Evolutionary Genomics of Western Poplar (*Populus trichocarpa*)

Rajesh K Bawa

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

Doctor of Philosophy
in
Forest Resources and Environmental Conservation

Jason A. Holliday, Chair
Joel W. McGlothlin
Xiaowei Wu
Liqing Zhang

June 14, 2017
Blacksburg, Virginia

Keywords: Demographic history, Population Genetics, Populus, Local adaptation
Copyright 2017, Rajesh K. Bawa

Evolutionary Genomics of Western Poplar (*Populus trichocarpa*)

Rajesh K. Bawa

ABSTRACT (academic)

Forest trees are an important pool of biodiversity at the gene, individual and an ecosystem level. This variation is a result of complex environmental interactions, as well as neutral and selective forces acting on populations. Patterns of standing genetic variation are the result of adaption to past and contemporary climate change, but also historical demographic events, and disentangling the role of these forces is a central problem in population genomics. The overall goal of this study is to characterize the relative effects of demography and selection in the genome of *Populus trichocarpa*, a riparian deciduous tree species of North America. Specifically, I used a variety of methods to summarize patterns of genetic diversity and population structure in *P. trichocarpa*, and to reconstruct its demographic history. I subsequently incorporated these demographic insights to guide the application of several methods to identify genome-wide targets of natural selection within and among rangewide populations adapted to heterogeneous selection regimes. Results of this study provide insights into the history of divergence and differentiation in *P. trichocarpa* populations and help us identify the functional genetic variants contributing to phenotypic divergence and fitness of the individuals in it.

This work was supported by the National Science Foundation Plant Genome Research Program (IOS:1054444), the USDA National Institute of Food and Agriculture PINEMAP project and Virginia Tech Department of Forest Resources and Environmental Conservation.

ABSTRACT (public)

The extant genetic variation in the forest tree populations is a product of evolutionary history and complex environmental interactions. This genetic variation can therefore be leveraged to identify both the adaptive genetic variation and understand the demographic events that might have caused these patterns. We, in this study, present and analyze the range-wide population genomic dataset of western poplar (*Populus trichocarpa*) and show that the current distribution of western poplar originated from the southern western poplar populations around hundred thousand years ago. Our study also identified large number of loci associated with environmental local adaptation. We identified number of genes involved in temperature sensing, light signaling pathways, and plant immunity, suggesting role of biotic and abiotic factors as major drivers of adaptation. We also detected significant overlap between highly divergent loci and adaptive genetic variation, indication the role of adaptation in species diversification. Overall, the results of this study advance our understanding of the demographic history and the major drivers of local adaptation in western poplar populations. It also gives us an observational and experimental evidence of role of environment and demographic history in shaping the genetic make-up of organism.

Dedication

To mom and dad for having conviction in my goals, and inspiring me to strive for the best.

Acknowledgments

Firstly, I would like to thank my advisor, Dr. Jason Holliday, for giving me an opportunity to work in the exciting field of genomics. Having him as an advisor has been an enriching experience and has shaped my approach towards research and teaching. I thank him for all his invaluable guidance, patience and support throughout my Ph.D. program. I would also like to thank my advisory committee: Dr. Joel McGlothlin, Dr. Xiaowei Wu and Dr. Liqing Zhang for their useful suggestions and help.

My research was funded by the National Science Foundation Plant Genome Research Program (IOS:1054444), the USDA National Institute of Food and Agriculture PINEMAP project and Virginia Tech Department of Forest Resources and Environmental Conservation. My sincere thanks to all of them. Without their precious support it would not be possible to conduct this research.

I also thank all my fellow labmates and colleagues: Lecong, Xiaoyan, Mandy, Haktan, Regis for all their help and support. Having them around both in lab and classes made all these experiences more pleasant and fun.

Lastly, I would to thank my family and friends: My parents and brother for their continuous love and support and to all my friends (Hunny, Amrit, Karthik, Ranjeet, Atul, Amit, Sandeep and Umesh) for all the fun times.

Contents

Dedication **iv**

Acknowledgements **v**

Table of Contents **vi**

List of Figures **xi**

List of Tables **xv**

1 Genomics of Speciation in Temperate and Boreal Angiosperm Trees **1**

Abstract 1

Modes of Speciation 2

Pre-zygotic Barriers to Gene Flow 4

Habitat preference/Eco-geographic isolation 5

Temporal isolation 6

Immigrant inviability 6

Pollinator preference and conspecific pollen precedence 7
Post-zygotic Barriers to Gene Flow 9
Hybrid incompatibility 9
Ecological and sexual selection against hybrids 10
Bateson-Dobzhansky-Muller incompatibilities (BDMI) 11
Chromosomal speciation 11
Genic speciation 12
Role of Polyploidy and Hybridization in Speciation 13
Polyploid speciation 13
When gene–flow promotes divergence 15
Genomics of Speciation 16
Conclusions 20
References 22

2 Genome–wide insights into the fine scale population structure and early demographic history of Western Poplar (*Populus trichocarpa*) 37

Abstract 37
Introduction 38
Materials and Methods 40
Sample collection and sequencing 40
Population structure 41

Linkage Disequilibrium and Genome-wide heterozygosity 41
Introgression analysis and Gene-flow 42
Effective population size estimation and Demographic Analysis 42
Divergence time estimation 44
Results 45
Genetic relationships between population 45
Patterns of Linkage Disequilibrium and Genome-wide heterozygosity 47
Introgression and Gene-flow 48
Effective population size and Divergence time estimation 49
Range Expansion 50
Divergence time estimation 51
Discussion 52
Patterns of population structure, linkage disequilibrium and gene-flow 52
Introgression with balsam poplar 53
Demographic history and effective population size 55
References 57
Figures and Tables 64
Supplementary Figures 69

3 Evolutionary Genomics of local adaptation in *Populus trichocarpa* 72

Abstract 72

Introduction 73

Material and Methods 75

Data collection, sequencing and population structure 75

Summary statistics 75

Signatures of selection 76

Genomic Islands of Divergence using Hidden Markov Model 76

Single locus genome scans 77

Enrichment analyses 78

Results 79

Summary statistics 79

Signatures of selection 81

Genomic islands of divergence 81

Single locus scans 82

Discussion 83

Summary Statistics 83

Genomic islands of divergence 85

Single locus scans 87

References 88

Figures and Tables 99

Supplementary Figures and Tables 106

List of Figures

2.1 Genetic structure of *P.trichocarpa* samples.

A. Geographic distribution of sample locations. B. Principal component analysis of the data, excluding the outgroups. The transition in colors from blue to red represent change in latitude from north to south. C. Admixture results for K = 3, 8. Colors in the graph represent different population clusters. . . 65

2.2 Linkage disequilibrium patterns and genome-wide heterozygosity.

A. LD measured for each population and each pair of SNPs within the 50 evenly spaced recombination distance categories. B. Box-plots for average genome-wide heterozygosity for each population 66

2.3 Introgression between *P.balsamifera* and *P. trichocarpa.*

A. Inferred tree showing population tree and historical admixture between balsam poplar and trichocarpa populations. The color gradient on the arrows indicate different migration weights at the gene-flow event B. Residual fit from the phylogenetic tree in A. 66

