

Genotyping for the study of population genetics and trait evaluation in *Picea sitchensis*

Tomás Byrne

February 2024



Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Thesis

Submitted in fulfilment of the requirements for the Doctor of Philosophy at
Trinity College Dublin

Supervisors:

Prof. Trevor R. Hodkinson

Botany, School of Natural Sciences

Trinity College Dublin

Dr. Susanne Barth

Crops Research Centre, Oak Park, Carlow

Teagasc

Dr. Stephen L. Byrne

Crops Research Centre, Oak Park, Carlow

Teagasc

Dr. Colin Kelleher

Plant Molecular Laboratory, Glasnevin, Dublin

National Botanic Gardens of Ireland

Project Lead

Dr. Niall Farrelly

Forestry Development Department, Athenry, Galway

Teagasc

Declaration

I hereby declare that I am the main contributor to this thesis. I also declare that the work presented in this thesis has not been previously submitted in part or whole for any degree at any other college or university. I agree that the library of Trinity College Dublin can lend or copy this thesis on request

A handwritten signature in black ink that reads "Tomás Byrne". The signature is written in a cursive style and is positioned above a solid horizontal line.

Tomás Byrne

February 2024

Acknowledgements

This work could not be carried out without the funding from the Department of Agriculture, Food and the Marine under grant award 17/C/29. I am very grateful to all the staff at the OPW in JFK arboretum, Co. Wexford for access to their Sitka spruce collection.

I am eternally grateful to my supervisors Susanne Barth, Trevor Hodkinson, Stephen Byrne and Colin Kelleher for giving me the opportunity to do this PhD and thankful for all the support along the way. Thanks for all the help with admin, corrections, contributions and sampling that were core to me finishing this PhD. I would like to thank the GENESIS project lead, Niall Farrelly for all the organisation, advice and guidance along the way.

I would like to thank all the staff at Teagasc Oak Park for their help, expertise and support. I would also like to thank all the friends I have made especially Katie, Niamh, Leona, Robyn, Tara and Conor.

To my family, I would like to thank all of them for the love, support and encouragement to pursue a PhD. To my partner, Keith, I couldn't have done this without you. To his family, thank you for all the support and to my closest friends, thank you for sticking by me.

Summary

Genotyping provides information on the genetic makeup of an individual or a population of individuals. Molecular markers are commonly used as an affordable method of genotyping individuals, and the most popular modern marker is the Single Nucleotide Polymorphism (SNP). SNPs have a wide range of applications, including their use in understanding the evolutionary history and genetics of populations. In breeding they can be used for discovery, prediction and selection of traits. Furthermore, they can be used in the comparison of populations and understanding of parentage. In this thesis there were two large populations genotyped using SNPs and study, compare and investigate the populations.

Sitka spruce, *Picea sitchensis*, is native to North America, occupying coastal regions along the Pacific North West and has become the predominant forestry species in Ireland since its introduction but little is known about the genetic makeup of these forests or the collections used to breed and select new spruce in Ireland (chapter 1). In chapter two the IUFRO (International Union of Forest Research Organisation) population was genotyped, a representative breeding population of the native range of Sitka spruce. The population consists of 80 provenances with 9-15 genotypes per provenance. The two genotyping platforms explored were Genotyping by Sequencing (GBS) and SNPseq. It was found that due to number of SNPs sequenced and price points that GBS was more efficient and less at risk of bias because it uses short read Whole genome sequencing (WGS). 1177 samples representing all 80 provenances were genotyped. 36567 SNPs were discovered after SNP filtration. Genetic diversity studies show the population is very diverse and has similar diversity levels to other spruces in this region that have been studied. Using PCA and DAPC it was discovered that the population was structured into one main cluster and

isolation of North Alaskan provenances, including two islands, Montague and Kodiak Island. There was an isolation effect based off these two islands. Admixture analysis reveals a genetic distribution pattern that correlates with the retreat of the cordilleran ice sheet. This suggests a recolonization pattern akin to some of the previous hypothesis.

Sitka spruce occupies various different ecological and environmental gradients. This wide range of climates allows for local adaptations at the provenance level. In chapter 3 local adaptation was investigated using various analyses. A Genome Wide Association Study (GWAS) and a Genotype Environment Analysis (GEA) was used to study the IUFRO population. From these analyses, loci involved with traits and various climatic measures were discovered. A Canonical Correlation (CANCOR) analysis was used which combines genotypes, environments and phenotypes together to study the interaction between the three. From the CANCOR analysis there were 121 loci positively correlated with an environmental trait and ten loci positively correlated to a phenotypic trait. There were two significant clusters of loci, one was correlated with adaptations to cold wintering conditions and the other was correlated with adaptations to conditions ideal for growing. The distribution of these loci involved with height are positively correlated below the 50th latitude and negatively correlated above the 50th latitude. Overall local adaptation across the range that appears to be characterised by subtle to moderate shifts in Minor Allele Frequency (MAF) in the traits tested. This is similar to what is seen in other species such as Loblolly Pine, *Pinus taeda*.

A second population was genotyped as part of this thesis. The Irish Sitka Spruce Tree Improvement Programme (ISSTIP) collection is a breeding population that originates from the native range of Sitka spruce, from Haida Gwaii to Oregon. In chapter 4, GBS was used to genotype 215 samples of Sitka spruce from the ISSTIP population to compare to the IUFRO population. There is a subtle loss in genetic diversity after restricting the

provenances used in the breeding program. The range selected for breeding was less diverse than the entire breeding range leading to loss of potentially interesting traits. The DAPC demonstrated that the ISSTIP population has become genetically differentiated from the IUFRO population. Admixture analyses reveal the ISSTIP population primarily shares no admixture with other populations apart from Haida Gwaii. This genotyping has captured and been used compared the population genetics of a breeding population.

Seed orchard dynamics were studied as part of this thesis. A custom KASP assay to inexpensively and efficiently genotype Irish Sitka spruce seed orchards was designed using a minimal set of markers. The markers were designed to distinguish between all genotyped material from the ISSTIP population. The assay was deployed to test the offspring of an Irish seed orchard and test for contamination and inbreeding. 54.41% of the offspring were from outcrossing parent pairs, 21.53% were as a result of inbreeding and 24.65% were as a result of contamination from fraternal sources. The presence of over contributing parental genotypes in the population, with some parents being responsible for 31.60% of offspring.

Overall this thesis aimed to establish a baseline for genetic diversity in Sitka spruce and compare it to a breeding population. The IUFRO population has also be used to understand the evolution of Sitka spruce and potential loci correlated with traits of interest. Two diverse populations have been genotyped in this study, which can act as resources for spruce research including analysis of traits and evolution. These resources will also be useful for breeders for progeny selection, genomic selection and deployment of trees in seed orchards.

Table of Contents

Declaration	3
Acknowledgements	4
Summary.....	5
Table of Contents	8
Chapter 1- Introduction	11
1.1 Distribution of Sitka Spruce.....	11
1.2 Genomics and Population Genetics of conifers	15
1.3 Detecting adaptation and population genetics.....	18
1.4 Breeding.....	20
1.5 Seed Orchards	21
1.6 Populations studied	23
1.7 Workflow and aims.....	25
Chapter 2- Genetic Diversity and Structure of a Diverse Population of <i>Picea sitchensis</i> Using Genotyping-by-Sequencing	26
2.1 Introduction	27
2.2 Materials and Methods.....	30
2.2.1 Sample Populations	30
2.2.2 DNA Extraction and Sequencing	32
2.2.4 Variant Calling.....	32
2.3.5 Genetic Diversity Statistics.....	33
2.2.6 Population Structure.....	33
2.2.7 Phylogenetic Analysis.....	34
2.2.8 Isolation by Distance	34
2.2.9 Admixture Analysis	35
2.3 Results and Discussion	35
2.3.1 Genotyping a Large Sitka Spruce Population, Covering Its Native Geographic Range.....	35
2.3.2 Geographic and Genetic Diversity.....	37
2.3.3 Genetic Structure	39
2.2.3 Genetic Admixture and Ancestry	46
2.4 Conclusions	50
Chapter 3- Environmental and landscape genomics of <i>Picea sitchensis</i> (Sitka spruce) reveals clusters of loci involved with both resistance to wintering and increases in height.	52
3.2 Materials and Methods.....	56
3.2.4 GWAS	59
3.2.5 CANCOR	59

3.2.6 Minor Allele Frequencies	60
3.2.8 Functional Analysis	60
3.3 Results	61
3.3.1 GWAS/GEA.....	61
3.3.2 CANCOR Analysis	62
3.3.3 Allele Frequencies and Loci Distribution.....	65
3.3.4 Functional Annotation	68
3.4 Discussion	70
3.4.1 Adaptation in Sitka Spruce.....	70
3.4.2 Adaptations to Northern Wintering Conditions	71
3.4.3 Adaptations to Southern Conditions	71
3.5 Conclusions	73
<i>Chapter 4- Comparison of the native North American Sitka Spruce and a Sitka Spruce breeding population.</i>	<i>74</i>
4.1 Introduction	75
4.2 Methods.....	77
4.2.1 Plant Collection and Genotyping	77
4.2.2 Variant Calling.....	78
4.2.3 Inferring Origins of Unknown Genotypes	78
4.2.4 Diversity Statistics	79
4.2.5 Clustering	79
4.2.6 Gene Flow between Populations	79
4.2.7 Admixture	80
4.3 Results	80
4.3.1 Variant Calling and Inferring Origins.....	80
4.3.2 Diversity Statistics	80
4.3.3 Population Clustering.....	81
4.3.5 D-statistics and Ancestry	84
4.4 Discussion	86
4.5 Conclusions	88
<i>Chapter 5- A panel of KASP markers for the accurate genotyping of Sitka spruce seed orchards reveals dominant parental genotypes.</i>	<i>89</i>
5.1 Introduction	90
5.2 Materials and methods	92
5.2.1 Seed orchard.....	92
5.2.2 Plant growth	94
5.2.3 DNA extractions	94
5.2.4 KASP assay	95
5.2.5 Data analysis	95
5.3 Results	96
5.3.1 Marker efficiency.....	96
5.3.2 Identity checks	97
5.3.3 Parentage.....	98
5.4 Discussion	99

5.5 Conclusions	103
Chapter 6- General Discussion.....	104
6.1 Discussion	105
6.2 Conclusions and future work.....	112
Appendices.....	114
Appendix 1: Introduction.....	114
Appendix 2: Developing SNPseq targets	114
Appendix 3: Developing PCR targets for genotyping Sitka spruce.	118
Appendix 4: Isoform sequencing	122
Appendix 5: Additional material from chapter 2.....	125
Appendix 6: Additional material from chapter 5.....	130
Bibliography.....	136

Chapter 1- Introduction

1.1 Distribution of Sitka Spruce

Picea sitchensis (Bong.) Carrière (Sitka spruce) is a conifer native to the coastal pacific northwest of North America (OECD, 2006). Sitka spruce is an evergreen conifer with greyish-brown to purplish thin scaly bark; its pungent needles are sharp, flattened and stand out in all directions; its cones are typically 5-9 cm long, reddish brown and its scales are rounded and irregularly toothed at the tip (Figure 1.1). It can live for 50 to 200 years but there is massive variation in lifespans depending on site.

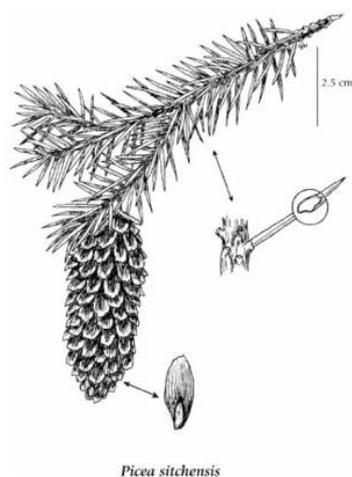


Figure 1.1- Sitka spruce needles and cone. Phenotypic representation of Sitka spruce.

(Douglas et al., 1998)

Its range begins in coastal Alaska, where it occupies several islands, principally Kodiak Island and Montague Island. It then stretches down into south Alaska where it occupies islands among the Alexander Archipelago. In Canada, it occupies coastal regions, along riverbeds and the islands Haida Gwaii and Vancouver Island. Its range ends in the United States in northern California (Figure 1.2). Along the range it overlaps, and hybridizes, with some of the seven *Picea* species native to North America (Hamilton &

Aitken, 2013). *Picea engelmannii* Parry ex. Engelm. (Englemann spruce) hybridizes with Sitka spruce in hybrid zones located in Northwestern British Columbia and South Alaska (Hamilton & Aitken, 2013). *Picea glauca* (Moench) Voss (White spruce) also hybridizes in this zone to form *Picea × lutzii* Little (Lutz’s spruce) (Menon et al., 2021). Some members of the range may also have hybridized with *Picea mariana* (Mill.) Britton, Sterns & Poggenburg (Black Spruce) which overlaps with the range in the northernmost areas (Fowler, 1983).

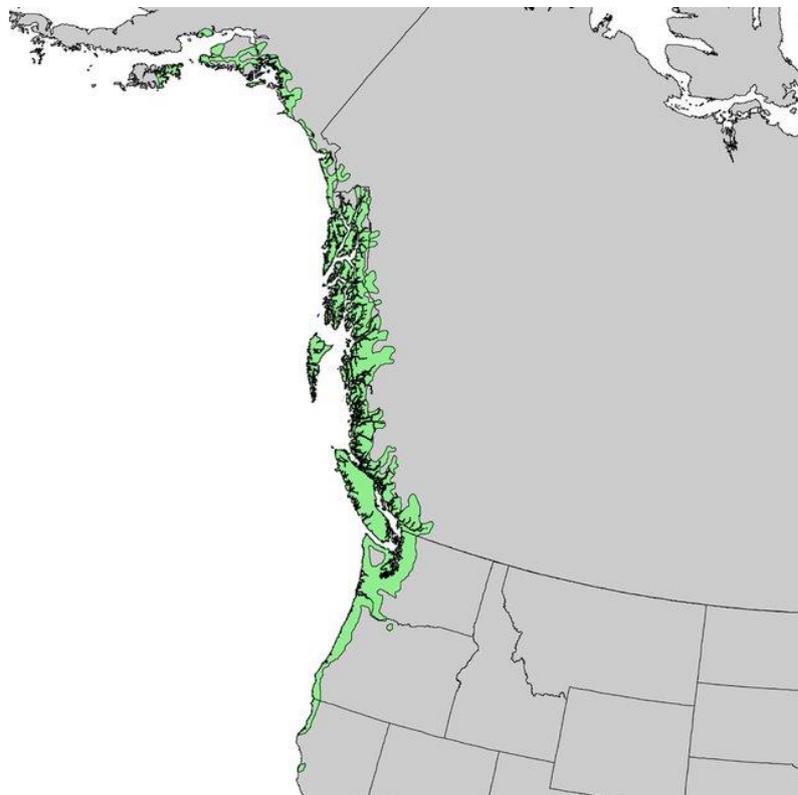


Figure 1.2- Distribution of Sitka spruce in North America. (Cvjetkovic et al., 2018)

Sitka spruce occupies many ecological and climactic niches along the Pacific Northwest. In the northern range, temperatures can reach -7.3°C with 100% snow cover throughout the winter seasons (Fick & Hijmans, 2017; Hall et al., 2006). Solar radiation is limited in these areas with the mean diurnal range in winter being 4.4hrs at the northernmost tip of the range (Fick & Hijmans, 2017). Amongst the range, the mean rainfall varies greatly

from 4015mm to 621mm per annum (Fick & Hijmans, 2017). The southern reaches of the range are relatively arid in comparison to the rest of the distribution. Sitka spruce survives this aridity by receiving moisture from water vapour coming from the warm coastal air. Southern water vapour is at a maximum 1.44kPa as compared to northern vapour being at a minimum of 0.2kPa in Southern Alaska (Fick & Hijmans, 2017). The landscapes Sitka spruce occupies are also diverse. The range is nestled between the Pacific Ocean and the Rocky Mountain range so there are varying altitudinal differences. The species occupies riverbeds notably Skeena and Naas Rivers which span inland along the hybrid zones.

Sitka spruce is most closely related to White spruce and Engelmann spruce (Lockwood et al., 2013). It is part of the North American clade of spruces. The fossil record suggests that *Picea* diverged within Pinaceae 135 million years ago in the peripherals of the Pacific basin (O'Driscoll, 1977). *Picea koyamae* Shiras is thought to be an ancestor species of Sitka spruce that migrated from Siberia to Alaska but there is no current fossil evidence to support this (O'Driscoll, 1977). Sitka spruce distribution was shaped by Pleistocene glaciation 18,000 years ago (Byrne et al., 2022). At the extent of the cordilleran ice sheet the current native range was covered by the ice sheet, apart from some of the peripheral islands like Kodiak Island and the southern range (Haro, 2017; Menounos et al., 2017). There are three theories of how Sitka spruce survived this glaciation (O'Driscoll, 1977). Firstly, it may have moved north and recolonized from the southern range. Secondly it could have occupied the peripheral areas that escaped glaciation, like Kodiak Island and the tips of Haida Gwaii. Thirdly, glaciation could have created isolated islands on mountain peaks surrounded by a glacier, termed Nunataks. This could have isolated the species and once glaciers retreated, they could recolonize the lowlands. There has been no definitive evidence to support these theories and no fossilized pollen evidence has been found.

Sitka spruce is a key forestry crop in Ireland, United Kingdom (UK), Denmark and Northern France. It was introduced to Ireland as a forestry crop in the early 1900s (Carey, 2010). Government policy reserved good land for crop and animal agriculture, driving forestry to poor quality marginal lands (Rohan, 2019). Many forestry species could not cope with upland soils, but Sitka spruce was well suited to this habitat due to its coastal North American origin (Farrelly et al., 2009). Sitka spruce became dominant in Irish forestry for this reason. By 1995, private forestry had overtaken public forestry with 60% of forestry cover being Sitka spruce. Sitka spruce is distributed similarly in the UK as it is a coastal island nation with similar climate which Sitka spruce thrives in. Similar distribution is seen in northern coastal France and Denmark.

The population used in this thesis was the IUFRO (International Union of Forest Research Organization) collection, which represents trees sourced from a diverse distribution of the native range of Sitka spruce. The establishment of the IUFRO collection was carried out with two seed-collecting phases in 1968 and 1970 (O'Driscoll, 1978). Seed was collected from 80. The provenances selected were at least 50–80 km from each other and represented commercial or scientific areas of interest. The provenances occupied 19 geographic regions, which were determined at collection. Initially, this collection was used to test the performance of provenances in non-native regions for forestry (O'Driscoll, 1978). The collection was planted across eleven countries in its entirety, or subsets of provenances were used, to assess their establishment and performance. The initial experiments found that provenances from Washington were best suited to Ireland, Haida Gwaii was most suited to the UK and Oregon was the best suited to France (O'Driscoll, 1978).

1.2 Genomics and Population Genetics of conifers

The genomes of conifers are comparably large, varying in size from ~6.5Gb to 37Gb and often have intraspecies variation (De La Torre et al., 2014). Most conifers have 12 ($2n=24$) chromosomes which are all similar in size (De La Torre et al., 2014). Their genome size is one of the contributing factors to low rates of sequencing projects in conifers. Other factors include generation length, heterozygosity and the outbreeding nature of conifers (Ahuja & Neale, 2005). Unlike angiosperms, there is no evidence of whole genome duplication events that would cause an increase in genome size, however there is evidence of genome size increase due to accumulation of transposable elements (Ahuja & Neale, 2005). Initial studies using bacterial artificial chromosomes indicated large genome sizes of conifers were determined by the large number of retrotransposons. 62% of the genome size of loblolly pine is composed of retrotransposons of which 70% are long terminal repeat retrotransposons (LTR-RT) mainly from the *Gypsy* and *Copia* superfamily's (Wegrzyn et al., 2014). A similar situation is seen in the genome of Norway spruce where the genome grew over the course of millions of years due to the insertion of LTR-RTs. 69% of the Norway spruce genome is composed of transposable elements (Nystedt et al., 2013). The *Gypsy* superfamily is far most abundant in Spruce, Pine, Fir and Yew, however the *Copia* superfamily is more abundant in Juniper (Ahuja & Neale, 2005; De La Torre et al., 2014; Prunier et al., 2016). Long introns are also characteristic of conifer genomes with conifers having some of the longest introns found in any plant species. In Norway spruce, introns of 68kb in length were discovered and introns of length 158kb were found in loblolly pine (Ahuja & Neale, 2005; De La Torre et al., 2014; Nystedt et al., 2013; Wegrzyn et al., 2014). Local adaption is characterized by the evolution of an organism's population to become more suited to its environment than other members of its species (Mimura & Aitken, 2010). Local adaption is commonly seen in conifers with species covering large areas and low

rates of speciation (Prunier et al., 2016). The adaptations found in spruces and the wide range of phenotype variants are due to many evolutionary processes at work over generations. These processes act on the population level, species level and between species. Within populations, mutations can be neutral or negatively or positively affect the individual. Sitka spruce has one of the highest per generation mutation rates, but its long lifespan means its populations have a low per year mutation rate (Hanlon et al., 2019). Mutations in spruce are most likely to be synonymous and have no overall phenotypic effect (Buschiazzo et al., 2012). Positive adaptations can be kept and passed down through a population through natural selection. Negative adaptations will be removed through natural selection; however neutral mutations are carried through a population at a random level through the process of genetic drift. Within a species and its populations, mutations can be distributed by gene flow. Given enough time, and sufficient gene flow, populations will become less differentiated from each other (Kremer et al., 2012). The inverse is also true with isolation events leading to differentiated populations (Cordeiro et al., 2019). The amount of gene flow between populations is very much dependent on mating systems of the species. Due to the wind pollinated nature of conifers, pollen can travel hundreds of kilometers allowing for extensive gene flow (Chen et al., 2018; Hamilton & Aitken, 2013; Jimenez-Ramirez et al., 2021; Kremer et al., 2012; O'Connell et al., 2007).

Genetic diversity refers to the variety of alleles in a population and variation in non-genic regions (Mahoney & Springer, 2009). The allelic variation confers differences in phenotypes and adaptive traits, with positive traits being passed down through natural selection. Genetic diversity is important for the conservation of biodiversity and breeding applications. Measurements of genetic diversity are varied and largely depend on population size and population type (ref, perhaps Holsinger & Weir, 2009). Measurements based on the frequency of variants include Nei's Genetic diversity, Fixation index and

allelic diversity. Nei's genetic diversity (or expected heterozygosity) is a widely used measure of diversity and refers to the probability of two alleles chosen at random to be different (Nei, 1973). Furthermore, nucleotide diversity is often used for sequence data, and nucleotide diversity is defined as the average number of nucleotide differences per site between two sequences in all possible pairs in the sample population (Korunes & Samuk, 2021). Fixation (F_{st}) refers to the measurement of coancestry where the probability of two individuals being identical by descent (Holsinger & Weir, 2009). Allelic diversity refers to the average number of alleles at a locus in a population (Caballero & García-Dorado, 2013). Genetic diversity can also be described in terms of number of variants with measures such as polymorphism rate, and allelic richness. Allelic richness refers to the number of variants per individual in a population. Inference of inter and intra-specific population genetic diversity is key to the conservation of forest genetic resources and findings can be implemented into breeding programmes, conservation and general tree research. A diverse range of genotypes can be used to understand a populations structure, adaptations, and evolution. Maintaining diversity is key in breeding programmes to avoid inbreeding depression (Williams & Savolainen, 1996).

Population structure describes the genetic variation within and between subpopulations. This structure is intrinsic to maintaining genetic diversity and is influenced by evolutionary processes. The overall population is referred to as the metapopulation which can contain subpopulations and local populations. Random mating within the entire population is referred to as the population being in panmixis (Cassiman, 2003). Conifers have low levels of species diversity and low levels of distinctive population structures as compared to angiosperms (Prunier et al., 2016). Conifers tend to occupy large geographical ranges and undergo local adaptation (De La Torre et al., 2019; Günther & Coop, 2013; Hornoy et al., 2015; Mimura & Aitken, 2010). Like all organisms, isolation in refugia can

cause genetic differentiation leading to increased genetic diversity and divergent evolutionary patterns (Cordeiro et al., 2019; Stojnic et al., 2019). The measurement and visualization of population structure can be completed in many ways. It can be estimated with F statistics ranging from zero to one, where high F_{st} demonstrates high population differentiation. Alternatively, modeling for K (number of populations) can be completed and evaluated using the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC) (Kamvar et al., 2014; Verity & Nichols, 2016). Visualization of population structure and clustering of subpopulations can be done using ordination methods such as Principal Component Analysis (PCA) on genetic variation. Discriminate Analysis of Principal Components (DAPC) is also used to visualize and understand population structure (Kamvar et al., 2014). Other software can be used to visualize structure and ancestry including STRUCTURE or ADMIXTURE, which both use maximum likelihood approaches with differing algorithms (Alexander et al., 2009; Pina-Martins et al., 2017; Pritchard et al., 2000; Raj et al., 2014).

1.3 Detecting adaptation and population genetics

The ability to detect adaptation, population structure and diversity can allow for improved breeding and greater understanding of populations. Genotyping is the use of molecular markers to understand the genetic make-up of an individual within a population. Restriction fragment length polymorphism (RFLP) is a genotyping technique that exploits variations in homologous DNA sequences, known as polymorphisms, to distinguish individuals (Tanksley et al., 1989). RAPD (Random Amplification of Polymorphic DNA) is a PCR method that amplifies random DNA segments by using arbitrary short primers (Williams et al., 1990). The most used genotyping method is Single Nucleotide Polymorphisms (SNP) genotyping which detects single base pair changes in the DNA. This involves the measurement of variation of SNPs amongst individuals. There are many variations and

types of SNP genotyping methods. Dynamic allele-specific hybridization (DASH) genotyping takes advantage of the differences in the melting temperature in DNA that results from the instability of mismatched base pairs (Howell et al., 1999) SNP microarrays have multiple probes with a fluorophore attached on a solid surface. To prevent misreads multiple probes are located on the microarray. Crop specific simple sequence repeats (SSR) have facilitated the development of high-density maps for wheat, barley, potato and many other crops (Rasheed et al., 2017). SSRs are frequently used for gene mapping and tagging, yet there are limitations to their use due to low density. Amplified fragment length polymorphism (AFLP) are fragments, typically 80-500 bp in size, that can be compared between individuals of a population (Mba & Tohme, 2005). This was a popular method for genotyping individuals but is becoming replaced by SNPs (Rasheed et al. 2017). When multiplexed AFLPs can be used to genotype with up to 50 loci at a time (Paun & Schönswetter, 2012). The advent of cheap Next Generation Sequencing (NGS) techniques has allowed for cost effective high throughput genotyping. The first stage of developing a genotyping platform is to compile current information on the genomic resources of a species (Elshire et al., 2011). The key resource for this stage will be access to a genome assembly for the species which can be used to inform marker selection or for species without a reference genome, a transcriptome can be used or markers can be clustered. Genotyping by Sequencing (GBS) has been used widely in crop species to identify SNPs (Elshire et al., 2011). A similar approach is called Restriction-site associated DNA (RAD) sequencing, and both rely on a reduction of genome size before sequencing. Depending on number of markers, cost per sample for SNP based genotyping can be as low as \$5 per sample (Hall et al., 2020). Along with cheaper cost, NGS based SNP genotyping has greater accuracy due to the higher density of markers.

1.4 Breeding

The breeding cycle of forestry tree crops begins with the selection of trees with desirable traits. Their progeny are deployed in a seed orchard where they cross and then the offspring are deployed in forestry. This cycle can reoccur to increase genetic gain over time. At every stage of this process, genotype data can be used. For progeny selection, genomic selection and genome wide association studies, markers are used to inform selections (Rasheed et al., 2017). Genomic selection uses markers to maximize the genetic gain in a breeding program (Kumar et al., 2012). Genome wide association studies can be used to relate desirable phenotypes to genotypes using SNPs. At this selection stage of the breeding program, markers are used on large populations. For this reason, markers need to be cost effective. After the selection stage, markers can also be used in the next stage of breeding, the field deployment in seed orchards (Yuan et al., 2016)

Several types of molecular markers have been used in the deployment and evaluations of seed orchards for breeding. Alloenzymes, cpSSRs, nuclear SSRs and RAPD have been used in *Pinus contorta*, *Picea glauca* and *Pinus thunbergii* respectively (Goto et al., 2002; Schoen & Stewart, 1986; Stoehr & Newton, 2002). SSRs are commonly used in parentage studies which allow for genetic distance-based deployment of seed orchards (Amico et al., 2019; Sun et al., 2017). SSRs have a high degree of polymorphism which also allows them to be used in evaluations such as pollen contamination, inbreeding and fertilization (Mason, 2015). They are also difficult to standardize and share reproducibly between laboratories. SNPs are now widely used as they are readily abundant, and they have a low frequency of genotyping errors. They are reproducible among laboratories, and they are widely used in the selection of desirable parents so they can be easily integrated into the deployment and evaluation of seed orchards. For long term breeding programmes,

SNPs can be discovered and implemented into a genotyping platform to assist with every stage of the program from selection to evaluation (Rasheed et al. 2017).

1.5 Seed Orchards

Seed orchards are key to delivering genetic gain in a cheap but effective manner. Genetic gain can be simplified as the heritability of a trait and the selection differential.

The selection differential in seed orchards is defined as the difference between the average genetic value of a selected individual or group and the average genetic value of the total base population (Dwivedi et al. 2015; Kang et al. 2001b). Seedling seed orchards are generally established from provenance trials or trials where progeny testing is limited (Giertych 1975). In this case, the selection differential is lower leading to reduced genetic gain. Clonal seed orchard populations are progeny tested leading to a larger selection differential, increasing genetic gain. Trait heritability can be defined phenotypic or genotypic methods. (Zas and Sampedro 2015). Traditionally heritability is estimated based on phenotyping trials however genotyping technology has increased the accuracy of heritability estimations. Genetic relatedness matrices can be constructed using molecular markers to reflect allele sharing between individuals in a population (Olsson et al. 2001). In seed orchards heritability can be managed through parental population size, with larger populations reducing selection intensity, which reduces genetic gain (Olsson et al. 2001). Using smaller populations would increase heritability but inbreeding is more likely to occur. Population size and layout is key to maximizing genetic gain while balancing breeding pitfalls.

The base population for the seed orchard can consist of individuals from a founder population or recruitment populations (Gullberg 1985). Individuals are chosen based on their phenotypic attributes and are put through progeny testing where the best individuals can be chosen for the breeding population. For seed orchards rouging, the removal of poor

performing clones, may occur when new progeny information becomes available (Riley 1964). New progeny can be added to the seed orchard as an infusion population to increase the total genetic gain (White 1987). The infusion population can replace underperforming individuals that have suffered from stress or have low reproductive success.

There is no set amount of parents to be included in a seed orchard as population size varies on a case by case basis depending on site and species. Large breeding populations reduce the potential for genetic gain of the offspring due to factors like admixture and the decrease in selection differential (Lindgren and Prescher 2005). Smaller populations without carefully chosen parental stock increase the potential for gain although the offspring would have a reduced genetic base leading to potential inbreeding depression causing problems such as reduced resistance to environmental changes over their lifespan. Seed orchards can vary widely in design, typically ranging between 5-100 parents (McKeand et al. 2003; Kang et al. 2001a). Theoretically seed orchards could also be made up of one parent that performed best in the progeny testing crossing with itself multiple times but this causes numerous negative effects due to inbreeding (Williams 1996). Calculations developed by Lindgren and Prescher (2005) tried to elucidate the optimal number of parents for a seed orchard but the range was too broad. Many seed orchards now use 10-30 parents with good returns in genetic gain and minimal inbreeding, yet effective standards have not yet been developed and many seed orchards are developed on a case by case basis depending on the number of parents and site availability (Schmidt 1993; Funda 2012). It should also be noted that parents do not equally contribute to the offspring and differences in reproductive success can lead to a skew in the representative population (Galeano et al. 2021).

Seed orchard design has been constantly improved to maximise genetic gain while reducing inbreeding and external contamination. Once the selection of optimal genotypes

has been completed through trials, the field deployment of seed orchards allows for the cheap and effective production of new seeds (Schmidt 1993). Seed orchard designs can be classed into four groups of design strategies. Group 1 does not involve any specific design strategy and parents are planted without any particular design. This method is nicknamed the “free love” orchard (van Buijtenen 1971). Early seed orchard design focused on complete randomisation design for maximising mating amongst unrelated clones. Another class, group 2, of seed orchard design was then developed, using blocking design. Blocking designs rely on grouping parents and arranging them in differing patterns with the goal of splitting parents of the same genotype. Common block designs are completely randomised block design, cyclical designs or incomplete balanced block design. Designs are generally picked to suit a site. Then a third layout strategy of seed orchard design (Group 3) led to a vast increase in design options for seed producers by using computational power to assess layout. This third generation of design strategies allows for the solving of problems unique to the designer's situation, allowing for the reduction of mating amongst relatives, maximising genetic gain and reducing inbreeding. Strategies like replicated randomised staggered clonal rows (R^2CR) and minimum inbreeding (MI) are associated with this generation of designs (El-Kassaby et al. 2014). As seed orchard research progresses there is an increase in design options, solving situation specific issues and increasing genetic gain. With the advent of well-priced genetic sequencing technologies, the fourth group of seed orchard design implementing marker-assisted seed orchard optimization is evolving.

1.6 Populations studied

There were two main populations used in this thesis. The population representing the native range of Sitka spruce was the IUFRO (International Union of Forest Research Organisation) collection (O'Driscoll, 1978). The establishment of the IUFRO collection was carried out with two seed-collecting phases in 1968 and 1970. Seed was collected from

81 provenances, with twenty representative trees sampled per provenance and seed being bulked per provenance for distribution. Information regarding this collection is sparse, with missing information regarding how many trees each provenance represents. There were also some questions regarding whether some tree collected were actually White Spruce or hybrid Spruce. This is one pitfall of this study. The collection was distributed amongst eleven known countries. The collection sent to Ireland was planted in John F Kennedy Arboretum in Wexford (52.315°N, -6.941°W). It is unclear when they were planted as records have been lost. Planting was done with 80 of the 81 provenances with 30 trees being planted per provenance. Periodic thinning was applied but there were no records of when thinning was done. The population as it stands has 80 provenances with 9-15 genotypes per provenance.

