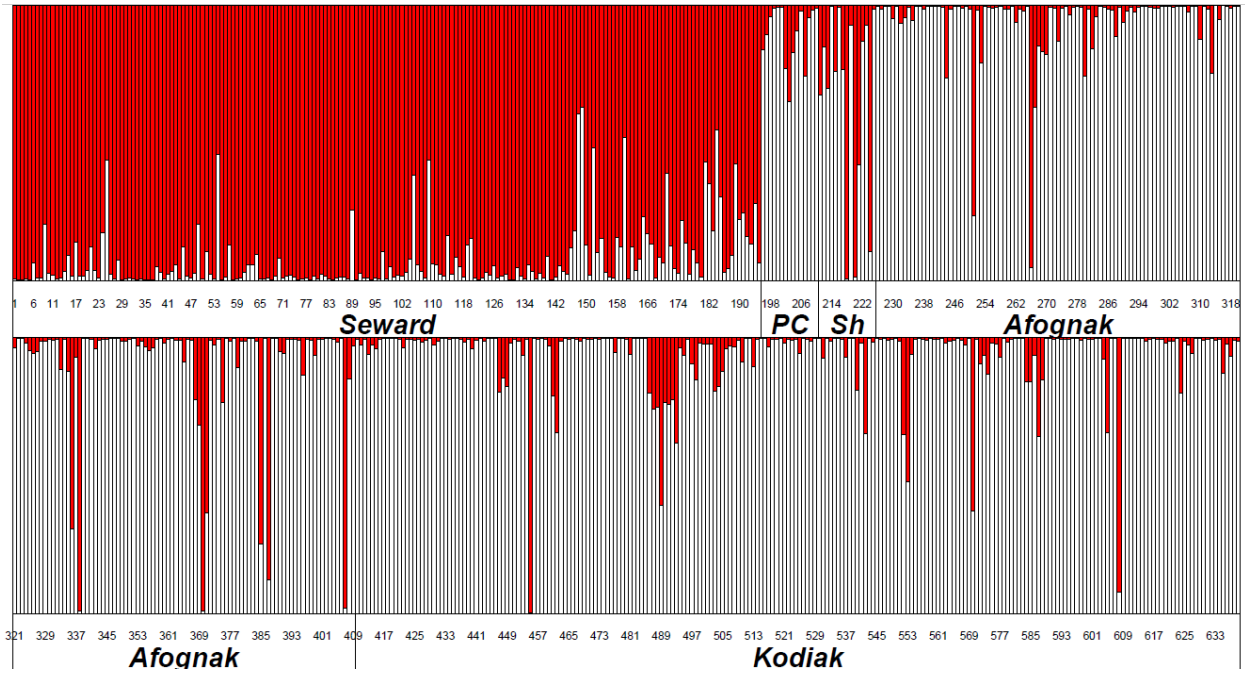
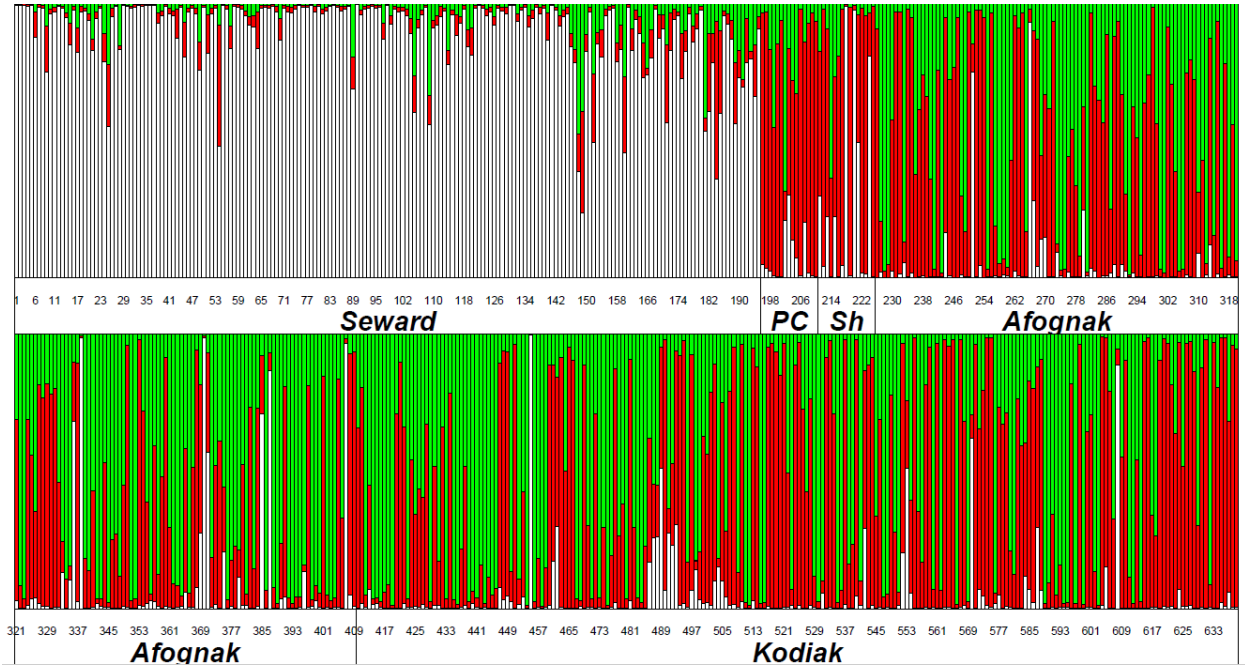


Figure B.2 Representation of PC 1 to 6 (a-c) and distribution of eigenvalues (d) for PCA

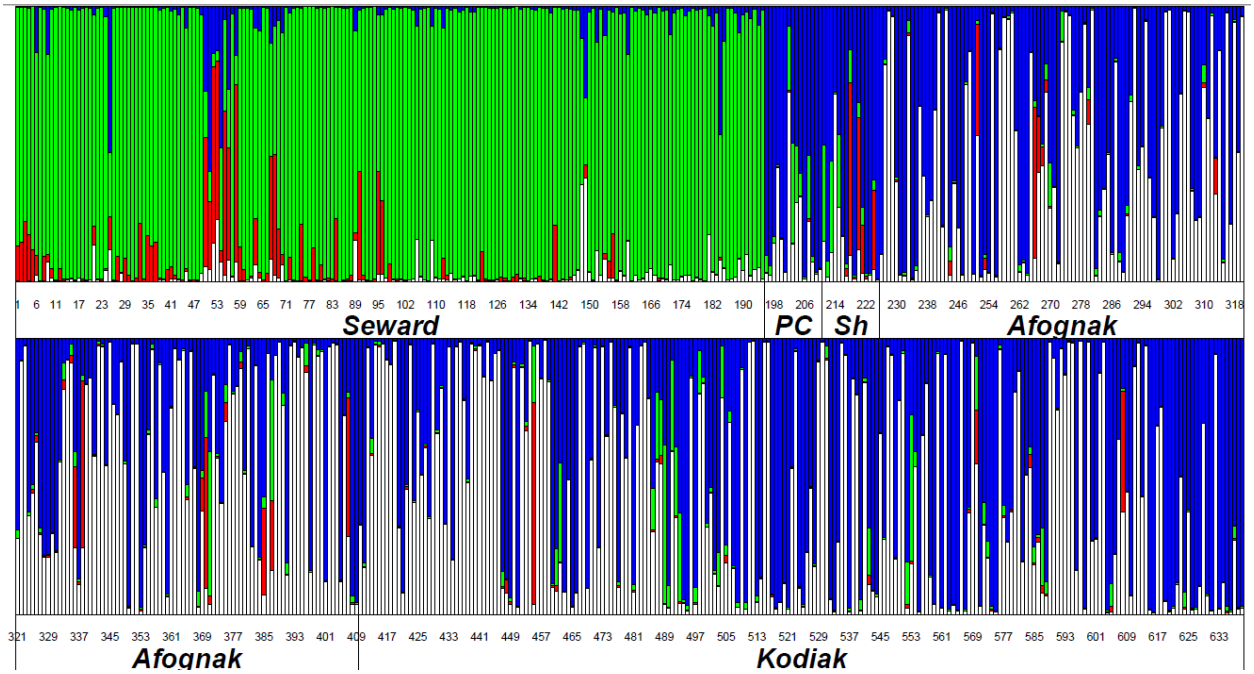
a.


Seward
PC
Sh
Afognak
Afognak
Kodiak

b.


Seward
PC
Sh
Afognak
Afognak
Kodiak

c.



d.

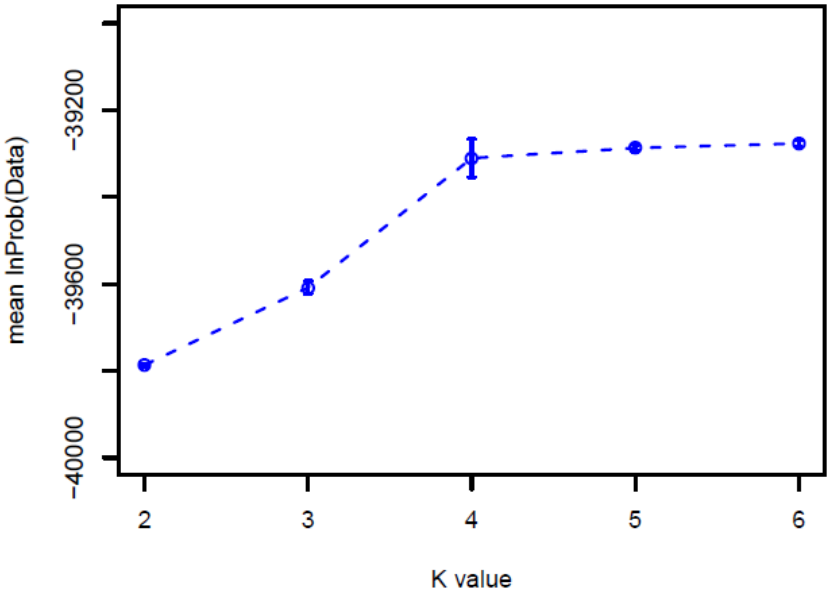


Figure B.3 *Structure* results exploring different K values. a-c: *Structure* barplots for K=2, K=3, K=4. d) Posterior probability of models with different K values and prior information. Error bars indicate standard error over three independent runs for each value of K.

Appendix C Supplemental figures for Chapter 3

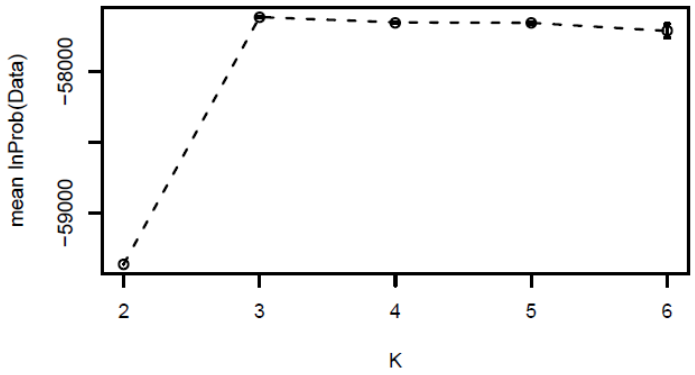


Figure C.1 Posterior probability of *Structure* models with different K values. Error bars represent standard error of the mean over three iterations.