2.4 EEMS analysis showing the migration and diversity rates between all populations.

A. Effective migration rates and gene-flow barriers between all populations. Color gradient from blue to orange represent change from high to low gene-flow regions B. Diversity surface plot showing distribution of genetic diversity in the populations. Color gradient from blue to orange represent high to low genetic diversity. 67

2.5 Demographic history as inferred by Linkage disequilibrium method.

A. Spatial and temporal variation in N_e estimates. N_e was calculated using LD estimates from 50 evenly distributed recombination classes. B. Harmonic mean of the effective population size for each population over all recombination distance classes. N_e was also calculated separately for each chromosome, which was used to generate 95% confidence intervals. 67

2.6 Effective population size estimates and origin of range expansion.

A. Historical changes in effective population size inferred using MSMC method. The time scale on the x-axis was calculated using 1×10^{-8} mutation rate B. Range expansion based estimates of origin of ancestral population using Peter and Slatkin (2013) method for all samples. Brighter yellow regions indicate more likely origin of expansion. Red and gray regions indicate unlikely origin. 68

S2.1 Cross-validation error from the admixture runs. 69

S2.2 Pairwise comparisons of Genome-wide heterozygosity among all populations. 69

S2.3 Treemix results without adding the migration events.

A. Inferred graph. B. Residuals. 70

S2.4 Neighbor joining tree showing relationship between populations using T_f divergence estimates. 70

S2.5 Neighbor joining tree showing relationship between populations using T_{LD} divergence estimates. 71

S2.1 Geographical distribution of the samples used in the study, with color codings corresponding to different populations determined in the study Bawa and Holliday (2017). The neighbor joining tree on the left represents relationship between all 10 populations, and clustering of them to south, center and north for some of the analysis. 99

S2.2 Differences in nucleotide diversity and haplotype diversity among different population of *P. trichocarpa*. 100

S2.3 Distribution of Tajima's D statistic for 10 kb sliding window in three bigger sub groups (North, Center and South). The dotted line shows the simulated distribution of D expected under different demographic models, and the green density plot shows distribution of D under neutral expectation. 101

S2.4 Patterns of genomic divergence between three populations (North, South and Center) of *P. trichocarpa*. Green color denote SNPs that occur in the highly divergent regions. The top manhattan plot is North vs. South, central manhattan plot is North vs. Center and last one is Center vs. South populations. 102

S2.5 Selection scan results. The figure shows the results from HapFLK scan, OutFlank method and H-statistics respectively performed on all 10 demographic populations. The genome-wide thresholds correspond to $P < 0.01$ (red dashed lines). 103

S3.1 The distribution of observed and simulated values of Tajima's D in all 10 demographic populations. 106

S3.2 **Overlap in the genetic regions in the genomic islands of divergence** A. Overlap in the genetic regions between different pairs of populations for highly differentiated regions B. Overlap in the genetic regions between different pairs of populations for low differentiated regions 106

List of Tables

3.1 Properties of high and low genomic islands of differentiation identified using HMM approach. 104

3.2 Gene Ontology enrichments for selection scan outliers (common between hapFLK and OUTFLANK). 105

S3.1 Gene-Ontology enrichments for variants in highly differentiated regions in North vs. South populations. 107

S3.2 Gene-Ontology enrichments for variants in highly differentiated regions in North vs. Center populations. 107

S3.3 Gene-Ontology enrichments for variants in highly differentiated regions in Center vs South populations. 108

S3.4 Gene-Ontology enrichments for variants from the hapFLK selection test . . . 109

S3.5 Gene-Ontology enrichments for variants from OUTFLANK selection test. . . 110

Chapter 1

Genomics of Speciation in Temperate and Boreal Angiosperm Trees

Abstract

Understanding how species form and persist has been a central objective of evolutionary biology since the time of Darwin. Species emerge where two or more populations develop reproductive barriers yielding phenotypic and genetic divergence. This definition has led some to question the reality of species boundaries in plants. In contrast to animals, in which behavioral traits may allow sympatric groups to establish and maintain reproductive isolation (RI), plants do not benefit from mate selection in quite the same way. As a result, the definition of plant species has long confounded evolutionary biologists, as have the ecological and genetic processes that underlie their formation. Temperate and boreal angiosperm trees make an interesting case study in this regard, where a lack of RI often leads to difficulties in defining closely related species. With widespread and overlapping distributions, many tree species must resist gene-flow from congeners, often in the absence of pre- and post-zygotic barriers, in order to maintain adaptations to their particular niche. While high rates of intraspecific gene-flow likely promote species cohesion, angiosperm trees also show extensive inter-specific hybridization in contact zones. In this chapter we highlight recent

work to understand the pattern and process of speciation in angiosperm trees, and how contemporary sequencing technologies are providing insights into the genomic mechanisms that underlie genomic porosity on the one hand, and isolation on the other.

The principle of benefit derived from divergence of character...will lead to most divergent variations...being preserved and accumulated by natural selection – Charles Darwin

Modes of Speciation

Geography drives local adaptation and is therefore one of the major forces leading to speciation. Differences in altitude, latitude, edaphic factors, and biotic interactions can trigger the development of reproductive barriers, but some underlying level of reproductive isolation (RI) is usually necessary to establish local adaptation. The extent of local adaptation depends on the balance between gene–flow and selection (Savolainen et al., 2007), and this interplay is therefore key to speciation. Wind dispersal of pollen is extensive in trees, ranging from 100 of meters to several kilometers (Austerlitz et al., 2000; Smouse and Sork, 2004), and animal–mediated pollen flow, while less common, can also extend several kilometers (Savolainen et al., 2007; Hamrick, 2000). In wind-pollinated trees, seed dispersal can be 20-200 times lower than pollen dispersal (Ennos, 1994; Bacles et al., 2006), but fossil based seed dispersal estimates, ranging from 50 to 500 m/year (Hewitt, 1999), suggest that seed migration may have a significant impact on overall population connectivity. For example, in tulip poplar (*Liriodendron tulipifera*) seed dispersal up to 10 km has been reported (Nathan et al., 2002). The degree of effective migration between populations distinguishes speciation in terms of allopatry, which is purely geographic isolation with very low number of migrants per generation ($N_m \sim 0$), parapatry, in which populations occupy adjoining regions with low to intermediate level of gene-flow ($0 < N_m < 0.5$), and sympatry, where the effective

migration is high (Nm $>$ 0.5) (Coyne and Orr, 2004; Mallet et al., 2009).

Allopatric speciation occurs when there is no geographical contact between populations and thus very low or non-existent gene–flow (Coyne and Orr, 2004). In this scenario, selection and drift can act freely and independently (Nosil, 2012). Geological processes have led to RI and diversification across many tree genera. For example, divergence of three closely related *Populus* species (*P. tremuloides*, *P. davidiana* and *P. tremula*) coincided with the tectonic split of the Bering land bridge and the rapid uplift of the Tibetan plateau, which generated a geographical divide (Du et al., 2015). Allopatric speciation may also have been a consequence of post–glacial history. Many northern trees migrated out of glacial refugia to colonize higher latitudes since the last glacial maxima, and associated long distance dispersal could have been followed by decreased gene–flow from the ancestral population, resulting in geographical isolation that facilitated formation of independent evolutionary lineages. An interesting example is the sister species *Populus trichocarpa* and *P. balsamifera*, which diverged $\sim$75 ka. This time period corresponded with the last glacial maxima (LGM) and modeling of the paleoecological niches for these species during the LGM suggests a scenario of allopatric speciation, with *P. trichocarpa* restricted to the coast, and refugial *P. balsamifera* populations to central Alaska (which was ice-free due to low precipitation) and the intermountain western United States (Levsen et al., 2012).