The second population used in this study was the Irish Sitka Spruce Tree Improvement Program (ISSTIP). This population is made of improved Sitka spruce selected from Irish seed stands however exact locations of these seed stands are lost with all of them being felled (Lee et al., 2013; Thompson et al., 2005). Additionally the origins of the seed are not fully known. Records show Irish Sitka spruce stands have material originating from Washington and Haida Gwaii but some breeders claim that Oregon stock is used as well (O'Driscoll, 1977; Thompson et al., 2005). Progeny testing has been carried out but details are limited due to record keeping issues. Clones of evaluated plus trees are currently maintained in a gene bank for reference (Lipow et al., 2002). The gene bank is located in the National Botanic Gardens in Kilmacurragh (52.929°N, -6.148°W). Some of the progeny of the ISSTIP population were used to establish two seed orchards in Ireland. Records show which parents were used however there is no record of a design strategy for either seed orchard or a rouging strategy. These two populations are studied in this thesis

but there are shortcomings in this study due to the lack of detailed records for these populations.

1.7 Workflow and aims

This thesis aimed to understand the evolution, biogeography, genetics, and adaptation of Sitka spruce natural populations and breeding populations. The populations used in this thesis were the IUFRO population, which represents the natural range of Sitka spruce, and the Irish Sitka Spruce Tree Improvement Program (ISSTIP) population which represents the Irish breeding range. The aim was to genotype plants from across the IUFRO range using GBS and the ISSTIP population. The data from the IUFRO range was used to investigate the biogeography of Sitka spruce. Using Genome Wide Association Studies (GWAS) and Genotype Environment Analysis (GEA) these data were used to investigate potential adaptation. Comparisons were made between the ISSTIP and IUFRO population to assess the baseline diversity of the breeding population.

The specific aims of each paper/chapter were to:

- I. Assess the biogeography, diversity and evolutionary history of the natural range of Sitka spruce.
- II. Investigate the nature of adaptation in Sitka spruce and discover loci correlated with specific traits of interest.
- III. Compare and contrast the Irish breeding population with the natural range of Sitka spruce.
- IV. Implement genotyping data to study seed orchard dynamics.

Chapter 2- Genetic Diversity and Structure of a Diverse Population of *Picea sitchensis* Using Genotyping-by-Sequencing

This chapter has been published as Byrne, T.; Farrelly, N.; Kelleher, C.; Hodkinson, T.R.; Byrne, S.L.; Barth, S. Genetic Diversity and Structure of a Diverse Population of *Picea sitchensis* Using Genotyping-by-Sequencing. *Forests* 2022, 13, 1511. <https://doi.org/10.3390/f13091511>

Contributions by Tomás Byrne: Designing and carrying out experimental sampling, laboratory sample preparation, data analysis, writing, reviewing.

2.1 Introduction

Picea sitchensis (Bong.) Carrière (Sitka spruce) is one of the seven *Picea* species native to North America, with a native range from Alaska 69° N to coastal California 39° N, occupying coastal areas, islands in the Alexander Archipelago and river regions (Griffith, 1992). *Picea sitchensis* is a species of considerable commercial importance in Atlantic Europe, where it is well adapted to the mild maritime conditions of Britain, Ireland and the coastal region from France to Denmark. In its native range, commercial interests are limited to Alaska, British Columbia and Haida Gwaii. Sitka spruces native range overlaps with the native ranges of *Picea engelmannii* Parry ex. Engelm. (Engelmann spruce) and *Picea glauca* (Moench) Voss (White spruce). White spruce and Sitka spruce can hybridize to form *Picea x lutzii* Little, in areas such as South Alaska and North-western British Columbia (Hamilton & Aitken, 2013). However, Sitka spruce and Engelmann spruce hybrids do not commonly occur. Hybrids of White and Engelmann spruce can also occur in this area, but neither hybrid is of economic importance (Degner, 2015). The geographical niche occupied by Sitka spruce reflects its adaptation to moist coastal areas provided by mild maritime conditions with high humidity. It continues inland until these conditions change and other species dominate. Peripheral zones are occupied largely by White spruce. This creates a large latitudinal spread and a thin longitudinal niche of Sitka spruce along coastal areas. As a result of this, genetic isolation is likely due largely to geographical distance along the north–south range. However, the isolation of spruce on three large islands, namely Kodiak Island, Montague Island and Haida Gwaii (50, 36 and 55 km from the coast respectively), may also present a more significant barrier to gene flow than isolation by distance processes occurring along the continuous stretch of Sitka spruce from Alaska to California (Gapare et al., 2005). The large population sizes of conifer species and their

efficient gene flow cause low frequencies of rare alleles, nucleotide diversity and genetic differentiation (Buschiazzo et al., 2012; Florin, 1964).

Phylogenetic and historical biogeographical relationships between the *Picea* species have been inferred from DNA analyses, crossing experiments and fossil evidence (Lockwood et al., 2013; Wright, 1955). The fossil record suggests that *Picea* diverged within Pinaceae 135 million years ago in the peripherals of the Pacific basin (Florin, 1964). *Picea koyamae* Shiras is thought to be an ancestor species of Sitka spruce that migrated from Siberia to Alaska; however, there is no current fossil evidence to support this (OECD, 2006). It has been inferred that many conifer populations were fragmented during the Pleistocene era due to glaciation, isolating populations in refugia (Critchfield, 1984). Due to an incomplete fossil record, this theory cannot be fully supported by fossils and requires molecular phylogenetic analyses. These inferences are leading to a better understanding of conifer evolution and the discovery of traits of interest for breeders.

Molecular markers, such as Single Nucleotide Polymorphisms (SNPs), have allowed for improved breeding, trait discovery, population genetics and phylogenetics (Galeano et al., 2021; Korecky et al., 2021; Rasheed et al., 2017). Genotyping-by-sequencing (GBS) allows for the discovery of genome-wide SNPs. For plant breeding, these can be specifically used to create linkage maps, discover traits through genome wide association studies and improve the selection of parental crosses through genomic selection (Rasheed et al., 2017). In conservation genetics and molecular ecology, SNPs can be used to investigate population diversity, structure and ancestry (Galeano et al., 2021; Pereira-Dias et al., 2019).

The population used in this study was the IUFRO (International Union of Forest Research Organisation) collection, which represents a diverse distribution of the native

range of Sitka spruce (O'Driscoll, 1978). The establishment of the IUFRO collection was carried out with two seed-collecting phases in 1968 and 1970. Seed was collected from 81 provenances, with twenty representative trees sampled per provenance and seed being bulked per provenance for distribution. The provenances selected were at least 50–80 km from each other and represented commercial or scientific areas of interest (O'Driscoll, 1977). The provenances occupied 19 geographic regions, which were determined at collection. Initially, this collection was used to test the performance of provenances in non-native regions for forestry. The collection was planted across eleven countries in its entirety, or subsets of provenances were used, to assess their establishment and performance. The initial experiments found that provenances from Washington were best suited to Ireland, Haida Gwaii was most suited to the UK, and Oregon was the best suited to France (O'Driscoll, 1972, 1978). Some of these initial field collections remain in multiple countries and have been used for breeding and scientific purposes. For example, in the United States, the genetic variability of enzymes within ten provenances from this collection was characterized (Syde, 1990), and in Canada, it has been used to study the resistance of Sitka spruce to white pine weevil (*Pissodes strobe*) (King et al., 2004). The IUFRO collection in the John F Kennedy Arboretum in Wexford, Ireland presents a large collection located in the same area, allowing for novel studies to be completed.

To preserve, build on and utilize the IUFRO collection, as many individuals of the population as possible were genotyped. Genotyping this population serves numerous functions. Firstly, this is a widely distributed collection grown in several countries and captures the diversity of the native population, along with areas of interest to breeders. Secondly, a collection this large will allow us to investigate fundamental questions about its ancestry, hybridization, clinal adaptation and post-glacial spread. Finally, with the effects of climate change, it is uncertain that our current breeding stock is resilient against

some of the foreseeable climatic changes, such as drought, and unforeseeable changes, such as the invasion of new insect pest species. The genotyping of this material will allow for the acceleration of breeding and genetic gain, hopefully combating productivity losses associated with climate change. This will act as a DNA bank that can be used to investigate traits and screen for resistant genotypes for breeding. This study aimed to build a key resource for Sitka spruce and North American Spruce research and will expand on existing resources that are available to research. This population was described by using the genotyping data to highlight areas of interest to scientists and breeders.

2.2 Materials and Methods

2.2.1 Sample Populations

The IUFRO population was planted in 1975 and 1978 at John F. Kennedy Arboretum in county Wexford, Ireland (52.315°N, -6.941°W) (Parra-Salazar et al., 2022). This population consisted of seeds collected from 80 provenances of the original 81 available in the IUFRO population (Figure 2.1) (Table A5.1), with 30 trees planted per provenance in a line. Each line was thinned sporadically, resulting in 1400 trees by 2021. The population consisted of 19 geographic locations, as defined by the original IUFRO collection, ranging from Alaska (152.53 W, 58.0 N) to California (121.45 W, 48.07 N), with an elevation range of 0 to 671 m (Figure 1) (Table A5.1).

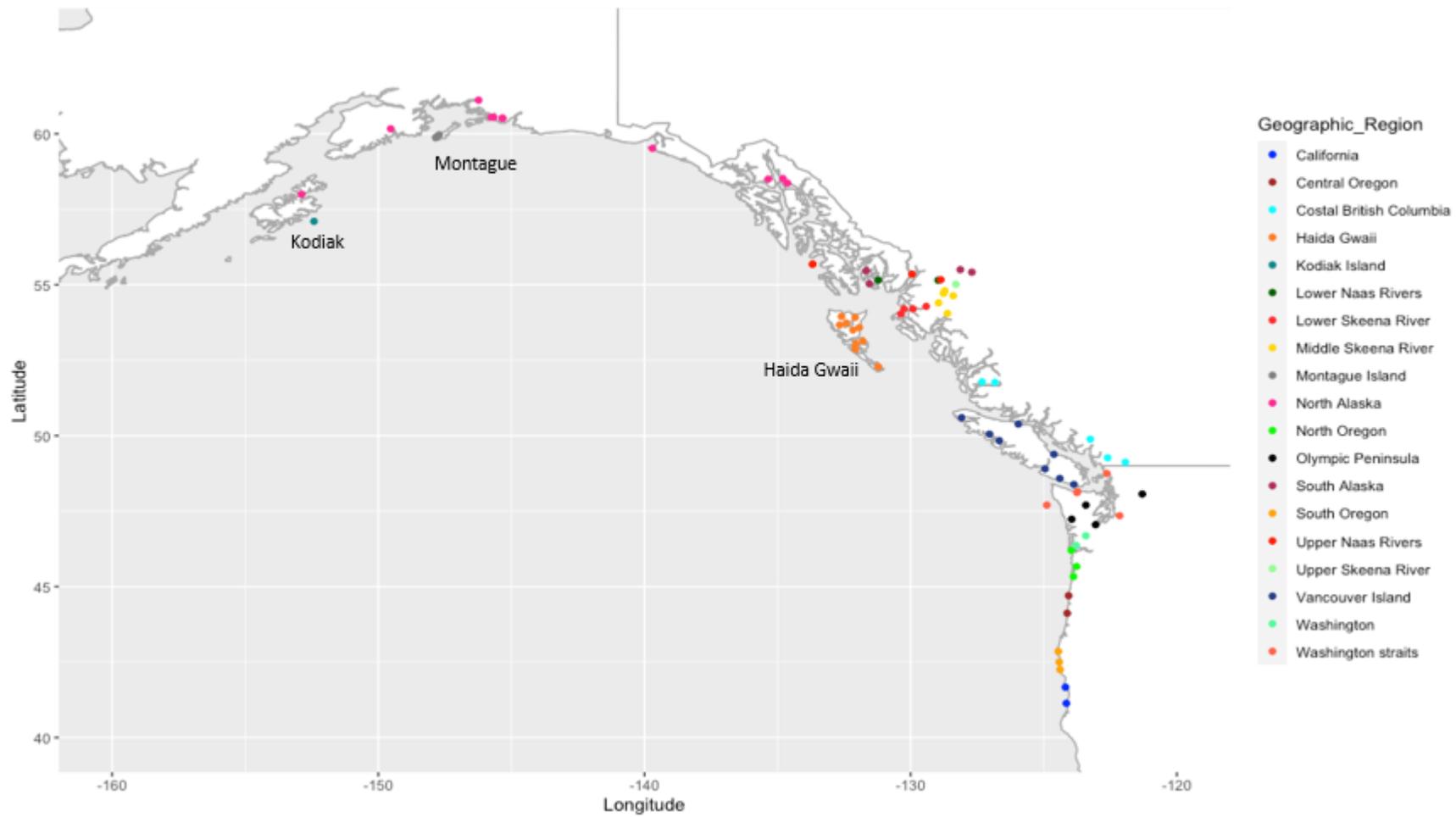


Figure 2.1. Origins of Sitka spruce in the IUFRO population. The main population in the database is the IUFRO population, which consists of trees planted from seed collected in locations marked

2.2.2 DNA Extraction and Sequencing

Vascular cambium was sampled for DNA extractions in May 2021, using a cork borer. The samples were then frozen at $-20\text{ }^{\circ}\text{C}$, which allowed the cambium layer to split from the bark layer. The cambial layer was removed and freeze-dried prior to DNA extraction. 30 mg of the freeze-dried cambial sample was transferred to deep-well plates, and three titanium beads were added to each well prior to milling. Cambium samples were milled for 1 h at 30 oscillations per minute on a Retsch MM400 mill (long milling time was used due to the woody nature of the samples). The samples were extracted by using the Machery and Nagel NucleoMag Plant DNA kit (744400.1) modified for the Kingfisher flex extraction system. The incubation time was increased to three hours to aid in the breakdown of woody tissue. A second elution step was added at the end of the Kingfisher protocol that increased the yield of DNA. Samples were quantified after extraction by using the Quant-iT PicoGreen dsDNA Assay Kit (P7589) and a BioTek Synergy HT to assess if samples were at a quantity of $30 \pm 5\text{ ng}/\mu\text{L}$. Samples were sent to LGC Genomics (Berlin, Germany), where genotyping-by-sequencing (GBS) libraries were prepared by using restriction enzymes to reduce genome complexity (Elshire et al., 2011). The restriction enzymes used for library preparation were PstI-ApeK1, and the resulting libraries were sequenced on an Illumina platform in paired-end mode and read lengths of 150 bp to achieve a target depth of 3 M paired-end reads per sample.

2.2.4 Variant Calling

The FASTQ sequences generated from the GBS were aligned to the Q903_v1_1000 plus Sitka spruce genome (GCA_010110895.2) (Agency, 2020), using BWA-mem with default parameters (Li & Durbin, 2009). Samtools (v1.9) mpileup generated a mpileup file, which was processed with bcftools (v1.9) to call variants (Li et al., 2009). A minimum site

mapping quality of 20 was required to retain an SNP. The VCF files were summarized and visualized by using *vcfR* (v1.12.0) and *vcftools* (v0.1.16) (Danecek et al., 2011) that allowed for the filtering of variants. We removed indels and only retained biallelic SNPs. The minimum genotype quality (GQ) was set to 20, and the minimum read depth was set to 5, as the majority of variants were above this threshold. The minimum allele frequency was set to 0.05. Sites with greater than 30% of missing data points were filtered out, leaving 108,606 SNPs (`—remove-indels —min-alleles 2 —max-alleles 2 —minDP 5 —maf 0.05 —max-maf 0.975 —minGQ 20 —max-missing 0.7`). SNPs were thinned so that no two SNPs would be within 10 bp of each other, using *vcftools* (v0.1.16). An investigation into the Loss of Heterozygosity (LoH_e) and Gain of Heterozygosity (GoH_e) was completed by using *vcftools* (v0.1.16)(`-hwe -het`) and *dplyr* (v1.0.8) to access filtering on HWE.

2.3.5 Genetic Diversity Statistics

The data analysis was completed by using R version 4.0.2, unless specified otherwise. The VCF file of 36567 variants was converted into a *genind* object by using the *adegenet* (v2.15) package (Jombart, 2008; Jombart & Ahmed, 2011). Observed heterozygosity (H_o), expected heterozygosity (H_e), gene diversity (H_t and D_{st}), allelic richness (A_R), fixation index (F_{st}) and population differentiation statistics (D_{est} and G_{st}) were calculated across the entire population, using the *hierfstat* (v0.5-10) function `basic.stats` (de Meeus & Goudet, 2007), and summarized per population, using *dplyr* (v1.0.8).

2.2.6 Population Structure

Investigation into the prior assignments of the geographic regions for discernible population structure was completed. An Analysis of Molecular Variance (AMOVA) was used to compare the given geographic regions of the IUFRO population and run by using *poppr* (v2.9.3) (Kamvar et al., 2014) on a *geneclone* object created from the VCF file, using

adegenet (Jombart, 2008; Jombart & Ahmed, 2011). The AMOVA was run with the geographic region being hierarchical to the provenance. A principal component analysis (PCA) was run on the dataset, using adegenet and ggplot2 (v3.3.5), retaining 1400 principal components. A discriminate analysis of principal components (DAPC) retaining 6 discriminate analyses was applied to analyse population structure, using supervised clustering with regions as prior (Kamvar et al., 2014; Miller et al., 2020). Analyses of the population structure while using undefined prior regional assignments were also conducted to investigate the population structure. Optimal genetic clusters (K) were analysed by using snapclust in adegenet, using the Bayesian information criterion (BIC) (Jombart, 2008). Values for each BIC model were compared, and the optimal K was determined by using the elbow method.

2.2.7 Phylogenetic Analysis

The relationships between the geographic regions were investigated to show the genetic distance between regions and also highlight any population structure. The IUFRO population was clustered by geographic region and transformed into a genind object by adegenet and dplyr (v1.0.8) (Jombart, 2008; Jombart & Ahmed, 2011). The aboot function in poppr was used to create an unweighted pair group method with arithmetic mean (UPGMA) clustering tree with 100 bootstrap replicates (Highton, 1993; Kamvar et al., 2014).

2.2.8 Isolation by Distance

The geographic distance amongst the populations was compared to their genetic distance to investigate gene flow and diversity. The pairwise Weir and Cockerham F_{st} were calculated for each pair of provenances, using the hierfstat package (de Meeus & Goudet, 2007; Weir & Cockerham, 1984). Geographic distances between provenances were

measured by using the package *geodist* (v0.0.7). The geographic distance (km) was plotted against $F_{st}/1-F_{st}$, and outliers were investigated.

2.2.9 Admixture Analysis

Admixture (v1.3) was used to investigate the optimal number of ancestor populations (Alexander & Lange, 2011). *Admixture* was run between $K = 2$ and $K = 50$, with 10 reps per K , showing an optimal K of 11 determined by the elbow method. The resulting Q matrix for $K = 11$ was plotted per geographic region, as defined in Figure 1, using *ggplot2*. *Admixture* on a provenance level was plotted on a geographic map, using *scatterpie* (v0.1.7), *ggplot* and *rnaturalearth* (v0.1.0) (South, 2022; Wickham, 2016; Yu, 2017).

2.3 Results and Discussion

2.3.1 Genotyping a Large Sitka Spruce Population, Covering Its Native Geographic Range

The continuous development of genotyping technologies, such as GBS, has resulted in a dramatic increase in sequencing density and reduction in cost, leading to the adoption of SNP genotyping across many forestry species (Rasheed et al., 2017). However, this uptake in SNP genotyping has produced a wealth of data which cannot be fully taken advantage of in most forestry tree species due to the absence of high-quality reference genomes and/or appropriate tools designed and benchmarked for large-genome species (He et al., 2014; Wang et al., 2020). The availability of a draft Sitka spruce genome, while being a fragmented assembly, allows for refined processing of the GBS data, resulting in high mapping rates. Encouragingly, the average mapping rate to the draft assembly was 84.9% across the 1184 samples (seven samples were removed due to poor mapping rates). In total, 81.9% of the reads were properly paired.

Putative SNPs were identified in the population, and standard filtering on read depth, minor allele frequency and genotype quality resulted in a final database of 108,608 variants (O'Leary et al., 2019). The variant filtration applied to the database was aimed at finding a balance between false negatives and false positives. The filtering of the SNP sets is dependent on the task with, for example, association studies requiring hard filtration of the dataset (Veeckman et al., 2019). Overly harsh filtering to reduce false positives would bias sampling toward frequent alleles impacting our ability to capture diversity, and population structure. This final dataset was formed after thinning to ensure that no two SNPs were within 10 bp of each other; this was performed to remove SNPs in strong Linkage Disequilibrium (LD), as conventional LD filtering is challenging in fragmented assemblies (Pavy et al., 2012; Qu et al., 2020). Deviations in Hardy–Weinberg equilibrium (HWE) were due to G_oH_e (Figure A5.1), indicating that deviations were due to sequencing and alignment errors rather than natural processes (Chen et al., 2017). HWE was filtered to $p > 0.001$, resulting in 36,567 variants that were used in a further analysis. The sequence data were submitted to NCBI (BioProject PRJNA852515). The resulting SNP database in this study includes 36,567 variants over 1177 genotypes and captures most of the native range of Sitka spruce, but future improvements in the genome assembly and bioinformatics tools can allow for the reanalysis of the GBS dataset (Pavan et al., 2020). This created a key genetic database for the IUFRO population grown in Ireland. For breeders, this database can provide a baseline of genetic diversity to compare against breeding stock and seed orchards. Most notably SNP sets have been used for evaluating the parental contributions and contamination in seed orchards, allowing for the better design and selection of parents (Galeano et al., 2021). For researchers, this database can allow for the investigation of traits of interest, using methods such as Environmental Association Analysis (EAA) (Blanco-Pastor et al., 2021; De La Torre et al., 2019). For example,

responses to climate change have been investigated by using EAA in both *Lolium perenne* and *Pinus taeda* (Blanco-Pastor et al., 2021; De La Torre et al., 2019), utilizing SNP sets.

2.3.2 Geographic and Genetic Diversity

The IUFRO collection that was used in this study primarily captures a representative sample of the entire native range of Sitka spruce, with a large diversity of habitats (OECD, 2006). There are large clusters of populations sampled from around the Vancouver region, Coastal British Columbia, Alaska and Haida Gwaii (Figure 2.1). These areas are more populous, with larger forestry industries, leading to more seed being included in the original collection. The more isolated populations, such as those of Montague Island, Afognack Island, Kodiak Island, South-eastern Alaska and Northern California, are not as well represented in this database. The high geographic diversity and isolation of these adjacent populations does not, however, result in high genetic differentiation due to the high amounts of gene flow between populations. The genetic diversity of the geographic regions measured by H_e ranged from 0.17 to 0.24, with an overall H_e of 0.21. In most cases, the H_o was greater than the H_e , but some regions had the same H_o and H_e . Heterozygosity is low compared to a study using SSRs, but it is similar to what is found in a study using SNPs, thus highlighting marker differences (A'Hara & Cottrell, 2004; Hamilton et al., 2013). Additionally, no private alleles were discovered within the geographic regions, and the overall allelic richness (A_R) was 1.198 (Table 2.1), further suggesting high amounts of gene flow among the provenances and geographic regions.

Table 2.1. Summary of the genetic diversity and differentiation of the IUFRO population.

	Definition	Overall
H_o	Observed Heterozygosity	0.21
H_s	Within Population Gene Diversity	0.198
H_t	Overall Gene Diversity	0.204
D_{st}	Gene Diversity among samples	0.006
F_{st}	Fixation index	0.029
F_{is}	Inbreeding Coefficient	0
$Dest$ *	Population Differentiation	0.0078
G_{st} **	Population Differentiation	0.0284

Notes: * Measures based off Jost (2008), ** Measures based off Hamrick and Godt (1989).

It is in the peripheral populations where genetic differentiation is seen, notably Kodiak Island, which is the most isolated region. However, there is effective gene flow across thousands of kilometres in non-isolated populations, similar to what is seen in other conifers (Holliday et al., 2012). The efficacy of this gene flow is reduced over larger distances, but physical barriers such as the ocean result in the larger genetic differentiation of those populations. Our results for genetic differentiation amongst provenances and regions are in accordance with those of another study (Gapare et al., 2005) with a reported G_{st} of 0.03 across eight sampling sites, from Alaska to California. H_o and H_e differed between studies, but it is difficult to compare with these studies due to differences in sample size, sampling range, marker type and marker number; however, the large sample size and distribution combined with detailed genotype data used in this study allows for a more complex analysis.

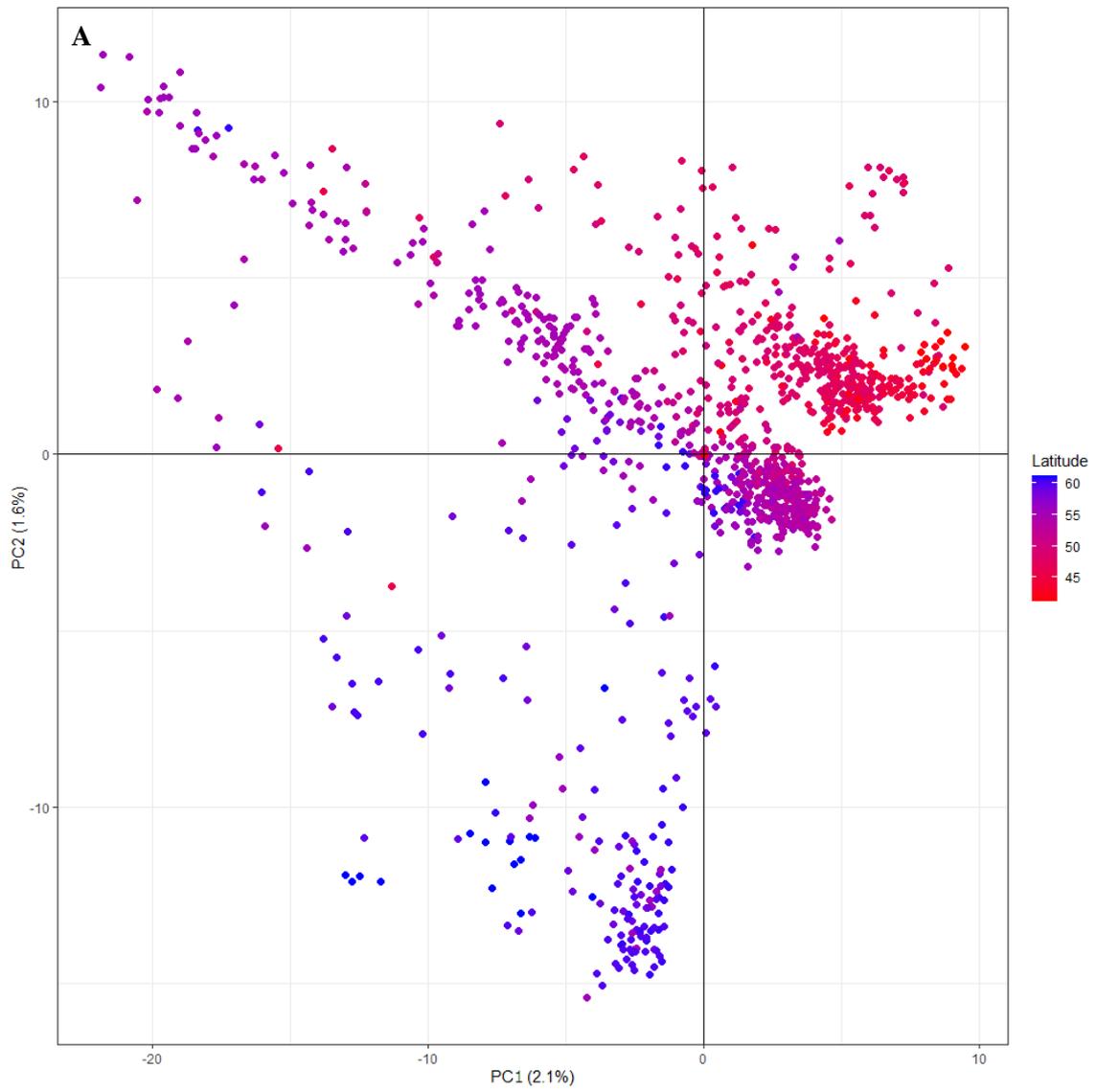
2.3.3 Genetic Structure

The AMOVA for the entire Sitka spruce population indicated that the majority of genetic variation occurred between samples with between-region genetic variation only accounting for 1.98% ($p > 0.01$) of the total genetic variation. The variation between provenances was 7.49% ($p > 0.01$), with within-provenance variation being 3.77%, yet not significant ($p > 0.31$). The low overall F_{st} value (0.023) shows low levels of differentiation within the entire population. These F_{st} values are similar to what were seen in previous studies and suggest an excess of heterozygosity within the populations (Holliday et al., 2010). Some regions have higher levels of differentiation (Table 2.2), notably Central Oregon, with an F_{st} of 0.16. This lack of distinct population structure is somewhat typical in conifer species with large distributions across environmental gradients when gene flow is not prohibited.

The PCA analysis revealed a high degree of uncertainty in population structure (Figure 2.2A); however, some general differentiation is apparent, notably in regard to the Kodiak and Montague islands. The axes of the PCA account for 2.1% and 1.6% of the variation, further indicating the presence of high genetic mixing within and among populations of this species, as described in other studies. The indication of some structure and differentiation is seen in the DAPC (Figure 2.2B), which also shows differentiation of Kodiak and Montague islands, along with Haida Gwaii, suggesting an isolation effect on these islands. The same is not seen on the other large island in the population, Vancouver Island.

Table 2.2. Differentiation and diversity statistics of the geographic regions within the IUFRO population.

Geographic Regions	H_o	H_e	Fst	Allelic Richness	Number of individuals
California	0.18	0.18	0.117	1.180	28
Central Oregon	0.17	0.17	0.162	1.171	29
Coastal British Columbia	0.23	0.21	-0.021	1.209	69
Haida Gwaii	0.19	0.19	0.091	1.186	151
Kodiak Island	0.24	0.18	0.086	NA	9
Lower Naas Rivers	0.22	0.21	-0.028	1.21	29
Lower Skeena River	0.23	0.21	-0.032	1.211	50
Middle Skeena River	0.23	0.22	-0.055	1.216	75
Montague Island	0.19	0.18	0.108	1.182	38
North Alaska	0.22	0.21	-0.025	1.209	159
North Oregon	0.23	0.2	0.003	1.204	33
Olympic Peninsula	0.2	0.19	0.072	1.19	79
South Alaska	0.2	0.19	0.050	1.194	84
South Oregon	0.18	0.18	0.128	1.178	38
Upper Naas Rivers	0.23	0.23	-0.117	1.228	55
Upper Skeena River	0.24	0.22	-0.079	1.22	22
Vancouver Island	0.21	0.2	0.028	1.199	119
Washington	0.2	0.19	0.074	1.189	34
Washington Straits	0.22	0.2	0.010	1.202	76



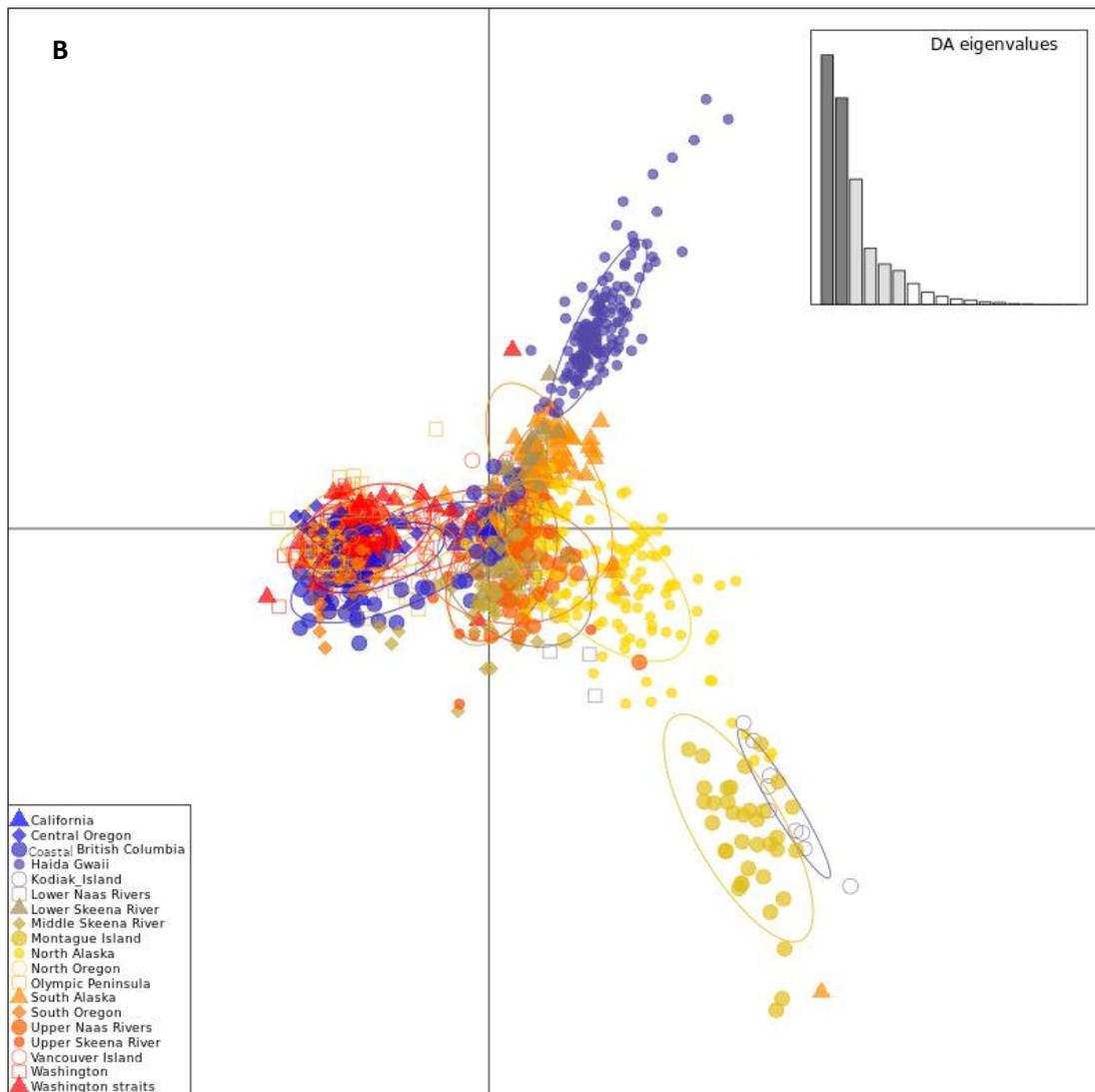


Figure 2.2. Genetic structure of the Sitka spruce IUFRO population. All 19 distinct geographic regions are represented here. **(A)** Principal component analysis (PCA) performed with adegenet, using 1400 principal components. **(B)** Discriminate analysis of principal components (DAPC) supervised clustering (with geographic regions as prior) performed with adegenet, using 20 principal components and 6 discriminate analyses.