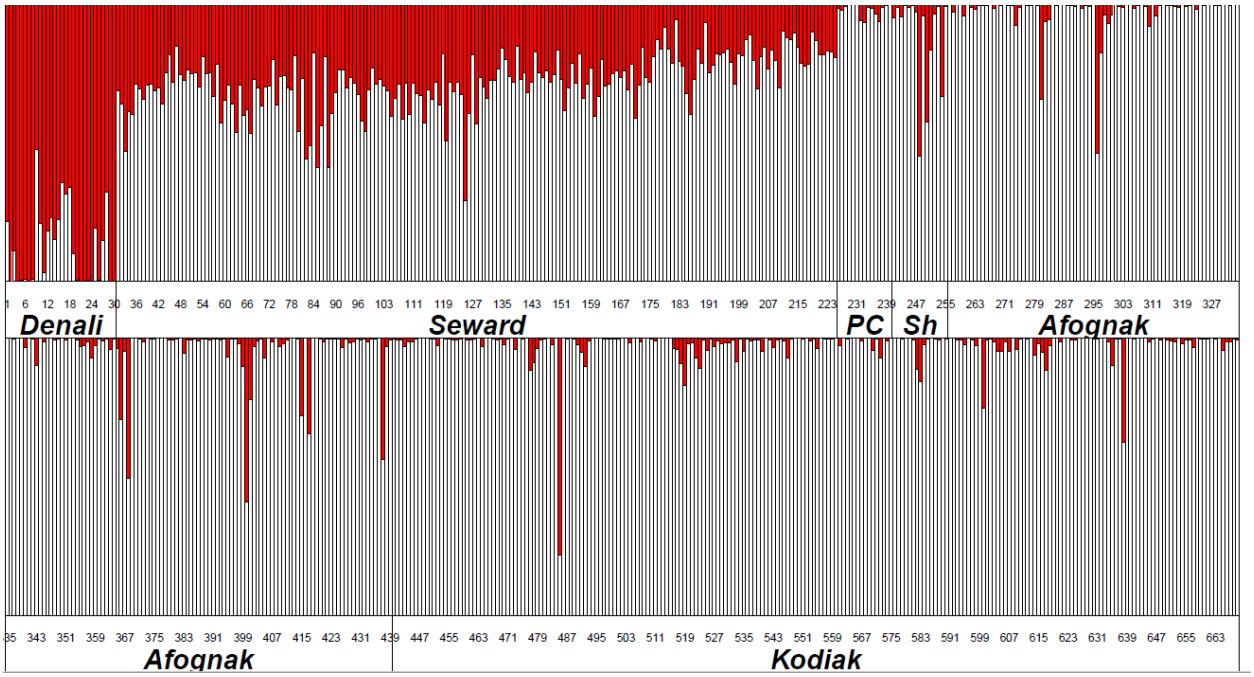


Figure C.2 *Structure* barplot for K=2.

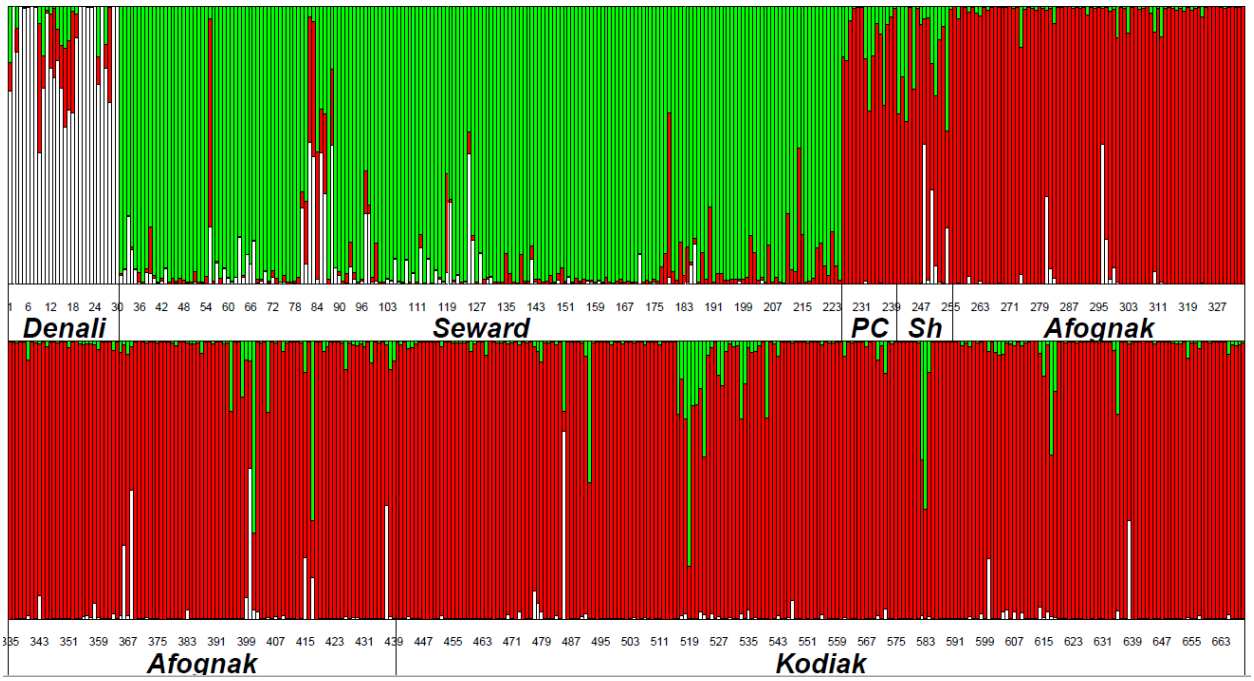


Figure C.3 ***Structure*** **barplot for K=3.**

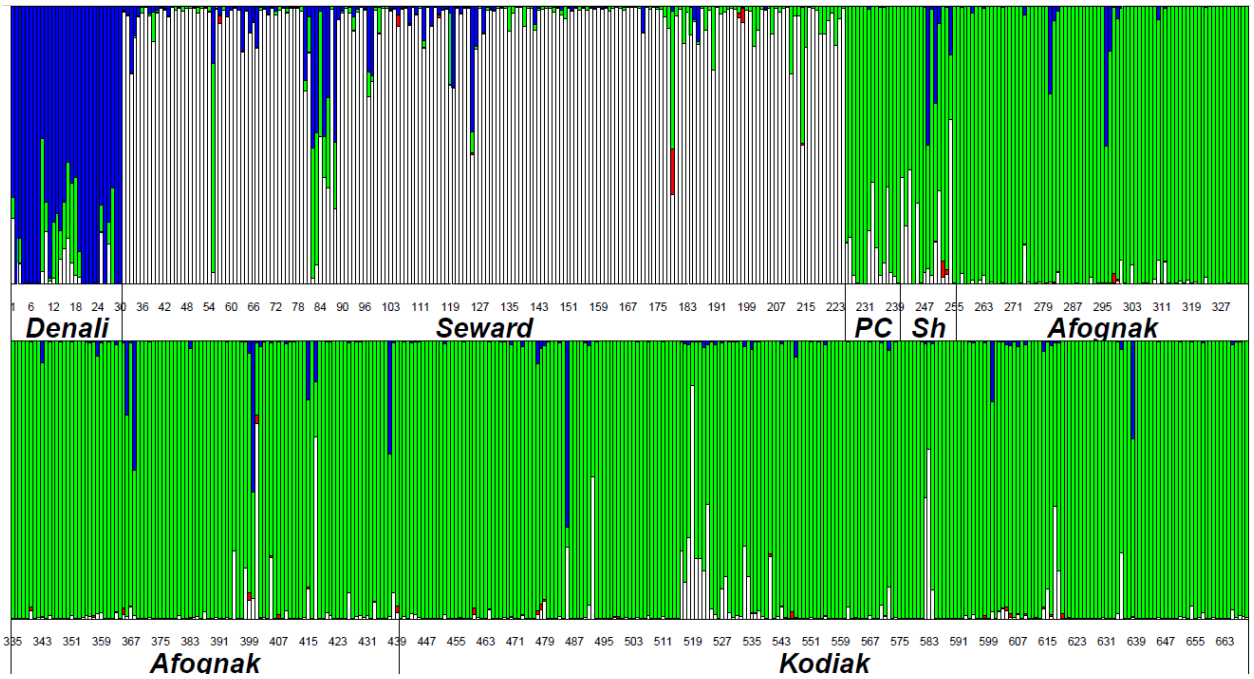


Figure C.4 ***Structure*** **barplot for K=4.**

Appendix D Supplemental materials and figures for Chapter 4

D.1 Reducing the number of summary statistics

For each set of 1M simulations, I used a two-step process to reduce the number of summary statistics. I first used a "training set" of 10,000 independent simulations to calculate pairwise correlations between summary statistics. Then, for each pair of statistics, if the absolute correlation coefficient was higher than 0.8, I discarded one of the two statistics. To decide which of the two to discard, I used the same training set of simulations to perform a linear regression of one of the two statistics onto the each model parameter, then repeated this process for the second statistic. I kept the statistic with the strongest association with most model parameters. After obtaining a set of uncorrelated statistics, I transformed the summary statistics using a partial-least-squares (PLS) regression (Boulesteix & Strimmer, 2007) to the training set of 10,000 independent simulations using R scripts provided with *ABCtoolbox* (Wegmann et al., 2010). The principal components obtained from the PLS transformation were used as new summary statistics. There is a consensus that the appropriate number of summary statistics is close to the number of model parameters. As the models have between 3 and 5 parameters we chose to keep 5-7 PLS for each ABC analysis performed, i.e., two more than the number of parameters.

D.2 Creating "imperfect" PODs

I performed coalescent simulations using *scrm* and the same method as described for standard PODs to generate imperfect PODs. Instead of directly calculating summary statistics from the *scrm* output, I subjected the latter to a process mimicking sequencing and variant calling. For each POD, the *scrm* output was converted to a fasta file using a modified version of a python script called *ms2fasta* (https://github.com/svohr/ms_utils). I then used a custom R script to perform in-silico sequencing and SNP calling. The process is as follows:

(i) sequences are fragmented in 100-bp reads;

(ii) reads are amplified following a negative exponential distribution with parameter lambda equal to 2/(chosen individual depth);

(iii) sequencing errors are randomly introduced into amplified reads at the chosen rate;

(iv) genotype likelihoods are calculated on all potentially polymorphic sites following the *GATK* method which uses the following equation:

$$P(D|G = \{A_1, A_2\}) \prod_{i=1}^{i=M} P(b_i|G = \{A_1, A_2\}) = \prod_{i=1}^{i=M} \left(\frac{1}{2}P(b_i|A_1) + \frac{1}{2}P(b_i|A_2)\right)$$

with

$$P(b|A) = \begin{cases} \frac{e}{3} \ \ if\ b \neq A \\ 1 - e \ \ if\ b = A \end{cases}$$

where M is the sequencing depth, b_i is the observed base in read i, e is the per-nucleotide error rate. This method implies that the sequencing error rate is accurately known. In most variant calling platforms, it is estimated from the raw genomic data.