While allopatric speciation has generated much extant diversity within tree genera, parapatric and sympatric speciation (collectively referred to as non-allopatric speciation) have seen a greater research focus in recent years due to their more complex genetic and ecological underpinnings. Non-allopatric divergence is characterized by high gene–flow, and its interplay with selection determines the pace of speciation (Coyne and Orr, 2004; Savolainen et al., 2007). Parapatric speciation, where there is balance between selection and gene–flow, is perhaps the most common mode of non-allopatric divergence. In parapatric speciation, newly

arisen adaptive mutations can become fixed in the face of gene–flow provided the strength of selection is greater than that of migration (Haldane 1948; Slatkin 1973). Parapatric divergence may be clinal, where populations are continuously distributed across a heterogeneous environment, or involve discrete populations (i.e., the stepping–stone model). Theoretical studies have shown that it is easier to achieve parapatric speciation in the context of discrete rather than continuous populations, but additional factors (e.g., assortative mating, pre– or post–zygotic isolation) are usually required. Cain et al. (1999) studied reinforcement dynamics in mosaic and clinal hybrid zones, and showed that the stepping-stone model is characterized by selection against both unfit homozygotes and heterozygotes, but clinal hybrid zones involve reinforcement selection only against heterozygotes, leading to faster divergence in the stepping-stone model. However, in trees, clinal variation between species across contact zones is more common (Lexer et al., 2007; Penaloza-Ramirez et al., 2010; Roe et al., 2014; Zeng et al., 2011). Hence, both pre– and post–zygotic barriers to gene–flow are expected to contribute to establishment and maintenance of RI where tree species diverge in the absence of geographical isolation.

Pre-zygotic Barriers to Gene Flow

Reproductive isolation (RI) involves the development of barriers to gene–flow among populations, which may be due to ecological or genetic factors. These barriers are key to the emergence of species, and normally are a by–product of diversifying selection and genetic drift (Rieseberg and Willis, 2007). Theory and empirical studies suggest RI is the most common way for species to form, as it does not require complete geographic isolation for genetic differences to accumulate (Agrawal et al., 2011; Lowry et al., 2008). RI causes an overall reduction in the effective migration rate, which leads to both neutral and adaptive

divergence among populations over time (Baack et al., 2015). These reproductive barriers can be differentiated based on their time of occurrence and can be caused both by external (ecological) forces or internal (genetic) forces. The first line of barriers are pre–zygotic (pre–pollination). Pre–pollination barriers restrict the transfer of pollen of one species to the stigma of another, and can be classified further as (i) eco–geographic isolation arising from differential habitat preference; (ii), mechanical isolation, caused by differences in reproductive structures; (iii) temporal isolation, such as differences in reproductive phenology; (iv) or competition based isolators (Baack et al., 2015).

Habitat preference/Eco-geographic isolation: Habitat isolation is defined as geographic isolation between populations, and arises due to local adaptation. This mode of RI is normally characterized by low fitness of individuals in the alternative habitat, reducing the likelihood of successful mating with individuals from different populations. Habitat isolation can be described in two forms: macrospatial and microspatial isolation. Macrospatial isolation is a form of allopatric speciation, where there is no gene–flow between populations due to geographical barriers. Microspatial isolation, on the other hand, is mostly driven by local adaptation, where populations occur in the same general area, but their reproductive encounters are reduced due to adaptations and preferences to different ecological niches (Coyne and Orr, 2004). Both macrospatial and microspatial isolation occur commonly in angiosperm trees. For example, a study in white oaks (*Quercus gambelii* and *Q. grisea*) showed that even though there is overlap in geographical distribution of both of these species, hybrids are rare even in areas of sympatry, and each of these lineages remains genetically distinct (Howard et al., 1997). Similarly, *Populus* species in the intermountain western United States (*P. angustifolia*, *P. fremontii* and *P. deltoides*) have overlapping geographical distributions and can freely hybridize in nature, but all the species maintain distinct niches, with few natural hybrids outside of contact zones. This could be attributed to distinct local adaptation of

these three lineages to different geographical environments that contribute to some level of reproductive isolation (Stettler, 1996).

Temporal isolation: Divergent developmental schedules among populations may facilitate RI, though in continuously distributed tree populations this mechanism may also contribute to a pattern of isolation–by–distance or isolation–by–adaptation. Temporal isolation in trees may arise due to climatic differentiation among populations that yields divergence in optimal timing of flowering. Among annual plants, extensive divergence in flowering time has been documented along environmental clines. For example, coastal and inland ecotypes of *Mimulus guttatus* are temporally isolated as a result of very little overlap in flowering time (Lowry et al., 2008). In the perennial sunflower *Helianthus maximilani*, strong differentiation was observed among populations for flowering time and other growth related traits. Due to their extended juvenile periods, coupled with the difficulty of studying flower phenology in reproductively mature (large) individuals, few studies of intra– or interspeci c variation in flowering have been undertaken in trees. In *Eucalyptus globulus*, dwarf coastal ecotypes evolved from tall inland ecotypes and differ in their flowering time. No introgression was found between these species in spite of their close proximity, which suggests temporal RI between them (Foster et al., 2007). Similar studies in other angiosperm trees, including *Platanus acerifolia*, *Ulmus minor*, and *Carpinus betulus*, showed divergence in flowering time in populations from different latitudes (Chuine et al., 2000), which suggests that RI as a function of differences in reproductive phenology may be an important route to RI in trees, but more research is needed in this area. While studies of vegetative phenology in trees has yielded extensive evidence for local adaptation in the timing of these transitions (Savolainen et al., 2007), more work is needed to understand the role of variation in the timing of flowering as a potential factor in the development of RI.

Immigrant inviability: This type of isolation normally arises when populations have re-

duced survival in the non–native habitat, and should not be confused with the habitat isolation where non–native populations are excluded in certain geographic regions or ecological niches. Immigrant inviability reduces effective gene–flow and fosters selection for habitat preferences through fitness trade–offs (Nosil, 2012). These trade-offs are associated with poor adaptation to non–native habitats and are caused by local adaptation to heterogeneous environments. In angiosperm trees, where pollen is usually wind–dispersed, distant populations subject to divergent selection frequently have lower fitness in their non–native habitat (Hardy et al., 2006; Kawecki and Ebert, 2004; Ducousso et al., 1996). Various factors – biophysical, physiological, and climatic – contribute to selection against immigrants. Many studies of immigrant inviability in trees have been conducted within species due to an interest in finding well-adapted provenances for reforestation (Morgenstern, 1996). Traits of particular importance to temperate and boreal trees are related to climatic adaptation, and include vegetative bud phenology (Luquez et al., 2008), fall, winter and spring cold hardiness(Howe et al., 2003), and drought tolerance (Street et al., 2006).