The BIC supports the clustering seen in the DAPC, with estimates of K at 3/4 (Figure A5.2). The BIC estimation of four optimal clusters does not take into account prior regional assignment, and the DAPC and PCA alone do not lead to any strong conclusions

on population assignment, but taken together, they show a core population of Sitka spruce on the North American mainland, with the separation of populations isolated on islands. This island effect is typical across species due to limitations on gene flow. Spruce seeds have a mean wind dispersal of 345 m and typically conifer pollen travels 4–6 km, which isolates Afognak, Haida Gwaii and Kodiak Island (Figure 2.1) however, some pollen may be introduced by strong winds or by human and animal interference, which can also generate seed-mediated gene flow (Chen et al., 2018; Ebrahimi et al., 2018; Korecky et al., 2021; O'Connell et al., 2007).

The overall isolation effect is not just one of distance (Figure 2.3A) but largely due to isolation based on geographical barriers, namely the open sea separating islands, and prevailing wind which is from the Northwest (Figure 2.3B). The correlation coefficient of isolation by distance was 0.46 and 0.41 without these outlier islands of Kodiak and Montague. Phylogenetic structure clustering shows some distinct clustering groups with high bootstrap confidence support (Figure 2.4). Again, the Kodiak and Montague islands are outliers, likely due to their geographic distance from the mainland. The regions surrounding Naas and Skeena Rivers cluster away from the main population. The AMOVA was rerun, taking into account the structure as signified by the PCA, BIC, DAPC and UPGMA, but this did not change the outcome.

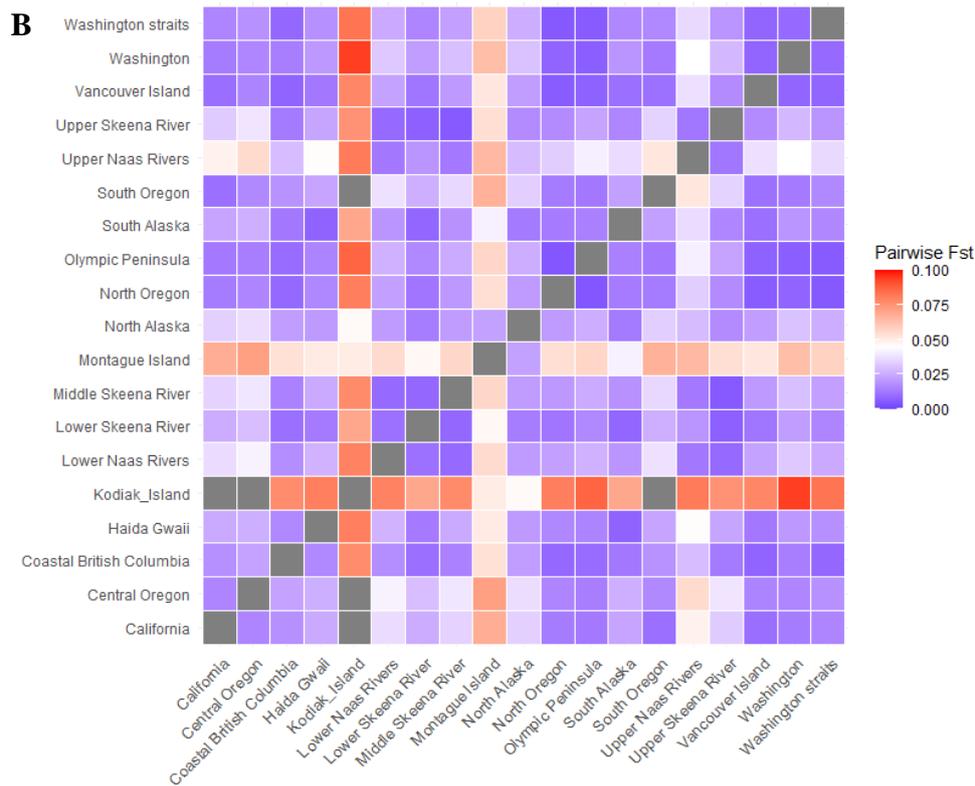
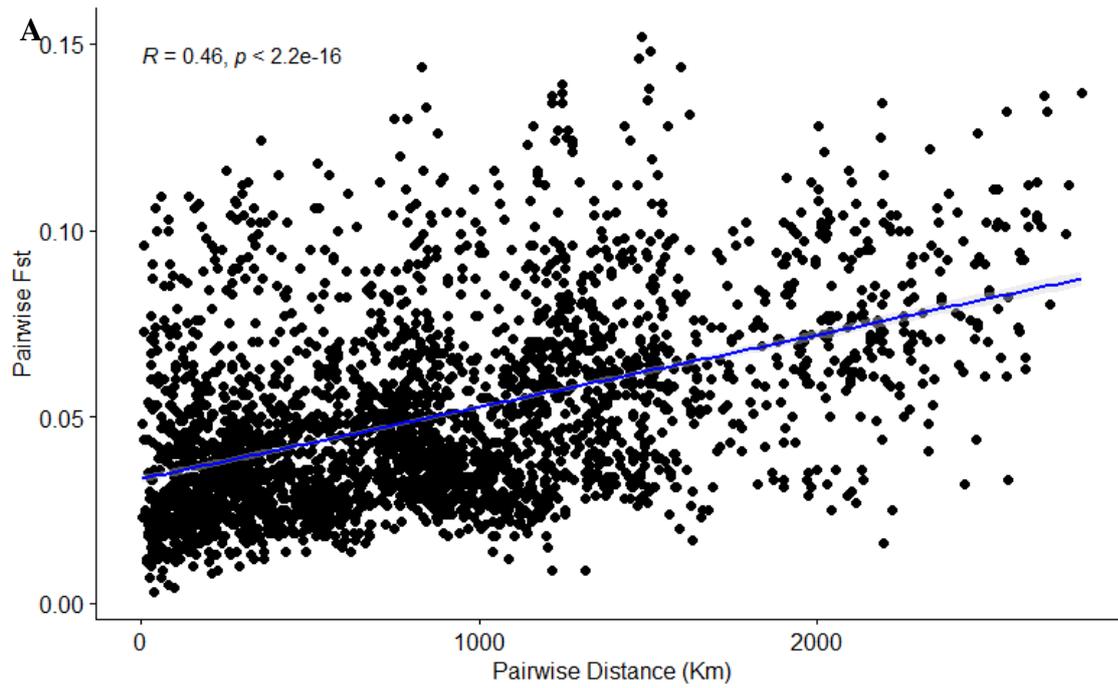


Figure 2.3. Isolation by distance of the Sitka spruce population. (A) Pairwise F_{st} values for each provenance have been plotted against geographic distance (km). (B) Pairwise F_{st} for each geographic region.

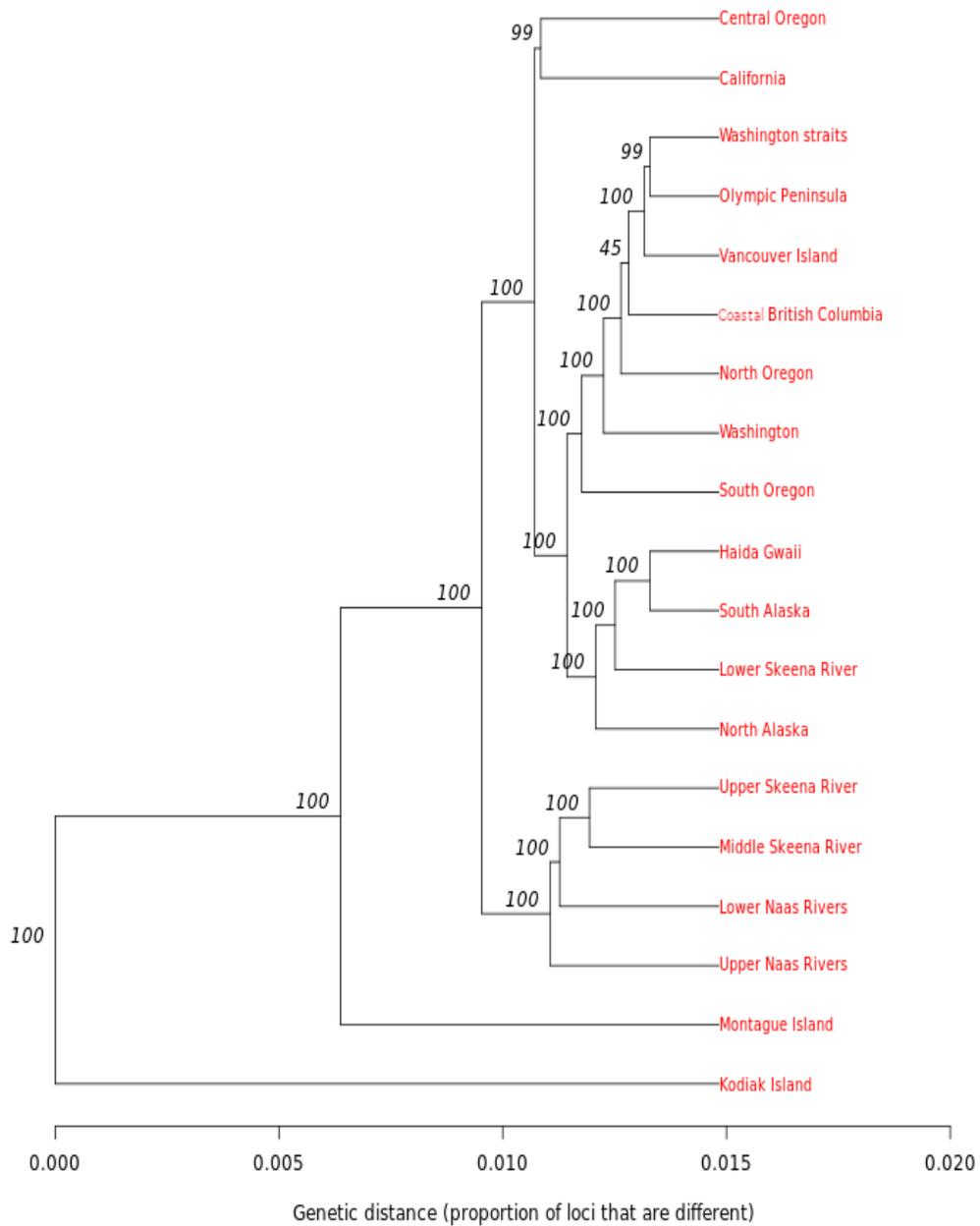


Figure 2.4. Genetic clustering of geographic regions. Tree was constructed by using pairwise genetic distance between loci, using the UPGMA method, which employed Nei's genetic distance. Numbers above branches represent bootstrap values (100 replicates).

On the mainland, genetic clustering indicates segregation of the genotypes originating from the Skeena and Naas rivers (Figure 2.4). This area is a known hybrid zone with White spruce, so the genetic influence is likely to be seen between the two closely related species (Degner, 2015; Hamilton & Aitken, 2013). Conifers have a slow rate of speciation, so the genetic differences between White spruce and Sitka spruce may be slight in many areas of this population, especially in the hybrid zone (Prunier et al., 2016). Identifying “pure” White spruce and “pure” Sitka spruce markers in our dataset would allow for these genetic differences to be fully quantified but the effect of gene flow within the entire population and the close relationship between the two species means the recognition of “pure” Sitka spruce or White spruce is not a realistic concept.

2.2.3 Genetic Admixture and Ancestry

The admixture model confirms the population structure and enlightens some of the questions about the evolutionary history of the species (Figure 2.5). This again confirms the island effects on Kodiak and Montague Islands, with no clear admixture events at all in the Kodiak region. Kodiak Island is primarily composed of the K7 ancestral population, indicating there has been little historical mixing with other populations. Montague Island is also primarily composed of a single ancestral population, K11, yet mixing has occurred with K2 and K3. This raises questions about when the isolation event occurred (Figures 2.5 and 2.6). The geological history of the Kodiak archipelago shows no recent land bridges between it and the mainland (Farris, 2020). The geological history of the area and the history of spruce evolution leads to the conclusion that the Pleistocene glaciation created a refugium on Kodiak Island which in turn recolonized the mainland (Critchfield, 1984). During the Pleistocene glaciation, the Cordilleran ice sheet was at its largest 18,500 years ago (Menounos et al., 2017). The ice sheet covered the area of the entire IUFRO collection range apart from Oregon and Washington states, with some debate on whether refugia

existed on islands such as Kodiak or Haida Gwaii (Haro, 2017). Coastal glaciation retreat occurred earliest, around 18,000 years ago, allowing for the recolonization of the species from islands, as is apparent in the admixture model with the K7 and K9 ancestral populations (Figure 2.6). The southern retreat of the ice sheet was slower with regions such as Washington and Vancouver with glacial retreat occurring 15,000–16,000 years ago (Haro, 2017). This data aligns with the Cordilleran ice sheet retreat. However no fossil evidence has been found to back up this data. There are other theories of recolonization, including the spread of Sitka spruce from Nunataks, but fossil evidence would be needed to support this.

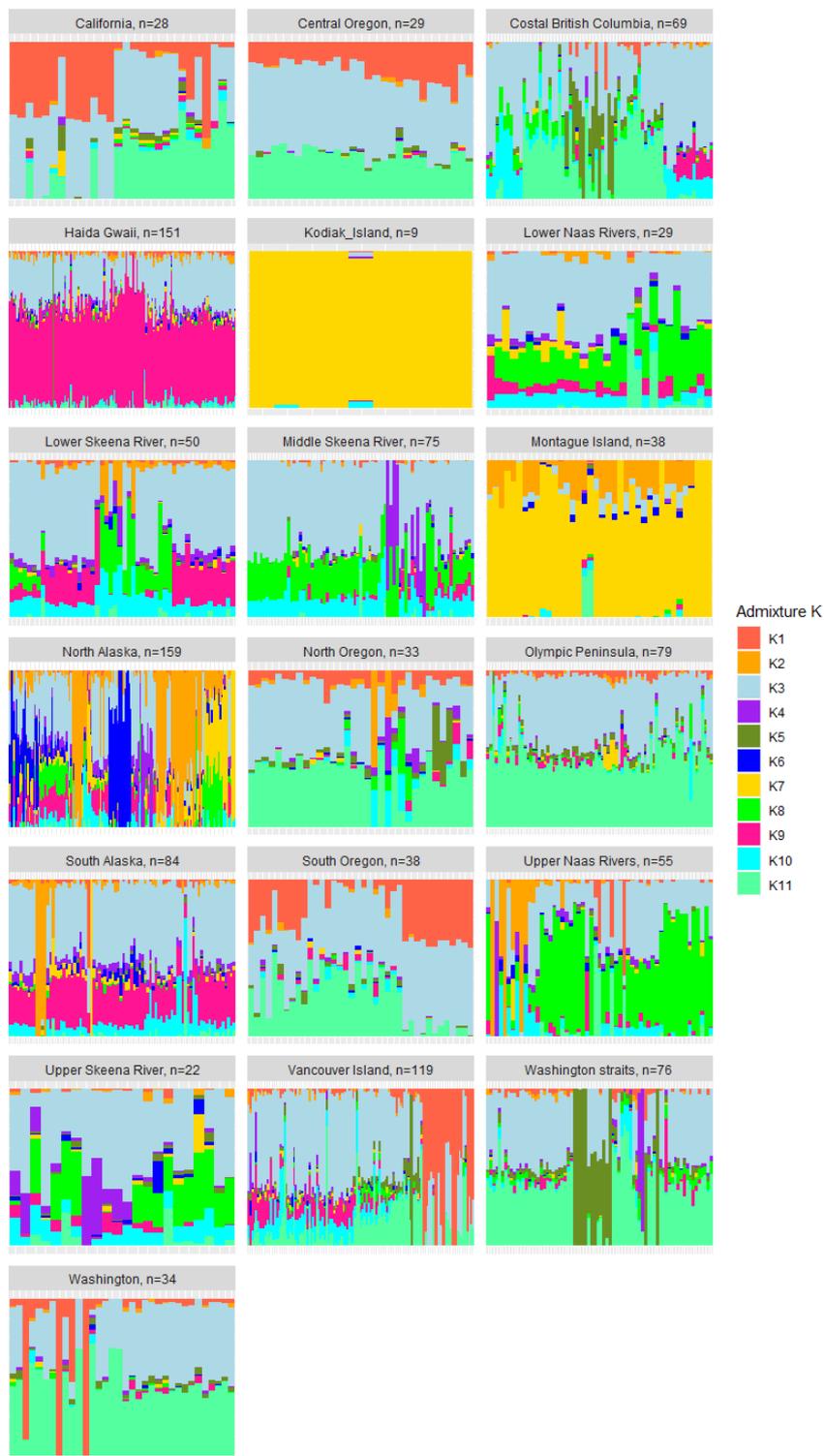


Figure 2.5. Genetic admixture of all 19 geographic regions of the IUFRO population for 11 ancestral populations. Optimal K (ancestral populations) for admixture is 11 based on 10 replicates of K 2-50, using a cross-validation method. Each geographic region is represented here, showing the ancestral makeup of the population.

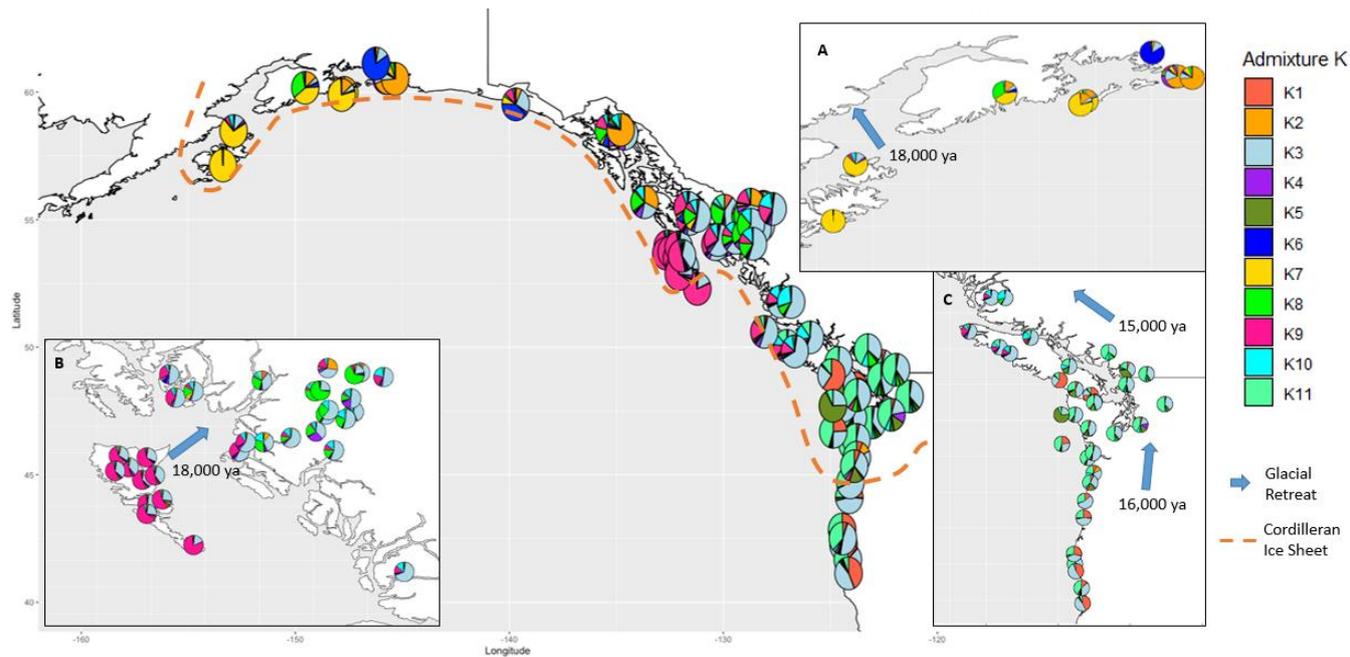


Figure 2.6. Geographic depiction of admixture of all 80 provenances of the IUFRO population for eleven ancestral populations. Optimal K (ancestral populations) for admixture is eleven based on 10 replicates of K 2-50. The map shows the largest extent of the Cordilleran ice sheet at the maximum extent of the Pleistocene glaciation 18,500 year ago (ya). Arrows represent key areas of frontal glacial retreat starting in coastal areas. **(A)** Represents the northern expanse of the range with admixture from Kodiak Island recolonizing the range. **(B)** Representation of Haida Gwaii recolonizing the mainland. **(C)** Southern expanses of the range showing admixture after the retreat of the Cordilleran ice sheet.

The K8 population occurs frequently in the hybrid zones, possibly due to hybridization events. This is also supported by the UPGMA tree, which indicates clustering of provenances located around the Skeena and Naas rivers. This area did not become fully clear of ice until approximately 12,000 years ago, suggesting that it could have been one of the last areas to be recolonized by both Sitka spruce and White spruce. The North Alaskan provenances have the most diverse genetic admixture; however, the region also has the

largest sample size ($n = 159$) (Figure 2.5). In the southern regions of the IUFRO collection, K3, K1 and K11 are most represented.

The population genetic structuring of the species has clearly been shaped by the Pleistocene glaciation and the retreat of the Cordilleran ice sheet, which has isolated areas such as Kodiak, but remixture of some of the refugia has occurred into the main population. This answers key questions about isolation events of Sitka spruce but raises additional questions about the evolutionary spread of ancestral populations and how they relate to East Asian spruce. A phylogenetic analysis of North American and East Asian species is required at a detailed level to piece together the puzzles of *Picea* ancestry. Unfortunately, such genetic resources are not all collected yet; however, here we have contributed an SNP dataset for Sitka spruce, a species which is at the forefront of the ancient land bridge between Asia and America. On the western reaches of this range, White spruce is well distributed and often hybridizes with Sitka. Future work should focus on comparing these closely related species and identifying patterns of speciation and adaptation between the two species. Carrying out similar work in the western range of White spruce may allow for the reclassification of some samples in our database as hybrids or White spruce.

2.4 Conclusions

The IUFRO collection has been a valuable resource for scientific research of Sitka spruce and for use within breeding programs and has been distributed to many organizations across different countries. This study built on the existing resource by genotyping 80 of the 81 provenances in the original collection. Genotype data was used to study population diversity and found a high degree of gene flow within the population on Mainland North America, with diverse clusters occurring on isolated islands. The IUFRO collection and the

genotyping data will allow for trait-association studies as the genomic resources available in Sitka spruce improve. This will be beneficial for breeders and enable further phylogenetic comparisons of spruce species of scientific interest.

Data Availability Statement: Raw sequence data were submitted to NCBI (BioProject PRJNA852515)

Chapter 3- Environmental and landscape genomics
of *Picea sitchensis* (Sitka spruce) reveals clusters of
loci involved with both resistance to wintering and
increases in height.

This chapter has been submitted to Ecology and Evolution

Contributions by Tomás Byrne: Designing and preparing out experimental data, data analysis, writing, reviewing.

3.1 Introduction

Low linkage disequilibrium (LD) is characteristic of conifer genomes and the independence of genes contributes to the adaptation patterns seen in conifers (Prunier et al., 2016). Angiosperms may undergo allopatric speciation by adaptation leading to reproductive isolation, however, conifers are less likely to do this due to low LD and large genome size (Ahuja & Neale, 2005; Pavy et al., 2012). For this reason, local adaptation to host environments is quite common in conifers with many alleles with small effects driving the adaptation (De La Torre et al., 2019; Hornoy et al., 2015). In *Picea sitchensis* (Bong.) Carr. (Sitka spruce) this has been demonstrated using phenotype data where bud set timing was adaptive to local climate types (Mimura & Aitken, 2010). In Loblolly pine, *Pinus taeda* L., climate adaptation at a local level is due to these subtle changes in the frequency of alleles of moderate/small effect (De La Torre et al., 2019) and in Lodgepole pine, *Pinus contortus* Douglas, climatic adaptations have been shown to be characterised by a slow rate of recombination (Lotterhos et al., 2018). Conifer adaptation is generally slow and characterised by low changes in frequency of many alleles, which is likely to affect the diversity of local adaptations found in these types of conifer studies. The long lifecycle of conifers results in low recombination cycles, contributing to slow adaptation.

Evergreen conifers have multiple mechanisms to survive colder harsh winter environments (Chang et al., 2021; Senser & Beck, 1984). Low temperatures can reduce the rate of enzymatic activity, disrupt metabolic processes and inhibit protein repair systems (Chang et al., 2021). To overcome this, some conifers have developed cold hardiness, which can stabilise and change membrane lipid composition which function to reduce the build-up of ice crystals in extracellular spaces (Chang et al., 2021; Senser & Beck, 1984). Cold hardiness is reported in some white spruce populations, a close relative of Sitka spruce (Sebastian-Azcona et al., 2018). Growth cessation during colder periods is an adaptation

linked to shorter diurnal ranges. This is seen in Sitka spruce with delayed bud set due to wintering suggesting growth cessation in buds (Mimura & Aitken, 2010).

Low water stress, heat stress and high solar radiation (kJ/m) all combine to create drought stress. Thus, drought stress is a complex environmental stressor which requires a suite of differing adaptations (Brodribb et al., 2014). Conifers exhibit two diverging pathways to adapt to low water stress, one involves raised levels of abscisic acid (ABA) to close stomata and another method involves leaf desiccation (Teskey et al., 2015). Transpiration cooling is important for heat stress however, and stomatal closure can increase the severity of heat stress (Teskey et al., 2015). It is known that reductions in total seasonal solar radiation correlate to a reduction in tree height, so generally it is not considered a stressor on its own unless it exacerbates drought stress through evapotranspiration (Strand et al., 2006).

Sitka spruce occupies a large geographic and climatic range along the Pacific Northwest of North America. This presumed large gradient in environments requires specific local adaptations across the population (Gapare et al., 2005). The IUFRO (International Union of Forest Research Organisations) population was established in 1968 and 1970 by Matthews and Fletcher, representing the natural range of Sitka spruce, collected at intervals of 50km and representing >80 provenances. (Syde, 1990). The collection includes samples over 3000 km in extent from Alaska to California. Elevation ranges from sea level to 671m, and an maximum distance from coast of 164km. Annual precipitation ranges from 4015 mm to 621 mm (Fick & Hijmans, 2017). Mean annual temperature ranges from -7.3°C to 17.4°C (Fick & Hijmans, 2017). Mean hours of sunlight range is 10.8hrs to 4.43hrs per day (Fick & Hijmans, 2017). This highlights the range of environments present in this data set making it key for the study of local adaptation in conifers. This diverse and large range of environments likely leads to adaptations at the

local level driven by shifts in allele frequencies, a common pattern in conifers (De La Torre et al., 2019). To discover the genetic basis of these adaptations, a number of methods have been developed. Genome wide association studies take trait data and genotype data and associated molecular markers with a trait (Casola, 2019; Chen et al., 2021; Hiraoka et al., 2018). This is useful for the discovery of the molecular function of traits or for the identification of molecular markers for the deployment in selection programs as has been done in *Picea abies* (L.) H. Karst. (Chen et al., 2021). Similar techniques can be used to associate genes with an environment using genotype-environment association. In conifers, trait association has been completed in Lodgepole pine and Loblolly pine by associating climatic variables with loci, elucidating mechanisms of adaptation (De La Torre et al., 2019; Lotterhos et al., 2018). The ability to associate genotype, phenotype and environment is important in these studies, as it indicates a pathway towards adaptation. While not previously undertaken in conifers, Canonical correlation analysis (CANCOR) has allowed for the discovery of loci that correlate with phenotypic growth due to environmental conditions in the forage grass *Lolium perenne* (Blanco-Pastor et al., 2021) (also outbreeding and wind pollinated). This analysis merges data to discover loci responsible for improved phenotypes for the environment, which is key for breeders.

In this chapter data from a diverse collection of Sitka spruce, available in an arboretum collection in Ireland, representing the entire natural range of the species was utilised. This collection consists of 80 provenances with roughly 15 trees per provenance. In total 1177 genotypes were available for the study. Genotype Environment Analysis (GEA), Genome Wide Association Study (GWAS) and Canonical Correlation (CANCOR) was completed to discover loci associated/correlated with specific traits and adaptations. The CANCOR analysis was used to see the phenotypic effects of environments on

genotypes. These loci can then be traced through populations, elucidating the driving forces of adaptation in Sitka spruce.

3.2 Materials and Methods

3.2.1 Genotype Data

Data had previously been generated by Genotype by Sequencing (GBS) to assess genetic diversity and structure of a diverse population of *Picea sitchensis* (Byrne et al., 2022) so it was utilized for this study. GBS data from 1177 genotypes of the IUFRO population, representing 80 sampling locations, were therefore obtained from NCBI BioProject PRJNA852515 (Byrne et al., 2022). These data were generated with 150bp paired-end GBS using Pst1-ApeK1 following (Elshire et al., 2011). Sequences were aligned to the Q903-v1-1000 plus Sitka spruce genome (GCA_010110895.2) (Gagalova et al., 2022) using BWA-mem with default parameters (Li & Durbin, 2009). Variant calling was completed using SAMtools v1.9 mpileup (Li et al., 2009). VCFtools was used for filtering (Danecek et al., 2021). Indels were removed, multi-allelic SNPs were removed, minimum genotype quality was filtered to 20, read depth was filtered to 5, max missing data was filtered to 0.7 and MAF was set to 0.05. This resulted in 66974 loci that were transformed into an allelic matrix for further study.

3.2.2 Phenotype Data

Phenotype data were collected in November 2022 from each of the 1177 genotypes in the IUFRO population at John F. Kennedy Arboretum in Wexford, Ireland (52.315°N, -6.941°W). Diameter at breast (DBH) height was taken at 1.3m from the base of the tree

using a steel DBH tape. Height was recorded for each tree using a vertex transponder (Hagloff V) to represent the distance from ground level to tree tip.

3.3.3 Environmental data

All data were downloaded in raster format and data for each location of the IUFRO population were extracted using the Point Sampling Tool in qGIS (Jurgiel, 2020; QGIS.org, 2022). Climatic data was similarly downloaded from worldclim.org (Table 3.1)(Fick & Hijmans, 2017). These data consisted of readings from 1970-2000 with a raster resolution of 1km². Annual snow cover data were obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS) on NASA’s Terra satellite for the years 2000-2020 (Hall et al., 2006). Soil data were obtained from the Unified North America Soil Map which is a combination of soil mapping from Canada and global soil maps (Liu et al., 2014). Topographical maps were obtained from United States Geological Survey (USGS) EROS (EROS, 2010). Forest fire risk maps were obtained from Genonode CSIRO (GFDRR, 2022).

Table 3.1. List of environmental traits used in GAPIT and CANCOR analysis

Landscape Features	Unit
Latitude	Decimal
Longitude	Decimal
Altitude	Meter
Forest fire risk	0-300 (300 is high risk)
Distance to coast	Kilometres
<u>Climate</u>	
Precipitation of coldest quarter	mm
Precipitation of warmest quarter	mm
Precipitation of driest quarter	mm
Precipitation of wettest quarter	mm
Precipitation seasonality (coefficient of variation)	mm
Precipitation of driest month	mm
Precipitation of wettest month	mm
Annual precipitation	mm
Mean temperature of coldest quarter	°C

Mean temperature of warmest quarter	°C
Mean temperature of driest quarter	°C
Mean temperature of wettest quarter	°C
Annual temperature range	°C
Minimum temperature	°C
Maximum temperature	°C
Temperature seasonality (amplitude between maximum and minimum)	°C
Isothermality (Measure of the day-to-night temperatures oscillate relative to annual oscillations)	°C
Mean diurnal range (length of daylight)	Hours
Annual mean temperature	°C
Mean wind speed	m/s
Min wind speed	m/s
Max wind speed	m/s
Mean solar radiation	kJ/m/day
Maximum solar radiation	kJ/m/day
Mean vapour pressure	kPa
Minimum vapour pressure	kPa
Max vapour pressure	kPa
Mean snow cover	%
Minimum snow cover	%
Maximum snow cover	%
Soil	
Dominant component	%
Soil depth	cm
Top soil sand	% weight
Top soil silt	% weight
Top soil clay	% weight
Top soil gravel	% volume
Top soil organic carbon	% weight
Top soil pH	scale
Top soil bulk density	g cm ⁻³
Top soil cation concentration	meq/100g
Subsoil sand	% weight
Subsoil silt	% weight
Subsoil clay	% weight
Subsoil gravel	% volume
Subsoil organic carbon	% weight
Subsoil pH	scale
Subsoil bulk density	g cm ⁻³
Subsoil cation concentration	meq/100g

3.2.4 GWAS

GAPIT3 was used to complete the GWAS to identify associated loci (Wang & Zhang, 2021). GAPIT3 model selection was used to select the appropriate model for performing the GWAS. Fixed and random model Circulating Probability Unification (FarmCPU) was chosen as it was the most effective for the size of the matrix used and is effective in the control of false positives (Wang & Zhang, 2021). The GAPIT3 FarmCPU model was run with an optimal 3 principal components to explain the variation. Regions 500bp from each loci were highlighted for future functional analysis.

3.2.5 CANCOR

The CANCOR test was based on scripts by Blanco-Pastor et al (2021). The CANCOR test was used to analyse both environmental and phenotypic data to find loci responsible for adaptations. Environmental and phenotypic data were scaled using scale function in R and then run against the genotypic data using the CCorA function in the R package vegan (Dixon, 2003). Then the significance of outlier loci was tested using a χ^2 test on Mahalanobis distances. Loci were considered an outlier if the FDR=0.1. The best represented loci in the first two canonical dimensions were selected. To simplify results, soil parameters were removed as their projection norms were not above 0.9 for any soil parameter, meaning few significant loci could be found. Loci with an alternative allele frequency that was relatively high ($|r|>0.25$) were kept. This resulted in lists of loci that were positively or negatively correlated with one of the phenotypic traits of environmental parameters tested. This algorithm is flexible and can take various differing phenotypic or environmental parameters

3.2.6 Minor Allele Frequencies

To investigate the Minor Allele Frequency (MAF) of the adaptive traits, VCFtools v0.1.16 was used to separate associated and non-associated loci and report on MAF statistics. VCFtools was also used to separate loci from traits of interest and report on their MAF. The provenances were split into Northern (north of 50th N latitude) and Southern (south of 50th N latitude). MAF for each trait of interest for was investigated for differences between the Southern and Northern populations.

3.2.7 Loci Tracing in Populations

In order to visualise the distribution of traits associated with height, loci were traced amongst provenances. Loci associated with tree height from the CANCOR analysis were selected and were searched for in the 80 geographic provenances using VCFtools v0.1.16. Each locus was labelled with a positive and negative effect on height based on the CANCOR analysis. Mean number of loci per provenance associated with a negative or positive change in height were counted using R dplyr. Number of positive or negative loci per provenance were geographically mapped using the R package naturalearth (South, 2022).