(v) Genotypes are called by selecting individual genotypes with the highest likelihood; (vi) the fasta file with original reads is modified to incorporate inferred genotypes at potentially polymorphic sites; (vii) fragments are re-assembled into sequences of original marker length (here 200bp). This step mimics the alignment process and reassembles sequences assuming perfect alignment (no mapping error);

(viii) the fasta file of inferred markers is converted to the *ms* input format;

(ix) Summary statistics are calculated from the inferred markers following the procedure described in the main methods.

D.3 Estimating model parameters using the SFS

I used *fastsimcoal2* (Excoffier et al., 2013; Excoffier & Foll, 2011) to simulate 100 pseudo-observed datasets (PODs) of type 1 (10,000 sequences of 100bp). I repeated this for each of the four models depicted in Figure 4.1. The SFS from these 100 PODs was then input into *fastsimcoal2*, which approximates a composite likelihood from a number of simulations set by the user (here 10,000 simulations) and iteratively performs a conditional maximization algorithm (ECM) to estimate the parameter values corresponding to the maximum likelihood. I allowed 20 to 40 ECM cycles with a stopping criterion (minimum relative difference in parameters between two iterations) of 10^{-3}. For each POD, I performed 50 iterations of simulations and ECM and retained the parameter estimates with highest maximum likelihood. I used estimates from 100 PODs to calculate the relative prediction error (RPE) of each parameter for each model. For 10 PODs, I included a parametric bootstrap step to the SFS inference: for each of the 10 PODs, I simulated 100 SFS using the maximum likelihood values obtained from the estimation as true values. Then, I ran the estimation again in an identical manner for these 100 SFS. 95% confidence intervals were calculated for each of the 10 PODs from the quantiles of the parameter estimates from the 100 SFS bootstraps, using custom R scripts (R Core Team, 2016).

D.4 Supplemental tables and figures

Table D.1 Comparison of performance of model 2 (4-parameter model including N_{02} as a parameter, with exponential growth of population 2) with a corresponding model where population 2 experiences a sudden size change at $T_{EXP}/10$. The simulated datasets are 10k sequences of length 100bp. The value 1 for haplotype phase means that the LD-based statistics are included in the summarization step, 0 means that they are excluded. Prediction error based on 1000 random simulations is displayed. 95% HDI was averaged over 100 random simulations and the corresponding standard error is shown in brackets. 95% HDI and standard error values are rounded to the nearest integer.

haplotype phase	parameter	Prediction error		mean 95% HDI [se]	
		growth	N change	growth	N change
1	N_1	0.018	0.013	13668 [411]	11547 [314]
1	N_2	1.509	1.393	81886 [63]	82763 [33]
1	N_{02}	0.682	0.635	704 [21]	668 [22]
1	T_{EXP}	0.604	0.697	343 [10]	325 [11]
0	N_1	0.001	0.002	3021 [94]	5541 [303]
0	N_2	1.57	1.474	81692 [151]	82688 [39]
0	N_{02}	0.577	0.623	628 [25]	685 [21]
0	T_{EXP}	0.327	0.579	275 [10]	294 [12]

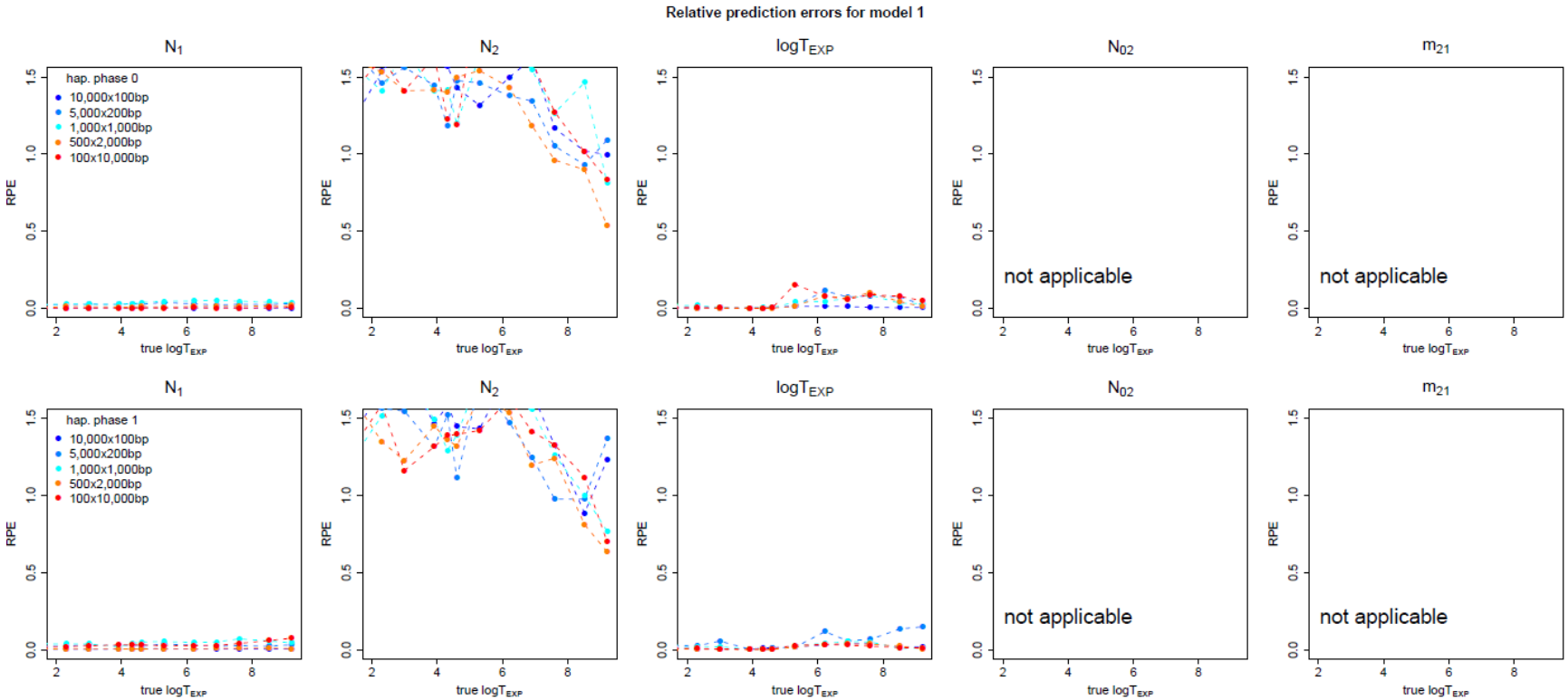


Figure D.1 Relative prediction error of model parameters for different fixed values of T_{EXP}: model 1. Note that the T_{EXP} values are represented on a log scale for better visibility of recent demographic events. Different colors represent different types of simulated datasets.

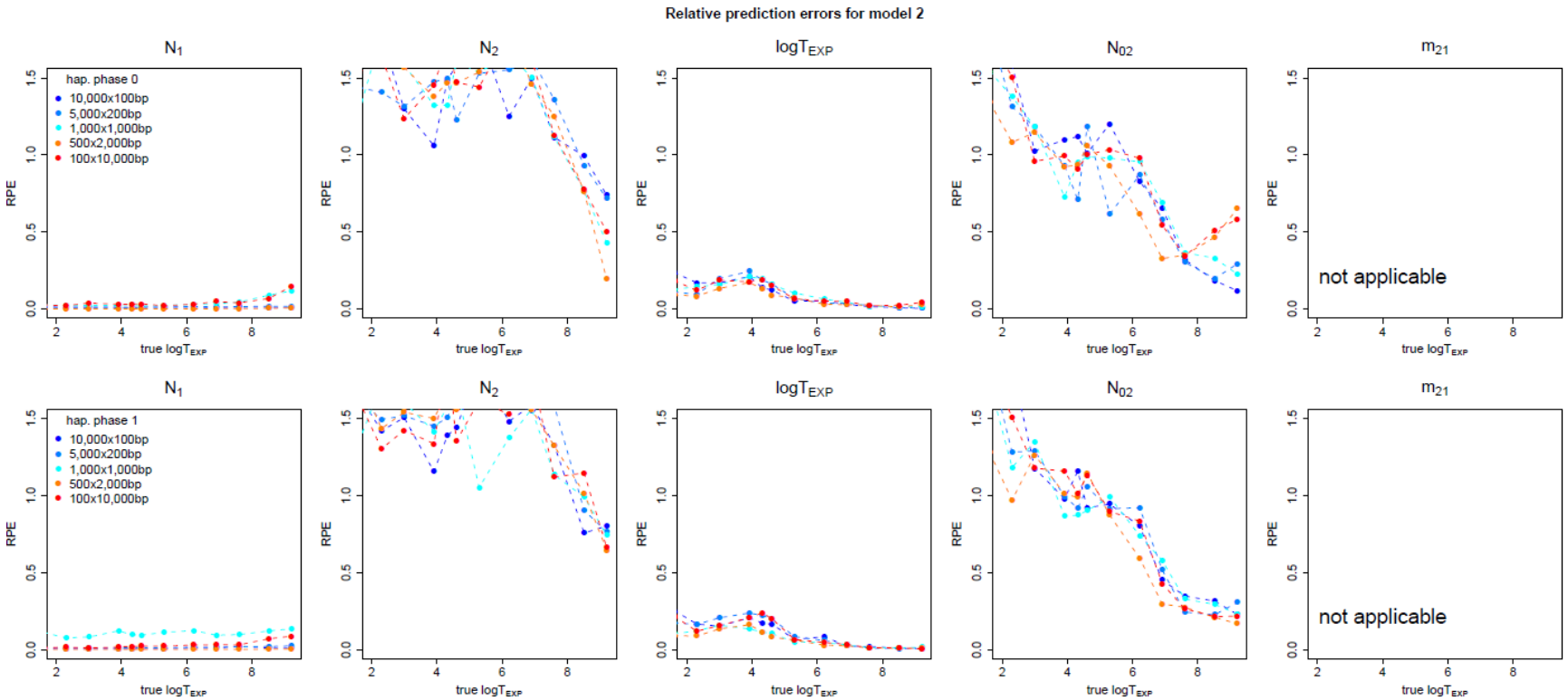


Figure D.2 Relative prediction error of model parameters for different fixed values of T_{EXP}: model 2. Note that the T_{EXP} values are represented on a log scale for better visibility of recent demographic events. Different colors represent different types of simulated datasets.