Pollinator preference and conspecific pollen precedence: While insect pollination is uncommon in temperate and boreal trees, sexual isolation in a few cases may be due to differential pollinator preferences that arise from co-evolution and/or habitat isolation. These barriers have been shown to be very effective mode to enforce RI among populations growing in close proximity (Coyne and Orr, 2004; Kirkpatrick and Ryan, 1991). While flower color is the most commonly identified pollinator divergence pattern in plants, other modes of pollinator divergence may be more important in angiosperm trees. In Joshua tree (*Yucca brevifolia*), a tree-like yucca endemic to the southwest United States, eastern and western subspecies are pollinated by different species of yucca moth (*Tegeticula synthetica* in the west and *T. antithetica* in the east) (Pellmyr and Segraves, 2003), which appears to be due to differences in oral morphology (Godsoe et al., 2008). Abiotic factors also play a role in

RI among populations. In European oaks, pollen fitness differed significantly for mesic and dry sites, leading to reduced cross-pollination among these two site types (Williams et al., 2001). While pollen-pistil interactions and differences in pollen fitness can act as absolute barriers, RI can be escalated by competition between inter– and intraspecific pollen, also known as conspecific pollen precedence (CPP). CPP is normally characterized by decreased fertilization rates interspecific pollen, and therefore is also referred to as a post-pollination pre–zygotic barrier. In oaks, fertilization and fruiting was reduced with inter–specific compared with intra–specific pollen (Varela et al., 2008), indicating pollen competition as a potential mode of RI (Rieseberg et al., 1995).

Pollinator preference and CPP can result from both ecological and genetic factors. While habitat preference or local adaptation might force plants to adapt to a new set of pollinators and pollination based morphological changes, an alternative reason for the development of pre–zygotic barriers is reinforcement (Hopkins, 2013). Recent studies in *Phlox spp.* and reweed (*Chamerion sp.*) have suggested development of these pollinator divergences as selection against hybridization (Hopkins and Rausher, 2012; Baldwin and Husband, 2011). Development of some of these pre–pollination barriers might have happened after development and establishment of post-zygotic barriers (Wallace et al., 1912). Since, most post–zygotic barriers involve wasted resources in terms of fewer gametes with lower fitness, populations might use pre–pollination barriers like flower color or conspecific pollen precedence to reduce interspecific pollination (Baack et al., 2015). Having said that, not many studies in angiosperm trees show this type of mating isolation, and other forms of RI (e.g., habitat preference or temporal isolation) that evolve early in the speciation process and reduce the need for sexual or mating isolation, may be more important (Nosil and Sandoval, 2008).

Post-zygotic Barriers to Gene Flow

Post-zygotic barriers are a second line of defense against inter–specific pollination, as they normally evolve after the establishment of pre–pollination barriers. These barriers can be divided into two forms. First, extrinsic or ecological barriers are caused by divergent selection or local adaptation and are dependent on genotype × environment interactions (Nosil, 2012). Extrinsic post–zygotic barriers include reduced hybrid viability, hybrid incompatibility, and sexual selection against hybrids. By contrast, intrinsic or genetic barriers are independent of the environment and are caused by genetic factors (Seehausen et al., 2014). Intrinsic post- zygotic barriers can take the form of Bateson-Dobzhansky-Muller incompatibilities, gene-duplications, mutations, genomic conflict, or chromosomal rearrangements.

Hybrid incompatibility: Post–mating isolation caused by incompatible genetic differences between loci of parental individuals is referred to as hybrid incompatibility. These incompatibilities are genetic in nature, but can be initiated by local adaptation. Such incompatibilities arise due to divergent selection that favors different combination of alleles in different geographical areas. When parents from different locally adaptive regions form a hybrid, the alleles from the different environments may not be compatible with each other. Various theoretical and empirical studies have shown that intrinsic genetic incompatibilities can be driven by ecological forces. A simulation study by Agrawal et al. (2011) showed that intrinsic hybrid incompatibility can originate via divergent selection and high linkage disequilibrium among adaptive loci, and suggested that these mechanisms could drive populations to RI even in the face of high gene–flow. This was corroborated by an experimental study in monkeyflower (*Mimulus sp.*) where adaptation to copper tolerance was linked to various loci responsible for hybrid incompatibility (Wright et al., 2013). Another interesting example where ecology drives hybrid incompatibility is cyto–nuclear interactions (Whittemore and Schaal, 1991; Fishman and Willis, 2006). These interactions are relatively common in plants,

and transplant experiments have shown lower hybrid fitness caused by a mismatch of the organelle and nuclear genomes from differentially adapted parents (Fields et al., 2014; Moison et al., 2010; Macaya-Sanz et al., 2011) showed extensive segregation distortion in backcross progeny arising from a cross between a female *Populus tremula* × *alba* F1 hybrid and a pure *P. alba* father. This distortion favored *P. tremula* haplotypes in the offspring, which may have been due to genomic conflict between interspecific cyto–nuclear combinations.

Ecological and sexual selection against hybrids: Similar to immigrant inviability, this mode of RI is characterized by lower fitness of the hybrid in either of its parents home environments. The basic premise behind such isolation is that hybrids formed by parents from two different environments will have an intermediate phenotype and highest fitness in the intermediate of the two 'extreme' niches. These intermediate niches are not very common in nature, leading to lower overall hybrid fitness (Nosil, 2012). In addition, intermediate hybrids may not have appropriate oral or reproductive traits, which could cause pollinator-based discrimination, leading to sexual isolation (Baack et al., 2015). This type of post–zygotic isolation should be very common among the trees, because of high gene–flow among populations. In the early stages of the speciation process, intraspecific gene–flow might lead to fertilization and formation of hybrids between ecotypes, but the net effective gene–flow is substantially reduced by ecological and sexual selection against hybrids, leading to RI (Varela et al., 2008; Roe et al., 2014). In European aspen (*Populus tremula*) and white poplar (*Populus alba*), preferential gene–flow was observed from European aspen to white poplar, even though these species can introgress with each other in either direction (Lexer et al., 2005). Similar studies of hybrid zones in the sympatric oaks *Quercus gambelii* and *Q. grisea*, showed preferential pollen movement and better fruit survival from Quercus gambelii to Q. grisea than in the other direction (Williams et al., 2001). This suggests ecological and sexual selection against hybrids as a form of RI in these systems.