3.2.8 Functional Analysis

BLASTX was used to search for genes positioned within 500bp of loci (Altschul et al., 1990). Blast results were migrated to OmicsBox where gene ontology was analysed (Götz et al., 2008).

3.2.9 Co-association Analysis

The co-association analysis was used to investigate groups of loci associated with traits of interest. The results from the CANCOR analysis were used as the input matrix for the co-

association analysis. The input matrix consisted of correlations of loci and environmental traits. The Euclidian pairwise distances between the correlations of loci and environmental traits were calculated using the `dist` function in R and clustered using hierarchical clustering in the `stats` package in R. Co-associations were visualised using the `ComplexHeatmap` package in R (Gu et al., 2016).

3.3 Results

3.3.1 GWAS/GEA

The GAPIT3 model selection was performed to compare the efficacy of five models, namely General Linear Model (GLM), Mixed Linear Model (MLM), Compressed MLM (CMLM), Fixed and random model Circulating Probability Unification (FarmCPU), and Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway (BLINK) (Wang & Zhang, 2021). FarmCPU was selected based off model selection and for its false positive control (Liu et al., 2016). It was then implemented to test both environmental and phenotypic parameters.

From the GEA analysis, 573 associated loci passed the significance threshold of a false discovery rate of 0.1. 139 of these involved soil parameters, 393 involved climatic conditions and 41 of these were landscape specific. All loci and positions have been reported. The GWAS discovered eight associated loci for height and one for DBH. Of the loci involved with height, three were also associated loci for environmental parameters, namely minimum water vapour, forest fire risk and maximum solar radiation.

3.3.2 *CANCOR Analysis*

The *CANCOR* analysis revealed positive and negative correlated loci that were correlated ($|r|=0.25$) with phenotypic traits (Figure. 3.1). Ten correlated loci with positive phenotypic effects on height were found but no correlated loci for positive increases in DBH were discovered. Incidentally, two of these loci for height were identified in the *GAPIT* analysis. The two loci were associated with environmental traits for maximum solar radiation and forest fire risk (aridity-based parameters). Twenty-six loci were associated with a reduction in height and one locus was associated with a reduction in DBH. One of these loci was shared between DBH and height and none of these loci were shared between *CANCOR* and *GAPIT* analysis. From the environmental analyses, 121 loci were positively adaptive towards one or more climatic variables. Of these, 62 were shared between the *GAPIT* and *CANCOR* analyses. 131 loci were negatively correlated with adaptation to one or more environmental parameters. None of these negatively correlated loci were shared with the *GAPIT* analyses.

Association between environmental and phenotypic traits was defined as sharing a position in the first two canonical dimensions. All positive associations related to height. Four associations were due to maximum solar radiation (*max_srad*) which is typical for spruce trees to grow taller in response to a higher incidence of solar radiation, however, these loci may indicate more efficient use of solar radiation. Two adaptations towards isothermality were discovered and two involved precipitation seasonality. There were twenty-six loci associated with a negative reduction in height. Of these, 15 loci were associated with maximum or mean snow cover indicating adaptive reduction in size to cope with high snow cover. 11 loci were involved with minimum or mean solar radiation (*mean_srad*)(Figure. 3.1).

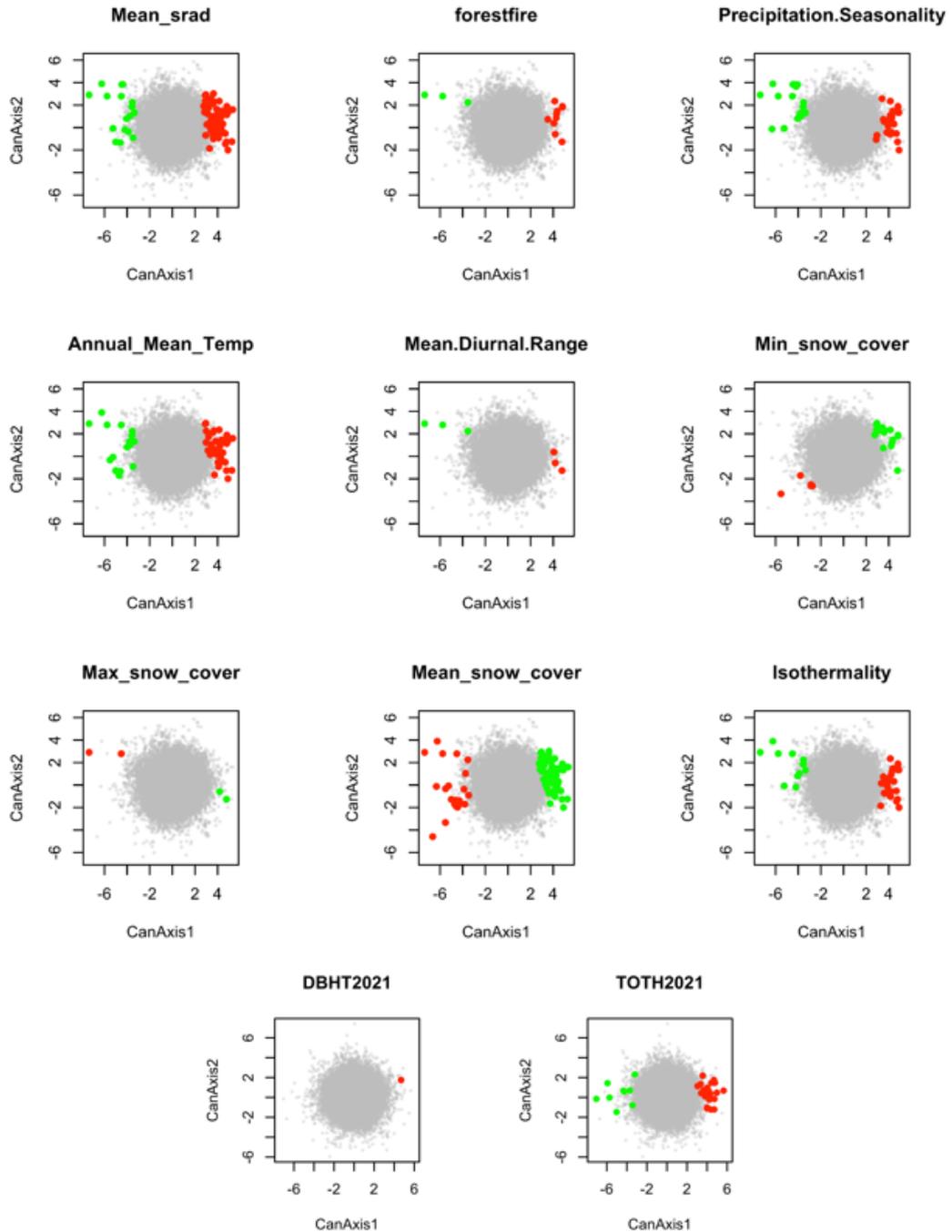


Figure 3.1. Outlier loci for specific environmental and phenotypic data. The outlier status of loci was determined by the Canonical Correlation (CANCOR) analysis where minor allele frequency was relatively highly correlated ($|r|=0.25$) with a trait. Loci coloured grey represent all the loci tested. Loci coloured red are considered negatively correlated and loci coloured green are considered positively correlated.

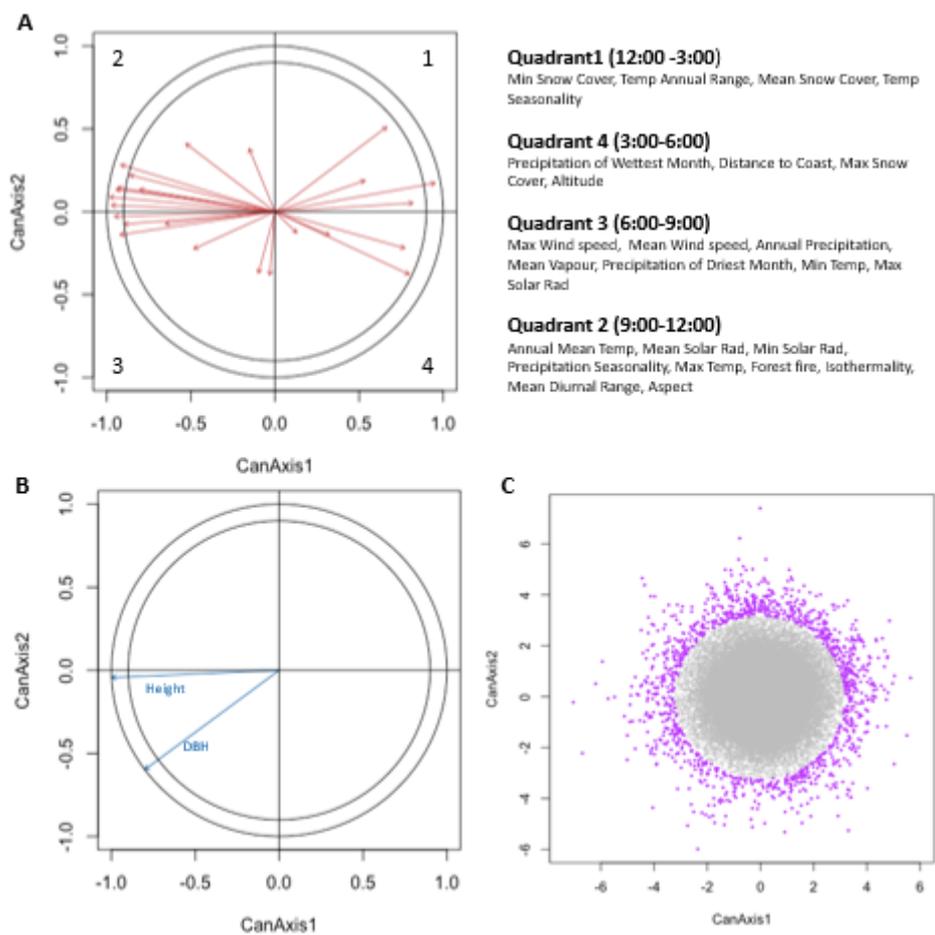


Figure 3.2. (A) Representation of environmental variables on canonical axis as correlated with alternate allele frequency. (B) Representation of phenotypic variables on canonical axis as correlated with alternate allele frequency. (C) Representation of loci on the first two canonical axis. In (A-B) the inner circle marks 0.9 on projection norm and the outer circle represents 1 on projection norm values.

There were approximately two environmental gradients discovered in the environmental analyses (Figure. 3.2). The first, and main cluster, is found between quadrants 2 and 3 with correlated loci being found for solar radiation (mean srad, min_srad, max_srad), aridity (forest fire, isothermality, precipitation seasonality, precipitation of driest month) and temperature extremes (max temp, min temp) (Figure. 3.2a). The second cluster,

which was smaller, is found between quadrants one and four with correlated loci for snowfall (min snow cover, mean snow cover, max snow cover) and temperature ranges (temperature seasonality, temperature annual range) being discovered (Figure. 3.2a). Height is mainly affected by the first cluster and DBH is located in quadrant three but is not directly affected by any environmental parameters (Figure. 3.2b).

3.3.3 Allele Frequencies and Loci Distribution

Loci with either a negative or positive correlation with height were traced through the population and mean number of loci per sample per provenance were represented geographically (Figure. 3.3). Northern populations contained loci with negative correlations with height mainly caused by high snow cover and low solar radiation. More negative loci were distributed in Alaska than further south. Loci with positive correlations towards height were distributed in the southern ranges of Sitka spruce. This split between northern and southern populations occurred at the 50th latitude.

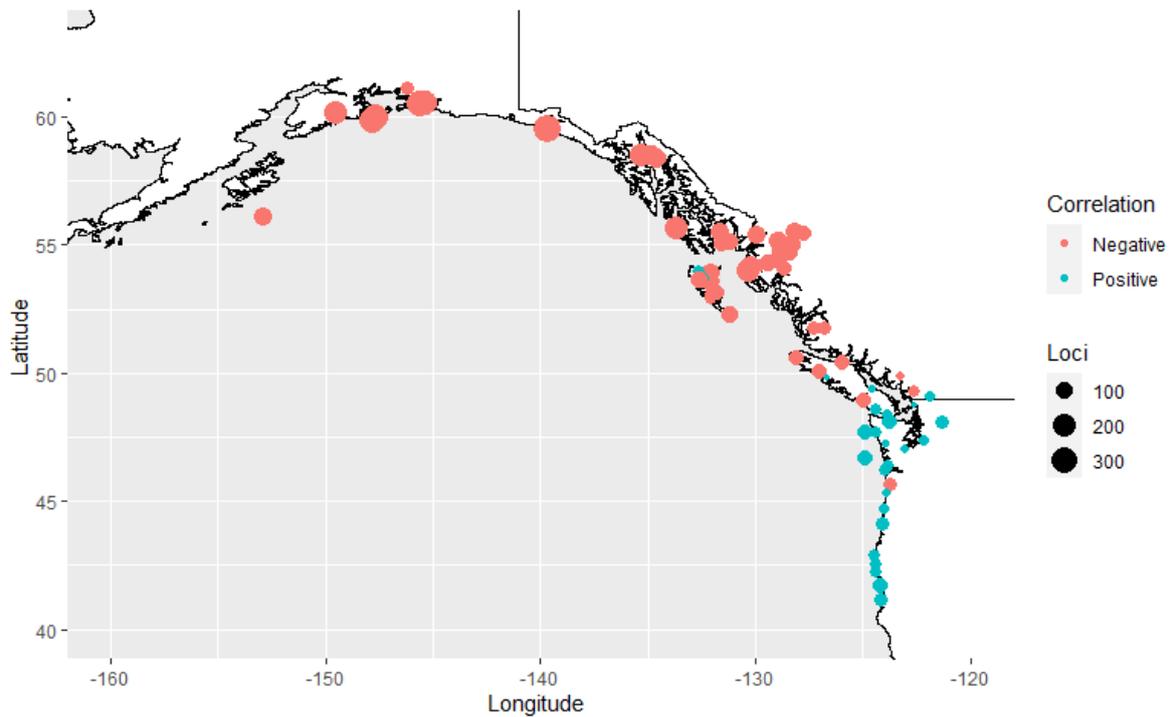


Figure 3.3. Geographic distribution of correlated loci positively and negatively correlated with height across the Pacific Northwest. Red circles refer to loci negatively correlated with height, whereas green circles refer to positive correlations. Circle size refers to number of loci associated with height.

To characterise adaptation, minor allele frequency (MAF) loci was measured. Mean MAF for possibly adaptive loci was 0.439 and mean MAF for non-adaptive loci was 0.117. In Figure 3.4 A and B there is a considerable shift in MAF between possible adaptive and non-adaptive loci indicating that environmental adaptation is characterised by high minor allele frequency. The same cannot be concluded from phenotypic adaptation however, because only nine associated loci were discovered from the phenotypic GAPIT analysis. However, the mean MAF for phenotypic associated loci was 0.403.

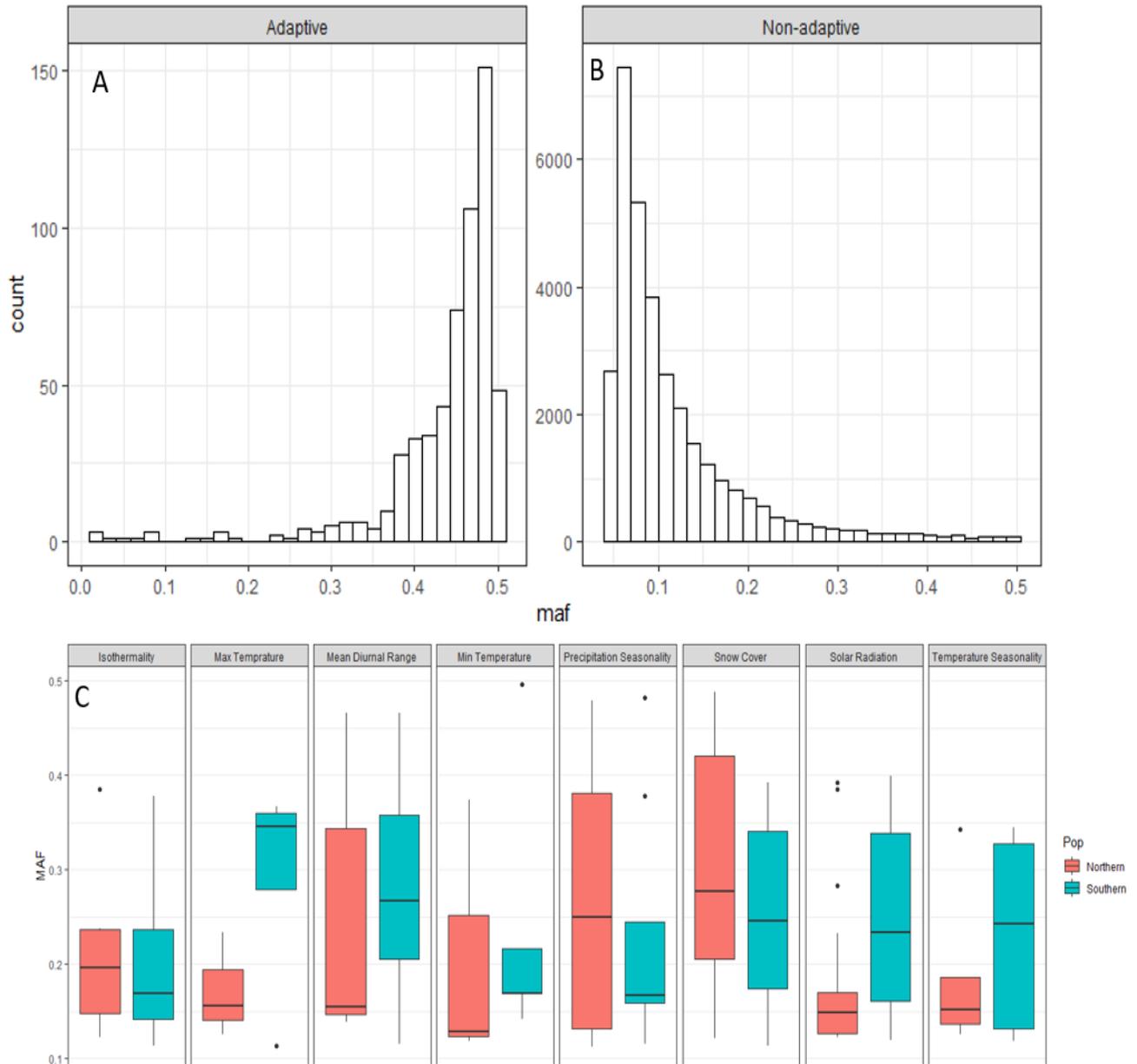


Figure 3.4. (A-B) Minor allele frequency of possibly adaptive alleles and non-adaptive alleles based on GAPIT3 analysis. **(C)** Shifts in minor allele frequency of correlated alleles between Northern and Southern ranges of the population based of GAPIT analysis.

3.3.4 Functional Annotation

Of the loci discovered from the GAPIT and CANCOR analyses, 169 of them could be functionally annotated as being located within a gene. 135 of these had gene ontology (GO) terms, 12 were unknown proteins and 20 were hypothetical proteins (Figure. 3.5). The loci which correlated with height or DBH and an environmental parameter were functionally annotated using BLAST. Of these 14 were annotated and 13 had GO terms.

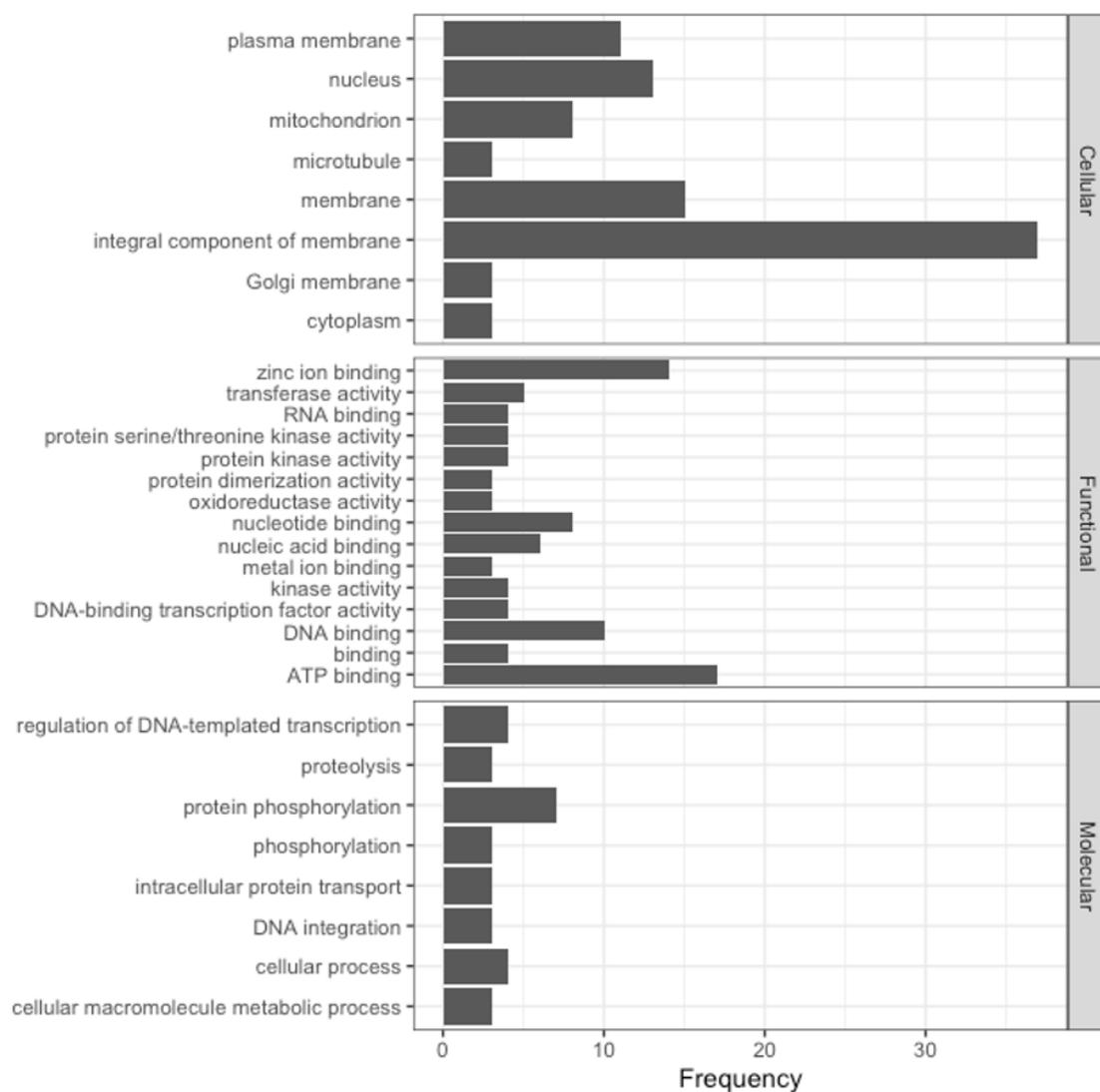


Figure 3.5. Gene ontology of regions associated with associated loci. Gene ontology terms shown have a frequency of greater or equal to three.

3.3.5 Co-association Analysis

The co-association analysis revealed clusters of loci that correlated with traits of interest. The loci with a relatively high correlation ($|r|=0.25$) from the CANCOR analysis could be grouped into three main clusters (Figure. 3.6). Cluster one and three are mainly associated with traits associated with good growing conditions in Sitka spruce. Cluster three is strongly associated with these traits. Cluster two is strongly associated with traits associated with freezing and snow cover. Cluster one contains 425 loci in total, cluster 2 contains 562 loci and cluster 3 contains 224 loci.

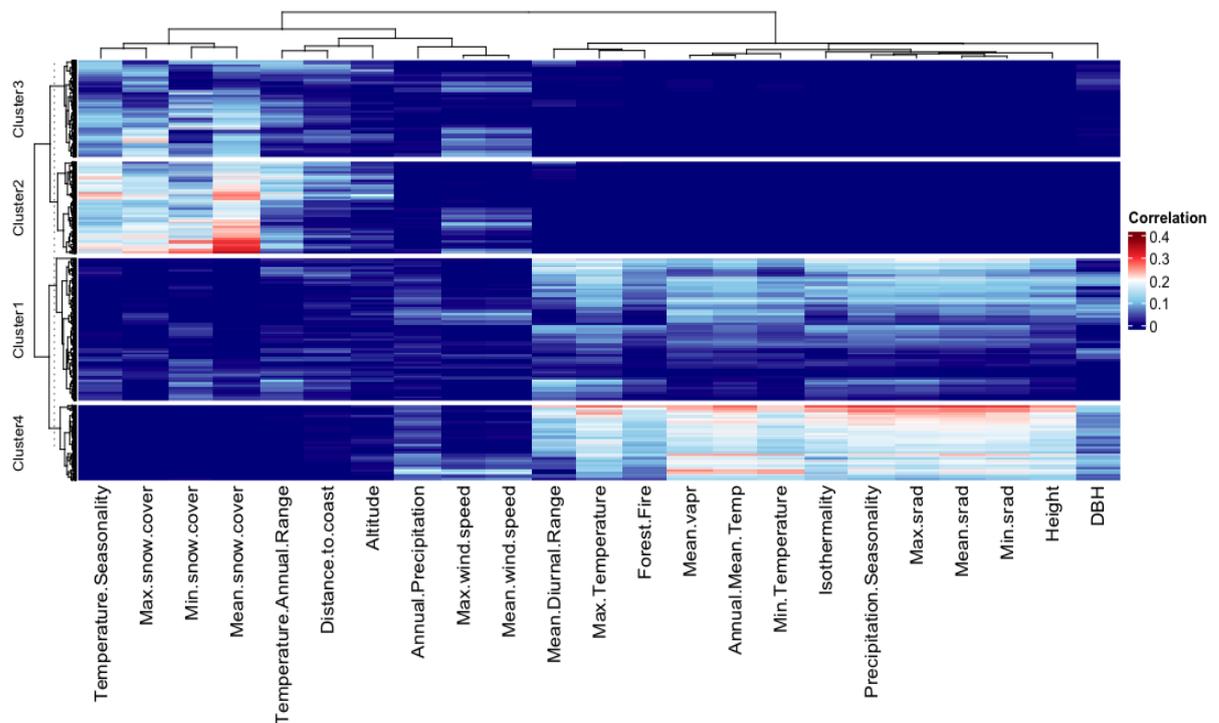


Figure 3.6. Environmental co-association analysis from the results of CANCOR analysis. Correlations between SNP and environmental factors were used for clustering analyses and pairwise comparisons.

3.4 Discussion

3.4.1 *Adaptation in Sitka Spruce*

Studies such as GWAS are ideally suited to conifers due to the characteristics of their genome which is marked by a slow rate of evolution and low diversity despite the large range of environmental gradients the populations occupy (Farjon, 2010). The low rates of LD slow the rate of evolution and lead to gradients of locally adapted genotypes. This, however, is ideal for GEA and GWAS studies, as high LD is problematic for the discovery of loci via GWAS, due to the linkage of associated loci causing false discovery (Christoforou et al., 2012). High LD does however necessitate very high marker density. However linkage of associated loci can lead to evolution by adaption through divergence, so harsh LD filtering is also not recommended. Here 694 loci that appear associated with growth parameters of Sitka spruce were discovered. A split between the northern and southern populations that shows loci associated with increases in height in the southern population was discovered. This is similar to what has been presented previously (Mimura & Aitken, 2010), with reductions in growth rate in northern populations and slowing of bud set. The allelic distribution pattern suggests local adaptation in the northern and southern populations. The distribution of these correlated loci suggests these alleles occurred in the farthest extent north and are naturally selected until it meets the southern range in British Columbia. In the southern range, natural selection favours loci with positive correlations toward height. This highlights the range below the 50th N latitude is ideal for breeding towards increased height.

Previous studies in conifers have shown subtle shifts in MAF responsible for adaptation (De La Torre et al., 2019). These data show possibly adaptive loci with very high MAF compared to overall MAF. This indicates recent local adaption due to the non-

conserved alleles (Günther & Coop, 2013). This would agree with what has been observed previously with post glaciation spread of Sitka spruce creating the need for local adaptation (Byrne et al., 2022; Gapare et al., 2005). Shifts in MAF were discovered as associated with some climatic correlated alleles. Traits identified as associated with southern conditions (Figure. 3.6) had subtle to moderate increases in MAF in southern ranges indicating recent adaptation. The same is seen in northern ranges where traits associated with cold and snowfall saw subtle increases in MAF. This subtle to moderate change in MAF is characteristic of adaptations in conifers, as shown in *Picea glauca* and *Pinus taeda* (De La Torre et al., 2019; Hornoy et al., 2015).

3.4.2 Adaptations to Northern Wintering Conditions

The two main wintering stressors on Sitka spruce are increased pressure involved with snowfall and freezing. Snowfall has the potential to result in shoot leader break. Northern conifers already have conserved adaptations to snowfall, with homeotic genes conferring a triangular form factor and thin needles to reduce the build-up of snow (Du et al., 2020; Uddenberg et al., 2015). This study has identified a reduction in size is correlated to snowfall. The high minor allele frequency of these possibly adaptive alleles indicates recent local adaptations. This is clear in the geographic distribution of these alleles with northern populations having more alleles negatively correlated with height. From the co-association analysis, the traits associated with freezing are grouping together as cluster 3. These traits are temperature seasonality, and minimum, maximum and mean snow cover. From the CANCOR analysis there is a negative correlation between height and these freezing traits.

3.4.3 Adaptations to Southern Conditions

The CANCOR analysis revealed a southern range of Sitka spruce that contains loci that correlate strongly with increases in height. From previous studies, Sitka spruce has been

shown to potentially have recolonised, post glaciation, from Kodiak Island, Haida Gwaii and the southern range (Byrne et al., 2022). The data presented in Figure 3.3 identified southern range specific adaptations that allow for optimal height growth in this range. This southern range has implications for the adaptive evolution of Sitka spruce, showing that southern recolonization led to correlated loci which spread through the populations to the 50th N latitude. The traits identified from the co-association study show that cluster 3 can be strongly linked to good growing conditions, with cluster 1 being mildly linked. Minimum, maximum and mean solar radiation are all associated with loci involved with increased height. This is common as an increase in solar radiation allows for increased tree growth (Strand et al., 2006). Increases in the daylight range allow greater amounts of solar radiation, increasing growth. The temperate climate of the southern range is optimal for growth, with freezing temperatures causing growth cessation, as seen in the northern range being less likely. Mean vapour pressure is grouped with these traits as Sitka spruce is well adapted to receiving moisture through sea fog. Forest fire risk is seen as linked to the same cluster of traits as it is based on the aridity index. This may indicate some adaptations in this loci cluster to aridity in these conditions. It may also be linked to forest fires themselves, as known adaptations in other conifers have been found. This would take additional phenotyping to elucidate further. For a breeding context a cluster of loci has been defined that should be included for better growing trees in ideal conditions. These data analyses have identified a range of trees that could be included in breeding programs to increase height of trees. These loci could be included through genomic selection models by adding weight to selected loci.

3.5 Conclusions

The methods employed here have allowed for the merging of phenotype, genotype and environmental data to elucidate the relationships between the three. The analyses have grouped loci into three clusters which relate to wintering traits such as snowfall, and southern growth conditions which seem optimal for maximising tree size. The spread of these loci correlated with height and discovered a north south divide with the southern population having a higher proportion of loci positively correlated to height. Many of these correlated loci have occurred recently and are not conserved throughout the population, as shown by the high MAF of possibly adaptive loci compared to non-adaptive loci. Figure 3.4 also shown that subtle shifts in allele frequency between northern and southern populations are responsible for adaptation to certain traits in Sitka spruce.

**Chapter 4- Comparison of the native North
American Sitka Spruce and a Sitka Spruce
breeding population.**

4.1 Introduction

The breeding of Sitka spruce is a relatively new practice, however selective breeding at any stage can result in a loss of genetic variability (Khoury et al., 2022). Genetic erosion is defined as the loss of genetic diversity of a species over a given time and area (Leroy et al., 2018). This is a threat to the genetic diversity of a species. Drivers of genetic erosion include breeding practices, habitat loss, climatic change and economic needs. Assessing this genetic variability and comparing the population to the wild type is essential to maintain diverse gene pools for continuing breeding. Direct comparisons between natural populations and breeding populations are not common in tree breeding, which leaves questions of the effects of breeding on the diversity and structure of the breeding population. Breeding populations have been shown to have a lower genetic diversity than wild type populations if effective population size is not met (Dawson & Goldringer, 2012; Fu, 2015; Leroy et al., 2018; Lynch & O'Hely, 2001). If the breeding population is too small this can result in inbreeding depression resulting from the loss of genetic variability (Charlesworth & Willis, 2009). Greater variability within populations increases the likelihood of adaptive alleles being passed through the population (Prunier et al., 2016).

Breeding of Sitka spruce in Ireland falls under the Irish Sitka Spruce Tree Improvement Programme (ISSTIP), an initiative to cultivate 'improved' material (understood to mean individuals with greater vigour as well as desirable form-related traits) for use in Ireland in place of seedlings acquired by conventional means such as imported seed lots or collection of seed from Irish stands (Lee et al., 2013; Thompson et al., 2005). The prominence of Sitka spruce within Ireland is itself the result of efforts to improve forest stand productivity (Fennessy et al., 2012), and amongst Irish Sitka spruce stands material originating from Washington and Haida Gwaii predominates (O'Driscoll, 1977; Thompson et al., 2005). The former source is the most suitable for Irish growing conditions

for the purposes of improving productivity and the latter source preferred for its relative resistance to frost damage while still offering improvement of phenotype (Horgan et al., 2003). Following on from improvement at the species and population level, improvement is now being pursued at the individual level. Material in the ISSTIP population, located in the National Botanic Gardens in Kilmacurragh (52.929°N, -6.148°W), is drawn from ‘plus trees’ (Lee et al., 2013; Thompson, 2013), trees displaying above-average phenotypic qualities among otherwise average material in Irish Sitka spruce stands. Multiple sets of progeny trials have been carried out, namely open-pollinated progeny trials examining the presence of additive variation for traits of interest among half-sibling progeny (Thompson, 2013), and full-sibling trials evaluating the combining ability of parents and parental genotype combinations that consistently generate progeny with improved traits (Glombik et al., 2015). Clones of evaluated plus trees are currently maintained in gene banks for reference (Lipow et al., 2002), while a small number of clones have also been deployed as seed orchard material for the cultivation of improved material based on assessment of the half-sibling progeny trials (Cahalane et al., 2007)

The original plus trees are known to have been drawn from stands planted using imported material rather than home collected seed. Much of what is known about the origin of these individuals is sparse beyond this, with records referring to origins in terms of provenance or larger geographic areas, and the use of seed from allotments made from combining multiple provenances complicates this. As this seed is exclusively from imported material as opposed to Irish seed, the expectation would be that genotypes within the ISSTIP would reflect the region from which seed for the original plus trees was collected. The production of improved seed and all experimental work done within the ISSTIP to date is ultimately centred around clones of these plus trees, with no second generation of breeding genotypes having yet been created. With genotype data available for a subset of the ISSTIP

population, along with prior genomic data that contextualised this material in terms of geographic origin (Byrne et al., 2022), it is now possible to evaluate the genetic composition of the ISSTIP population.