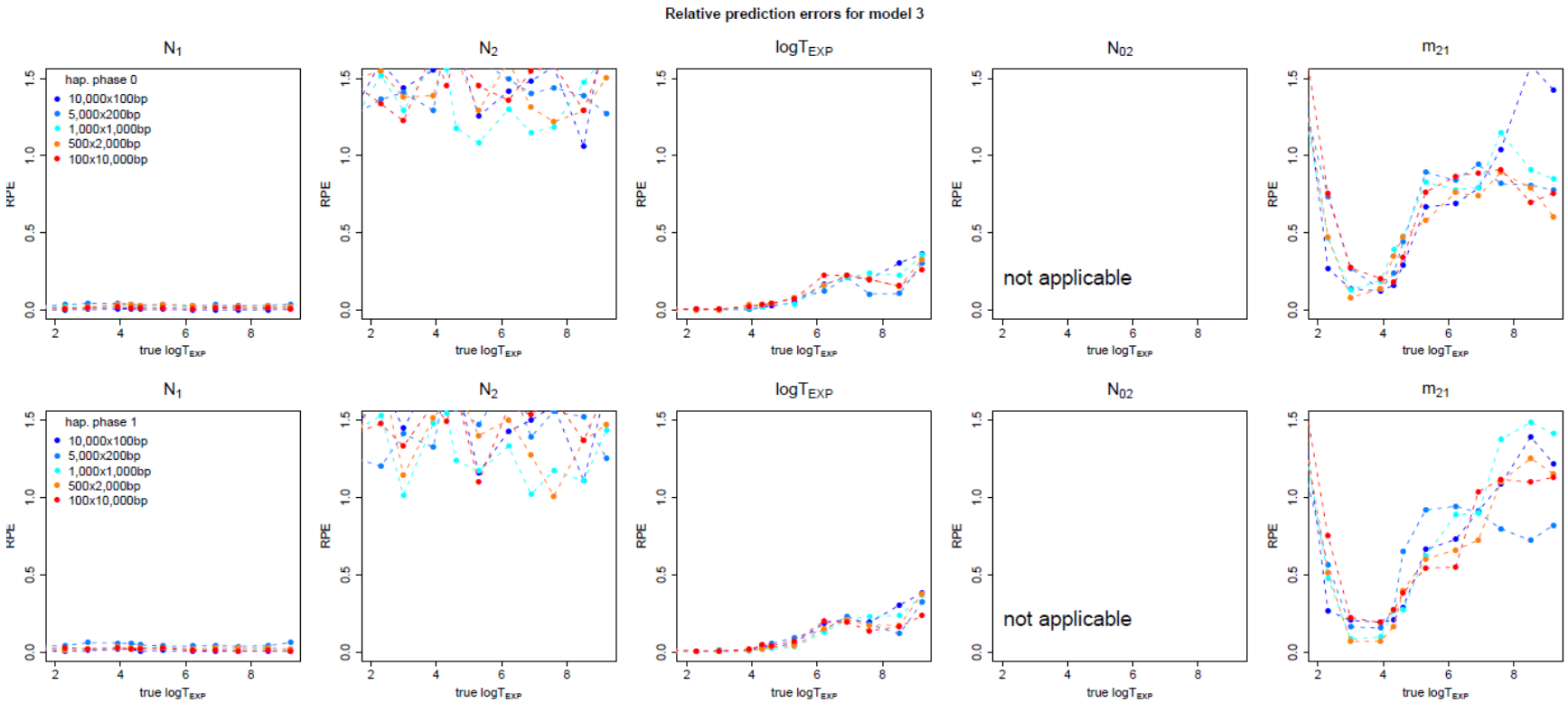


Figure D.3 Relative prediction error of model parameters for different fixed values of T_{EXP}: model 3. Note that the T_{EXP} values are represented on a log scale for better visibility of recent demographic events. Different colors represent different types of simulated datasets.

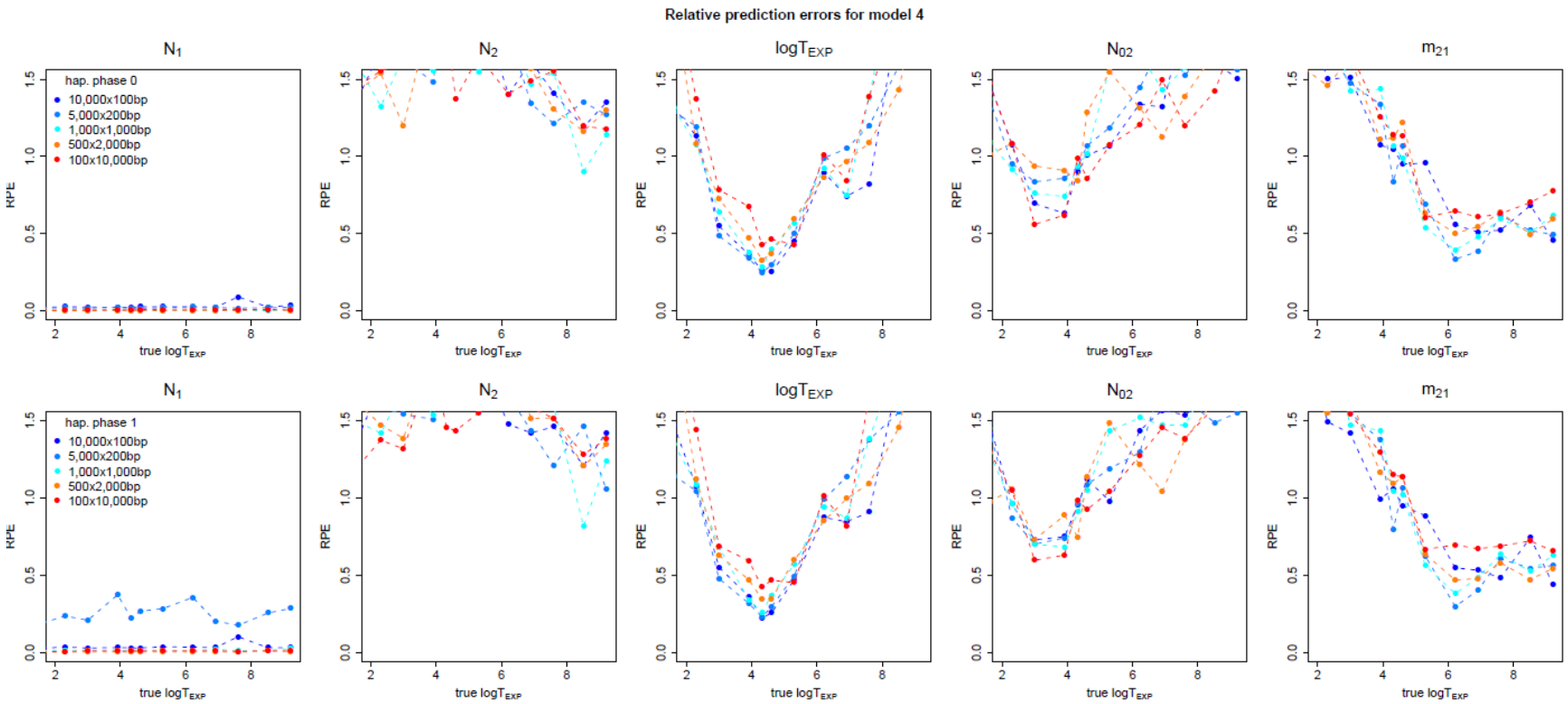


Figure D.4 Relative prediction error of model parameters for different fixed values of T_{EXP}: model 4. Note that the T_{EXP} values are represented on a log scale for better visibility of recent demographic events. Different colors represent different types of simulated datasets.

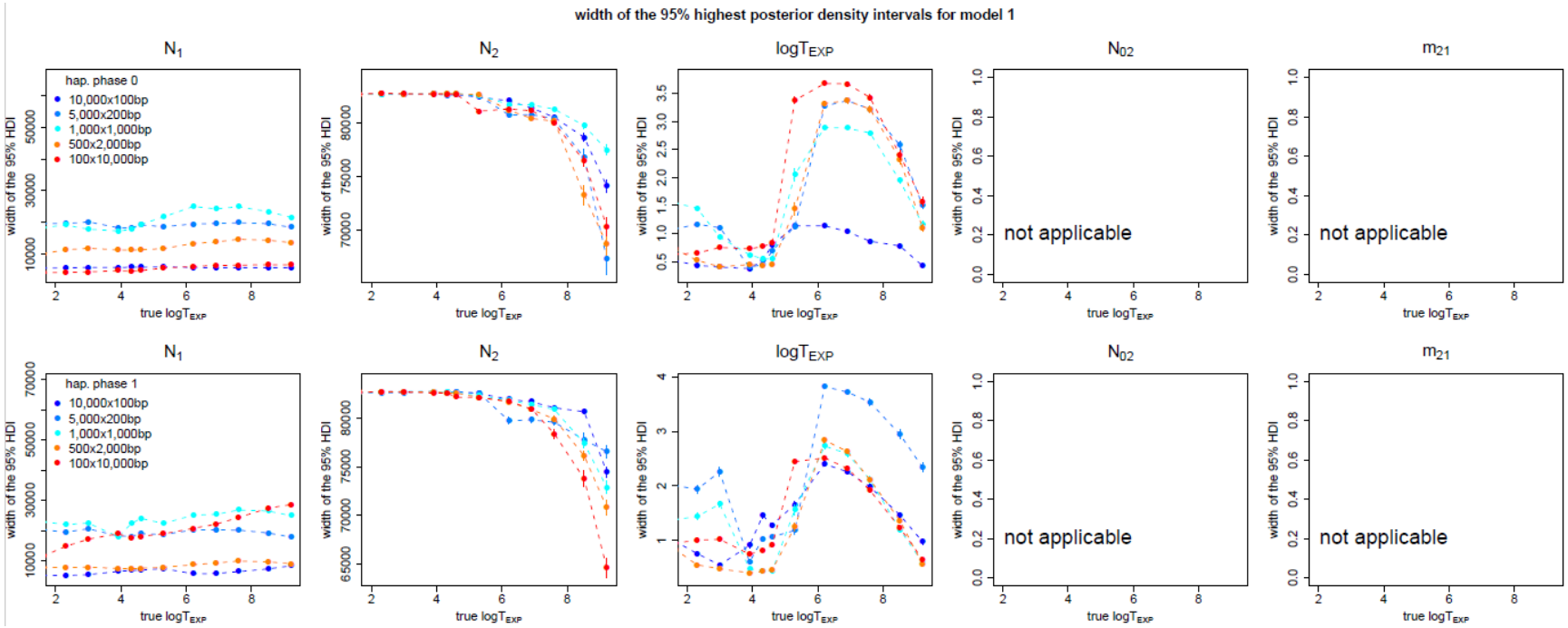


Figure D.5 Mean 95% highest posterior density intervals of model parameters for different fixed values of T_{EXP}: model 1. Note that the T_{EXP} values are represented on a log scale for better visibility of recent demographic events. Different colors represent different types of simulated datasets. Error bars represent standard errors of the estimates (N=100 PODs).