Bateson-Dobzhansky-Muller incompatibilities (BDMI): BDMIs are an intrinsic post–zygotic barrier arising from epistatic interactions between two or more loci, which reduce fitness in the hybrids (Bateson, 1909; Muller, 1940; Dobzhansky et al., 1937). BDMIs were at first thought to be caused by genetic drift, but drift may be too slow and dependent on N_e to explain BDMIs, and meiotic drive, transposons, or selfish genetic elements are the more likely causes of such incompatibilities (Crespi and Nosil, 2013; Nei et al., 1983). No studies in angiosperm trees have empirically demonstrated the role of BDMIs in RI, but various types of admixture and hybrid isolation behaviors have been attributed to such epistatic interactions among loci. A review on admixture mapping and hybrid zones highlighted the importance of epistatic interactions in RI (Buerkle and Lexer, 2008). Since RI is normally effective when there is a synergistic relationship between multiple phenotypes, there is a good chance that epistatic interactions such as BDMI might play an important role in determining these effects (Buerkle and Lexer, 2008). Moreover, the phenotypes observed in hybrid zones might not be due to additive effects, but rather to epistatic interactions. Larcombe et al. (2015) performed cross-pollination experiments among eucalyptus species and fitted three different models of the relationship between hybrid fitness and genetic distance. The results suggest a decrease in hybrid fitness with increase in genetic distance and that the relationship between genetic distance and hybrid compatibility was exponential in nature. This result supports the snowball hypothesis of BDMI in generating RI.

Chromosomal speciation: Rearrangements,inversions,and translocations can cause disruptions in the meiotic paring of chromosomes, leading to sterility or incompatible hybrids (Livingstone and Rieseberg, 2004). In addition, these rearrangements decrease the frequency of recombination due to mechanical interference with crossing over, and therefore can lead to accumulation of hybrid incompatibilities in a relatively short period of time (Noor et al., 2001). The impact of these arrangements could also be due to the genes harbored in rear-

ranged regions (Navarro and Barton, 2003). A study in monkeyflower (Lowry and Willis, 2010) supports a role for chromosomal inversion in flowering time and other traits related to RI. Different ecotypes of monkeyflower were found to have two inversions, which when introgressed into different ecotypic backgrounds led to significant differences in adaptive traits. Recent studies in *Populus* and *Salix* have revealed the presence of chromosomal rearrangements in sex chromosomes, and their role in divergence. Hou et al. (2015) studied the evolution of sex chromosomes from autosomes in the *Salicaceae*, and found that in *Populus* sex chromosomes have recently evolved, following divergence from *Salix*, and that these chromosomes contain highly divergent haplotypes with severe suppression in recombination, suggesting the importance of sex chromosomes in RI between species. Coyne and Orr (2004) suggest that while chromosomal speciation may cause RI, it can only be effective in the presence of meiotic drive and small effective population size. More studies are needed to test the importance of chromosomal speciation in trees to determine if it can alone be an effective mode of RI, and this topic is further discussed in the section entitled 'Genomics of Speciation'.

Genic speciation: Locus–based incompatibilities originating by mutations, including duplications and deletions, fixed through selection or drift, are referred to as genic speciation. While mutations in a single gene could enable speciation, provided those mutations have a strong effect on RI (Nosil and Schluter, 2011), it is unlikely that most speciation events are caused by single mutations. The effect of single or multiple genes can be manifested either in the form of pleiotropy or high linkage disequilibrium with the genes causing RI (Coyne and Orr, 2004). Various empirical studies have demonstrated the role of genic speciation in RI. For example, the *yup* flower colour gene leads to pollinator-based isolation in monkeyflower (Moyle et al., 2012; Bradshaw and Schemske, 2003). While the *yup* case is a striking example, characterizing the genetic architecture of local adaptation may more generally lead

to a better understanding of the loci that underlie interspecific adaptive divergence. Recent population genomic studies have highlighted the role of divergent selection in speciation.

Role of Polyploidy and Hybridization in Speciation

Speciation until now has been mostly defined as divergence between populations, which has progressed long enough for the populations to diverge into species. Having said that, homoploid divergence is not the only way speciation progresses: polyploidization and hybridization are additional routes, very common in plants, that are known to play role in speciation (Coyne and Orr, 2004).

Polyploid speciation: Polyploidy is defined as the condition where the individual carries three or more set of complete chromosomes, and polyploidization is very important in context to speciation because it can be a rapid route to reproductive isolation (Ramsey and Schemske, 1998). Polyploidy may persist as stable sets of homeologous pairs of chromosomes, or go through a process of re–diploidization (i.e., paleopolyploidy). Being far more common in plants than animals, polyploidy is one of the major distinguishing features of plant speciation (Wood et al., 2009). Polyploids can be classified into three different types: autopolyploids, which arise from intraspecific crosses, allopolyploids, which arise from interspecific crosses, and autoallopolyploids, which combine the characteristics of both the auto and allopolyploids. Polyploids can be formed through somatic doubling, meiotic non–reduction, or polyspermy (Grant, 1981; Ramsey and Schemske, 1998). Somatic doubling involves failure of the chromosomes to separate during anaphase; meiotic non–reduction occurs when the cell membrane fails to form during gametogenesis, giving rise to diploid gametes; and polyspermy occurs when two male gametes (sperms) fertilize an egg (Grant, 1981; Levin, 2002). Polyspermy is uncommon in plants as compared to somatic doubling

and meiotic non–reduction, with the latter thought to be the most common mechanism of polyploidization (Ramsey and Schemske, 1998; Thompson and Lumaret, 1992). Meiotic non–reduction can be triggered by the presence of major genes, which include deleterious mutations at low frequency (Levin, 2002). Moreover, polyploidization may be involve an intermediate stage where a triploid individual backcrosses with one of the diploid parent to form the stable polyploid (Ramsey and Schemske, 1998). The frequency of polyploidy varies significantly among plants, with ∼31 % of ferns being polyploid, but only ∼15 % of angiosperms (Wood et al., 2009). Polyploidy is common in some tree genera (Levin and Wilson, 1976; Stebbins et al., 1950), which has been attributed to the greater degree of clonal/vegetative propagation compared with other taxa, which increases the incidence of somatic doubling (Grant, 1981). Self–fertilization, which is common in plants, increases the chances of polyploid establishment as compared to outcrossing species, where polyploid individuals may remain at low frequency for a long period of time (Grant, 1981; Stebbins et al., 1950). However, in trees, while selfing does occur, inbreeding depression is a significant factor (Petit and Hampe, 2006) and may limit the persistence of polyploids arising from selfing.

The role of polyploidy in speciation depends on the frequency with which different kinds of polyploidization occur. Autopolyploidy occurs at a much higher rate than allopolyploidy: 10-5 events per generation, which is similar to the genic mutation rate (Ramsey and Schemske, 1998). Ramsey and Schemske (1998) highlighted various factors that could favor one type of polyploidy over another in terms of speciation. First, autopolyploids are not easy to identify because they are morphologically similar to their parents, which suggests that in general autopolyploidy may not lead to diversification. In addition, autopolyploids are mostly sterile, so although common, autopolyploids are not persistent (Ramsey and Schemske, 1998). This is consistent with the rarity of in situ allopolyploidization on oceanic islands, wherein autopolyploidy is the primary route to polyploidization due to the higher level of divergent

species found on such islands (Schluter et al., 2000). While the rate of formation of allopolyploids may be similar to that of autopolyploids on a per meiosis basis, the former also depends on the rate of interspecific hybridization, which may be rare, and hence the overall number of allopolyploidization events is lower (Ramsey and Schemske, 1998).