The IUFRO (International Union of Forest Research Organisation) population of Sitka spruce represents samples from the entire native range of Sitka spruce (O'Driscoll, 1972, 1978; Sype, 1990). The native range of Sitka spruce stretches along the coast from Alaska to Northern California (Lee et al., 2013; OECD, 2006). It occupies coastal regions, riverbeds and islands. It thrives across varying ecological and climactic niches, having undergone local adaptation along the range. 81 provenances are represented in this population with genotype data available for 80 of these provenances. These provenances represented 19 geographic regions. This study utilizes previous research where the IUFRO population was genotyped and investigated for population diversity and structure (Byrne et al., 2022). In this study the IUFRO population is used as a baseline for the comparison of the ISSTIP population (Irish breeding). This comparison is possible due to the same GBS genotyping method used. This study aims to assess the diversity of the breeding population, potential genetic erosion, population structure, gene flow potential and demography of the ISSTIP population compared to IUFRO population.

4.2 Methods

4.2.1 Plant Collection and Genotyping

Needles from 215 genotypes of the ISSTIP population were harvested and dried and stored using silica gel. Samples were also freeze-dried before grinding on a Retsch MM400 mill. Samples were ground with two titanium beads for 10 minutes at 30 oscillations per second. Ground samples were then extracted using the Machery and Nagel NucleoMag Plant DNA kit (744400.1) on the Kingfisher flex extraction system. A second elution step was added

to the Kingfisher flex system. Samples were normalised to 30 ± 5 ng/ μ L using the Quant-iT PicoGreen dsDNA Assay Kit (P7589). Samples were sent to LGC (Berlin, Germany) where they underwent library prep for Genotyping-by-Sequencing (GBS) using PstI-AepK1 (Elshire et al., 2011). The libraries were sequenced on an Illumina platform with paired end sequencing to achieve a target depth of 3M paired-end reads per sample.

4.2.2 Variant Calling

Data from the IUFRO population had previously been generated to compare the breeding population (ISSTIP) to a population drawn from across the full native range. The IUFRO population acted as a baseline for comparison against the ISSTIP population. Reads from the ISSTIP genotyping were combined with the downloaded reads from NCBI BioProject PRJNA852515. The combined sequences were aligned to the Q903-v1-1000 plus Sitka spruce genome (GCA_010110895.2)(Gagalova et al., 2022) using BWA-mem with default parameters (Li & Durbin, 2009). Variant calling was completed using SAMtools v1.9 mpileup (Li et al., 2009). VCFtools was used for filtering (Danecek et al., 2011) where indels were removed, only biallelic variants were kept, minimum genotype quality was filtered to 20, read depth was filtered to 5, maximum missing data was filtered to 0.7 and MAF was set to 0.05.

4.2.3 Inferring Origins of Unknown Genotypes

Of the 215 trees genotyped, 85 were marked as having an unknown origin. To decipher the origin a Genomic Relationship Matrix (GRM) was created using PLINK (Purcell et al., 2007). The matrix contained all 1177 IUFRO genotypes and 215 ISSTIP genotypes. The closest relations in the IUFRO population were determined for the 85 unknown samples using the GRM.

4.2.4 Diversity Statistics

Analysis was completed using R version 4.0.2. The package adegenet (v2.15) (Jombart, 2011) was used to convert the VCF file to a genind object. The function basic.stats from the hierfstat (v0.5-10) (de Meeus, 2007) package was used to calculate observed heterozygosity (H_o), expected heterozygosity (H_e), gene diversity (H_t and D_{st}), and fixation index (F_{st}).

4.2.5 Clustering

To visualise the clustering of the ISSTIP population in comparison to the IUFRO population a Principal Component Analysis (PCA) was used. A PCA was run on a genlight object using the glpca function in adegenet retaining 1400 principal components. The PCA was visualised using ggplot2(v3.3.5). From this PCA it was clear that the ISSTIP population was clustered into two distinct populations. Using dplyr, the ISSTIP population was segregated from the IUFRO population and plotted as a PCA using ggplot2. A Discriminate Analysis of Principal Components (DAPC) was also run using the dapc function in adegenet. This was run using 6 discriminate axes and 1400 principal components.

4.2.6 Gene Flow between Populations

To assess potential gene flow between the ISSTIP population and the regions with the IUFRO population, Patterson's D statistics and f_4 -ratio were used (Malinsky et al., 2021). The D statistic, also known as the ABBA-BABA statistic, The D statistic considers ancestral alleles (A) and derived alleles (B) and tests for introgression between them. The related f_4 -ratio tests for the related admixture fraction between a trio of populations and an outgroup. Kodiak island was used as an outgroup as previous evidence has shown it to be

the most genetically distant population. The program D-suite was used to calculate D-statistics and f4-ratios. Ggplot2 was used to visualise pairwise D-statistics and f4-ratios.

4.2.7 Admixture

To investigate ancestry of the ISSTIP population, with the IUFRO population as background, Admixture (v1.3) was used (Alexander & Lange, 2011; Alexander et al., 2009). Optimal number of clusters in the IUFRO population was previously determined by the elbow method and combining data from the DAPC. Admixture was run for K=2-20 with 10 reps. Admixture was plotted using ggplot2.

4.3 Results

4.3.1 Variant Calling and Inferring Origins

The variant calling combined the downloaded IUFRO population with the ISSTIP population and aligned it with the Sitka spruce genome and called variants. From this a VCF file of 66974 variants and 1392 individuals was produced. There was 11.23% missing data in the VCF file. The VCF file was firstly used to infer unknown origins in the ISSTIP population. All 85 of the unknown samples from the ISSTIP population could be inferred back to a specific geographic region using a GRM. This allowed for further analysis on the regions of origin.

4.3.2 Diversity Statistics

The diversity statistics of the IUFRO and the ISSTIP population are summarised in Table 4.1. Overall the ISSTIP population is less diverse than the IUFRO population and differentiated from it. This is clear from the heterozygosity statistics and from the *Fst* and *Dst/Dest* values, which are measures of population differentiation. The ISSTIP population

is smaller than the IUFRO population, with the ISSTIP population containing 215 samples spread over a geographic area of Haida Gwaii to Northern Oregon. This is in comparison to the IUFRO population which has 1177 genotypes spread over a range from Alaska to Californian. The difference in geographic distance a contributing factor to lower genetic diversity. The breeding range of the IUFRO population comes from Washington, Haida Gwaii and Oregon. When the breeding range was isolated from the IUFRO population (Table 4.1), it appears that the ISSTIP population has similar genetic diversity to the natural population based off heterozygosity. Negative Inbreeding Coefficients are generated by analysis type and essentially represent no inbreeding.

Table 4.1. Genetic Diversity statistics of the ISSTIP and IUFRO populations.

	Definition	ISSTIP	IUFRO	IUFRO Breeding Range
Ho	Observed Heterozygosity	0.1342	0.21	0.147
Hs	Within Population Gene Diversity	0.1302	0.198	0.134
Ht	Overall Gene Diversity	0.1309	0.204	0.136
Dst	Gene Diversity among samples	0.0007	0.006	0.0016
Fst	Fixation Index	0.0054	0.0292	0.0116
Fis	Inbreeding Coefficient	-0.0307	-0.0604	-0.0552
Dest	Population Differentiation	0.0011	0.0078	0.0021

4.3.3 Population Clustering

Clustering can be seen in the PCA where the Irish breeding population occurs across two distinct groups (Figure 4.1). The DAPC shows a different situation with the entire ISSTIP population being segregated from the IUFRO population (Figure 4.2) likely due to DAPC overestimating population structure.



Figure 4.1. Principal Component Analysis (PCA) of the IUFRO population and the ISSTIP (Irish Breeding) population. (A) Combined PCA of the IUFRO (Green) and ISSTIP populations (Red). The PCA was run with 1400 principal components. Colour labels include origins inferred using a GRM.

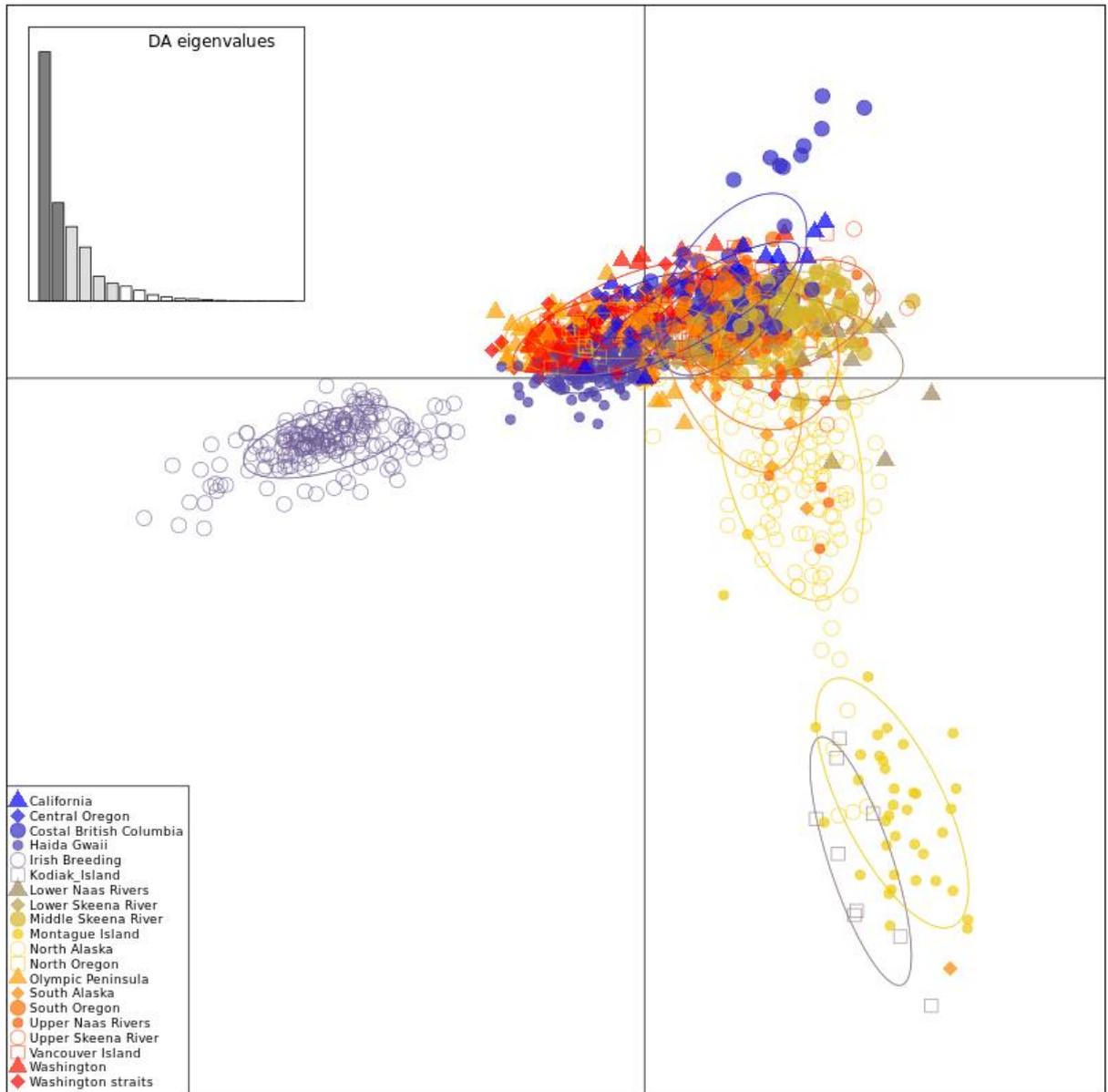


Figure 4.2. Discriminate Analysis of Principal Components (DAPC) of the IUFRO and ISSTIP (Irish Breeding) population. DAPC was run on 20 principal components with 6 discriminate axes.

4.3.5 D-statistics and Ancestry

To elucidate relationships between the geographic regions of the IUFRO population and the ISSTIP population, gene flow statistics were employed. The pairwise D-statistics revealed genetic dissimilarity between the regions of the IUFRO population and the ISSTIP (Irish Breeding) population. The highest genetic dissimilarity was found between Coastal British Columbia and the Irish Breeding population (Figure 4.3a). The f4-ratio show the admixture fraction between the geographic regions. As is expected there is admixture between the Irish Breeding and Washington, California and Oregon based off the f4-ratio (Figure 4.3b). The admixture graph shows 3 clusters with the Haida Gwaii being more differentiated than the mainland (Figure 4.4).

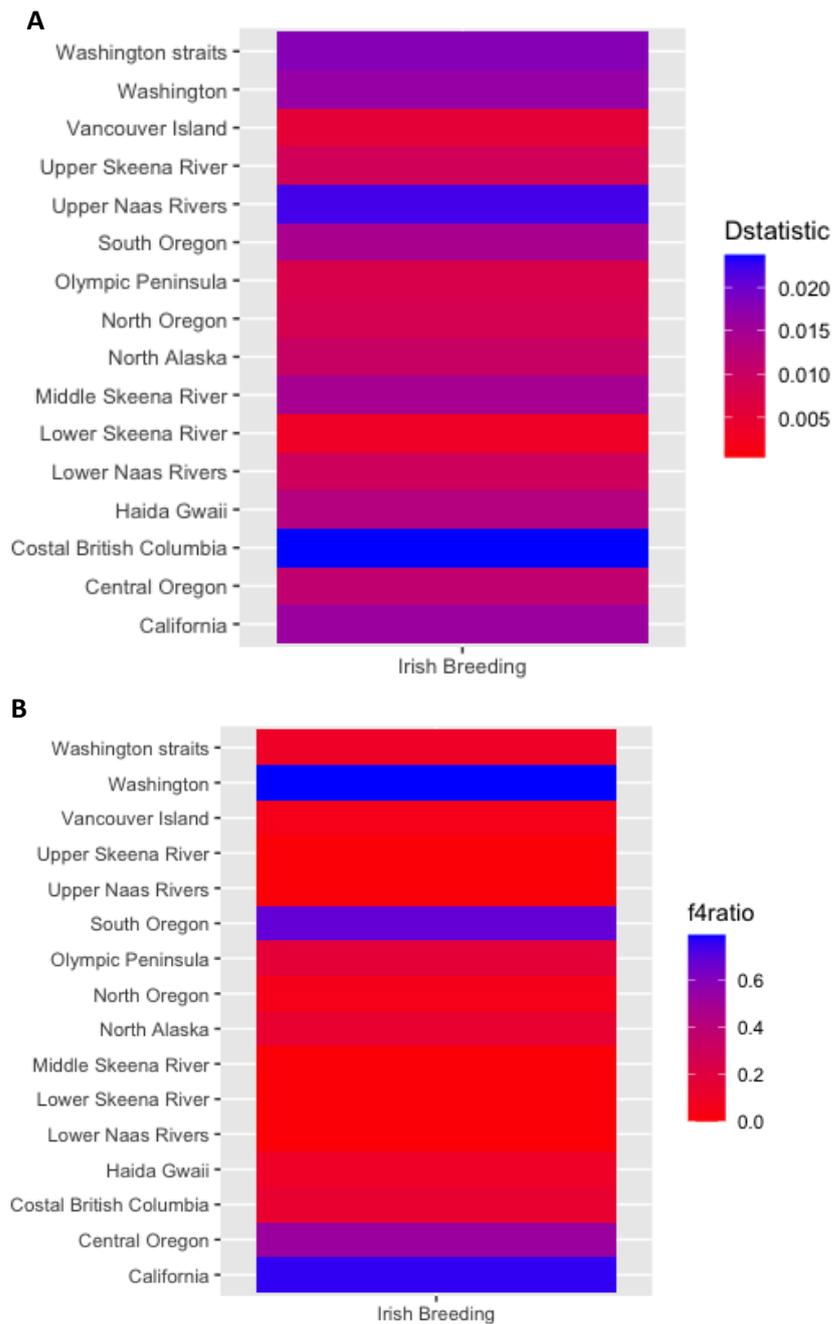


Figure 4.3. (A) Gene flow measured between geographic regions and the ISSTIP (Irish Breeding) population using Patterson’s D-statistic a measure of genetic dissimilarity. (B) Admixture fraction between the geographic regions and the ISSTIP populations as measured by f4-ratio.

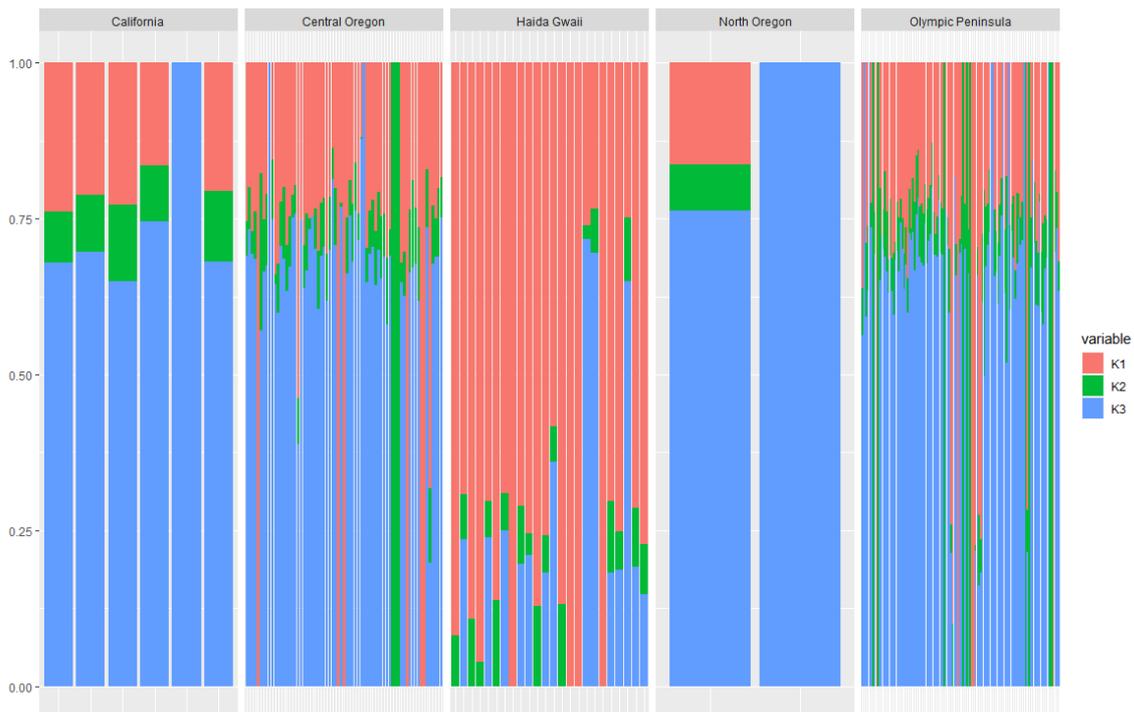


Figure 4.4. Admixture of the ISSTIP (Irish Breeding). K=3 was used as predetermined by the results from the DAPC and the elbow method for best K.

4.4 Discussion

The IUFRO population is a key resource for the comparison of natural populations and breeding populations (Byrne et al., 2022). It contains 80 provenances that range from coastal Alaska to northern California. The genetic diversity and structure of the population was previously assessed and can act as a baseline for assessing other populations against. This is key for breeders and scientists studying the genetics of conifers. For breeders the IUFRO population can act as a baseline to compare against and to introduce new provenances into breeding programs. With careful monitoring and planning, breeders can maximise genetic gain while avoiding genetic erosion (Williams & Savolainen, 1996).

Genetic diversity of the ISSTIP population was assessed against the IUFRO population. The IUFRO population was much more diverse, however this is not surprising given the large geographical range of its source. The diversity of the IUFRO population is related to isolation by distance and by island refugia effects (Byrne et al., 2022). The ISSTIP population spreads over a smaller geographical range than the IUFRO population. The origins of the ISSTIP population are largely from Haida Gwaii, Oregon and Washington. However detailed origins are not known due to record keeping flaws. The selection criteria is not clear also, which is a flaw of this study and future breeding programs. When comparing these locations from the IUFRO to the ISSTIP population, there are very similar results. The IUFRO breeding range is slightly more diverse ($H_s=0.134$) compared to the ISSTIP population ($H_s=0.1302$). This difference in genetic diversity is indicative of potential genetic erosion based off progeny selection however continual monitoring of the breeding program is needed to elucidate genetic erosion.

In the DAPC the entire breeding population is separated away from the main cluster. With the selection of the founding population, a gradient in population structure can be observed similar to what is seen with Kodiak Island in the main population. However, this population structural change is not associated with F_{st} isolation as is the case for Kodiak Island (Byrne et al., 2022). From the PCA the structural change is not apparent. The population structure according to the PCA, shows the ISSTIP population being segregated into two groups. Overall it appears there is a structural shift in the population due to the early stages of breeding, as seen in the DAPC.

Gene flow, as measured by genetic dissimilarity, between the Irish breeding population and the rest of the regions is quite ample according to the high D statistics between regions. Of note, genetic dissimilarity is high between Upper Naas River and the Irish breeding population, two populations that are genetically and geographically different.

When investigating the f_4 ratio this is not seen. As a measure of co-ancestry the f_4 ratio shows relationships between the Irish breeding population and California, Central and South Oregon and Washington. This however is contradictory to the admixture plot which shows the main ancestral population (K3) being dominant in the Irish Breeding population, but not being shared with any other geographic region. The ISSTIP population is reported to be made up of material from mainly Haida Gwaii and Washington. Some samples in the ISSTIP population share ancestry with Haida Gwaii, but not Washington, as would be expected. To elucidate this further, additional genotyping would be needed to be complete after subsequent breeding, to see further shifts in ancestry.

4.5 Conclusions

The ISSTIP population is still relatively genetically diverse, but there is some loss of genetic variation as compared to the IUFRO range. The ISSTIP population is structurally differentiated from the IUFRO population based off a DAPC analysis, but not the PCA or the *F_{st}*. The DAPC still points to potential structural isolation of the ISSTIP population due to early breeding. The gene flow between populations is ample with gene flow between neighbouring regions to the origins of the ISSTIP population, being highest. Genetic work can help identify unknowns in a breeding program however it does come with some uncertainty. It cannot replace good record keeping which is imperative to all breeding programs, but especially forestry breeding where generations are long.

Chapter 5- A panel of KASP markers for the accurate genotyping of Sitka spruce seed orchards reveals dominant parental genotypes.

This chapter will be submitted.

Contributions by Tomás Byrne: Designing and preparing experimental sampling, laboratory sample preparation, data analysis, writing, reviewing.

5.1 Introduction

Seed orchards are an effective way of bulk generating seed through open pollinated improvement of material. Improvement is when favourable traits are selected through progeny selection. In the case of Sitka spruce favourable traits are height, DBH and form however other favourable traits could include disease resistance, wood quality or climate resistance. Seed orchards comprise selected superior genotypes that are arranged in specific spatial manners to allow for optimal pollination amongst genotypes. Seedling seed orchards are established from seedlings, leading to a broader genetic base but a lower genetic gain. In contrast, clonal seed orchards, which are more common, include multiple clones leading to increased genetic gain (Giertych, 1975). These clonal seed orchards contain replications of improved parents from breeding programs to maximize potential crosses (Schmidt, 1993). In both types of orchards, trees are often capped in height to allow for ease of management and harvest of cones. This systematic canopy control reduces the crown dominance of certain parents, which can lead to unequal parental representation in the offspring (Funda, 2012). Pollen production can also be induced through hormonal means or the application of abiotic stressors. For example, gibberellic acid can be injected into the trunk of the tree to induce pollen production (Pharis, 1976) or a similar result can also be achieved, at a lower cost, by inducing drought stress (Schmidt, 1993). In warmer, drier climates this is achieved through the withholding of irrigation, but in wetter climates, it can be induced by girdling. Girdling involves removing rings of bark from the base of the tree, reducing the flow of nutrients and water (Schmidt, 1993). Root pruning can also achieve this but specialised equipment is needed, and the lack of specificity may result in damage to the tree, including some stability issues (Schmidt, 1993).

Effective clonal seed orchard design should maximize the distance between ramets of the same clone and genetically related clones (van Buijtenen, 1971). Current design

principles of seed orchards rely on randomisation, replication and distance (Funda, 2012). Randomisation of clones maximises potential crosses between the clone and its neighbouring trees. Replication of ramets allows for the production of larger cone crops. Replication strategies vary from increasing the total number of ramets per clone and randomising the orchard as one, or creating block designs and replicating the block (El-Kassaby et al., 2014). Distance measures are used to segregate clones and reduce inbreeding (Williams, 1996). Physical distance maximises the space between ramets of the same clone. This is often maximised through block designs. Genetic distance is now being implemented in orchards with the consideration of genetic relationships between all clones and maximising the physical distance based on closely related clones (Chaloupková et al., 2019). This is similar to diversity index breeding which also seeks to maximise the genetic diversity among progeny.

With the introduction of genetic distance based design, genotyping is being employed to accurately investigate genetic relationships within breeding populations. Genotyping uses molecular markers to identify individuals based on their genetic variation (Rasheed et al., 2017). The development of array and sequence based genotyping has made genotype data available to breeders at moderate cost. This allows for the calculation of genetic distance based measures for improving seed orchard design (Yuan et al., 2016b). It can also be used to monitor existing seed orchards to calculate the parentage of the offspring. This can reveal imbalances in reproductive success of clones within the orchard and unidentified parentage can indicate contamination of the orchard from external sources (Ebrahimi et al., 2018). Parentage assays can also monitor the seed orchard and reveal total inbreeding in the stand, which can then be used to inform future design and selection (Galeano et al., 2021).

This paper aims to use competitive allele specific PCR (KASP) assay to develop a custom inexpensive assay for monitoring Irish Sitka spruce seed orchards. Genotyping the offspring of the seed orchard can determine the efficiency of said orchard. The efficiency of the orchard refers to how much outbreeding there is, how much selfing/inbreeding there is and how much contamination there is. Furthermore this study investigates the over contribution of genotypes contributing to offspring.

5.2 Materials and methods

5.2.1 Seed orchard

The seed orchard sampled in this study was located in Ballintemple, Co. Carlow Ireland (52.74N, -6.67W). The seed orchard followed no particular design. It consisted of 60 parents taken from the Irish Sitka Spruce Tree Improvement Program (ISSTIP) population which was based on superior performance from progeny trials. The parents were not equally represented, with 10 parents occurring only once and one occurring up to 18 times. The design pattern of the seed orchard was not known but seemed to focus on randomisation and did not include distance based measures. There were a total of 342 trees in the seed orchard (Table A6.1). Trees were spaced 5 meters apart. The trees were 12 years old at the time of seed sampling. The trees came from grafted stock so they were clones and not part sibs. The trees were girdled to allow for seed production. Cone harvest and seed randomisation was done by Coillte and seed was provided to Teagasc after normal commercial operations. Cones were harvested from each of the trees and bagged with no effort to divide cones by maternal genotype, dried and seed was extracted. During the extraction and drying process, cones were mixed, then seeds were mixed, randomising seeds. Seeds were harvested in 2020.

Table 5.1. Summary of parents in the seed orchard and the occurrence of parents. The parents are clonally propagated through grafting. The layout of the seed orchard is in Table A6.1

Parental Genotype	Count in seed orchard
190	18
455	16
574	13
280	13
43	12
577	11
595	10
575	10
150	9
619	8
600	8
588	8
583	8
559	8
374	8
291	8
266	8
230	8
608	7
579	7
527	7
520	7
339	7
243	7
226	7
698	6
689	6
578	6
206	6
183	6
587	5
264	5
219	5
210	5
582	4
580	4

570	4
519	4
377	4
218	4
719	3
547	3
541	3
321	3
217	3
589	2
546	2
545	2
209	2
191	2
599	1
597	1
535	1
445	1
261	1
251	1
233	1
225	1
184	1
116	1

5.2.2 Plant growth

Randomised seed lots were collected from the pooled seed samples. Approximately 288 seeds were planted in trays, in John Innes 2 type compost and topped with perlite to reduce water loss. They were grown in glasshouses for 12 weeks until seedlings appeared. The needles from the seedlings were removed and placed in micro-centrifuge tubes where they were freeze dried for 12 hours.

5.2.3 DNA extractions

DNA was extracted from 20mg of freeze dried tissue for 288 samples. Tissue was powdered using a Retsch MM400 mill system with titanium 6mm beads and samples milled for 5 minutes at 30 oscillations per minute. Tissue was extracted using the Machery and Nagel

NucleoMag Plant DNA kit (744400.1) modified for the Kingfisher flex extraction system. Samples were quantified after extraction by using the Quant-iT PicoGreen dsDNA Assay Kit (P7589) and a BioTek Synergy HT to assess if samples were at a quantity of 60 ± 5 ng/ μ L.

5.2.4 KASP assay

The ISSTIP population was genotyped using GBS and all 60 parents were included in this (Chapter 4). Genotypes were processed using the pathway described in Chapter 4. VCFtools was used to segregate parents in the VCF file. KASP markers were designed using the Perl script MinimalMarkers (Winfield et al., 2020). The process involves the iterative processing of all markers to identify markers which can discriminate trees. 21 markers were identified and provided to LGC Genomics where primer pairs were designed and the KASP assay was run (Table A6.2 for primers targets).

5.2.5 Data analysis

Data was provided by LGC Genomics. SNPViewer was used to initially assess the quality of calls per marker (Technologies, 2023). Then call rate was calculated using Microsoft Excel and markers with a missing rate of more than 30% were removed. The offspring file was generated by transforming the KASP file to GenePop format (Rousset, 2008). The parents file was created by using VCFtools to filter the ISSTIP genotyping by sequencing (GBS) data to selected markers and then using PDGspider to transform the VCF to Genpop format (Danecek et al., 2011; Lischer & Excoffier, 2011). Cervus software was used to calculate allele frequencies for all markers (Kalinowski et al., 2007). It was also employed to perform identity analysis on the offspring. Identity mapping was set to a maximum of 18 matching loci to be considered for identity analysis and zero mismatching loci were allowed. The R package ‘apparent’ was used to calculate parentage of the loci (Melo &

Hale, 2019). MaxIdent was set to 0.10, increasing computation time but increasing accuracy, and nLoci was set to 10 to allow for a greater range of potential parents. Most likely parent was determined if it had a p value of less than 0.05, a low Gowers Genetic Dissimilarity (GD) and a high number of typed loci. GD is measured by the genetic identity of the expected parents i and j and of all the potential offspring. The apparent package tested a triad of parent 1 parent 2 and offspring. Most likely parent was calculated by the ratio of GD to number of loci typed.

5.3 Results

5.3.1 Marker efficiency

The 21 markers were all evaluated using SNPviewer and Microsoft Excel to elucidate call rates. One marker had a missingness rate of above 30%, leading to it being removed from the panel. Using Cervus, the markers were evaluated for heterozygosity, homozygosity and null allele frequency (Table 5.1).

Table 5.2. Summary statistics of markers used to genotype the 288 offspring. Observed Heterozygosity (HObs), Expected Heterozygosity (HExp), Polymorphic Information Criterion (PIC) and Null Allele Frequency (F(Null)).

Locus	HObs	HExp	PIC	F(Null)
Ps-1r180608s00000518	0.55	0.47	0.36	-0.08
Ps-1r180608s00002395	0.34	0.64	0.56	0.28
Ps-1r180608s00005263	0.08	0.29	0.27	0.63
Ps-1r180608s00007303	0.59	0.49	0.41	-0.11
Ps-1r180608s00015943	0.55	0.47	0.39	-0.09
Ps-1r180608s00016186	0.56	0.63	0.55	0.02
Ps-1r180608s00025480	0.44	0.57	0.53	0.14
Ps-1r180608s00037345	0.22	0.26	0.24	0.07
Ps-1r180608s00038506	0.25	0.52	0.47	0.34
Ps-1r180608s00051678	0.19	0.19	0.18	0.00
Ps-1r180608s00056477	0.02	0.03	0.03	0.25
Ps-1r180608s00098651	0.20	0.48	0.45	0.45
Ps-1r180608s00174019	0.01	0.01	0.01	0.00
Ps-1r180608s00182550	0.07	0.06	0.06	-0.01
Ps-1r180608s00219415	0.01	0.28	0.25	0.91
Ps-1r180608s00240604	0.07	0.15	0.14	0.35
Ps-1r180608s00255114	0.04	0.29	0.26	0.79
Ps-1r180608s00271298	0.10	0.32	0.29	0.56
Ps-1r180608s00274982	0.02	0.24	0.21	0.80
Ps-1r180608s01075355	0.10	0.10	0.10	0.00

5.3.2 Identity checks

The shared identity of the offspring was checked using Cervus. The first validation check was to compare positive controls to see if they had shared identity. The identity of the positive controls matched. A total of 45 offspring had shared identities with one or more of the offspring. These data can be found in Table A6.3.

5.3.3 Parentage

The R package apparent was used to test for outcrossing, selfing and contamination. A total of 156 offspring were produced through outcrossing and 62 were produced through selfing. Gowers Genetic Dissimilarity (GD) (Gower, 1971) was used to identify parents (Figure 5.1). 71 samples could not be parentage verified which could potentially be contamination. From the outcrossed samples, the frequency of parental occurrence is summarised in Table 5.2. Parentage for each of the offspring is summarised in Table A6.4.