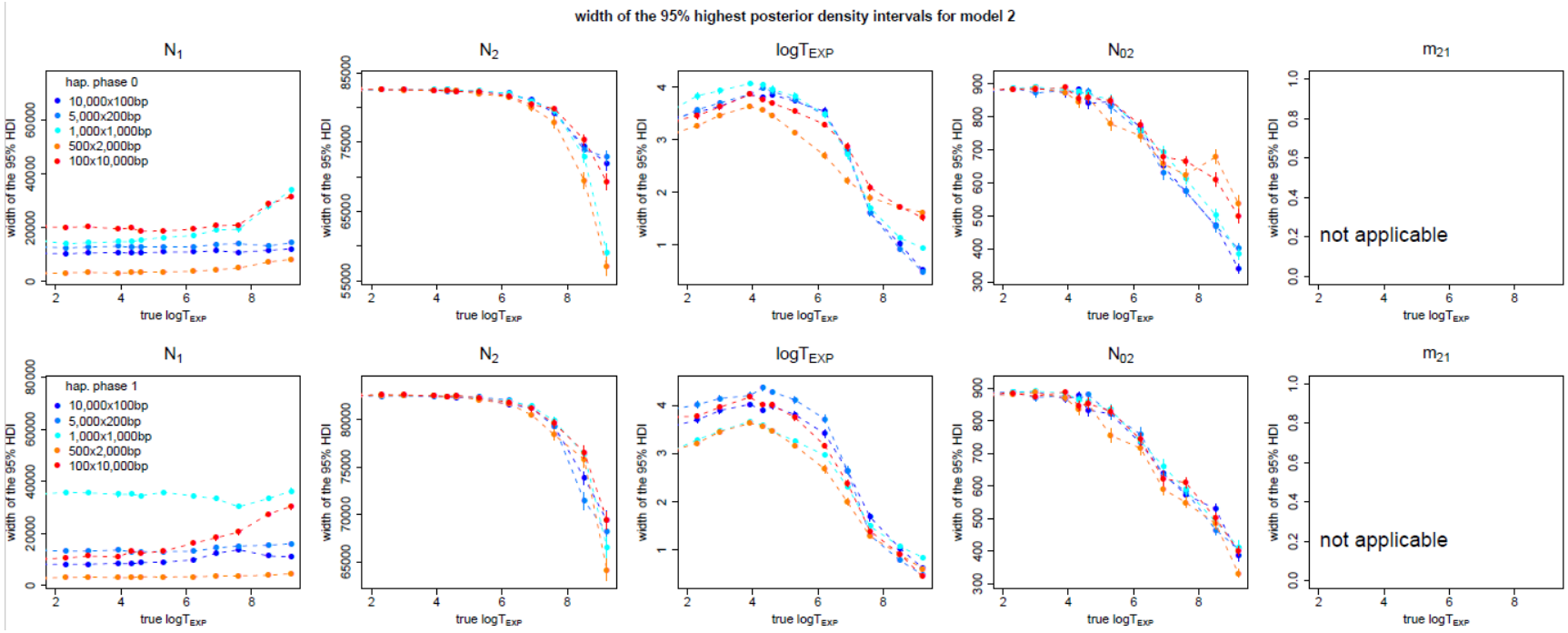


Figure D.6 Mean 95% highest posterior density intervals of model parameters for different fixed values of T_{EXP}: model 2. Note that the T_{EXP} values are represented on a log scale for better visibility of recent demographic events. Different colors represent different types of simulated datasets. Error bars represent standard errors of the estimates (N=100 PODs).

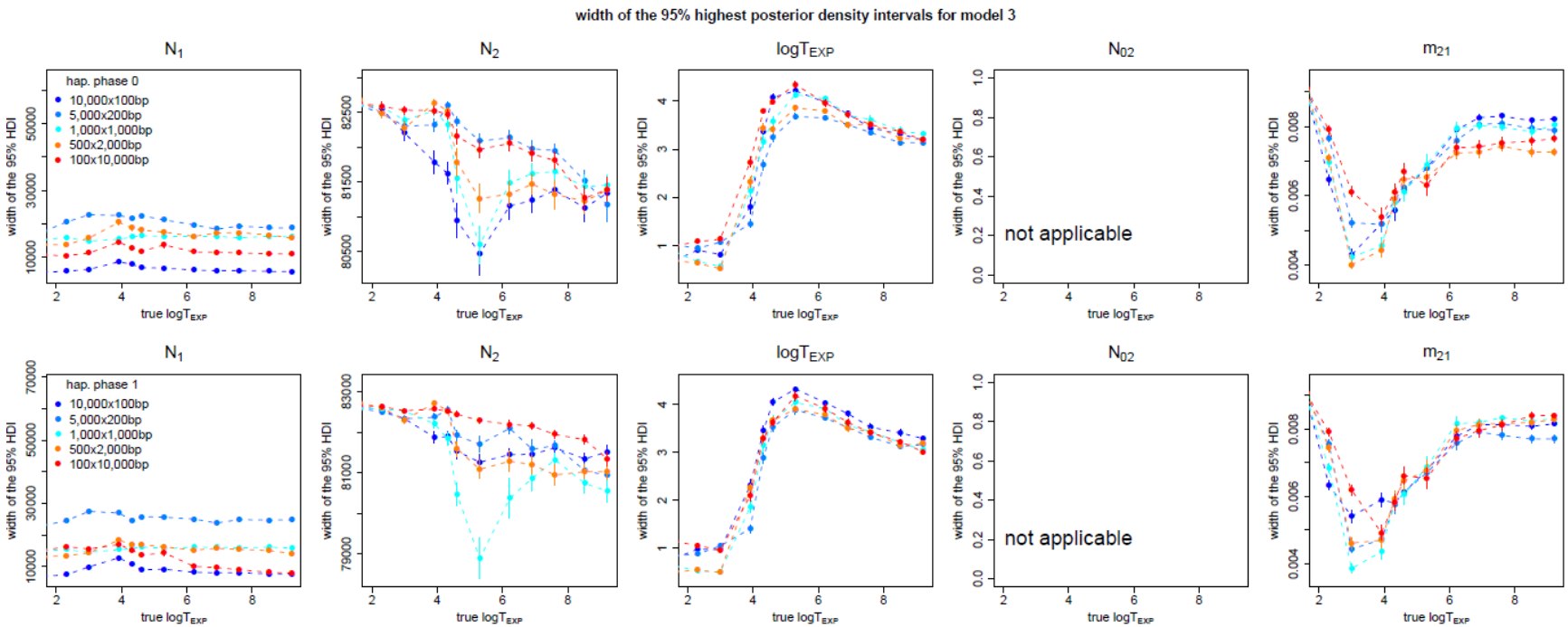


Figure D.7 Mean 95% highest posterior density intervals of model parameters for different fixed values of T_{EXP}: model 3. Note that the T_{EXP} values are represented on a log scale for better visibility of recent demographic events. Different colors represent different types of simulated datasets. Error bars represent standard errors of the estimates (N=100 PODs).

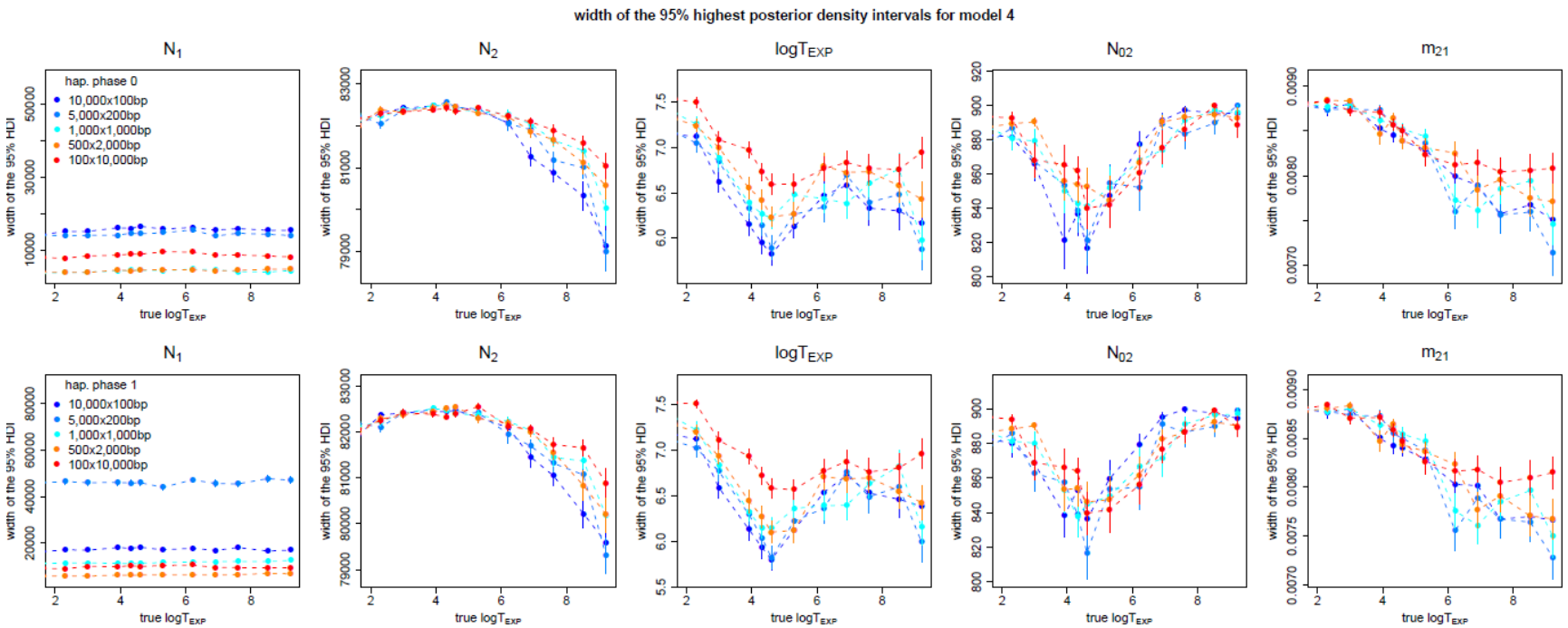


Figure D.8 Mean 95% highest posterior density intervals of model parameters for different fixed values of T_{EXP}: model 4. Note that the T_{EXP} values are represented on a log scale for better visibility of recent demographic events. Different colors represent different types of simulated datasets. Error bars represent standard errors of the estimates (N=100 PODs).

Appendix E Applying approximate Bayesian computation to *Picea sitchensis* postglacial colonization

E.1 Introduction

Evolutionary and ecological responses of tree species to new environments is a topic of increasing concern in the current context of rapid human-induced climate change. Highly precise predictions about future climate envelopes are now available (Mahony et al., 2017), but whether trees will be able to track their climatic niche spatially or adapt to new biotic and abiotic conditions remains an open question (Aitken et al., 2008). Temperate and boreal tree species have expanded their ranges northward since the last ice age ended around 18,000 years ago. Studying the rates and patterns of postglacial colonization in the northern hemisphere can help guide predictions about species' responses to current and future climates (Petit et al., 2008). The concomitant use of phylogeography, paleoecology and species distribution modelling has allowed for the inference of postglacial refugia and subsequent colonization routes for most widely distributed temperate and boreal tree species in America and Europe (Gavin et al., 2014). From this knowledge, average postglacial rates of migration from refugia were estimated using mechanistic modelling approaches (Feurdean et al., 2013). In this appendix, I attempt to estimate the pace of postglacial tree population migration using genomic data in a phylogeographic framework, as an empirical application of Chapter 4.