There are many cases of diversification in plants resulting from polyploidization. For example, autotetraploid reweed (*Chamerion sp.*) flowers 1 week later compared to its diploid progenitors, causing reproductive isolation, and although there was some overlap in flowering between parental (diploid) and tetraploid reweed, this asynchrony decreased hybridization by at least 50 % (Husband and Schemske, 2000). In the perennial herb *Heuchera grossularifolia*, autotetraploid plants larger oral parts compared to the diploid parents, which altered the number and type of pollinators on the polyploid plant, leading to reproductive isolation (Segraves and Thompson, 1999). Angiosperm tree taxa harboring polyploid species include *Alnus* (alder), *Betula* (birch), *Fraxinus*(ash), *Tilia* (basswood or linden), *Acer* (maple), *Salix* (willow) and *Magnolia* (Wright, 1976). Phylogenetic analysis of birch, maple, ash, and magnolia reveals multiple polyploid lineages in each of these genera, which suggests independent polyploidization events led to speciation and diversification in these genera (Grimm et al. 2006; Jarvinen et al. 2004; Jeandroz et al. 1997; Parris et al. 2010). In addition to extant polyploids, there is evidence for paleopolyploidy in angiosperm trees, the best example of which being the salicoid duplication revealed by sequencing of the *Populus trichocarpa* genome (Tuskan et al., 2006).

When gene–flow promotes divergence: Gene–flow is one of the most important factors determining the pace of adaptive divergence and potential for speciation in trees. While it usually acts as a homogenizing force, decreasing phenotypic diversity among populations, an increase in geographical contact can also trigger speciation in some scenarios. Regions where two species geographical ranges overlap (i.e., hybrid zones) are characterized by strong

selection against maladapted hybrids. This selection against hybrids can cause asymmetric hybridization, and lead to assortative mating (Nosil, 2012; Lexer et al., 2007). If subsequent chromosomal rearrangements lead to sterility between the hybrid offspring and the parent species, recombinational speciation may ensue (Rieseberg, 1997). The most extensively studied example of potentially nascent hybrid speciation is in central Europe, where the ranges of *Populus alba* and *P. tremula* overlap. *P. alba* and *P. tremula* are morphologically distinct in their leaf characteristics, and occur in very different habitats, with the former mostly occurring in lowland flood plain forests and latter occurring in mixed upland habitats (Lexer et al., 2007). Extensive hybridization between the two parental species suggests weak reproductive isolation, and hybrids mostly occur in the contact zone and have an intermediate leaf phenotype (Lexer et al., 2009). Interestingly, substantial differences in ancestry have been observed between the seedling and adult stages, with the former comprised of the full range of hybrid classes, while the latter are mostly comprised of intermediate (especially F1) individuals. A defined genotypic discontinuity between the hybrids and parental species suggests this may be an example of incipient hybrid speciation (Lindtke et al., 2012). However, range shifts in the parental species may be necessary to reinforce RI with the hybrids for speciation to be realized (Lexer et al., 2010). Natural hybrid zones between tree species are common, and future work in these settings should shed new light on the role of gene–flow in promoting divergence.

Genomics of Speciation

The advent of next generation sequencing methods has led to a recent focus on the genomic regions underlying speciation (Hohenlohe et al., 2010; Hudson, 2008; Ellegren, 2008). Genetic differentiation is highly heterogeneous across the genome both within– and among–species,

and potential factors that affect interspecific genome–wide divergence include selection, drift, structural variation (Noor et al., 2001), variable mutation and recombination rates (Noor and Feder, 2006), and genomic conflict (Rice and Holland, 1997). The genomics of divergence are expected to differ between allopatric and sympatric speciation. In allopatric divergence, disruptive gene–flow is not an important factor – all regions of the genome therefore have an equal chance of differentiation, and even though this differentiation can be heterogeneous, clustering is not favored by migration–selection dynamics (Noor and Bennett, 2010; Turner and Hahn, 2010). Therefore, divergence in allopatry is expected to result smaller regions of differentiation, which are much more dispersed across the genome as compared to in the cases with gene–flow. For example, a study of genomic divergence in allopatric populations of ferns reported small, diffuse regions of differentiation (Nakazato et al., 2007). In contrast, gene–flow does impede both neutral and adaptive divergence in cases of sympatric speciation. Due to this homogenizing effect, tight linkage among adaptive genes should be favored (Gavrilets, 2004). Theoretical studies corroborate this expectation. Yeaman and Whitlock (2011) used individual-based simulations to examine the effects of low, intermediate, and high migration rates on the genetic architecture of adaptation. They found that as migration rates approach a critical threshold, above which local adaptation collapses, increasingly concentrated genetic architectures are favored. When migration was just below the critical threshold the effect was maximized, and divergence could be explained by a single large–effect haplotype. The magnitude of this effect depended on the mutation rate, effect size of individual mutations, and strength of selection, and was most pronounced when effect sizes of individual alleles within the 'island of divergence' was small. (Yeaman, 2013) further showed that these islands may be the result of genomic rearrangements that bring adaptive alleles into close physical proximity, rather than divergence hitchhiking alone.

Where species diverge in sympatry, a particular focus in recent years has been identifying

these regions of the genome subject to divergent selection, which may resist recombination between nascent or recently diverged species and therefore contribute to RI. F_{ST} is commonly used in this context as a measure of genetic differentiation, and unusual loci or regions are identified by comparison with a model–based or empirical null distribution (Beaumont and Balding, 2004; Foll and Gaggiotti, 2008). Another framework that has been employed in the context of natural hybrid zones uses multinomial regression to identify genomic clines that deviate from the genome–wide pattern of introgression, and are indicative of selection (Gompert and Buerkle, 2009). Regions of exceptional differentiation have been referred to as genomic islands or genomic continents of divergence, depending on their size (NOSIL et al., 2009; Via, 2011).

Studies in animal taxa provided the first evidence of genomic islands of divergence (Harr, 2006; Nadeau et al., 2011; Turner et al., 2005). Among plants, Andrew and Rieseberg (2013) studied wild, hybridizing sunflowers (*Helianthus sp.*) and reported varied levels of genetic differentiation and size of genomic islands between ecotypes within and between species. In contrast to interspecific comparisons between *H. annuus* and *H. petiolaris*, fewer and more clustered genomic islands were identified within species when ecologically divergent populations that occupy sand dune and non–sand dune habitats were compared. This result supports the work of (Yeaman and Whitlock, 2011) and suggests that early in the speciation process, when gene–flow is a significant impediment to genomic divergence, clustering of ecologically relevant loci promotes local adaptation. The findings of Andrew and Rieseberg (2013) are consistent with a recent study in black cottonwood (*Populus trichocarpa*), both altitudinal and latitudinal clines were sampled. The altitudinal clines, which entail higher migration relative to selection, had more numerous and larger genomic islands of divergence compared with the latitudinal case, which reinforces the idea that gene–flow promotes clustering of adaptive loci that resist the deleterious effects of recombination with maladapted

migrants (Holliday et al., 2016). A genome scans of two red oak species (*Quercus robus* and *Quercus petraea*) revealed multiple loci on different linkage groups as targets of divergence, and some of these targets were closely clustered and showed high correlation within distances < 2 cM (Scotti-Saintagne et al., 2004). This suggests the presence of multiple islands of divergence across the genome and the importance of linkage disequilibrium (genetic hitchhiking) in elevating the levels of divergence.