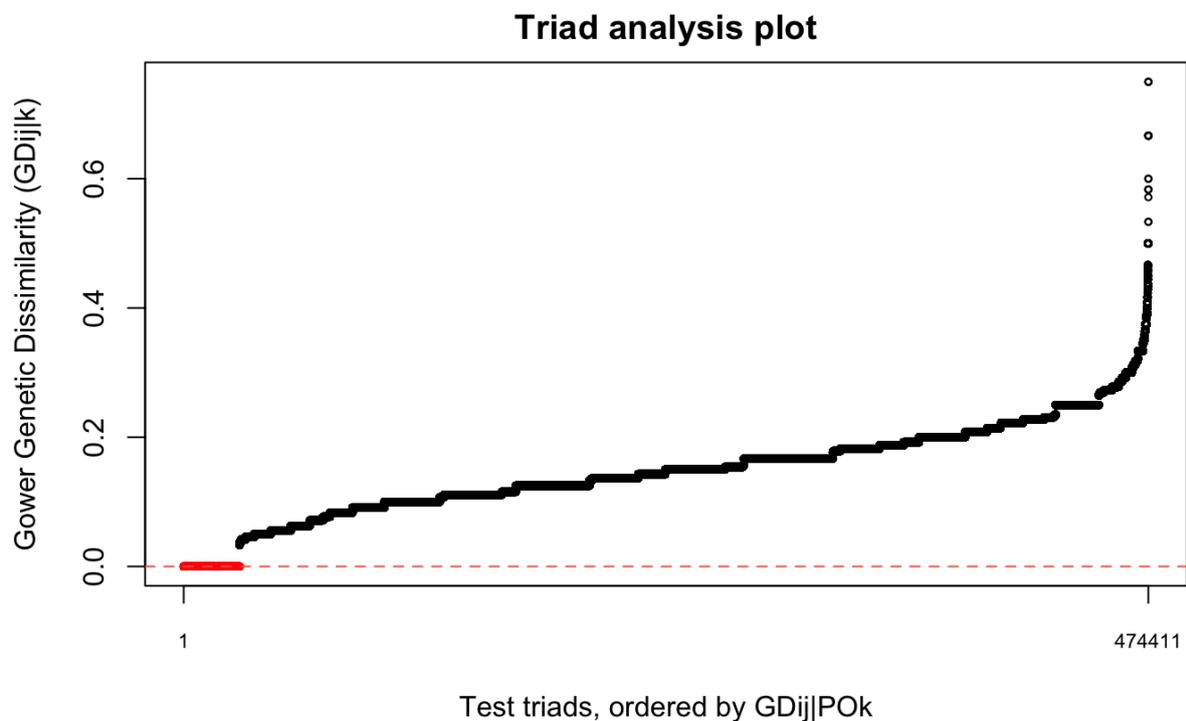


Figure 5.1. Distribution of triads ordered by lowest genetic dissimilarity of parents and potential offspring. Triads of Parent1(i)/Parent2(j)/Offspring(k) were tested for Genetic Dissimilarity (GDijk). Samples in red have a GD of 0 showing they are potentially identical.

Table 5.3. Frequency (f) of parental genotypes and their occurrence in the seed orchard.

Parent	f in Offspring	f in Seed orchard
S43	71	12
S578	60	6
S589	27	2
S243	18	7
S190	14	18
S698	11	6
S226	10	7
S233	9	1
S264	8	5
S519	7	4
S520	6	7
S191	5	2
S580	5	4
S321	5	3
S575	4	10
S251	4	1
S719	3	3
S619	3	8
S150	2	9
S579	2	7
S230	2	8
S218	2	4

5.4 Discussion

Seed orchards remain one of the best methods for bulk propagation in conifers. Historically seed orchard design relies heavily on randomisation and maximising distance between clones (Funda, 2012). However more modern designs have focused on implementing genetic distance into the design. SNP molecular markers, like microsatellites, are commonly used to evaluate the parents to aid with seed orchard design (Yuan et al., 2016a).

Currently seeds are considered improved if progeny are selected carefully and seed orchards are designed effectively. However, testing of the seeds to see if they are actually from improved stock is not common. A set of molecular markers (KASP assay) were used to accurately genotype Sitka spruce and to test the efficiency of the seed orchard in producing offspring from outcrossing parents.

The KASP assay developed for seed orchard evaluation is simple and cost effective. It can be used in cases where one of the parents are known, no parents are known or to confirm known parents. This assay is based off 21 SNP markers. The markers were designed using ‘minimal markers’ which aims to identify the minimum number of markers needed for a KASP assay (Winfield et al., 2020). The minimal markers needed to identify all the parents was 13. However, 21 markers were used to counteract low call rate or failing markers. This assay could in future studies be expanded to 30 markers to cheaply genotype the entire breeding range of Sitka spruce, which spans from Washington to Haida Gwaii, based off ‘minimal markers’ predictions, however it is more accurate to use larger marker sets. Or it could be expanded to 65 markers to genotype the entire IUFRO population based off ‘minimal markers’ (Byrne et al., 2022). This study was restricted to the 21 markers and the seed orchard testing here because of costs and the aim of our work.

From the parentage data there are four main breeding phenomena; outcrossing, selfing and over contributing parental genotypes in the offspring. Firstly outcrossing resulted in 156 of the 288 genotypes (54.41%). Outcrossing is important for seed orchards to create improved material. The spacing between trees of 5 meters allows for ample pollination without crowding the trees. The maximum distance between two parents was 111.8 meters. 62 offspring were generated through selfing (21.53%) meaning that parental trees are crossing with themselves or one of the other same clones (ramets). Reducing selfing can be done with the help of design. The most popular design in this case is

Minimum Inbreeding design which takes n clones and distributes them in a grid based on genetic relationships (Lstiburek et al., 2015; Lstiburek, 2010). That leaves 71 samples that could not be parentage verified (24.65%). This high percentage indicates potential contamination from outside pollen sources. The nearest Sitka spruce forest to the seed orchard is 1.1 kilometres away in the prevailing wind pathway. It is known that pollen can travel tens of kilometres making pollen contamination highly likely (Ebrahimi et al., 2018; Kremer et al., 2012; O'Connell et al., 2007). To elucidate the source of potential male contamination, offspring with known maternity could be sampled and potential sources of contamination (other Sitka spruce forests) could also be sampled and genotyped using this KASP assay. Maternity could easily be evaluated by collecting cones from specific trees.

Our results indicate the presence of over contributing parental genotypes which over contribute to the production of the offspring population. This term relates to the frequency of parental genotypes occurring in the genotyped offspring set. 71 of the offspring were assigned to parent S43, and 60 of the offspring being assigned to parent S578. The first reason why this may have occurred is seed orchard design. The parental genotype S43 occurs 12 times in the seed orchard whereas many parental genotypes only occur once. This irregularity in frequency of parental genotypes in the seed orchard is one of the contributing factors to over contributing parental genotypes. The design of this seed orchard followed no particular pattern. There are parents of the same genotype occurring side by side and parents overrepresented in the seed orchard. It may have been possible that all parents were planted equally and some had to be rouged but due to lack of clear record keeping this is unknown. Pollen dynamics may be another contributing factor to the detection of over contributing parental genotypes (Ebrahimi et al., 2018). Not all genotypes release the same amount of viable pollen at the same time nor are maternal genotypes pollinated equally. To further understand over contributing parental genotypes in the offspring, pollen viability

and cone timing tests could be done on this seed orchard over a number of years (Flores-Rentería et al., 2018). Wind speed and direction also plays a critical role at the time of pollen release. The results here indicate that some of the seed stock originates from improved parents, however having dominant parents could lead to inbreeding overtime if these offspring are used in breeding. It can also concluded that modern seed orchard designs should now implement both genetic distance, prevailing wind patterns and pollen dynamics into seed orchard designs. A study by Wang *et al.*, (2003) revealed fertility variation in Chinese Fir seed orchards has a significant effect on parental makeup of the offspring. In this study flowering phenology and fertility variation were measured. In future work, new seed orchard designs could account for this variation. By monitoring these attributes and structuring these data like a matrix, a new design based of matrix multiplication could be devised. Combining a flowering matrix, a fertility variation matrix and a genetic distance matrix, a spatial matrix could be devised that would give the optimal placement of each tree in a seed orchard.

Clonal seed orchards are very effective in the bulk production of seed, however they are not guaranteed to produce improved material. Furthermore these data have shown that the bulked progeny material may not always be from a diverse range of genotypes, either due to design or pollen dynamics. This however may not apply to all seed orchards or all seed orchard design types. The design of the seed orchard here seems to just focused on randomisation and not distance between clones, however a specific design was not taken into consideration. This lack of design resulted in closely related or the same parent being adjacent to each other. This likely resulted in higher rates of selfing. Care must be taken in the layout and management of seed orchards, especially when rouging or replacing dead parents. Detailed records on design and management should be kept. Further studies should focus on more types of seed orchards to see if dominant parental genotypes persist. If it

does, pollen dynamics of the progeny should be studied by for example using simple measures such as cone count and cone timing. Understanding these data would allow for the implementation of weighted distribution of progeny based on pollen and cone production (Giertych, 1975). Position of a progeny could be determined by a factor of genetic distance and amount of potential pollen production.

One pitfall of this study was the inability to collect material directly from parents. This resulted in the maternal genotype being unknown. By knowing the maternal genotype, software like CERVUS and apparentR can more accurately predict paternal genotype. This reduces the uncertainty of the results. More data could have been collected, including cone phenology, pollen dynamics and pollen release date. This could lead to better insights on why some parents are over contributing to offspring.

The KASP marker panel has only been tested on the progeny of a single seed orchard. It would be desirable to test it also on the progeny of another seed orchard and across seedlings from multiple cone years too. The results presented here indicate more work needs to be done to design, manage and monitor Irish seed orchards.

5.5 Conclusions

The dynamics of the seed orchard have been presented here including outcrossing, inbreeding and potential contamination. In this seed orchard there are over contributing parental genotypes that may have occurred because of design choices and pollen dynamics. New designs should be implemented to account for the pollen dynamics of the progeny. This assay can be used for the evaluation of Irish seed orchards. This cost effective solution can allow for continual monitoring of seed orchards and the improved deployment of seed orchards.

Chapter 6- General Discussion

6.1 Discussion

SNP sequencing forms the backbone of this thesis and serves as a tool for exploration of Sitka spruce populations. This study focuses on two populations, namely the IUFRO population and the ISSTIP population. The IUFRO population represents the entire native range of Sitka spruce (O'Driscoll, 1978). Our genotyping of this population represents the largest database of genotype data available for Sitka spruce. Genetic data has been collected from Sitka spruce from 7 sampling sites across the native range in a previous study (Gapare et al., 2005). SNP sequencing has become cheaper and more efficient, with higher marker density's available. Structural variants are becoming more popular, however, and this may overtake SNPs as sequencing gets cheaper. For the purpose of this study a high density SNP set is more than adequate. The presence of the IUFRO collection all in one place allowed for ease of sampling. The development of semi-automated DNA extraction systems allowed for hundreds of samples to be processed in a short space of time. Advancements in genotyping technologies have reduced the costs of genotyping for large scale projects such as this. Genotyping such a large population has given us insights into its evolution and adaptation. The genotyping of the IUFRO population also creates a baseline of population genetics to compare other species and other sub populations like the ISSTIP population.

Genotyping these populations opens numerous avenues to study Sitka spruce in this thesis and outside of this thesis. In this thesis genotyping information has been used to understand the population structure and diversity of the native range of the species. Genotyping data has allowed for the calculation of multiple genetic diversity parameters. The number of SNPs discovered by GBS allows for greater accuracy in these calculations as compared to lower SNP counts (Turakulov & Easteal, 2003). Understanding this baseline diversity is important so populations can be compared (Chapter 4). Monitoring

genetic diversity of the ISSTIP population is important in order to reduce the likelihood of inbreeding depression, assist in breeding programs and to aid in genomic selection/prediction amongst other reasons (Williams & Savolainen, 1996). This is especially important for breeders. Understanding population structure and phylogenetics is of interest to conservation scientists, breeders and evolutionary biologists (Pereira-Dias et al., 2019). What population structure reveals is divergence or convergence of populations based on their evolutionary history. This is important from a conservation standpoint to protect structurally distinct trees and maintain diversity of populations. For breeders, adding structurally distinct trees into progeny trials may introduce new traits as is seen in chapter three. Phylogenetics shows the relationships between individuals, provenances and geographic origins (Lockwood et al., 2013).

Breeding of forestry trees requires understanding of the progeny and relationships between them. In this thesis, the breeding population was examined with the context of the IUFRO population behind it. The genotyping of the ISSTIP population allows for the understanding of genetic erosion in early stage breeding populations. Genotyping the ISSTIP population also opens the opportunity for genomic prediction, genomic selection and progeny evaluation (Grattapaglia, 2022; Lebedev et al., 2020). These data are directly being used in parallel with this thesis to preform genomic selection and prediction which is key for marker assisted breeding.

The diversity of both populations has been calculated in chapters two and four. Multiple measures of diversity have been used to fully describe both populations. The IUFRO population ($H_o=0.21$) is more diverse than the ISSTIP population ($H_o=0.13$) however this is due to the greater source range of the IUFRO population being over 3000km as compared to the smaller extent of the ISSTIP range (O'Driscoll, 1978). This can be seen with population differentiation within IUFRO ($F_{st}=0.0292$) being more differentiated than

within ISSTIP ($F_{st}=0.0054$). It is difficult to compare these figures to other conifers because observed heterozygosity is a function of marker count and population size, and our population and number of markers are far larger than other studies. In comparison to work done by Gapare et al. (2005), which used SSRs and less dense markers, Sitka spruce diversity ranged from $H_o= 0.57$ to 0.45, this study shows Sitka spruce being less diverse. This difference is likely due to molecular marker count and type rather than population size, however the patterns seen in this study are comparable to what is seen in chapter 2. Greater diversity in a population can lead to adaptation.

The structure of the populations has been determined in three ways, with clustering dendrograms, with PCA and with DAPC. The dendrogram structure of the population is similar to what is seen in Gapare et al. (2005). The islands Kodiak and Montague appear to be isolated from the mainland populations especially Kodiak. This is confirmed by the PCA and DAPC. The cause of this isolation is likely due to the distance to the core mainland population. This isolation could also be caused by the prevailing wind in the region being North-westerly, reducing the spread of pollen to northern areas. When looking at the admixture of the entire IUFRO population, there is a correlation with the patterns of glacial retreat from the Pleistocene glaciation (Menounos et al., 2017). The pattern of glaciation shows expansion and then retreat at three fronts. On the first front, glaciation patterns moved and retreated from along Alaska, extending as far as Kodiak Island. Secondly, the ice sheet stretched across British Columbia, covering the islands of the Alexander archipelago including Haida Gwaii. On the final front, the ice sheet extended into Washington. This could indicate refuges for Sitka, in the south and on islands. As the glaciers retreated there may have been recolonization from the islands, principally Kodiak and Haida Gwaii and from the southern range. This is reflected in the modern day population structure where there is a differentiation in the islands that are significantly

distanced from the coast. This is indicative of a founder effect from these regions. This theory is missing appropriate fossil records to back it up. So far no studies on fossilised Sitka spruce or pollen studies have been completed. Another prevailing hypothesis is that Sitka spruce could have recolonised for Nunutaks, which are refuges on mountains above glaciers. There is no evidence for this happening in conifers. In fact most modern day Nunutaks seem devoid of life. In some cases smaller cold weather adapted plants can survive on Nunutaks but it is unlikely that these mountain tops could support trees (Schönswetter and Schneeweiss, 2019).

Diversity and structural evolution accumulate to drive adaptation of the population to its host environment or ecological niche (Prunier et al., 2016). There are many parameters effecting adaptation, but at its core adaptation is driven by positive natural selection (Prunier et al., 2016). Natural selection acts on genetic variation of a population, favouring individuals with traits that increase survivability and reproduction. In this study, changes in minor allele frequency (MAF) between populations have been captured. In other studies these mild to moderate shifts in MAF are characteristic of recent local adaptation (De La Torre et al., 2019; Hornoy et al., 2015). These shifts in MAF represent a new way of studying selection that is more precise than other methods such as selective sweeps (Burke, 2012). Mutation is also a driver of adaptation. In chapter three genome wide association studies (GWAS), genotype environment analysis (GEA) and canonical correlations (CANCOR) were used to detect mutations in the form of correlated loci. Conifers are known to have slow mutation rates but it is not unexpected to find so many associated loci in chapter three (Buschiazzo et al., 2012). The GWAS and GEA studies use similar techniques however the CANCOR analysis is a new technique which has only been previously used on perennial ryegrass. The algorithm is robust but the loci discovered to be correlated with traits have not been tested to be adaptive. While the algorithm tests for false

discovery rates, without phenotypic testing it cannot be clear that correlated loci will lead to adaptation. However CANCOR analysis is still very useful for combining data.

Understanding adaptation is key for the breeding of new traits. Phenotypic trials are the primary method of discovering new traits. Multiple provenances and origins of the trees of interest are planted in the same environment and then phenotyped (Chen et al., 2021; Hamilton et al., 2013; Skråppa & Steffenrem, 2021). Depending on the trait of interest, different parameters can be taken. In our study, height and diameter at breast height, were the only two phenotypic traits collected due to the large population, however more traits can be collected from the IUFRO population (Zarzosa et al., 2021). Traits like wood quality and growth phenology can also be of importance to breeding. When traits of interest are discovered, they can be implemented into breeding programmes through traditional breeding or molecular marker assisted breeding (He et al., 2014). Markers associated with traits of interest can be discovered through association studies, principally GWAS and GEA.

Conifers occupying colder climates are well adapted to survive harsh winter conditions. A combination of physiological and morphological strategies are employed by conifers to cope with harsh wintering conditions. One such adaptation is the ability to enter a state of dormancy to conserve energy and resources (Sebastian-Azcona et al., 2018). Conifers enter dormancy by changing metabolic states or through needle shedding. In chapter three there is a reduction in height associated with wintering conditions and snowfall. It may be that the Sitka spruce in Northern regions, such as Alaska, are more prone to enter dormancy, reducing overall height. There are clusters of loci associated with milder more favourable growing conditions. These clusters of loci are positively associated with solar radiation, precipitation and temperature. The loci involved with increases in height were distributed south of the 50th N latitude, showing this area should be the focus

of breeding, if height was a trait of interest. In chapter four, Haida Gwaii is included in the breeding programme, which according to chapter three has loci correlated with a reduction in height. Removing Haida Gwaii may increase height, but it would reduce the genetic variation of the breeding population. However, breeding programs focusing on just one trait can be detrimental. Many commercially important traits are polygenic and can be linked to other traits. Any breeding program needs to balance improving one trait without the loss of others. This is incredibly difficult to do as important traits will change with climate change and consumer needs. Replacing Haida Gwaii with provenances from California could be more favourable and would also reduce the impact of drier weather due to climate change.

The design of seed orchards has centred around three factors; randomisation, replication and distance. These early designs often only specified that ramets of the same clone could not be adjacent to each other, to reduce inbreeding. This approach to design maximised production and ease of management over the quality of offspring. With the aim of improving offspring quality and reducing inbreeding, the alternative block design aims to segregate parents by distance. Some designs, such as rotating block designs can fail to achieve this due to the arrangement of ramets on the edge of the blocks. Other designs such as cyclical balanced incomplete block design use systematic changes in each imbalanced block to guarantee ramets of the same clone are not adjacent. With the increased availability of computational power, the design was permuted to maximise the distance between ramets of the same clone. This eventually led to designs like Minimum Inbreeding which optimises the genetic distance between clones. Here the advent of cheaply available genotyping can be used to improve design and monitor orchards. With genetic distance based designs, the physical distance between ramets of the same clone and ramets of related clones is necessary, but a global physical distance of the entire orchard is necessary to

reduce contamination. Reduction of contamination is important in seed orchards. This can be done by using indoor seed orchards or locating the seed orchard away from other forests of the same species. While genotyping offers a solution for inbreeding and monitoring of contamination, differences in reproductive success still reduce the orchards overall efficiency. Reproductive phenology differs from tree to tree. Some trees from differing provenances may release pollen earlier. Orchards can monitor this now with genotyping, but no design solution has been proposed. The systematic application of hormones or alternative techniques to delay pollen production in over dominant genotypes could be used to balance parental contribution, however this is time consuming and resource intensive. Genotyping may produce a solution however, by treating reproductive success as a trait and selecting a population based on similarities in reproductive success, orchards could be balanced. This would add another selection criteria for forestry breeding, however with investment into GWAS this could be achieved, which could usher in a new design in seed orchards which balances parental contribution. A design solution to the problem of over contributing parents could be to create a spatial matrix design based of a relationship matrix and a phenology matrix. The relationship matrix could be derived from genetic data and the phenology matrix can be derived from phenotypic data.

One of the main limitations to this study is the lack of records kept about the populations and trials. The IUFRO collection has the best record keep yet there are still gaps in the knowledge. It is known when and where the seed was collected however it is not clear if the seed represents a pooled average from multiple trees in the area or one specific tree. This is especially concerning when very heavily planted zones are taken into account, as some seed may be imported. It is also not clear if this seed was taken from old growth forests, which would represent the native range, or from plantations. Sampling around the introgression zone is particularly tricky as some of the trees may be White

spruce or Lutz's spruce. Information from the ISSTIP population is sparse as it is a consolidation of different trials and research done by different parties. Data are lost because of this and some of the origins are unknown in chapter 4, a genomic relationship matrix was used to interpret some of the unknowns, however genetic data is not a substitute for good record keeping. The seed orchard sampled also has some issues due to no record keeping on the design and management. While setting up these trials and populations was not the responsibility of this research thesis, poor record keeping has affected some of the outcomes. One outcome of this thesis is to improve on record keeping by making genotyping data publicly available. Future trials should be properly recorded using digital methods, data should be backed up and updated every time new software versions are released. This will help with the continuity of data. Forestry organizations have a responsibility to safeguard and maintain datasets and records.

6.2 Conclusions and future work.

Two major populations have been genotyped in this thesis. The IUFRO population was used to reveal insights into the diversity and structure of the population. There is a correlation between genetic data and the extent of the ice sheet that suggests that after the retreat of the cordilleran ice sheet, Sitka spruce recolonised from the south, Haida Gwaii and Kodiak Island. The IUFRO population has been used to study the adaptation of Sitka spruce. There are potential local adaptation is characterised by shifts in MAF as is seen in other conifers. Loci correlated with increases of height have been discovered and they are localised to south of the 50th N latitude. The ISSTIP population was compared to the IUFRO population and revealed show early stage genetic erosion due to breeding.

The IUFRO population shows evolutionary trends in Sitka spruce, however this is just one Spruce species. Using the same genotyping methods, other North American spruces and Asiatic spruces could be genotyped cheaply and compared to create a phylogenetic tree. This will give more detailed studies on the spread of spruces. The arboretum where the IUFRO population also contains multiple of these North American and Asiatic spruces. Sampling White spruce, Lutz's spruce and Engelmann spruce could also reveal details regarding the hybrid zone. The CANCOR analysis is expandable and can be changed to take in new traits. A parallel study is investigating the adaptability of the IUFRO genotypes to climate change, so merging these data with the genotype data could give more detail on correlated loci. Genotype data available for the ISSTIP population has allowed for parallel studies focusing on genomic prediction and genomic selection. Regarding seed orchard studies, the other Irish seed orchard could be genotyped but collection strategy could be altered to know the maternal genotypes. Future work on seed orchards should also consider differences in reproductive phenology.

Appendices

Appendix 1: Introduction

The appendices outline work that was attempted but not completed due to time constraints or not fitting into the scope of the thesis. The first section outlines the set up for SNPseq, which is a targeted method of sequencing SNPs. This work was going to be the baseline for this thesis but due to the price drop in GBS it was determined that GBS was the better option. The next piece of work was Isoform sequencing of the Sitka spruce transcriptome. This was going to be used as a draft assembly as it is cheaper than Whole Genome Sequencing (WGS) and more accurate than SNP clustering. However the genome assembly of Sitka spruce became available during the project making this assembly redundant. The last piece of work was originally to be used as a genotyping platform for the seed orchard. While this assay worked *in silico* it did not work in practice. Due to time constraints it was determined that KASP was a better alternative. The other two sections in the appendices detail data that did not fit into the chapters either due to its reduced necessity or its length.

Appendix 2: Developing SNPseq targets

Introduction

SNPseq is a targeted high throughput approach to Single Nucleotide Polymorphism (SNP) genotyping based on Genotyping in Thousands (GT) seq (Campbell et al., 2015). This is a PCR based assay that targets preselected regions for genotyping. A first round PCR amplifies the pre-defined targets and then the PCR products are enriched. Barcodes are added to the PCR products in a second round of PCR. The libraries are then prepped and sequenced. The genotypes can then be called using custom tools. SNPseq is often used as

an alternative for GBS (Elshire et al., 2011). It is cheaper than GBS, the targets can be specified and the data can be processed faster. However, it requires a good reference genome, it has a lower marker density and it often requires prior knowledge of common SNP positions. Targeted SNPseq is a lower density genotyping method that can be used to replace KASP markers (competitive allele specific PCR) or SSR genotyping (Zhang et al., 2020).

A SNP panel has been developed that can be used to genotype a diverse range of Sitka spruce, the IUFRO population, using SNPseq. A GBS pilot study was employed to identify SNPs that can be filtered into the SNP panel (Elshire et al., 2011). The GBS was also used to indicate any rare alleles or markers from peripheral populations which should be conserved in the SNP panel. The panel was originally anticipated to be used on the 1177 Sitka spruce of the IUFRO population to genotype and be used in genetic diversity studies but was not completed as GBS became cheaper.

GBS Pilot Study

Due to the high canopy of this forest plantation at JFK Arboretum (~50m), standard needle sampling was not achievable/practical. Sampling of the cambium tissue was therefore used as a suitable alternative to leaves. Sampling occurred on two occasions, on the 4th September 2019 where 1 tree from each provenance was sampled and during the last week in May 2020 where every tree was sampled. For sampling, the outer layer of bark was scraped off and any lichen or moss was removed to prevent cross contamination. A cork borer was then pressed through the bark about 1-5 cm in depth until the borer reached the sapwood. The sample was then removed and stored in a 50ml centrifuge tube. It was kept at -20°C until it was processed by separating the cambium layer from the bark, freeze drying and grinding to a fine powder using tungsten beads. Samples were extracted using

Machery and Nagel NucleoMag DNA extraction which was modified to include a longer lysis time of 3 hours, a stronger magnetic field for the bead binding steps and a double elution step. The samples were quantified using Promega Quant-it PicoGreen kit and a Biotek Plate Reader, normalised and stored at -20°C.

44 of the samples that were extracted in September 2019 were genotyped using GBS with PstI-ApeK1 restriction enzymes (Elshire et al. 2011) by LGC group, Germany. The resulting data was demultiplexed, adapters were clipped and the resulting sequence aligned to the Sitka spruce Q903_v1_1000plus genome using BWA (Li and Durbin 2009). Mpileup and BCFtools were used to remove indels, keep biallelic SNPs only and restrict minor allele frequency (MAF) from 0.01 to 0.975 (Danecek et al. 2011; Narasimhan et al. 2016). This resulted in 294,692 variants which were explored further for filtering.

Marker selection

A matrix of SNPs was assembled to allow for the selection of unique and common SNPs. Only one SNP per scaffold was selected. Unique SNPs per sample were extracted and a maximum of 190 unique SNPs per sample were extracted. This resulted in a total of 7212 SNPs making up 72% of the total panel. SNPs unique to populations were investigated. SNPs common to samples were included. This included 348 common to 100% (heterozygous amongst all 44 samples) of the samples, 679 common to 100>90% of the samples and 812 common to 90>80% of the samples. This totalled 9148 SNPs out of a 10K panel. The panel, captures unique SNPs from each region which should allow for detailed diversity analysis. The inclusion of the common variants is useful to look at shared traits and compare populations. However, there are very few population specific variants available to include. Larger sample sizes will be needed to investigate population structure.

To fill the rest of the panel, the 58 parent samples of the seed orchard were processed exactly as the IUFRO population and selected based on common and unique samples. 10 SNPs unique to each parent were chosen to total 580 SNPs. There were 272 SNPs selected common to all parents. The IUFRO and seed orchard SNP panels were compared removing overlapping SNPs. This resulted in a 10K SNP panel that 91.48% represents the IUFRO population and 8.52% represents the ISSTIP seed orchard parents (Table A1.1).

Table A1.1. SNPs selected to create a 10K panel. Common 100% = Heterozygous SNPs in all samples

	Selected	Total variants available	Representation
IUFRO			91.48%
Unique	~190 per sample 7211	59351 (over 44 samples)	~1.9% per sample 72.1%
Common 100%	348	348	3.48%
Common 100>90%	679	679	6.79%
Common 90>80%	812	857	8.57%
IUFRO populations			
Skeena	46	46 (3 samples)	0.43%
Oregon	1	1 (4 samples)	0.01%
North Alaska	0	0 (8 samples)	
South Alaska	0	0 (4 samples)	
Mainland BC	0	0 (3 samples)	
QCI	0	0 (4 samples)	
Skeena/Naas	0	0 (5 samples)	
Vancouver	0	0 (5 samples)	
Washington	0	0 (7 samples)	
ISSTIP			
Unique	10 per sample 580	18096 (58 samples)	0.1% per sample 5.8%
Common	272	312	2.72%

Reason for not completing

While the SNP panel would have been effective, the price per sample of SNPseq had risen to 45 euros per sample for a 10K SNP panel. This is in comparison to GBS which was at 60 euro per sample at the time. GBS has a much higher density and is less biased than pre choosing SNPs, even if the SNPs were chosen carefully. For this reason GBS was used on the IUFRO population and the ISSTIP population.

Appendix 3: Developing PCR targets for genotyping Sitka spruce.

Introduction

Multi Allelic Scanning Haplotag (MASH) assay, is a relatively new technology to identify SNPs for a low cost (Leyva-Pérez et al., 2022). This assay is based on Genotyping in Thousands sequencing (GT-seq) (Campbell et al., 2015). The process involves multiplex PCR to amplify targets. Amplified samples undergo a second round of PCR to attach well specific barcodes. The PCR products are then purified and undergo size selection. This study attempted to develop SpruceMASH, a custom MASH assay for genotyping Sitka spruce.

Materials and Methods

Primer design

Primers were designed based off GBS data from the Irish Sitka Spruce Tree Improvement Program (ISSTIP) genotyped in chapter four. SNP density was calculated using VCFtools (version 0.1.16) for every 500 bp and regions with a density of 7-30 SNPs per 500 bp were selected. One region per scaffold was selected. Regions with a high SNP density that were 90-150 bp in size were selected, regions with conserved sequences across all samples were

selected and regions that occur at the start or end of the scaffold were removed. This resulted in a set of high density SNPs within a 90-150 bp window, which were flanked by conserved sequences. These regions were run through primer3plus to create a product size of 150-250 nucleotides with a primer size of 15-35 nucleotides, a melting temperature of 60-65°C and a GC content of 40-60% (Untergasser et al., 2007). Primerpooler was run to check for interactions between these primers and any primers that could not be multiplexed into a single pool were removed (Brown et al., 2017). Primerpooler was run where the lowest annealing temperature was 57°C, Magnesium was 3mM, ΔG was -6, and maximum amplicon length was 2000bp. Sequence tags were added with the R1 Illumina tag being added to the 5 prime end of the forward primer (TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG - FW primer) and the R2 Illumina tag being added to the 5 prime end of the reverse primer (GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG-REV primer). Once the tags were added the primers were rerun on primerpooler using the same settings. The primers ranged from 54-60 nt in length. Primer pairs were ordered from Integrated DNA technologies (IDT, USA) at 500 μ M each.

Well and plate specific barcodes were also ordered. Well specific forward primers were made up of an Illumina P5 adapter tag, a unique 6nt barcode and the first 14 bases of the R1 tag (AATGATACGGCGACCACCGAGATCTACAC-i5-TCGTCGGCAGCGTC). The plate specific reverse primers were made up of the Illumina P7 tag, a unique 6nt barcode and the first 15 bases of the R2 tag (CAAGCAGAAGACGGCATACGAGAT-reverse complementary sequence of i7 index - GTCTCGTGGGCTCGG). These were ordered at 100 μ M each and diluted to a working concentration of 10 μ M.

Genotyping panel

To test the efficacy of this assay, a diverse panel was used to genotype. 228 of these samples were offspring from seed orchards, 54 samples were from a diverse IUFRO range, three samples were from *Picea x lutzii* and three samples were from *Picea glauca*. The addition of *Picea x lutzii* and *Picea glauca* was for SNP comparisons on species. DNA was extracted using Marchery and Nagel NucleoMag DNA extraction which was modified to include a longer lysis time of three hours and a double elution step. The samples were quantified using Promega Quant-it PicoGreen kit and a Biotek Plate Reader and normalised to 40ng/μl.

MASH library construction

Library construction followed the protocol (<https://www.protocols.io/view/potatomash-library-construction-e6nvw53zdvmk/v1>) with some minor changes. The first PCR was used to amplify targets using a PCR cocktail and conditions as specified in the protocol. PCR 2 was used to add barcodes and was carried out as specified in the protocol. Samples could then be pooled and undergo PCR clean-up and gel size extraction as per protocol.

Troubleshooting

When running the PCR product it was found that the size of the band was at 150 bp rather than the expected 300 bp. It was found that primer dimers were forming during PCR 1 and then the barcodes were adding onto the primer dimer. To troubleshoot this, multiple parameters were tested as shown in Table A2.1. Primer concentration was varied in order to reduce primer dimerization and encourage primer binding. For PCR 2, the volume of PCR 1 product was varied. Initial DNA concentration was varied to reduce PCR inhibition or to increase the number of genome copies. Alternative barcodes that were ordered were also tested. PCR1 conditions were also varied where PCR 1: 98 °C 8 min; 8 cycles × (98

°C 1 min, 0.2 °C/s ramp down to 56 °C annealing 30 s, 72 °C 2 min); 16 cycles × (95 °C 30 s, 65 °C 30 s, 72 °C 30 s); 10 °C hold.

Table A2.1. Parameters tested for GenoSpruce assay.

Parameter	Original	Changes
Primer concentration	25nM	10nM, 5nM, 2.5nM
PCR1 volume used in PCR2	1/15	1/5, 1/10, 1/20, 1/30
DNA concentration (ng/μl)	40	10,20,30,50,60
Barcodes	Original Barcodes	Alternative Barcodes

Results and Discussion

The main issue was the formation of primer dimer after PCR1. *In-silico* analysis reveals no primer dimerization should theoretically be forming. However laboratory testing shows the presence of primer dimer. Judging by the lengths of the fragments cross primer dimerization is occurring. This can generally be caused due to contamination, to low annealing temperature or primer design. The primer design passed *in-silico* testing and annealing temperature was changed, so the problem may be an issue with the template DNA. Phenolic compounds and polysaccharides, which are found in high quantities in conifer species, are known inhibitors of PCRs (Schrader et al., 2012). Phenolic compounds can cross-link with nucleic acids changing their chemical properties, thus inhibiting DNA extraction or PCRs. This may be one of the reasons why such primer dimer is seen.

Common methods of dealing with PCR inhibitors include longer lysis times, proteinase K treatment and RNAase all of which have been applied to this DNA (Schrader et al., 2012).

Reason for not completing

The library preparation could not be completed due to primer dimer and time constraints.