The use of phylogeography to address the question of tree species migration has traditionally involved single-locus approaches from organelle markers to geographically characterize distinct refugial genetic groups (Cottrell et al., 2005; Lauterjung et al., 2018), and population genetic tests of simple demographic hypotheses using nuclear or organelle markers (Heuertz et al., 2006; Lauterjung et al., 2018). As sequencing advances have increased the number of available nuclear markers for all types of organisms (Garrick et al., 2015), it has become possible to test complex historical demographic models and estimate parameters with statistical phylogenetic methods involving the coalescent (Carstens et al., 2013; Cornille et al., 2013). This has allowed for a refinement of our knowledge of historical patterns of postglacial

migrations through testing alternative demographic models (Gao et al., 2012; Gugger et al., 2010; Holliday, Yuen, et al., 2010), but more rarely addressed actual estimates of demographic parameters (Carstens et al., 2013; Holliday, Yuen, et al., 2010). The growing availability of large, genome-wide datasets for nonmodel organisms through reduced-representation libraries (RRL) might provide enough power to precisely estimate colonization rates with approximate Bayesian computation (ABC), provided the use of a correct demographic model. In particular, these techniques provide data that is more representative of genome-wide variation and free of ascertainment bias. They also provide actual nucleotide diversity estimates. The use of restriction site-associated DNA sequencing (RADseq) has been tested in combination with a variety of demographic inference methods for other clades of organisms (Nadachowska-Brzyska et al., 2013; Shafer et al., 2015), but there is still a lot to discover about the potential of sequence capture methods in phylogeographic studies (Harvey et al., 2016). Here, I investigate the accuracy of migration rate estimates during the postglacial recolonization of *P. sitchensis* using sequence capture data and a popular inference method, ABC, and applying a demographic model investigated in Chapter 4.

P. sitchensis is an interesting system to perform this analysis, for several reasons. First, strong prior knowledge is already available about the general temporal and spatial pattern of postglacial colonization for this species. The historical northward colonization of *P. sitchensis* following the retreat of the Cordilleran ice sheet has been investigated in several paleofossil studies (reviewed in Tae, 1997). According to fossil records, the species closely followed the retreat of the ice sheet along the coast to Puget Sound 13,400 years BP, to Vancouver 12,350 years BP, and reached the Alexander Archipelago in Alaska around 10,500 years ago (see Figure E.1 for the location of landmarks). The pace of advance of *P. sitchensis* forests then decreased considerably, likely due to the presence of glaciers and large glacial rivers hindering colonization along the coast, as well as frequent droughts (Peteet, 1986). Coastal areas close to Anchorage became forested 2,680 years BP. The forest on the Kodiak Archipelago possibly started on Shuyak Island and North-East Afognak Island (Tae, 1997; Stacy Studebaker 2014, pers.comm.) and expanded south-westward to reach Kodiak Island around 500 years ago (Griggs, 1937; Tae, 1997). This appendix aims to provide an empirical application of Chapter 4, and therefore focuses on exploring the powers and limitations of ABC in parameter estimation rather than

model selection. The precise prior knowledge of *P. sitchensis* postglacial migration is therefore an asset, facilitating the choice of a single demographic model of expansion history. Holliday et al. (2010) conducted a reconstruction of the range-wide colonization history of *P. sitchensis* based on the allelic frequency spectrum of expressed sequence tag (EST) markers. This approach estimated a series of successive bottleneck times from Redwood (CA) to Kodiak Island (AK), but the limitations of the method and genetic data used resulted in absolute bottleneck times that were 4 to 10 times more ancient than expected for all sampled populations. These results show that there is a genetic signature of the northward recolonization of *P. sitchensis* in populations all along the expansion route. As larger genomic datasets are now available for the species, a new attempt at estimating colonization times could lead to more accurate results. This appendix asks whether using sequence capture data in an ABC framework can help resolve the issue of imprecise or biased parameter estimates. It provides an exploratory empirical demographic inference analysis, using a 2-population demic expansion model to approach successive postglacial migration events.

E.2 Materials and methods

Sampling, sequencing and alignments

I selected 4 to 6 individuals from each of 5 provenances in the northern part of the range of *P. sitchensis* (Figure E.1). These sample sizes are slightly smaller than those implemented in Chapter 4 (N=10 per population). Population A was sampled during 2013 field collections. All other populations were sampled in a Vancouver common garden planted by Mimura and Aitken (2007a). DNA was extracted from needle samples using a modified CTAB protocol. Libraries were prepared using the exact same method as Suren et al. (2016), with 23k probes developed from the interior spruce genome (Birol et al., 2013). Libraries were sequenced using the Illumina HiSeq 2000 system with paired-end reads. The *fastx* command *fastq_quality_trimmer* (Blankenberg et al., 2010) was applied to the sequences to remove sequences shorter than 20 nucleotides. Next, the *fastx* command *fastq_quality_filter* was applied to the selected fastq sequences to remove sequences with less than 90% of nucleotides with a PHRED quality score

higher than 10. Sequences were paired using custom R scripts developed by Kay Hodgins. Sequences were then mapped to the February 2013 version of the interior spruce genome (Birol et al., 2013) using the *bwa mem* alignment tool (Li & Durbin, 2009). Alignment files were sorted and indexed using *samtools* (Li et al., 2009). I used *Picard MarkDuplicates* (http://picard.sourceforge.net) to identify and remove PCR duplicate sequences. I identified indels and conducted local realignments around them using *GATK RealignerTargetCreator* and *IndelRealigner* (DePristo et al., 2011).

Genotype calling

The variant calling pipeline below was conducted using version 3.7 of the *GATK* program (McKenna et al., 2010). The *HaplotypeCaller* command was first invoked using the *-ERC GVCF* mode with a PHRED-scaled minimum confidence threshold of 20 for calling variants. Unfiltered vcf files including invariant sites were produced using *GenotypeGVCFs* with the *--allSites* option. Including invariant sites allowed keeping tracks of the number and lengths of sequences output. I created vcf files each containing a pair of populations to be tested. Hard filters were then applied using *GATK VariantFiltration* and *SelectVariant*. Following *GATK*'s recommendations as of June 2017, sites with QD < 2.0, FS > 60.0, MQ < 40.0, MQRankSum < -12.5, ExcessHet > 10.0 or ReadPosRankSum < -8.0 were discarded, as well as non-SNP variants and variants of more than 2 alleles. The minimum individual depth was set to 5. Finally, I discarded sequence markers that were missing in one or more individuals.

ABC procedure

A single model of postglacial colonization involving all sampled populations would have to be defined by many parameters (at least 12). Chapter 4 tested the power of ABC on 2-population models of spatial expansion, using simulated datasets equivalent to the one developed here. One of the main findings of Chapter 4 was the rapid loss of estimation power as model complexity increased to as few as 5 demographic parameters. I therefore performed independent inference of demographic history for pairs of adjacent populations, using a classic 2-population demographic model of spatial expansion with four parameters: effective population sizes N_1 and N_2, time of spatial expansion from population 1 to 2 (T_{EXP}), and subsequent per-generation migration rate from population 1 to 2 (m_{21}). This model is identical to model 3 in Chapter 4 (Figure 4.1c). For

all pairs of populations, I used the same priors as for simulation analyses in Chapter 4 (Table 4.1) with a few exceptions: I set the nucleotide mutation rate to 2×10^{-8}, which is the most recent estimate for *P. sitchensis* (Hanlon, 2018); I reduced the prior range for T_{EXP} from [2 : 10^3] to [2:500], which is more adequate considering the timeframe of postglacial recolonization in relation to generation times in tree species; finally I increased the population size priors for N_1 and N_2 from [10^3 : 10^4] to [10^3 : 3×10^4] to accommodate the generally high effective population sizes in conifers (Chen, 2012). To develop summary statistics, I selected all 28 summary statistics available in *msABC* that did not involve linkage information and knowledge of ancestral allelic states, as the datasets built for the analysis did not contain such information. After calculating summary statistics for each markers, the mean and variance of each summary statistic were calculated over polymorphic sequence markers. I reduced the number of obtained mean and variance statistics by using a PLS transformation (see section D.1 for details). The 6 first PLS components were used as summary statistics.

Building the datasets

I defined a valid marker as a DNA sequence with at least 100 continuous basepairs, or with no more than 10-bp gaps, filling the latter with NAs. To convert vcf files into *msABC* input files, I developed custom scripts involving shell scripts, one perl script written by Jon Degner, custom R scripts, and the *fastaconvtr* command (Ramos-Onsins & Vera, unpublished). A preliminary analysis of population structure using the complete dataset of sampled individuals was performed with PCA using with a random subset of 20k polymorphic SNPs extracted from the filtered sequence capture data, and using the *ade4* R package. Populations P, IB, V and RB clustered together on PCs 1, 2 and 3 but started to separate at higher dimensions (Figure E.2). A and K formed a separate, more variable group. IN formed a cluster on its own on PC1 and 2. The ABC analysis was performed on geographically adjacent population pairs, excluding IN and K because of uncertainties about the origin of IN and the very recent origin of K (see Chapter 2). As the model used here is a 2-population model, I created 4 pairs of populations following the expected colonization sequence: PR-IB, IB-V, V-RB, RB-A (Figure E.1). One dataset was created per pair of populations. Each dataset was highly complete (each marker was genotyped in all individuals) and contained >10k sequences of lengths ranging from 100 to 3350 bp (Figure

E.3). These observed datasets are slightly larger in the total number of nucleotides than simulated datasets tested in Chapter 4.

E.3 Results

The ABC inference results for the colonization history of pairwise *P. sitchensis* populations are displayed on Figure E.4. Theta for population 1 (source population) shows a narrow peak of posterior probability in all pairwise analyses. After conversion (N_i=theta$_i$/4μ), I obtained effective population size estimates ranging from N_1= 7.55×10^4 for V to N_1=9.4×10^4 for P, the southernmost population. Estimation of N_2 failed, which is not surprising as Chapter 4 shows that effective sizes of growing populations could not be estimated, even when inferring the simplest demographic models. However, estimation of m_{21} and T_{EXP} also seems to have failed in most cases, showing the highest posterior probabilities at the edge of the prior range. This suggests that there is a mismatch between simulated and observed datasets. An obvious reason for this would be that the demographic model was not appropriately chosen. There is a discrepancy between estimated posterior distributions and the distribution of parameter values of associated retained simulations (Figure E.4). This is also diagnostic of the inadequacy of the model. As a linear regression between statistics and model parameters is performed in the retained simulations (simulations with sets of summary statistics closest to the observed values), the posterior distribution will be slightly different from the distribution of parameter values in retained simulations. A discrepancy between the two distributions indicates that the observed set of statistics lies outside of the hyperdimensional cloud of statistics of retained simulations, making the results from the linear regression adjustment (and from the whole ABC analysis) unreliable. For example, for the pair V-RB, most T_{EXP} parameter values in the retained simulations lie below 50 generations, but the mode of the posterior is at the maximum possible value: 500 generations. Similarly, for P-IB, IB-V and RB-A analyses, the distribution of migration rates in retained simulations follows the prior distribution, suggesting that this parameter could not be estimated, but the posterior still shows an artefactual peak at the lower limit of the prior range. Figure E.5 confirms this issue clearly: it represents PLS-transformed summary statistics of the simulated and observed datasets in a pairwise manner. In the case of an

adequate demographic model and prior range, it is expected that the observed dataset falls into the hyperspace formed by the summary statistics of the simulations retained for the posterior estimation. Here, the observed data fall into the range of statistics of retained simulations up to PLS 3. Beyond the third dimension, the observed dataset has a completely different set of statistic values than the cloud of retained simulations.