These studies and various others (Chapman et al., 2013; Horton et al., 2012; Hudson et al., 2015) highlight mechanisms through which divergence and RI can be achieved at the genomic level. One question of interest is the importance of the size of differentiated region in sympatric speciation. Larger genomic islands may be due to stronger divergence hitchhiking, but also have a higher probability of harboring additional adaptive alleles that promote divergence and resist gene–flow. However, it is difficult to tease apart the effects of linked selection from direct selection on multiple adaptive alleles (Brger and Akerman, 2011). (Nosil and Sandoval, 2008) showed that reduced gene–flow within islands of divergence could increase the role of genetic drift, causing additional neutral differentiation and increasing the overall genetic divergence. Large genomic islands could also evolve due to their location in an area of the genome with naturally has a low recombination, e.g. centromeres or chromosome inversion, and such regions have been shown to promote divergence in some cases (Noor et al., 2007; Feder et al., 2003; Kirkpatrick and Barton, 2006). Finally, genome–wide patterns of divergence could also be affected by the origin of adaptive genetic variation, whether new mutations or standing variation. While the theory above assumes adaptation from de novo mutation, adaptation probably frequently proceeds from already segregating variants, which may be conditionally neutral (Barrett and Schluter, 2008; Savolainen et al., 2013). Because standing genetic variation has no wait period for mutation establishment, such variants rise in frequency much more rapidly than new mutations (Barrett and Schluter, 2008). However,

detecting such regions is difficult, as historical recombination has broken linkage disequilibrium with other genomic regions, and selection produces a less pronounced signature of hitchhiking, a phenomenon known as a soft sweep (Przeworski, 2003). Various studies have highlighted the role of pre–existing variation in divergence. For example, a genome–wide scan of Arabidopsis thaliana showed partial and soft selective sweep between populations (Horton et al., 2012). Trees in particular harbor a great deal of standing variation, and it is likely that adaptation and speciation proceeds at least in part from standing variation (Holliday et al., 2010, 2016).

Conclusions

We have discussed both ecological and genetic factors that impinge on the potential for speciation. Which factor is most critical to the buildup of RI in angiosperm trees? Nosil et al. (2005) and Lowry et al. (2008) suggest that pre–pollination barriers are most common in general as compared to post–pollination and post–zygotic barriers, with habitat isolation and immigrant inviability having the strongest effect. These factors clearly contribute to the formation and maintenance of angiosperm tree species. Perhaps the most important factor for the maintenance of species is immigrant inviability (local adaptation). In spite of high rates of intra– and interspecific gene–flow (and concomitantly high genetic connectivity among populations), local adaptation, particularly with respect to climate-related traits, is very high in trees (see Fetter, Gugger & Keller, this volume). While most studies of local adaptation in trees involve intraspecific common gardens or reciprocal transplants, they clearly demonstrate a likely role for immigrant inviability in the maintenance of species boundaries. On the other hand, high gene–flow, coupled with continuous species distributions, makes it more difficult for ecological speciation to occur in trees. Multiple reproductive

barriers, acting simultaneously, are therefore likely required to establish RI. Where a single reproductive barrier can generate some level of isolation in presence of gene–flow, multiple reproductive barriers can cause complete or nearly complete RI (Sobel and Chen, 2014). Adaptive divergence may be the most common initial barriers to gene–flow in angiosperm trees, and intrinsic post–zygotic isolation may lead to reinforcement. The presence of chromosomal rearrangements and gene–duplications in the *Populus trichocarpa* genome suggests some role of intrinsic post–zygotic RI, but little is known about structural rearrangements in other species that may facilitate recombinational speciation. Recent advances in sequencing technologies (e.g., the Pacific Biosciences RSII platform) bring studies of structural variation within reach, and such data should shed additional light on the contribution of genome lability to speciation in trees.

References

Agrawal, A. F., Feder, J. L., and Nosil, P. (2011). Ecological divergence and the origins of intrinsic postmating isolation with gene flow. *International Journal of Ecology*, 2011:15.

Andrew, R. L. and Rieseberg, L. H. (2013). Divergence is focused on few genomic regions early in speciation: Incipient speciation of sunflower ecotypes. *Evolution*, 67(9):2468–2482.

Austerlitz, F., Mariette, S., Machon, N., Gouyon, P.-H., and Godelle, B. (2000). Effects of colonization processes on genetic diversity: Differences between annual plants and tree species. *Genetics*, 154(3):1309–1321.

Baack, E., Melo, M. C., Rieseberg, L. H., and Ortiz-Barrientos, D. (2015). The origins of reproductive isolation in plants. *New Phytologist*, 207(4):968–984.

Bacles, C. F., Lowe, A. J., and Ennos, R. A. (2006). Effective seed dispersal across a fragmented landscape. *Science*, 311(5761):628.

Baldwin, S. J. and Husband, B. C. (2011). Genome duplication and the evolution of conspecific pollen precedence. *Proceedings of the Royal Society B-Biological Sciences*, 278(1714):2011–2017.

Barrett, R. D. and Schluter, D. (2008). Adaptation from standing genetic variation. *Trends Ecol Evol*, 23(1):38–44.

Bateson, W. (1909). *Heredity and variation in modern lights.*, page 1 online resource. s.n. ,, S.l.

Beaumont, M. A. and Balding, D. J. (2004). Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, 13(4):969–980.

Bradshaw, H. D. and Schemske, D. W. (2003). Allele substitution at a flower colour locus produces a pollinator shift in monkeyflowers. *Nature*, 426(6963):176–178.

Buerkle, C. A. and Lexer, C. (2008). Admixture as the basis for genetic mapping. *Trends in Ecology & Evolution*, 23(12):686–694.

Brger, R. and Akerman, A. (2011). The effects of linkage and gene flow on local adaptation: A two-locus continentisland model. *Theoretical Population Biology*, 80(4):272 – 288.

Cain, M. L., Andreasen, V., and Howard, D. J. (1999). Reinforcing selection is effective under a relatively broad set of conditions in a mosaic hybrid zone. *Evolution*, 53(5):1343–1353.

Chapman, M. A., Hiscock, S. J., and Filatov, D. A. (2013). Genomic divergence during speciation driven by adaptation to altitude. *Molecular Biology and Evolution*, 30(12):2553–2567.

Chuine, I., Belmonte, J., and Mignot, A. (2000). A modelling analysis of the genetic variation of phenology between tree populations. *Journal of Ecology*, 88(4):561–570.

Coyne, J. A. and Orr, H. A. (2004). *Speciation*. Sinauer Associates, Sunderland, Mass.

Crespi, B. and Nosil, P. (2013). Conflictual speciation: species formation via genomic conflict. *Trends in Ecology & Evolution*, 28(1):48–57.

Dobzhansky, T., Dobzhansky, T., on Evolution, W. B. P. C., and Genetics. (1937). *Genetics and the origin of species*. Columbia biological series,. Columbia University Press, New York,.

Du, S. H., Wang, Z. S., Ingvarsson, P., Wang, D. S., Wang, J. H., Wu, Z. Q., Tembrock, L. R., and Zhang, J. G. (2015). Multilocus analysis of nucleotide variation and speciation in three closely related populus (salicaceae) species. *Molecular Ecology*, 24(19):4994–5005.