Appendix 4: Isoform sequencing

Introduction

The majority of genes in higher eukaryotes are alternatively spliced. This allows for isoforms of the gene to be produced. Isoforms are formations of different transcripts from the same gene. An example of this is in the human genome where Bcl-xL inhibits cell death but Bcl-xs activates cell death (Stevens & Oltean, 2019). RNA-seq allows for short read sequencing of transcripts. Short read technologies are insufficient for understanding how exons are joined together, hence there is isoform uncertainty. Iso Seq gives full cDNA sequences from poly-A to 5' with no isoform uncertainty (S.-Y. Chen et al., 2017; Wang et al., 2016). Iso Seq process starts with mRNA, either PolyA selected or Total RNA. It is reversed transcribed to cDNA, size portioned and PCR amplified. cDNA synthesis is done using Clontech SMARTER PCR cDNA synthesis kit. Size selection allows for longer FLcDNAs. SMRTbell ligation attaches hairpins to the ends of the read of interest. Sequencing on the Pac Bio starts on the hairpins and the seq polymerase goes around the molecule multiple times until a consensus sequence (read of interest) is generated. Full length cDNA are sequence primer and PolyA framed. Reads of interests are then run through a custom bioinformatics pipeline.

Here Multithroughput Iso-seq was used to assemble a transcriptome of Sitka spruce. This transcriptome was originally to be used for the alignment of GBS reads (Ali et al., 2021). However the genome of Sitka spruce was made available before this was needed.

Materials and Methods

Plant material and RNA extraction

Needle samples from three trees of Washington origin were taken and flash frozen in liquid nitrogen. Samples were pooled and RNA was extracted using the Qiagen RNeasy Plant Mini Kit (74904). The RNA samples were checked on an AATI Fragment analyser using the RNA HS Kit (DNF-472-0500). Samples were shipped to BGI Hong Kong where they underwent multithroughput Iso-Seq and transcriptome assembly using BGIs custom pipelines.

Quality controls

To search for potential overlapping reads in the assembly, minimap2 (version 2.15) was used using the `ava-pb` function (Li, 2018). Contaminants were checked for using mash (version 2.1) `screen` function against common plasmids and genomes (Ondov et al., 2019). Benchmarking Universal Single-Copy Orthologue (BUSCO) was run using a custom python script `run_BUSCO` (version 3.0.1)(Manni et al., 2021).

Results and Discussion

A total of 512,066 reads were sequenced giving a total of 27.62 G bases. The max length of assembled reads were 277,532 bp, the mean length was 53.937.70 bp and the N50 length was 113,114. The distribution of read lengths is shown in Figure A3.1. There was 14.1% complete BUSCOs, 18% fragmented and 67.9% missing. BUSCOs represent the completeness of the assembly and the high degree of missingness in this assembly is

common with what is seen in spruce genome assemblies. In other spruce genome assembly's single copy BUSCO's range from 29.1% to 41.1%. There were no significant contaminants found in this assembly. This assembly is created with long read transcriptomics and can be compared to other transcriptome assemblies of spruce to discover isoforms of interest.

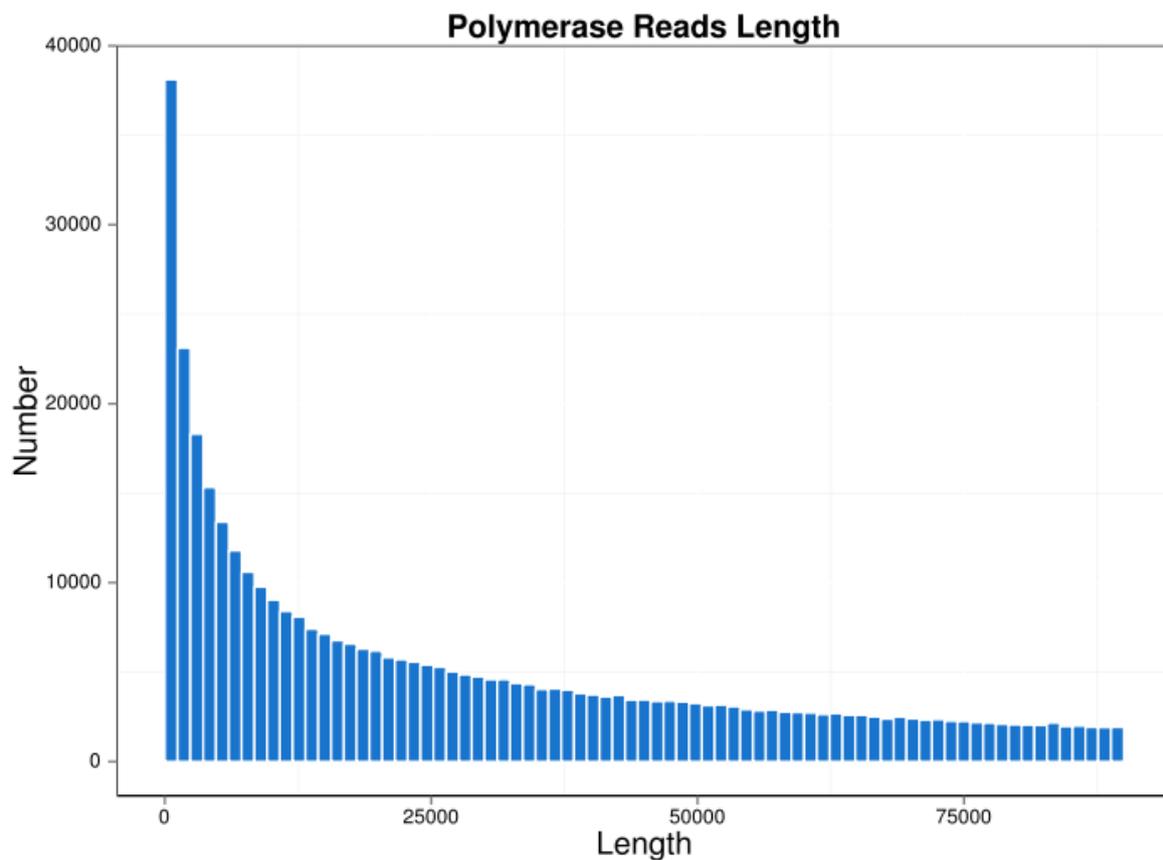


Figure A3.1 Frequency of reads of differing lengths.

Reason for not using in thesis

By the time our sequencing was completed the Sitka spruce genome had been published making GBS alignment much easier.

Appendix 5: Additional material from chapter 2

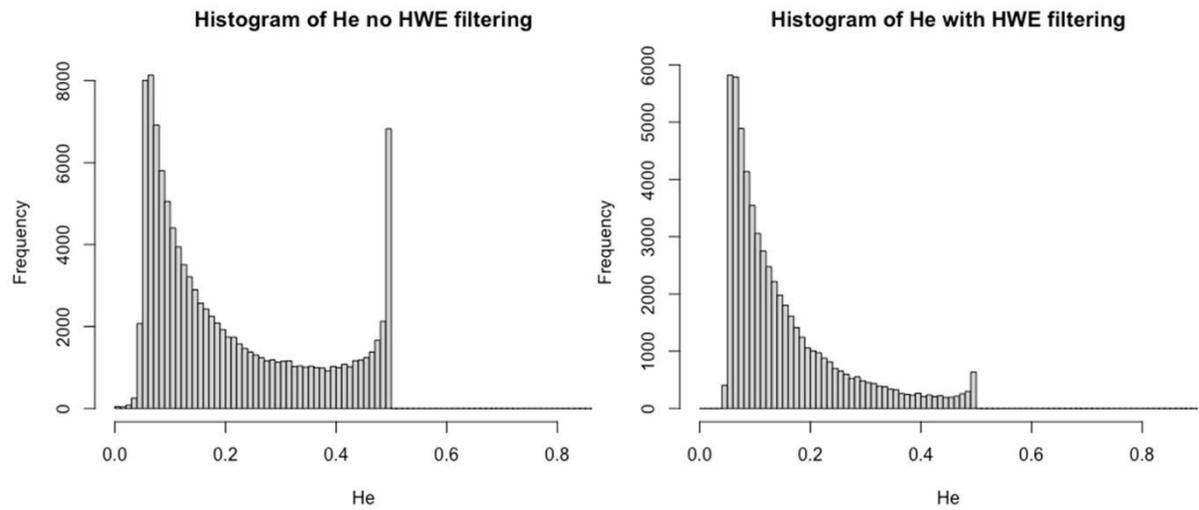


Figure A5.1. Gain of Heterozygosity (GoHe) associated with sequencing based errors causes deviations from Hardy Weinberg Equilibrium. (A) The GoHe in the IUFRO population without HWE filters and (B) elimination of GoHe with filtering on HWE ($p > 0.0001$).

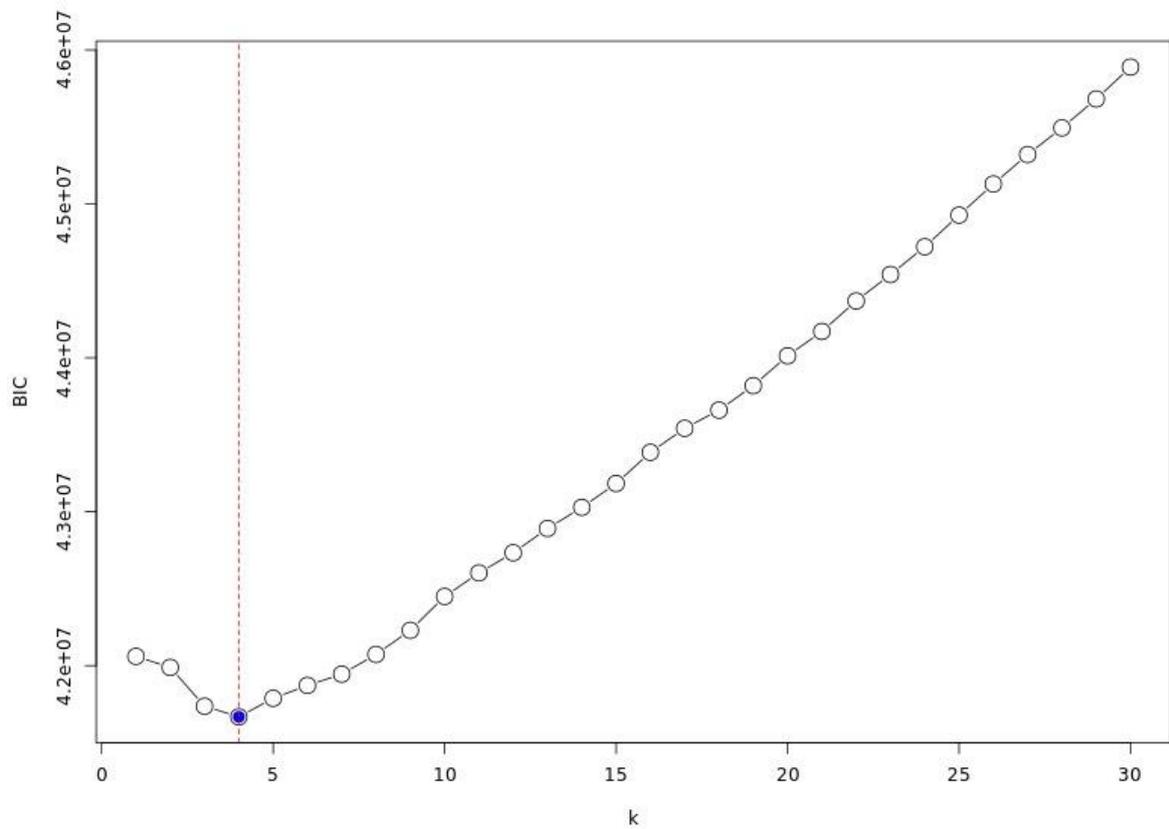


Figure A5.2. Optimal clusters as determined by Bayesian Information Criterion (BIC). Number of clusters within the population (K) was determined from a sample size of 1177 using 36567 SNPs for Bayesian clustering.

Table A5.1. Provenances and their respective geographic regions. Provenances are ordered from North to South based off their latitudinal coordinates. Coordinates recorded in decimal.

Provenance	IUFRO_NO	LatDMS	LongDMS	Elevation (m)
Valdez	3074	61.1167	-146.23	30
Eyak Lake	3077	60.5500	-145.67	6
Cordova	3076	60.5500	-145.75	30
Kenai	3075	60.5167	-151.27	18
Sheridan Glacier	3078	60.5167	-145.33	24
Seward	3079	60.1667	-149.53	30
Haning Bay	3080	59.9667	-147.72	22
McLeod Harbor	3081	59.8833	-147.83	10
Yakutat	3021	59.5167	-139.70	12
Dyea	3022	58.5000	-135.35	1
Kodiak Island	3082	58.7667	-152.42	106
Eagle River	3023	58.5167	-134.80	15
Duck Creek	3024	58.3667	-134.63	30
Afognack Island	3084	56.1000	-152.88	1
Ohmer Creek	3025	55.5833	-128.73	15
Derrick Lake	3026	55.6833	-133.68	224
Craig	3027	55.5000	-128.13	1
Old Hollis	3028	55.4667	-131.67	1
Cranberry River	3029	55.4667	-128.23	518
Ward Lake	3030	55.4167	-127.70	15
Dragon Lake	3031	55.3500	-129.95	259
Kitwanga	3032	55.1667	-128.87	671
Zolap Creek	3033	55.1500	-131.22	15
Fulmar Creek	3034	55.1500	-128.97	396
Moss Point	3035	55.0333	-131.55	1
Cedarvale	3036	55.0167	-128.30	244
Vesach Creek	3037	54.8000	-128.72	366
Pacific	3038	54.7667	-128.25	107
Kitsum Kalum	3039	54.7167	-128.77	137
Usk Ferry	3040	54.6333	-128.40	137
Shames	3041	54.4000	-128.95	30
Kasiks River	3042	54.2833	-129.42	30
Inverness	3044	54.2000	-130.25	30
Aberdeen	3045	54.2000	-129.92	1
Weden River	3046	54.0500	-128.62	168

Humpback Creek	3047	54.0333	-130.37	305
Chittenden Point	3068	53.9500	-132.60	30
Masset Sound	3048	53.9167	-132.08	1
Ain River	3069	53.7333	-132.42	30
Dinan Bay	3070	53.6667	-132.67	30
Tlell	3071	53.5833	-131.93	61
Link Road	3049	53.5000	-132.17	91
Copper Creek	3050	53.1333	-131.80	76
Moresby Camp	3051	53.0500	-132.07	61
Tasu Creek	3052	52.8667	-132.08	15
Jedway	3053	52.2833	-131.22	15
Noeick River	3054	51.7667	-126.83	198
Kilbella river	3072	51.7833	-127.32	61
Holberg	3056	50.6000	-128.08	122
Salmon Bay	3058	50.3833	-125.95	1
Fair Harbor	3059	50.0500	-127.03	30
Squamish River	3060	49.8833	-123.25	30
Thasis Inlet	3061	49.8333	-126.67	1
Big Qualicum River	3062	49.3833	-124.62	1
Haney	3063	49.2667	-122.60	305
Vedder	3064	49.1167	-121.93	30
Blenhiem Mountain	3073	48.9000	-124.95	244
Bellingham	3001	48.7500	-122.63	30
Port Renfrew	3065	48.5833	-124.40	15
Muir Creek	3066	48.3833	-123.87	1
Port Angeles	3002	48.1500	-123.73	107
Stillaguamish road	3067	48.1167	-123.75	366
Forks	3003	48.0667	-121.30	152
Kalaloch	3004	47.7000	-124.42	30
Brinnon	3005	47.7000	-124.88	3
Shelton	3006	47.3500	-122.15	6
Humptulips	3007	47.2333	-123.95	61
Hoquiam	3008	47.0500	-123.05	6
Raymond	3009	46.6833	-124.87	30
Naselle	3010	46.3667	-123.78	15
Astoria	3011	46.2000	-123.97	15
Necanium	3012	45.6667	-123.77	46
Tillamook	3013	45.3333	-123.88	122
Newport	3014	44.7000	-124.07	30
Florence	3015	44.1167	-124.12	152

Denmark	3016	42.8500	-124.45	152
Gold Beach	3017	42.5000	-124.42	30
Brookings	3018	42.2500	-124.38	91
Crescent City	3020	41.6667	-124.18	15
Big Logoon	3019	41.1333	-124.15	15

Appendix 6: Additional material from chapter 5

Table A6.1. Seed orchard map. Numbers represent clone in the seed orchard. Trees were spaced 5 meters apart.

	574		280	43	698	455	190	266	230	150	599	575	577	574	280	43	698	455	190
	577	150	559	575	577	574	280	43	698	445	190	206	578	183	608	527	520	243	588
				600	619	583	595	579	374	291	266	230	150	559	575	577	574	339	600
	575			577	574	280	43	698	455	190	719	547	541	321	217	582	280	226	619
	190	261		575	574	280	43	698	455	190	589	546	595	545	209	580	43	588	583
		266	588	559	577	619	583	595	579	374	291	266	230	150	191	570	689	600	595
190	455	291	226	150	575	600	547	541	321	217	582	580	570	559	719	519	455	619	597
455	698	374		230	559	588	719	226	588	600	619	583	519	575	547	377	190	583	374
190	43	579	339	226	150	226	191	339	577	574	280	595	377	577	541	218	582	595	291
455	280	595	243	291	230	339	209	243	575	190	43	579	218	574	321	587	580	579	266
190	574	583	520	374	266	243	545	520	559	455	689	374	587	280	217	264	570	374	230
455		619	527	579	291	520	595	527	150	230	266	291	264	43	582	219	519	291	150
		600	608	595	374	527	546	608	183	578	206	210	219	689	580	210	377	266	559
			183	583	579	608	589	184	535	225	233	251	190	455	570	206	218	230	575
			578	619	595	183	578	206	210	219	264	587	218	377	519	578	587	150	577
				600	583		619	600	588	226	339	243	520	527	608	183	264	559	574
					588		116	339	243	520	527	608	183	578	206	210	219	575	280
							206	210	219	264	587	190	455	190	455	280	574	577	43
										588	226	339	243	520	527	608	190	455	689
												577	190	455	689	43	280	574	190
														190	574	280	43	689	455

Table A6.2. Targets for KASP assay

LOCUS	ALLELE Y	ALLELE X	SEQUENCE
PS-1R180608S00000518	T	A	CTTACATTGTTTTTTT[A/T]TATCTGGAAGAAA
PS-1R180608S00002395	A	C	TTCACGCCACATATA[C/A]TCAAACATTA ACTG
PS-1R180608S00005263	C	T	CTGCCAAGCAATGTG[T/C]AAGGAAGGA GAATG
PS-1R180608S00007303	T	A	TCGTGCAGATGCTGT[A/T]GCTACAGCAA GAAG
PS-1R180608S00011572	C	A	TTACTGCAGACCAAAA[A/C]TCTGCCCTCA AGAG
PS-1R180608S00015943	G	A	CGAAGGAGCGAGGAT[A/G]GTTTTTCGAA ATTGA
PS-1R180608S00016186	T	G	AAAGTGTTTTTCCTT[G/T]AAGGAGCTGA CATG
PS-1R180608S00025480	C	T	GGCAGACGCCGATCT[T/C]GGTGAATTCC TTCT
PS-1R180608S00037345	A	C	GTCTGCAGTTCACAA[C/A]CTATCTCTGC ACCG
PS-1R180608S00038506	T	C	TCAATTGCTCAATGG[C/T]GCAGGAGTTC GTCC
PS-1R180608S00051678	T	A	TCTTTTATGTAAACA[A/T]TGGAGTATTGT TTA
PS-1R180608S00056477	A	G	AACTTACCATCCAAA[G/A]AGATCATTTA CATT
PS-1R180608S00098651	T	G	AGTGGAAGCAGGGGT[G/T]ATTCAACCTT TGGT
PS-1R180608S00174019	G	A	ATTTACAAAAAGACT[A/G]CGCATAAATT TCCC
PS-1R180608S00182550	G	A	CCAACAGGCTGTCCA[A/G]GAAAAAGGT CTTCA
PS-1R180608S00219415	T	G	TTGGGATGGATGGAA[G/T]ATCAGACAG CATGG

PS-1R180608S00240 604	A	G	TTGGGTGGAAACTGT[G/A]GAGGGACGT TCATG
PS-1R180608S00255 114	T	C	CCGAAAGAAACCATA[C/T]TCATGACACC TAGG
PS-1R180608S00271 298	T	C	TCAACCATGCAAGCA[C/T]AGTGCTCCAT TTTT
PS-1R180608S00274 982	A	C	CCAACGGTGATCTAG[C/A]TTGCCTTGCG TTCG
PS-1R180608S01075 355	G	A	ATTTGCATCTGCATC[A/G]GCACTAACTT GTCC

Table A6.3. Shared identities of offspring discovered in the seed orchard.

First ID	Loci typed	Second ID	Loci typed	Matching loci	Mismatching loci
SO_005	19	SO_006	18	18	0
SO_005	19	SO_008	19	18	0
SO_005	19	SO_019	19	18	0
SO_005	19	SO_020	19	18	0
SO_005	19	SO_021	19	18	0
SO_006	18	SO_007	20	18	0
SO_006	18	SO_019	19	18	0
SO_007	20	SO_012	18	18	0
SO_007	20	SO_019	19	19	0
SO_007	20	SO_024	18	18	0
SO_007	20	SO_217	19	19	0
SO_008	19	SO_009	18	18	0
SO_008	19	SO_020	19	19	0
SO_008	19	SO_021	19	18	0
SO_009	18	SO_020	19	18	0
SO_012	18	SO_019	19	18	0
SO_012	18	SO_024	18	18	0
SO_012	18	SO_193	20	18	0
SO_012	18	SO_205	20	18	0
SO_012	18	SO_217	19	18	0
SO_019	19	SO_024	18	18	0
SO_019	19	SO_217	19	18	0
SO_020	19	SO_021	19	18	0
SO_024	18	SO_193	20	18	0

SO_024	18	SO_205	20	18	0
SO_024	18	SO_217	19	18	0
SO_027	20	SO_262	19	19	0
SO_038	19	SO_133	19	18	0
SO_038	19	SO_206	20	19	0
SO_038	19	SO_218	18	18	0
SO_039	19	SO_177	20	19	0
SO_047	19	SO_208	19	18	0
SO_053	19	SO_169	18	18	0
SO_054	19	SO_087	19	18	0
SO_056	20	SO_057	20	20	0
SO_060	18	SO_260	20	18	0
SO_065	20	SO_068	20	20	0
SO_065	20	SO_069	19	19	0
SO_068	20	SO_069	19	19	0
SO_080	20	SO_081	20	20	0
SO_084	18	SO_166	19	18	0
SO_092	20	SO_093	20	20	0
SO_097	18	SO_098	18	18	0
SO_097	18	SO_109	19	18	0
SO_097	18	SO_193	20	18	0
SO_097	18	SO_205	20	18	0
SO_098	18	SO_109	19	18	0
SO_098	18	SO_193	20	18	0
SO_098	18	SO_205	20	18	0
SO_109	19	SO_193	20	19	0
SO_109	19	SO_194	18	18	0
SO_109	19	SO_205	20	19	0
SO_109	19	SO_217	19	18	0
SO_109	19	SO_241	18	18	0
SO_109	19	SO_251	18	18	0
SO_113	18	SO_168	19	18	0
SO_122	18	SO_206	20	18	0
SO_123	20	SO_267	19	19	0
SO_133	19	SO_206	20	19	0
SO_133	19	SO_252	18	18	0
SO_133	19	SO_264	18	18	0
SO_133	19	SO_276	18	18	0
SO_168	19	SO_251	18	18	0
SO_177	20	SO_194	18	18	0
SO_177	20	SO_241	18	18	0
SO_178	20	SO_284	18	18	0

SO_179	18	SO_193	20	18	0
SO_179	18	SO_205	20	18	0
SO_193	20	SO_194	18	18	0
SO_193	20	SO_205	20	20	0
SO_193	20	SO_217	19	19	0
SO_193	20	SO_241	18	18	0
SO_193	20	SO_251	18	18	0
SO_194	18	SO_195	19	18	0
SO_194	18	SO_196	19	18	0
SO_194	18	SO_197	19	18	0
SO_194	18	SO_198	19	18	0
SO_194	18	SO_205	20	18	0
SO_194	18	SO_241	18	18	0
SO_195	19	SO_196	19	19	0
SO_195	19	SO_197	19	19	0
SO_195	19	SO_198	19	19	0
SO_195	19	SO_200	18	18	0
SO_195	19	SO_241	18	18	0
SO_196	19	SO_197	19	19	0
SO_196	19	SO_198	19	19	0
SO_196	19	SO_200	18	18	0
SO_196	19	SO_241	18	18	0
SO_197	19	SO_198	19	19	0
SO_197	19	SO_200	18	18	0
SO_197	19	SO_241	18	18	0
SO_198	19	SO_200	18	18	0
SO_198	19	SO_241	18	18	0
SO_205	20	SO_217	19	19	0
SO_205	20	SO_241	18	18	0
SO_205	20	SO_251	18	18	0
SO_206	20	SO_210	18	18	0
SO_206	20	SO_218	18	18	0
SO_206	20	SO_252	18	18	0
SO_206	20	SO_264	18	18	0
SO_206	20	SO_276	18	18	0
SO_252	18	SO_264	18	18	0
SO_252	18	SO_276	18	18	0
SO_257	20	SO_273	20	20	0
SO_264	18	SO_276	18	18	0
SO_265	19	SO_271	19	18	0

Table A6.4. Occurrence of offspring from parental genotypes

	S190	S233	S233 .1	S243	S261	S264	S280	S321	S519	S520	S577	S578	S580	S589	S595	S597	S619	S698	S719
S150												2							
S190	17											1					3	10	
S191												2		3					
S218								1						1					
S226				2				2				2		4					
S230				1										1					
S233			45			6													
S243					1	2	1		7	6	1								
S251												4							
S266														1					
S321																1			
S43		9		4				2				42	5	9					
S445												1							
S455														1					
S519														1					
S520												1		1					
S541														1					
S546												1							
S559												1							
S575												3		1					
S577														1					
S578																		1	2
S579														2					
S589																1			1

Bibliography

- A'Hara, S. W., & Cottrell, J. E. (2004). A set of microsatellite markers for use in Sitka spruce (*Picea sitchensis*) developed from *Picea glauca* ESTs. *Molecular Ecology Notes*, 4, 4.
- Agency, B. C. (2020). *Picea sitchensis* (assembly Q903_v1)
- Ahuja, M. R., & Neale, D. B. (2005). Evolution of Genome Size in Conifers. *Silvae Genetica*, 54(1-6), 126-137. <https://doi.org/doi:10.1515/sg-2005-0020>
- Alexander, D. H., & Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, 12, 246. <https://doi.org/10.1186/1471-2105-12-246>
- Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9), 1655-1664. <https://doi.org/10.1101/gr.094052.109>
- Ali, A., Thorgaard, G. H., & Salem, M. (2021). Pacbio iso-seq improves the rainbow trout genome annotation and identifies alternative splicing associated with economically important phenotypes. *Frontiers in Genetics*, 12. <https://doi.org/10.3389/fgene.2021.683408>
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal Molecular Biology*, 215(3), 403-410. [https://doi.org/10.1016/s0022-2836\(05\)80360-2](https://doi.org/10.1016/s0022-2836(05)80360-2)
- Amico, I., Vilardi, J., Saidman, B., Ewens, M., & Bessega, C. (2019). Pollen contamination and mating patterns in a *Prosopis alba* clonal orchard: impact on seed orchards establishment [Pollen contamination and mating patterns in a *Prosopis alba* clonal orchard: impact on seed orchards establishment]. *iForest - Biogeosciences and Forestry*, 12(3), 330-337. <https://doi.org/10.3832/ifor2936-012>
- Blanco-Pastor, J. L., Barre, P., Keep, T., Ledauphin, T., Escobar-Gutierrez, A., Roschanski, A. M., Willner, E., Dehmer, K. J., Hegarty, M., Muylle, H., Veeckman, E., Vandepoele, K., Ruttink, T., Roldan-Ruiz, I., Manel, S., & Sampoux, J. P. (2021). Canonical correlations reveal adaptive loci and phenotypic responses to climate in perennial ryegrass. *Molecular Ecology Resources*, 21(3), 849-870. <https://doi.org/10.1111/1755-0998.13289>

- Brodribb, T. J., McAdam, S. A. M., Jordan, G. J., & Martins, S. C. V. (2014). Conifer species adapt to low-rainfall climates by following one of two divergent pathways. *Proceedings of the National Academy of Sciences*, *111*(40), 14489-14493. <https://doi.org/doi:10.1073/pnas.1407930111>
- Brown, S. S., Chen, Y.-W., Wang, M., Clipson, A., Ochoa, E., & Du, M.-Q. (2017). PrimerPooler: automated primer pooling to prepare library for targeted sequencing. *Biology Methods and Protocols*, *2*(1). <https://doi.org/10.1093/biomethods/bpx006>
- Burke, M. K. (2012). How does adaptation sweep through the genome? Insights from long-term selection experiments. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1749), 5029-5038. <https://doi.org/doi:10.1098/rspb.2012.0799>
- Buschiazzo, E., Ritland, C., Bohlmann, J., & K, R. (2012). Slow but not low: genomic comparisons reveal slower evolutionary rate and higher dN/dS in conifers compared to angiosperms. *BMC Evolutionary Biology*, *12*.
- Buschiazzo, E., Ritland, C., Bohlmann, J., & Ritland, K. (2012). Slow but not low: genomic comparisons reveal slower evolutionary rate and higher dN/dS in conifers compared to angiosperms. *BMC Evolutionary Biology*, *12*(1), 8. <https://doi.org/10.1186/1471-2148-12-8>
- Byrne, T., Farrelly, N., Kelleher, C., Hodkinson, T., Byrne, S., & Barth, S. (2022). Genetic diversity and structure of a diverse population of *Picea sitchensis* using Genotyping-by-Sequencing. *Forests*, *13*(9), 1511. <https://www.mdpi.com/1999-4907/13/9/1511>
- Caballero, A., & García-Dorado, A. (2013). Allelic diversity and its implications for the rate of adaptation. *Genetics*, *195*(4), 1373-1384. <https://doi.org/10.1534/genetics.113.158410>
- Cahalane, G., Doody, P., Douglas, G., Fennessy, J., O'Reilly, C., & Pfeifer, A. (2007). Sustaining and developing Irelands Forest Genetic Resources: An outline strategy. *COFORD*, 1-38.
- Campbell, N. R., Harmon, S. A., & Narum, S. R. (2015). Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing. *Molecular Ecology Resources*, *15*(4), 855-867. <https://doi.org/10.1111/1755-0998.12357>
- Carey, M. (2010). *The Avondale Initiative 1905* (COFORD Connects, Issue).

- Casola, C. (2019). Resequencing of massive pine genomes helps to unlock the genetic underpinning of quantitative traits in conifer trees. *New Phytologist*, 221(4), 1669-1671. <https://doi.org/10.1111/nph.15655>
- Cassiman, J.-J. (2003). A Dictionary of Genetics. *European Journal of Human Genetics*, 11(1), 105-105. <https://doi.org/10.1038/sj.ejhg.5200912>
- Chang, C. Y., Brautigam, K., Huner, N. P. A., & Ensminger, I. (2021). Champions of winter survival: cold acclimation and molecular regulation of cold hardiness in evergreen conifers. *New Phytologist*, 229(2), 675-691. <https://doi.org/10.1111/nph.16904>
- Charlesworth, D., & Willis, J. H. (2009). The genetics of inbreeding depression. *Nature Reviews Genetics*, 10(11), 783-796. <https://doi.org/10.1038/nrg2664>
- Chen, B. W., Cole, J. W., & Grond-Ginsbach, C. (2017). Departure from Hardy Weinberg Equilibrium and genotyping error. *Frontiers in Genetics*, 8. <https://doi.org/ARTN16710.3389/fgene.2017.00167>
- Chen, S.-Y., Deng, F., Jia, X., Li, C., & Lai, S.-J. (2017). A transcriptome atlas of rabbit revealed by PacBio single-molecule long-read sequencing. *Scientific Reports*, 7(1), 7648. <https://doi.org/10.1038/s41598-017-08138-z>
- Chen, X., Sun, X., Dong, L., & Zhang, S. (2018). Mating patterns and pollen dispersal in a Japanese larch (*Larix kaempferi*) clonal seed orchard: a case study. *Science China Life Science*, 61(9), 1011-1023. <https://doi.org/10.1007/s11427-018-9305-7>
- Chen, Z.-Q., Zan, Y., Milesi, P., Zhou, L., Chen, J., Li, L., Cui, B., Niu, S., Westin, J., Karlsson, B., García-Gil, M. R., Lascoux, M., & Wu, H. X. (2021). Leveraging breeding programs and genomic data in Norway spruce (*Picea abies* L. Karst) for GWAS analysis. *Genome Biology*, 22(1), 179. <https://doi.org/10.1186/s13059-021-02392-1>
- Christoforou, A., Dondrup, M., Mattingsdal, M., Mattheisen, M., Giddaluru, S., Nöthen, M. M., Rietschel, M., Cichon, S., Djurovic, S., Andreassen, O. A., Jonassen, I., Steen, V. M., Puntervoll, P., & Le Hellard, S. (2012). Linkage-disequilibrium-based binning affects the interpretation of GWAS. *Animal Journal Human Genetics*, 90(4), 727-733. <https://doi.org/10.1016/j.ajhg.2012.02.025>
- Cordeiro, E. M. G., Campbell, J. F., Phillips, T., & Akhunov, E. (2019). Isolation by distance, source-sink population dynamics and dispersal facilitation by trade routes: Impact on population genetic structure of a stored grain pest. *G3-Genes Genomes Genetics*, 9(5), 1457-1468. <https://doi.org/10.1534/g3.118.200892>

- Critchfield, W. B. (1984). Impact of the Pleistocene on the genetic structure of North American conifers. *Proceedings of the 8th North American Forest Biology Workshop*, Utah State University, Logan, USA.
- Cvjetkovic, B., Konnert, M., Holliday, J., & Anna-Maria, S.-L. (2018). Molecular markers used for genetic studies in Sitka spruce (*Picea sitchensis* (Bong.) Carr.).
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15), 2156-2158. <https://doi.org/10.1093/bioinformatics/btr330>
- Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *Gigascience*, 10(2). <https://doi.org/10.1093/gigascience/giab008>
- Dawson, J. C., & Goldringer, I. (2012). Breeding for Genetically Diverse Populations: Variety Mixtures and Evolutionary Populations. *Organic Crop Breeding*, 77-98. <https://doi.org/https://doi.org/10.1002/9781119945932.ch5>
- De La Torre, A. R., Birol, I., Bousquet, J., Ingvarsson, P. K., Jansson, S., Jones, S. J., Keeling, C. I., MacKay, J., Nilsson, O., Ritland, K., Street, N., Yanchuk, A., Zerbe, P., & Bohlmann, J. (2014). Insights into conifer giga-genomes. *Plant Physiology*, 166(4), 1724-1732. <https://doi.org/10.1104/pp.114.248708>
- De La Torre, A. R., Wilhite, B., & Neale, D. B. (2019). Environmental genome-wide association reveals climate adaptation is shaped by subtle to moderate allele frequency shifts in loblolly pine. *Genome Biology Evolution*, 11(10), 2976-2989. <https://doi.org/10.1093/gbe/evz220>
- de Meeus, T., & Goudet, J. (2007). A step-by-step tutorial to use HierFstat to analyse populations hierarchically structured at multiple levels. *Infection, Genetic Evolution*, 7(6), 731-735. <https://doi.org/10.1016/j.meegid.2007.07.005>
- Degner, J. (2015). Spruce hybridization in British Columbia. *Forest Genetics Council of BC*, 2.
- Florin, R. (1964). The distribution of conifer and taxad genera in time and space. *Annales De Geographie*, 73(400), 712-713.
- Douglas, G., Straley, G., Meidinger, D., & Pojar, J. (1998). *Illustrated Flora of British Columbia. Volume 1: Gymnosperms and Dicotyledons (Aceraceae Through Asteraceae)* (Vol. 1).