E.4 Discussion

What is wrong with the model?

The aim of this appendix was to test conventional and simple demographic models for inferring successive postglacial migrations in a tree species, using the ABC framework. The part of the range of *P. sitchensis* covered in this analysis is continuous at the regional level but fragmented at the landscape level, with many small islands as well as snow-capped mountains and wide bays separating habitat. I therefore assumed that model 3 from Chapter 4, with 2 demes and only 4 model parameters, was an extreme simplification of the patterns of colonization that occurred between pairs of sites along the expansion route of *P. sitchensis*, and similar to many temperate tree species. However, as "all models are wrong but some are useful" (Box, 1979), it seemed necessary to test whether discrete demographic modelling could be of any use in inferring continuous, large-scale processes such as forest tree migration. The northernmost population pair sampled in the *P. sitchensis* range, RB-A, was the most likely to be accurately described by model 3, based on prior knowledge and geographic layout. Indeed, the two populations are not only adjacent but also separated by a 70km-wide ocean strait, giving biological meaning to the 2-deme model and to model parameters such as N_{02} (number of founders in A) and m_{21} (subsequent migration from RB to A). Also, population A has been established recently enough (Griggs, 1937; Tae, 1997) that its demographic growth to the present would be closely approximated by the growth rate parameter modelled as exponential and ongoing to the current population size (Figure 4.1c). As model 3 would depict demographic history less faithfully in pairs of populations more distant from the current northern range edge, the results of this multiple analysis would help draw the limits for the use of simple discrete models in more distant and older populations.

Results were unambiguous: the observed datasets, for all pairs of populations considered including RB-A, did not occur in the same multidimensional space as the one created by coalescent simulations of model 3, a pattern diagnostic of the use of an inadequate demographic model. This is not entirely surprising in the case of PR-IB, IB-V, and V-R. First, these pairs of populations feature two sites hundreds of kilometers apart with no obvious demic separation. The coalescent process following the discrete model is therefore likely to shape genetic data differently from the colonization process between the two sites. Also, population sizes probably became relatively stable many generations ago at distant sites to the south of the current expansion front, in contrast with the model, which characterizes continuous population growth to the present for one of the two populations of the pair.

However, the failure of the analysis for the RB-A pair is more difficult to interpret. There are a few possible reasons for this result. First, many demographic processes specific to spatial expansion, such as allele surfing (Klopfstein et al., 2006) and occasional long-distance dispersal (Austerlitz & Garnier-Géré, 2003) would be unaccounted for in the demographic model and could have a long-lasting effect on the genetic composition of populations. Secondly, sampling a few individuals per populations might not allow an accurate representation of the complex genetic make-up of spatially expanding populations (Excoffier et al., 2009). Especially, *P. sitchensis* is known to hybridize with *P. glauca* at multiple sites in British Columbia and Alaska, including on the Kodiak Archipelago (see Chapter 3). The contribution of unsampled populations, potentially from different species, to the gene pool of sampled populations would also contribute to the differences between simulated and observed datasets. This illustrates how crucial it is to have a high level of confidence in the chosen demographic model, based on other solid lines of evidence. Otherwise, alternative demographic models need to be tested against the chosen model following appropriate procedures to statistically test model fit (Csilléry et al., 2010; Fearnhead & Prangle, 2012).

Future improvements

Because long-distance dispersal (Kremer et al., 2012) and hybridization with sister species (Jaramillo-Correa et al., 2009) are common in tree species during postglacial colonization, knowing the source population(s) of any given forest site and sampling accordingly for

demographic inference presents great challenges. One direction of research could involve similar models as the one used in this appendix, tested on different combinations of potential source populations. Model selection techniques could then be applied to determine which populations were the most likely genetic sources for the site under study. This would necessitate an appropriate choice of summary statistics, following a different procedure from the one implemented in this appendix (Marin et al., 2014). However, this would not solve the issues associated with using a discrete population model for populations on a continuous landscape. A more effective solution, potentially applicable in ABC, would be to create a spatially explicit model with a large number of interconnected demes over a landscape, and focus the estimation on demes corresponding to the geographic location of sampled sites. Such an approach was developed by Hamilton et al. (2005), who implemented ABC in a bi-dimensional stepping stone model of range expansion with migration, and obtained relatively accurate estimates for colonization times (but not for inter-deme migration rates). Further development of inference techniques addressing demographic models of individuals distributed along a spatial continuum and that are applicable to genomic data is currently needed.

Alternatively, demographic inferences of single-population models avoid having to select and sample potential source populations, and the validity of their results rely entirely on the accurate interpretation of estimated changes in population size, which can be challenging in the presence of continuous migration between populations. With many individuals and a large, informative genomic dataset, a study following similar ABC procedures as Holliday et al. (2010) could perhaps lead to accurate model parameter estimates, in addition to appropriate model selection. A set of seemingly powerful methods focusing on single-population models, including PSMC (Li & Durbin, 2011) and MSMC (Schiffels & Durbin, 2014), have recently been developed and rely on whole-genome sequences from a single (PSMC) or a few (MSMC) individuals. These methods have not been widely applied to inference in nonmodel organisms (Nadachowska-Brzyska et al., 2013), but could soon provide potential advances in demographic inference for tree population history, as whole-genome sequencing is becoming more available in tree species (Neale et al., 2013).

Conclusion

The ABC analysis implemented in this appendix failed to estimate historic colonization times and their associated demographic patterns of subsequent migration along the postglacial expansion route of *P. sitchensis*. The main problem seems to be the use of a model that poorly describes the actual demographic history of the species. This result was expected in analyses of more distant pairs of populations but disappointing in the analysis involving the two northernmost populations, RB and A, where prior knowledge seemed to support the 2-deme demographic model that was implemented. Whether this failure is due to unsampled source populations or spatial expansion processes that were not accounted for, it is informative for future phylogeographic studies making demographic inferences in tree populations. High migration rates among populations, typical of widespread forest tree species, have always made the demographic inference of population history difficult (Hamilton et al., 2005; Robinson et al., 2014). More appropriate methods for the inference of tree migration parameters would ideally involve an inference framework allowing continuous spatial diffusion models to be tested. As such methods are not fully developed in a user-friendly format yet, readily available solutions would be to increase sample sizes for each population considered, and to implement model selection. As ABC has its limitations when applied to spatial expansion models (see Chapter 4), testing and comparing different inference techniques should also be beneficial. Current and future advances in inference techniques and genomic datasets might overcome some of the current limitations of phylogeographic inference and make state-of-the-art methods more and more available in nonmodel species.

E.5 Figures

Figure E.1 Map of populations selected for principal component analysis of genetic data and approximate Bayesian computation, with location of landmarks mentioned in the introduction. The range of *P. sitchensis* is depicted in green and the colour code for sampled population matches the one used for PCA data points in Figure E.2.

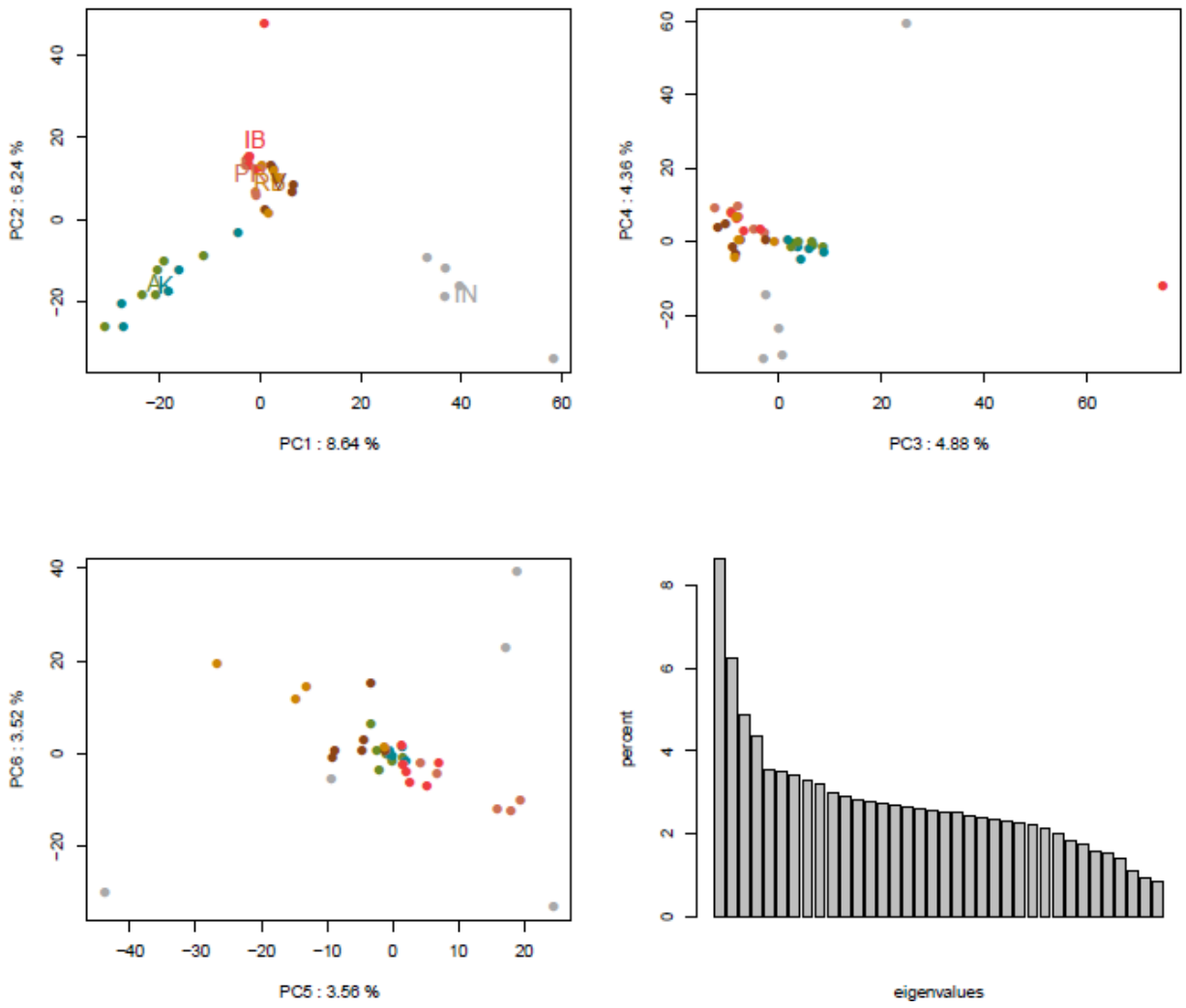


Figure E.2 PCA of genotypes sampled. PC 1 to 6 are represented, as well as the eigenvalue profile.