- Du, H., Ran, J.-H., Feng, Y.-Y., & Wang, X.-Q. (2020). The flattened and needlelike leaves of the pine family (Pinaceae) share a conserved genetic network for adaxial-abaxial polarity but have diverged for photosynthetic adaptation. *BMC Evolutionary Biology*, 20(1), 131. <https://doi.org/10.1186/s12862-020-01694-5>
- Ebrahimi, A., Lawson, S. S., Frank, G. S., Coggeshall, M. V., Woeste, K. E., & McKenna, J. R. (2018). Pollen flow and paternity in an isolated and non-isolated black walnut (*Juglans nigra* L.) timber seed orchard. *PLoS One*, 13(12). <https://doi.org/10.1371/journal.pone.0207861>
- 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(5). <https://doi.org/10.1371/journal.pone.0019379>
- Farjon, A. (2010). *A Handbook of the World's Conifers*. <https://doi.org/https://doi.org/10.1163/9789047430629>
- Farrelly, N., Dhubháin, Á., Nieuwenhuis, M., & Grant, J. (2009). The distribution and productivity of Sitka spruce (*Picea sitchensis*) in Ireland in relation to site, soil and climatic factors. *Irish Forestry* Vol 66. 66.
- Farris, D. W. H., P.J. (2020). Selected geologic maps of the Kodiak batholith and other Paleocene intrusive rocks, Kodiak Island, Alaska. *U. S. G. S. S. Investigations (Ed.)*, (Vol. 3441, pp. 10).
- Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37(12), 4302-4315. <https://doi.org/https://doi.org/10.1002/joc.5086>
- Fowler, D. P. (1983). The hybrid black × Sitka spruce, implications to phylogeny of the genus *Picea*. *Canadian Journal of Forest Research*, 13(1), 108-115. <https://doi.org/10.1139/x83-016>
- Fu, Y. B. (2015). Understanding crop genetic diversity under modern plant breeding. *Theorithacl Applied Genetics*, 128(11), 2131-2142. <https://doi.org/10.1007/s00122-015-2585-y>
- Gagalova, K. K., Warren, R. L., Coombe, L., Wong, J., Nip, K. M., Yuen, M. M. S., Whitehill, J. G. A., Celedon, J. M., Ritland, C., Taylor, G. A., Cheng, D., Plettner, P., Hammond, S. A., Mohamadi, H., Zhao, Y., Moore, R. A., Mungall, A. J., Boyle, B., Laroche, J. (2022). Spruce giga-genomes: structurally similar yet distinctive

- with differentially expanding gene families and rapidly evolving genes. *Plant Journal*, 111(5), 1469-1485. <https://doi.org/10.1111/tpj.15889>
- Galeano, E., Bousquet, J., & Thomas, B. R. (2021). SNP-based analysis reveals unexpected features of genetic diversity, parental contributions and pollen contamination in a white spruce breeding program. *Science Reports*, 11(1), 4990. <https://doi.org/10.1038/s41598-021-84566-2>
- Gapare, W. J., Aitken, S. N., & Ritland, C. E. (2005). Genetic diversity of core and peripheral Sitka spruce (*Picea sitchensis* (Bong.) Carr) populations: implications for conservation of widespread species. *Biological Conservation*, 123(1), 113-123. <https://doi.org/10.1016/j.biocon.2004.11.002>
- GFDRR. (2022). *WF-GLOBAL-CSIRO-30*. https://www.geonode-gfdrrlab.org/layers/hazard:csiro_wf_max_fwi_rp30#more
- Glombik, P., O'Reilly, C., & Grant, O. M. (2015). Early-height variation between full-sibling families of Sitka spruce growing in Ireland. *Irish Forestry*
- Goto, S., Miyahara, F., & Ide, Y. (2002). Identification of the male parents of half-sib progeny from Japanese black pine (*Pinus thunbergii* Parl.) clonal seed orchard using RAPD markers. *Breeding Science*, 52(2), 71-77. <https://doi.org/10.1270/jsbbs.52.71>
- Grattapaglia, D. (2022). Twelve years into genomic selection in forest trees: climbing the slope of enlightenment of marker assisted tree breeding. *Forests*, 13(10), 1554. <https://www.mdpi.com/1999-4907/13/10/1554>
- Griffith, R. (1992). *Picea sitchensis*. *Fire Effects Information System*. U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, Fire Sciences Laboratory. <https://www.fs.fed.us/database/feis/plants/tree/picsit/all.html>
- Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 32(18), 2847-2849. <https://doi.org/10.1093/bioinformatics/btw313>
- Günther, T., & Coop, G. (2013). Robust identification of local adaptation from allele frequencies. *Genetics*, 195(1), 205-220. <https://doi.org/10.1534/genetics.113.152462>
- Hall, D., Zhao, W., Wennstrom, U., Andersson Gull, B., & Wang, X. R. (2020). Parentage and relatedness reconstruction in *Pinus sylvestris* using genotyping-by-sequencing. *Heredity*, 124(5), 633-646. <https://doi.org/10.1038/s41437-020-0302-3>

- Hall, D. K., Riggs, G. A., & Salomonson, V. V. (2006). *MODIS/Terra Snow Cover 5-Min L2*. Version 5. <https://doi.org/http://dx.doi.org/10.5067/ACytyzb9BEOS>.
- 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(8), 1651-1662. <https://doi.org/10.3732/ajb.1200654>
- Hamilton, J. A., Lexer, C., & Aitken, S. N. (2013). Differential introgression reveals candidate genes for selection across a spruce (*Picea sitchensis* x *P. glauca*) hybrid zone. *New Phytologist*, 197(3), 927-938. <https://doi.org/10.1111/nph.12055>
- Hamilton, J. A., Lexer, C., & Aitken, S. N. (2013). Genomic and phenotypic architecture of a spruce hybrid zone (*Picea sitchensis* × *P. glauca*). *Molecular Ecology*, 22(3), 827-841. <https://doi.org/https://doi.org/10.1111/mec.12007>
- Hanlon, V. C. T., Otto, S. P., & Aitken, S. N. (2019). Somatic mutations substantially increase the per-generation mutation rate in the conifer *Picea sitchensis*. *Evol Lett*, 3(4), 348-358. <https://doi.org/10.1002/evl3.121>
- Haro, H. (2017). Animating the temporal progression of cordilleran deglaciation and vegetation succession in the pacific northwest during the late quaternary period. Western Cedar, Western Washington University.
- He, J., Zhao, X., Laroche, A., Lu, Z. X., Liu, H., & Li, Z. (2014). Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Frontiers Plant Science*, 5, 484. <https://doi.org/10.3389/fpls.2014.00484>
- Highton, R. (1993). The relationship between the number of loci and the statistical support for the topology of UPGMA trees obtained from genetic distance data. *Molecular Phylogenet Evolution*, 2(4), 337-343. <https://doi.org/10.1006/mpev.1993.1033>
- Hiraoka, Y., Fukatsu, E., Mishima, K., Hirao, T., Teshima, K. M., Tamura, M., Tsubomura, M., Iki, T., Kurita, M., Takahashi, M., & Watanabe, A. (2018). Potential of genome-wide studies in unrelated plus trees of a coniferous species, *Cryptomeria japonica* (Japanese Cedar). *Frontiers Plant Science*, 9, 1322. <https://doi.org/10.3389/fpls.2018.01322>
- 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*). *New Phytologist*, 188(2), 501-514. <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 Science*, 279(1734), 1675-1683. <https://doi.org/10.1098/rspb.2011.1805>
- Holsinger, K. E., & Weir, B. S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nature Reviews Genetics*, 10(9), 639-650. <https://doi.org/10.1038/nrg2611>
- Horgan, T., Keane, M., McCarthy, R., Lally, M., & Thompson, D. (2003). *A Guide to Forest Tree Species Selection and Silviculture in Ireland*. COFORD.
- Hornoy, B., Pavy, N., Gerardi, S., Beaulieu, J., & Bousquet, J. (2015). Genetic adaptation to climate in white spruce involves small to moderate allele frequency shifts in functionally diverse genes. *Genome Biology Evolution*, 7(12), 3269-3285. <https://doi.org/10.1093/gbe/evv218>
- Howell, W. M., Jobs, M., Gyllensten, U., & Brookes, A. J. (1999). Dynamic allele-specific hybridization. *Nature Biotechnology*, 17(1), 87-88. <https://doi.org/10.1038/5270>
- Jimenez-Ramirez, A., Grivet, D., & Robledo-Arnuncio, J. J. (2021). Measuring recent effective gene flow among large populations in *Pinus sylvestris*: Local pollen shedding does not preclude substantial long-distance pollen immigration. *PLoS One*, 16(8). <https://doi.org/ARTNe025577610.1371/journal.pone.0255776>
- Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, 24(11), 1403-1405. <https://doi.org/10.1093/bioinformatics/btn129>
- Jombart, T., & Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, 27(21), 3070-3071. <https://doi.org/10.1093/bioinformatics/btr521>
- Jurgiel, B. (2020). *Point sampling tool*. Retrieved 07/09/2022 from <https://github.com/borysiasty/pointsamplingtool>
- Kamvar, Z. N., Tabima, J. F., & Grunwald, N. J. (2014). Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, 2, e281. <https://doi.org/10.7717/peerj.281>
- Khoury, C. K., Brush, S., Costich, D. E., Curry, H. A., de Haan, S., Engels, J. M. M., Guarino, L., Hoban, S., Mercer, K. L., Miller, A. J., Nabhan, G. P., Perales, H. R., Richards, C., Riggins, C., & Thormann, I. (2022). Crop genetic erosion: understanding and responding to loss of crop diversity. *New Phytologist*, 233(1), 84-118. <https://doi.org/https://doi.org/10.1111/nph.17733>

- King, J., Alfaro, R., & Cartwright, C. (2004). Genetic resistance of Sitka spruce (*Picea sitchensis*) populations to the white pine weevil (*Pissodes strobi*): distribution of resistance. *Forestry* 77, 7.
- Korecky, J., Cepl, J., Stejskal, J., Faltinova, Z., Dvorak, J., Lstiburek, M., & El-Kassaby, Y. A. (2021). Genetic diversity of Norway spruce ecotypes assessed by GBS-derived SNPs. *Scientific Reports*, 11(1). <https://doi.org/ARTN2311910.1038/s41598-021-02545-z>
- Korunes, K. L., & Samuk, K. (2021). pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Molecular Ecology Resources*, 21(4), 1359-1368. <https://doi.org/10.1111/1755-0998.13326>
- Kremer, A., Ronce, O., Robledo-Arnuncio, J. J., Guillaume, F., Bohrer, G., Nathan, R., Bridle, J. R., Gomulkiewicz, R., Klein, E. K., Ritland, K., Kuparinen, A., Gerber, S., & Schueler, S. (2012). Long-distance gene flow and adaptation of forest trees to rapid climate change. *Ecology Letters*, 15(4), 378-392. <https://doi.org/10.1111/j.1461-0248.2012.01746.x>
- Kumar, S., Chagné, D., Bink, M. C. A. M., Volz, R. K., Whitworth, C., & Carlisle, C. (2012). Genomic selection for fruit quality traits in apple (*Malus×domestica* Borkh.). *PLoS One*, 7(5), e36674. <https://doi.org/10.1371/journal.pone.0036674>
- Lebedev, V. G., Lebedeva, T. N., Chernodubov, A. I., & Shestibratov, K. A. (2020). Genomic selection for forest tree improvement: methods, achievements and perspectives. *Forests*, 11(11), 1190. <https://www.mdpi.com/1999-4907/11/11/1190>
- Lee, S., Thompson, D., & Hansen, J. K. (2013). Sitka Spruce (*Picea sitchensis* (Bong.) Carr). *Forest Tree Breeding in Europe: Current State-of-the-Art and Perspectives* (pp. 177-227). https://doi.org/10.1007/978-94-007-6146-9_4
- Leitch, A. R., & Leitch, I. J. (2012). Ecological and genetic factors linked to contrasting genome dynamics in seed plants. *New Phytologist*, 194(3), 629-646. <https://doi.org/10.1111/j.1469-8137.2012.04105.x>
- Leroy, G., Carroll, E. L., Bruford, M. W., DeWoody, J. A., Strand, A., Waits, L., & Wang, J. (2018). Next-generation metrics for monitoring genetic erosion within populations of conservation concern. *Evolutionary Applications*, 11(7), 1066-1083. <https://doi.org/https://doi.org/10.1111/eva.12564>
- Leyva-Pérez, M. d. I. O., Vexler, L., Byrne, S., Clot, C. R., Meade, F., Griffin, D., Ruttink, T., Kang, J., & Milbourne, D. (2022). PotatoMASH; A low cost, genome-scanning

- marker system for use in potato genomics and genetics applications. *Agronomy*, 12(10), 2461. <https://www.mdpi.com/2073-4395/12/10/2461>
- Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18), 3094-3100. <https://doi.org/10.1093/bioinformatics/bty191>
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), 1754-1760. <https://doi.org/10.1093/bioinformatics/btp324>
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), 2078-2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Lipow, S. R., Clair, J. B. S., & Johnson, G. R. (2002). Ex situ gene conservation for conifers in the Pacific Northwest
- Liu, S., Y. Wei, W.M. Post, R.B. Cook, K. Schaefer, & Thornton., M. M. (2014). NACP MsTMIP: Unified North American Soil Map. <https://doi.org/https://doi.org/10.3334/ORNLDAAC/1242>
- Liu, X., Huang, M., Fan, B., Buckler, E. S., & Zhang, Z. (2016). Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genetics*, 12(2), e1005767. <https://doi.org/10.1371/journal.pgen.1005767>
- Lockwood, J. D., Aleksic, J. M., Zou, J., Wang, J., Liu, J., & Renner, S. S. (2013). A new phylogeny for the genus *Picea* from plastid, mitochondrial, and nuclear sequences. *Molecular Phylogenetic Evolution*, 69(3), 717-727. <https://doi.org/10.1016/j.ympev.2013.07.004>
- Lotterhos, K. E., Yeaman, S., Degner, J., Aitken, S., & Hodgins, K. A. (2018). Modularity of genes involved in local adaptation to climate despite physical linkage. *Genome Biology*, 19(1), 157. <https://doi.org/10.1186/s13059-018-1545-7>
- Lynch, M., & O'Hely, M. (2001). Captive breeding and the genetic fitness of natural populations. *Conservation Genetics*, 2(4), 363-378. <https://doi.org/10.1023/A:1012550620717>
- Mahoney, C. L., & Springer, D. A. (2009). *Genetic Diversity*. Nova Science Publishers. <https://books.google.ie/books?id=YROsAQAACAAJ>
- Malinsky, M., Matschiner, M., & Svardal, H. (2021). Dsuite - Fast D-statistics and related admixture evidence from VCF files. *Molecular Ecology Resources*, 21(2), 584-595. <https://doi.org/https://doi.org/10.1111/1755-0998.13265>

- Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution*, 38(10), 4647-4654. <https://doi.org/10.1093/molbev/msab199>
- Mason, A. S. (2015). SSR genotyping. *Methods Molecular Biology*, 1245, 77-89. https://doi.org/10.1007/978-1-4939-1966-6_6
- Mba, C., & Tohme, J. (2005). Use of AFLP markers in surveys of plant diversity. *Methods in Enzymology* (Vol. 395, pp. 177-201). [https://doi.org/https://doi.org/10.1016/S0076-6879\(05\)95012-X](https://doi.org/https://doi.org/10.1016/S0076-6879(05)95012-X)
- Menon, M., Bagley, J. C., Page, G. F. M., Whipple, A. V., Schoettle, A. W., Still, C. J., Wehenkel, C., Waring, K. M., Flores-Renteria, L., Cushman, S. A., & Eckert, A. J. (2021). Adaptive evolution in a conifer hybrid zone is driven by a mosaic of recently introgressed and background genetic variants. *Communications Biology*, 4(1), 160. <https://doi.org/10.1038/s42003-020-01632-7>
- Menounos, B., Goehring, B. M., Osborn, G., Margold, M., Ward, B., Bond, J., Clarke, G. K. C., Clague, J. J., Lakeman, T., Koch, J., Caffee, M. W., Gosse, J., Stroeven, A. P., Seguinot, J., & Heyman, J. (2017). Cordilleran ice sheet mass loss preceded climate reversals near the pleistocene termination. *Science*, 358(6364), 781-784. <https://doi.org/10.1126/science.aan3001>
- Miller, J. M., Cullingham, C. I., & Peery, R. M. (2020). The influence of a priori grouping on inference of genetic clusters: simulation study and literature review of the DAPC method. *Heredity*, 125(5), 269-280. <https://doi.org/10.1038/s41437-020-0348-2>
- Mimura, M., & Aitken, S. N. (2010). Local adaptation at the range peripheries of Sitka spruce. *J Evolution Biology*, 23(2), 249-258. <https://doi.org/10.1111/j.1420-9101.2009.01910.x>
- Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences*, 70(12), 3321-3323. <https://doi.org/doi:10.1073/pnas.70.12.3321>
- Nystedt, B., Street, N. R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., Scofield, D. G., Vezzi, F., Delhomme, N., Giacomello, S., Alexeyenko, A., Vicedomini, R., Sahlin, K., Sherwood, E., Elfstrand, M., Gramzow, L., Holmberg, K., Hällman, J., Keech, O., Klasson, L., Jansson, S. (2013). The Norway spruce genome sequence and conifer

- genome evolution. *Nature*, 497(7451), 579-584.
<https://doi.org/10.1038/nature12211>
- O'Connell, L. M., Mosseler, A., & Rajora, O. P. (2007). Extensive long-distance pollen dispersal in a fragmented landscape maintains genetic diversity in white spruce. *J Hered*, 98(7), 640-645. <https://doi.org/10.1093/jhered/esm089>
- O'Driscoll, J. (1972). Working plan for international ten provenance experiment. *Forest and Wildlife Service, Dublin, Ireland*.
- O'Driscoll, J. (1978). Sitka spruce international ten provenance experiment. <https://www.fao.org/3/11807e/L1807E06.htm>
- O'Leary, S. J., Puritz, J. B., Willis, S. C., Hollenbeck, C. M., & Portnoy, D. S. (2019). These aren't the loci you'e looking for: Principles of effective SNP filtering for molecular ecologists (vol 27, pg 3193, 2018). *Molecular Ecology*, 28(14), 3459-3459. <https://doi.org/10.1111/mec.14955>
- O'Driscoll, J. (1977). Sitka Spruce, its distribution and genetic variation. *Irish Forestry* 2, 11.
- O'Driscoll, J. (1978). Sitka spruce international ten provenacne experiment: results to end of nursery stage. 35-46.
- OECD. (2006). *Section 5 - Sitka Spruce (PICEA SITCHENSIS (BONG.) CARR.)*. OECD Publishing. <https://doi.org/https://doi.org/10.1787/23114622>
- Ondov, B. D., Starrett, G. J., Sappington, A., Kostic, A., Koren, S., Buck, C. B., & Phillippy, A. M. (2019). Mash Screen: high-throughput sequence containment estimation for genome discovery. *Genome Biology*, 20(1), 232. <https://doi.org/10.1186/s13059-019-1841-x>
- Parra-Salazar, A., Gomez, J., Lozano-Arce, D., Reyes-Herrera, P. H., & Duitama, J. (2022). Robust and efficient software for reference-free genomic diversity analysis of genotyping-by-sequencing data on diploid and polyploid species. *Molecular Ecology Resourse*, 22(1), 439-454. <https://doi.org/10.1111/1755-0998.13477>
- Paun, O., & Schönswetter, P. (2012). Amplified fragment length polymorphism: an invaluable fingerprinting technique for genomic, transcriptomic, and epigenetic studies. *Methods Molecular Biology*, 862, 75-87. https://doi.org/10.1007/978-1-61779-609-8_7
- Pavan, S., Delvento, C., Ricciardi, L., Lotti, C., Ciani, E., & D'Agostino, N. (2020). Recommendations for choosing the genotyping method and best practices for

- quality control in crop genome-wide association studies. *Frontiers Genetics*, *11*, 447. <https://doi.org/10.3389/fgene.2020.00447>
- Pavy, N., Namroud, M. C., Gagnon, F., Isabel, N., & Bousquet, J. (2012). The heterogeneous levels of linkage disequilibrium in white spruce genes and comparative analysis with other conifers. *Heredity*, *108*(3), 273-284. <https://doi.org/10.1038/hdy.2011.72>
- Pereira-Dias, L., Vilanova, S., Fita, A., Prohens, J., & Rodriguez-Burruezo, A. (2019). Genetic diversity, population structure, and relationships in a collection of pepper (*Capsicum* spp.) landraces from the Spanish centre of diversity revealed by genotyping-by-sequencing (GBS). *Horticulture*, *6*, 54. <https://doi.org/10.1038/s41438-019-0132-8>
- Pina-Martins, F., Silva, D. N., Fino, J., & Paulo, O. S. (2017). Structure_threader: An improved method for automation and parallelization of programs structure, fastStructure and MaverickK on multicore CPU systems. *Molecular Ecology Resources*, *17*(6), e268-e274. <https://doi.org/10.1111/1755-0998.12702>
- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945-959. <https://doi.org/10.1093/genetics/155.2.945>
- Prunier, J., Verta, J. P., & MacKay, J. J. (2016). Conifer genomics and adaptation: at the crossroads of genetic diversity and genome function. *New Phytologist*, *209*(1), 44-62. <https://doi.org/10.1111/nph.13565>
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Animal Journal Human Genetics*, *81*(3), 559-575. <https://doi.org/10.1086/519795>
- QGIS.org. (2022). *QGIS Geographic Information System*. QGIS Association. <http://www.qgis.org>
- Qu, J., Kachman, S. D., Garrick, D., Fernando, R. L., & Cheng, H. (2020). Exact distribution of linkage disequilibrium in the presence of mutation, selection, or minor allele frequency filtering. *Frontiers Genetics*, *11*, 362. <https://doi.org/10.3389/fgene.2020.00362>
- Raj, A., Stephens, M., & Pritchard, J. K. (2014). fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*, *197*(2), 573-589. <https://doi.org/10.1534/genetics.114.164350>

- Rasheed, A., Hao, Y., Xia, X., Khan, A., Xu, Y., Varshney, R. K., & He, Z. (2017). Crop breeding chips and genotyping platforms: progress, challenges, and perspectives. *Molecular Plant*, *10*(8), 1047-1064. <https://doi.org/10.1016/j.molp.2017.06.008>
- Rohan, S. (2019). Why Sitka spruce? a history of Irish forestry. *Political ecologies of Ireland*.
- Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., & Feldman, M. W. (2002). Genetic structure of human populations. *Science*, *298*(5602), 2381-2385. <https://doi.org/10.1126/science.1078311>
- Salazar Zarzosa, P., Diaz Herraiz, A., Olmo, M., Ruiz-Benito, P., Barrón, V., Bastias, C. C., de la Riva, E. G., & Villar, R. (2021). Linking functional traits with tree growth and forest productivity in *Quercus ilex* forests along a climatic gradient. *Science of The Total Environment*, *786*, 147468. <https://doi.org/https://doi.org/10.1016/j.scitotenv.2021.147468>
- Schoen, D. J., & Stewart, S. C. (1986). Variation in male reproductive investment and male reproductive success in White spruce. *Evolution*, *40*(6), 1109-1120. <https://doi.org/10.1111/j.1558-5646.1986.tb05737.x>
- Schönswetter P, Schneeweiss GM. (2019) Is the incidence of survival in interior pleistocene refugia (nunataks) underestimated? Phylogeography of the high mountain plant *Androsace alpina* (Primulaceae) in the European Alps revisited. *Ecol Evol*. 4078-4086. doi: 10.1002/ece3.5037.
- Schrader, C., Schielke, A., Ellerbroek, L., & Johne, R. (2012). PCR inhibitors – occurrence, properties and removal. *Journal of Applied Microbiology*, *113*(5), 1014-1026. <https://doi.org/https://doi.org/10.1111/j.1365-2672.2012.05384.x>
- Sebastian-Azcona, J., Hacke, U. G., & Hamann, A. (2018). Adaptations of white spruce to climate: strong intraspecific differences in cold hardiness linked to survival. *Ecology Evolution*, *8*(3), 1758-1768. <https://doi.org/10.1002/ece3.3796>
- Senser, M., & Beck, E. (1984). Correlation of chloroplast ultrastructure and membrane lipid composition to the different degrees of frost resistance achieved in leaves of spinach, ivy, and spruce. *Journal Plant Physiology*, *117*(1), 41-55. [https://doi.org/10.1016/S0176-1617\(84\)80015-2](https://doi.org/10.1016/S0176-1617(84)80015-2)
- Shen, R., Fan, J.-B., Campbell, D., Chang, W., Chen, J., Doucet, D., Yeakley, J., Bibikova, M., Wickham Garcia, E., McBride, C., Steemers, F., Garcia, F., Kermani, B. G., Gunderson, K., & Oliphant, A. (2005). High-throughput SNP genotyping on universal bead arrays. *Mutation Research/Fundamental and Molecular*

Mechanisms of Mutagenesis, 573(1), 70-82.

<https://doi.org/https://doi.org/10.1016/j.mrfmmm.2004.07.022>

Skrøppa, T., & Steffenrem, A. (2021). Performance and phenotypic stability of Norway spruce provenances, families, and clones growing under diverse climatic conditions in four nordic countries. *Forests*, 12(2), 230. <https://www.mdpi.com/1999-4907/12/2/230>

South, S. (2022). *rnaturalearth: World Map Data from Natural Earth*. . <https://github.com/ropensci/rnaturalearth>.

Stevens, M., & Oltean, S. (2019). Modulation of the apoptosis gene Bcl-x function through alternative splicing. *Frontiers in Genetics*, 10. <https://doi.org/10.3389/fgene.2019.00804>

Stoehr, M. U., & Newton, C. H. (2002). Evaluation of mating dynamics in a lodgepole pine seed orchard using chloroplast DNA markers. *Canadian Journal of Forest Research*, 32(3), 469-476. <https://doi.org/10.1139/x01-222>

Stojnic, S., Avramidou, E. V., Fussi, B., Westergren, M., Orlovic, S., Matovic, B., Trudic, B., Kraigher, H., Aravanopoulos, F. A., & Konnert, M. (2019). Assessment of genetic diversity and population genetic structure of norway spruce (*Picea abies* (L.) Karsten) at its southern lineage in europe. Implications for conservation of forest genetic resources. *Forests*, 10(3). <https://doi.org/ARTN25810.3390/f10030258>

Strand, M., Löfvenius, M. O., Bergsten, U., Lundmark, T., & Rosvall, O. (2006). Height growth of planted conifer seedlings in relation to solar radiation and position in Scots pine shelterwood. *Forest Ecology and Management*, 224(3), 258-265. <https://doi.org/https://doi.org/10.1016/j.foreco.2005.12.038>

Sun, W., Yu, D., Dong, M., Zhao, J., Wang, X., Zhang, H., & Zhang, J. (2017). Evaluation of efficiency of controlled pollination based parentage analysis in a *Larix gmelinii* var. *principis-rupprechtii* Mayr. seed orchard. *PLoS One*, 12(4), e0176483. <https://doi.org/10.1371/journal.pone.0176483>

Suontama, M., Klápště, J., Telfer, E., Graham, N., Stovold, T., Low, C., McKinley, R., & Dungey, H. (2019). Correction to: Efficiency of genomic prediction across two *Eucalyptus nitens* seed orchards with different selection histories. *Heredity*, 122(3), 380. <https://doi.org/10.1038/s41437-018-0159-x>

- Sype, H. R.-A., B. (1990). Genetic variability of Sitka spruce of the IUFRO collection. *Quebec, Canada*.
- Tanksley, S. D., Young, N. D., Paterson, A. H., & Bonierbale, M. W. (1989). RFLP mapping in plant breeding: New tools for an old science. *Bio/Technology*, 7(3), 257-264. <https://doi.org/10.1038/nbt0389-257>
- Teskey, R., Wertin, T., Bauweraerts, I., Ameye, M., McGuire, M. A., & Steppe, K. (2015). Responses of tree species to heat waves and extreme heat events. *Plant Cell Environ*, 38(9), 1699-1712. <https://doi.org/10.1111/pce.12417>
- Thompson, D., Lally, M., & Pfeifer, A. (2005). Washington, Oregon or Queen Charlotte Islands? Which is the best provenance of Sitka spruce (*Picea sitchensis*) for Ireland? *Irish Forestry* 62.
- Turakulov, R., & Easteal, S. (2003). Number of SNPS loci needed to detect population structure. *Human Hereditary*, 55(1), 37-45. <https://doi.org/10.1159/000071808>
- Uddenberg, D., Akhter, S., Ramachandran, P., Sundström, J. F., & Carlsbecker, A. (2015). Sequenced genomes and rapidly emerging technologies pave the way for conifer evolutionary developmental biology. *Frontiers Plant Science*, 6, 970. <https://doi.org/10.3389/fpls.2015.00970>
- Untergasser, A., Nijveen, H., Rao, X., Bisseling, T., Geurts, R., & Leunissen, J. A. (2007). Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Research*, 35, W71-74. <https://doi.org/10.1093/nar/gkm306>
- Veeckman, E., Van Glabeke, S., Haegeman, A., Muylle, H., van Parijs, F. R. D., Byrne, S. L., Asp, T., Studer, B., Rohde, A., Roldan-Ruiz, I., Vandepoele, K., & Ruttink, T. (2019). Overcoming challenges in variant calling: exploring sequence diversity in candidate genes for plant development in perennial ryegrass (*Lolium perenne*). *DNA Research*, 26(1), 1-12. <https://doi.org/10.1093/dnares/dsy033>
- Verity, R., & Nichols, R. A. (2016). Estimating the number of subpopulations (K) in structured populations. *Genetics*, 203(4), 1827-+. <https://doi.org/10.1534/genetics.115.180992>
- Wang, B., Tseng, E., Regulski, M., Clark, T. A., Hon, T., Jiao, Y., Lu, Z., Olson, A., Stein, J. C., & Ware, D. (2016). Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nature Communications*, 7(1), 11708. <https://doi.org/10.1038/ncomms11708>

- Wang, F., Zhang, S., Zhu, P. (2003). The effects of fertility and synchronization variation on seed production in two Chinese fir clonal seed orchards. *Science Reports*, 13, 62. <https://doi.org/10.1038/s41598-022-27151-5>
- Wang, J., & Zhang, Z. (2021). GAPIT Version 3: boosting power and accuracy for genomic association and prediction. *Genomics, Proteomics & Bioinformatics*, 19(4), 629-640. <https://doi.org/https://doi.org/10.1016/j.gpb.2021.08.005>
- Wang, N., Yuan, Y., Wang, H., Yu, D., Liu, Y., Zhang, A., Gowda, M., Nair, S. K., Hao, Z., Lu, Y., San Vicente, F., Prasanna, B. M., Li, X., & Zhang, X. (2020). Applications of genotyping-by-sequencing (GBS) in maize genetics and breeding. *Science Reports*, 10(1), 16308. <https://doi.org/10.1038/s41598-020-73321-8>
- Wegrzyn, J. L., Liechty, J. D., Stevens, K. A., Wu, L. S., Loopstra, C. A., Vasquez-Gross, H. A., Dougherty, W. M., Lin, B. Y., Zieve, J. J., Martínez-García, P. J., Holt, C., Yandell, M., Zimin, A. V., Yorke, J. A., Crepeau, M. W., Puiu, D., Salzberg, S. L., Dejong, P. J., Mockaitis, K., . . . Neale, D. B. (2014). Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. *Genetics*, 196(3), 891-909. <https://doi.org/10.1534/genetics.113.159996>
- Weir, B. S., & Cockerham, C. C. (1984). Estimating F-Statistics for the analysis of population structure. *Evolution*, 38(6), 1358-1370. <https://doi.org/10.1111/j.1558-5646.1984.tb05657.x>
- Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. *Springer-Verlag New York*. <https://ggplot2.tidyverse.org>
- Williams, C. G., & Savolainen, O. (1996). Inbreeding depression in conifers: Implications for breeding strategy. *Forest Science*, 42(1), 102-117. <https://doi.org/10.1093/forestscience/42.1.102>
- Williams, J. G., Kubelik, A. R., Livak, K. J., Rafalski, J. A., & Tingey, S. V. (1990). DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Research*, 18(22), 6531-6535. <https://doi.org/10.1093/nar/18.22.6531>
- Wright, J. (1955). Species crossability in spruce in relation to distribution and taxonomy. *For. Sci*(1), 30.
- Yu, G. (2017). *scatterpie: Scatter Pie Plot*. <https://CRAN.R-project.org/package=scatterpie>
- Yuan, H., Niu, S., El-Kassaby, Y. A., Li, Y., & Li, W. (2016). Simple genetic distance-optimized field deployments for clonal seed orchards based on microsatellite

markers: As a case of chinese pine seed orchard. *PLoS One*, 11(6), e0157646.

<https://doi.org/10.1371/journal.pone.0157646>

Zhang, J., Yang, J., Zhang, L., Luo, J., Zhao, H., Zhang, J., & Wen, C. (2020). A new SNP genotyping technology Target SNP-seq and its application in genetic analysis of cucumber varieties. *Scientific Reports*, 10(1), 5623.

<https://doi.org/10.1038/s41598-020-62518-6>