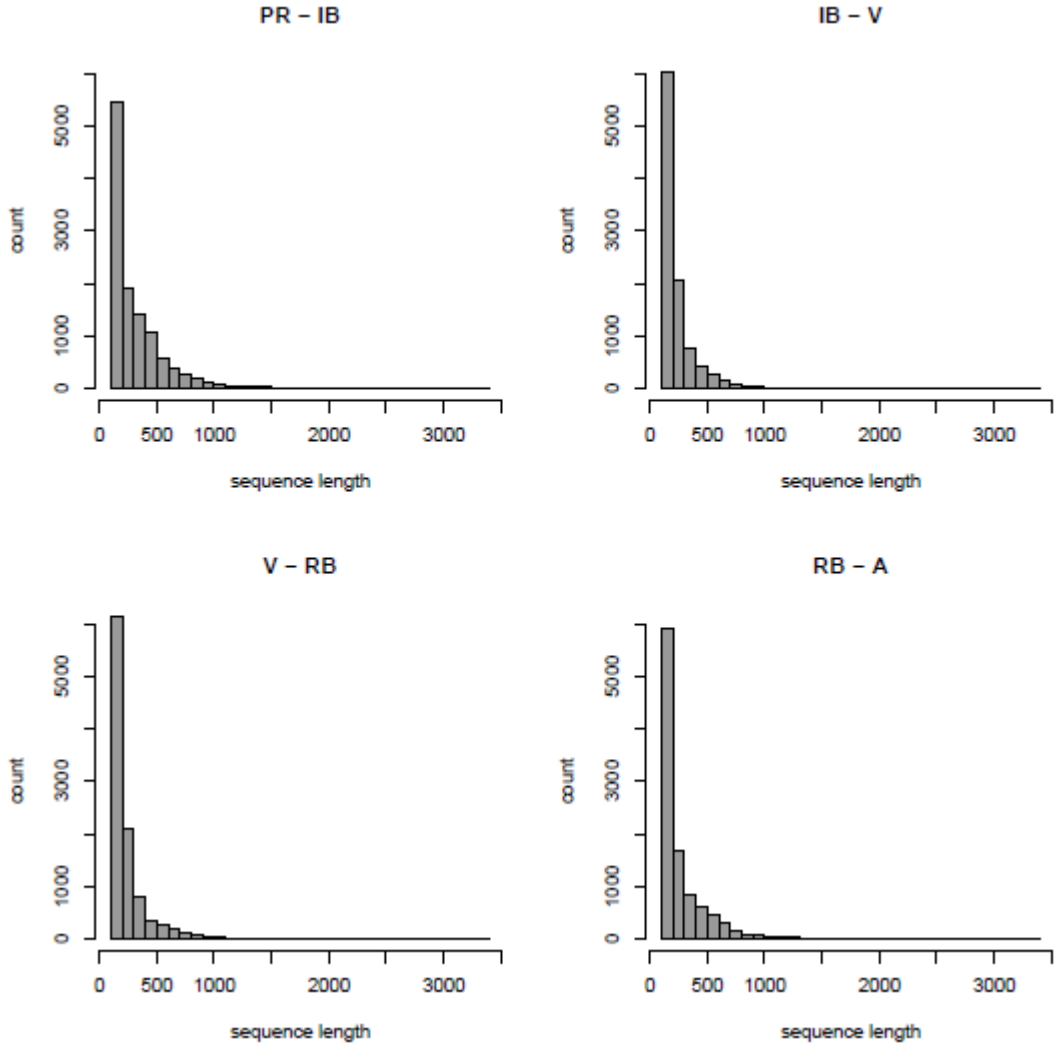


Figure E.3 Distribution of sequence lengths for different datasets used in empirical ABC analysis

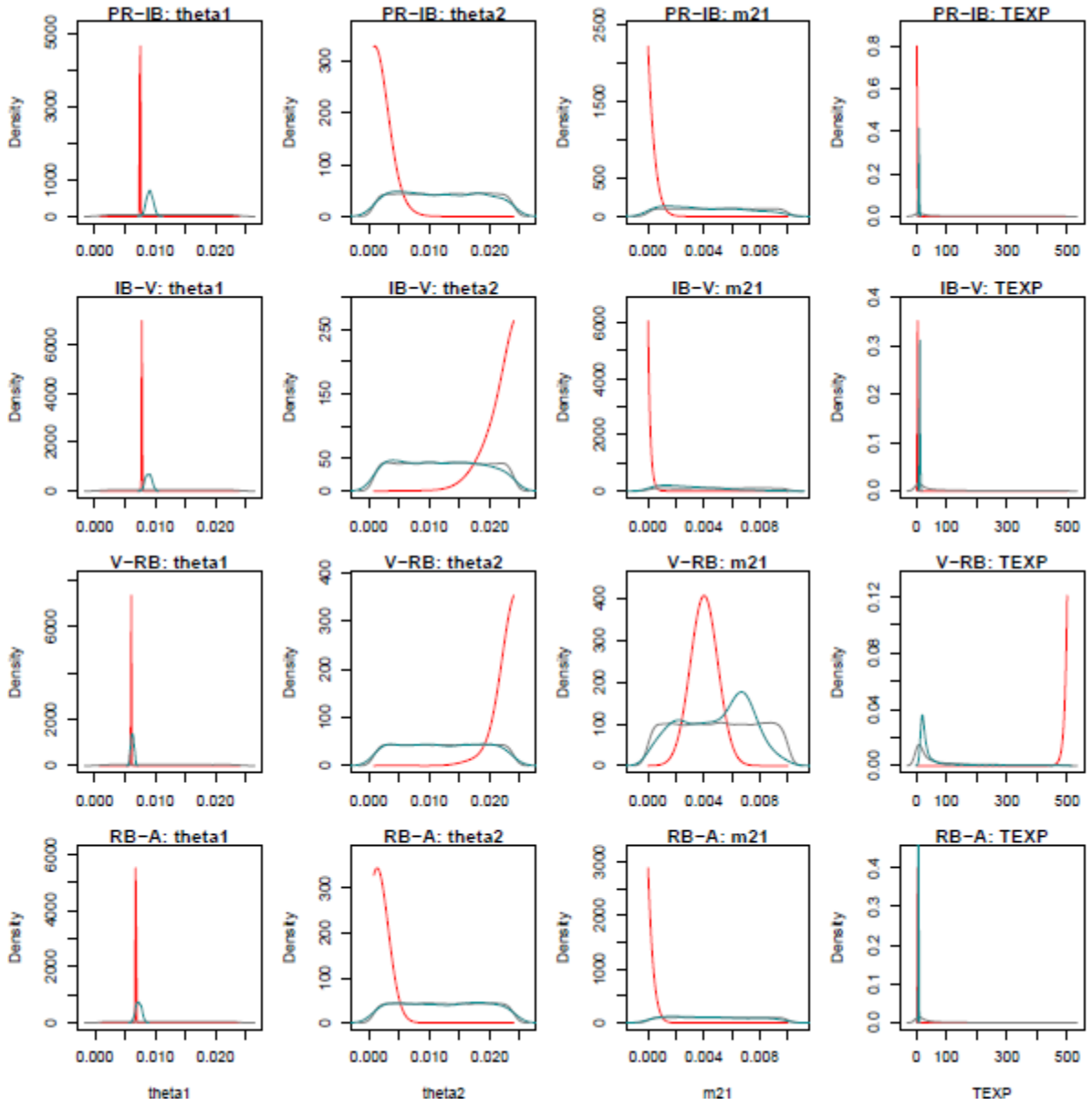


Figure E.4 ABC parameter estimates for 4 pairs of populations. Grey lines represent parameter priors, red lines correspond to the estimated posterior, and blue lines are the distribution of parameter values of retained simulations used to estimate the posterior. theta$_i$= 4N$_i$μ, with μ=2.10^{-8}, i being the population index (1 or 2).

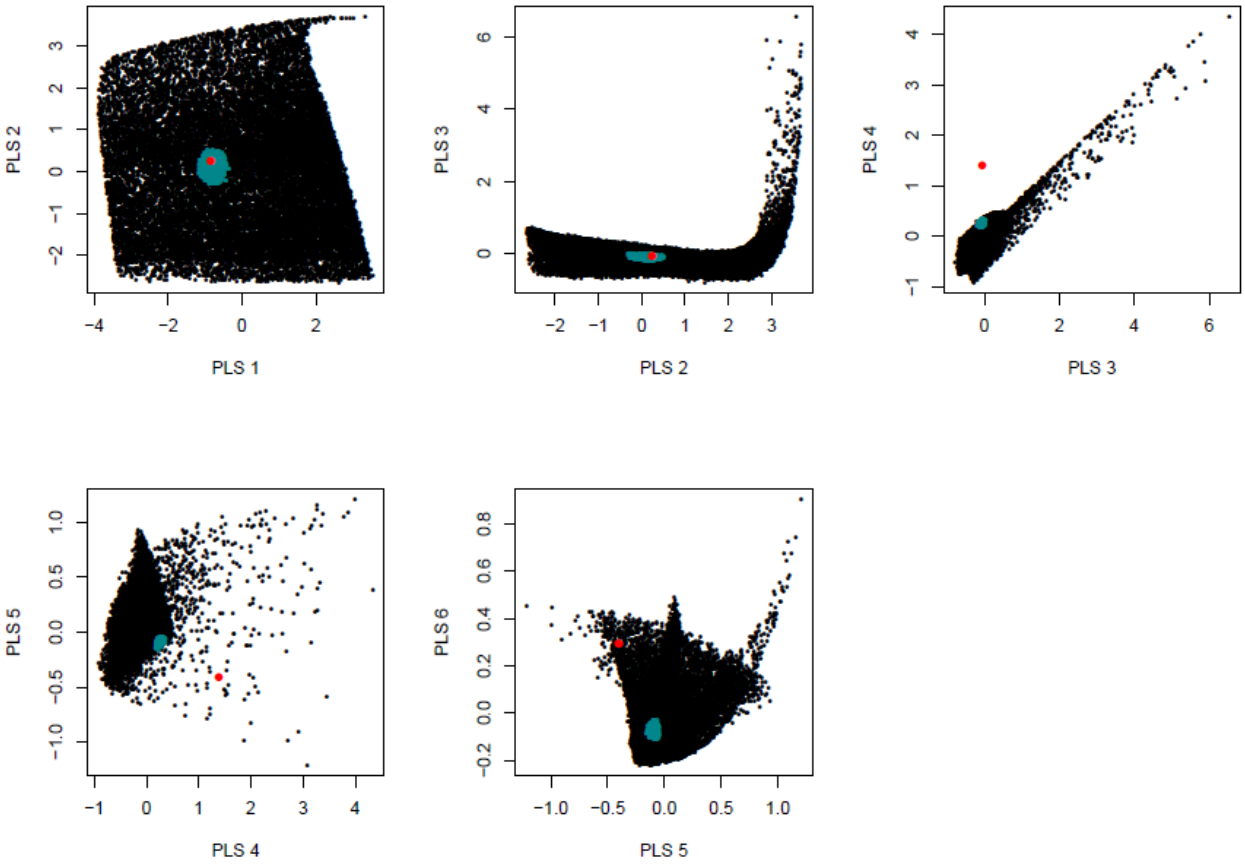


Figure E.5 Pairwise representation of PLS-transformed summary statistics for observed and simulated datasets. Black points represent a random sample of 20k simulations from the complete set of 1M simulations. Blue points represent the set of 1000 retained simulation from the posterior estimation. The red point corresponds to the position of the observed dataset.