Ducousso, A., Guyon, J. P., and Kremer, A. (1996). Latitudinal and altitudinal variation of bud burst in western populations of sessile oak (quercus petraea (matt) liebl). *Annales Des Sciences Forestieres*, 53(2-3):775–782.

Ellegren, H. (2008). Sequencing goes 454 and takes large-scale genomics into the wild. *Molecular Ecology*, 17(7):1629–1631.

Ennos, R. A. (1994). Estimating the relative rates of pollen and seed migration among plant-populations. *Heredity*, 72:250–259.

Feder, J. L., Roethele, J. B., Filchak, K., Niedbalski, J., and Romero-Severson, J. (2003). Evidence for inversion polymorphism related to sympatric host race formation in the apple maggot fly, rhagoletis pomonella. *Genetics*, 163(3):939–953.

Fields, P. D., McCauley, D. E., McAssey, E. V., and Taylor, D. R. (2014). Patterns of cyto-nuclear linkage disequilibrium in silene latifolia: genomic heterogeneity and temporal stability. *Heredity*, 112(2):99–104. Original Article.

Fishman, L. and Willis, J. H. (2006). A cytonuclear incompatibility causes anther sterility in mimulus hybrids. *Evolution*, 60(7):1372–1381.

Foll, M. and Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A bayesian perspective. *Genetics*, 180(2):977–993.

Foster, S. A., McKinnon, G. E., Steane, D. A., Potts, B. M., and Vaillancourt, R. E. (2007). Parallel evolution of dwarf ecotypes in the forest tree eucalyptus globulus. *New Phytologist*, 175(2):370–380.

Gavrilets, S. (2004). *Fitness landscapes and the origin of species*. Monographs in population biology. Princeton University Press, Princeton, N.J. ; Oxford, England.

Godsoe, W., Yoder, J. B., Smith, C. I., and Pellmyr, O. (2008). Coevolution and divergence in the joshua tree/yucca moth mutualism. *The American Naturalist*, 171(6):816–823. PMID: 18462130.

Gompert, Z. and Buerkle, C. A. (2009). A powerful regression-based method for admixture mapping of isolation across the genome of hybrids. *Molecular Ecology*, 18(6):1207–1224.

Grant, V. (1981). *Plant speciation*. Columbia University Press, New York, 2nd edition.

Hamrick, J. L. & Nason, J. (2000). *Gene flow in forest trees*, pages xiv, 352 p. CSIRO ; CABI Pub., Collingwood, Vic. Wallingford, Oxon.

Hardy, O. J., Maggia, L., Bandou, E., Breyne, P., Caron, H., Chevallier, M. H., Doligez, A., Dutech, C., Kremer, A., Latouche-Halle, C., Troispoux, V., Veron, V., and Degen, B. (2006). Fine-scale genetic structure and gene dispersal inferences in 10 neotropical tree species. *Mol Ecol*, 15(2):559–71.

Harr, B. (2006). Genomic islands of differentiation between house mouse subspecies. *Genome Research*, 16(6):730–737.

Hewitt, G. M. (1999). Post-glacial re-colonization of european biota. *Biological Journal of the Linnean Society*, 68(1-2):87–112.

Hohenlohe, P. A., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. A., and Cresko, W. A. (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced rad tags. *Plos Genetics*, 6(2).

Holliday, J. A., Yuen, M., Ritland, K., and Aitken, S. N. (2010). Postglacial history of a widespread conifer produces inverse clines in selective neutrality tests. *Molecular Ecology*, 19(18):3857–3864.

Holliday, J. A., Zhou, L., Bawa, R., Zhang, M., and Oubida, R. W. (2016). Evidence for extensive parallelism but divergent genomic architecture of adaptation along altitudinal and latitudinal gradients in populus trichocarpa. *New Phytol*, 209(3):1240–1251.

Hopkins, R. (2013). Reinforcement in plants. *New Phytologist*, 197(4):1095–1103.

Hopkins, R. and Rausher, M. D. (2012). Pollinator-mediated selection on flower color allele drives reinforcement. *Science*, 335(6072):1090–1092.

Horton, M. W., Hancock, A. M., Huang, Y. S., Toomajian, C., Atwell, S., Auton, A., Muliyati, N. W., Platt, A., Sperone, F. G., Vilhjalmsson, B. J., Nordborg, M., Borevitz, J. O., and Bergelson, J. (2012). Genome-wide patterns of genetic variation in worldwide arabidopsis thaliana accessions from the regmap panel. *Nature Genetics*, 44(2):212–216.

Hou, J., Ye, N., Zhang, D. F., Chen, Y. N., Fang, L. C., Dai, X. G., and Yin, T. M. (2015). Different autosomes evolved into sex chromosomes in the sister genera of salix and populus. *Scientific Reports*, 5.

Howard, D. J., Preszler, R. W., Williams, J., Fenchel, S., and Boecklen, W. J. (1997). How discrete are oak species? insights from a hybrid zone between quercus grisea and quercus gambelii. *Evolution*, 51(3):747–755.

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

Hudson, C. J., Freeman, J. S., Myburg, A. A., Potts, B. M., and Vaillancourt, R. E. (2015). Genomic patterns of species diversity and divergence in eucalyptus. *New Phytologist*, 206(4):1378–1390.

Hudson, M. E. (2008). Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources*, 8(1):3–17.

Husband, B. C. and Schemske, D. W. (2000). Ecological mechanisms of reproductive isolation between diploid and tetraploid chamerion angustifolium. *Journal of Ecology*, 88(4):689–701.

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

Kirkpatrick, M. and Barton, N. (2006). Chromosome inversions, local adaptation and speciation. *Genetics*, 173(1):419–434.

Kirkpatrick, M. and Ryan, M. J. (1991). The evolution of mating preferences and the paradox of the lek. *Nature*, 350(6313):33–38.

Larcombe, M. J., Holland, B., Steane, D. A., Jones, R. C., Nicolle, D., Vaillancourt, R. E., and Potts, B. M. (2015). Patterns of reproductive isolation in eucalyptus-a phylogenetic perspective. *Molecular Biology and Evolution*, 32(7):1833–1846.

Levin, D. A. (2002). *The role of chromosomal change in plant evolution*. Oxford series in ecology and evolution. Oxford University Press, Oxford ; New York.

Levin, D. A. and Wilson, A. C. (1976). Rates of evolution in seed plants - net increase in diversity of chromosome-numbers and species numbers through time. *Proceedings of the National Academy of Sciences of the United States of America*, 73(6):2086–2090.

Levsen, N. D., Tiffin, P., and Olson, M. S. (2012). Pleistocene speciation in the genus populus (salicaceae). *Systematic Biology*, 61(3):401–412.

Lexer, C., Buerkle, C. A., Joseph, J. A., Heinze, B., and Fay, M. F. (2007). Admixture in european populus hybrid zones makes feasible the mapping of loci that contribute to reproductive isolation and trait differences (september 20, 2006). *Heredity*, 98(3):183–184.

Lexer, C., Fay, M. F., Joseph, J. A., Nica, M. S., and Heinze, B. (2005). Barrier to gene flow between two ecologically divergent populus species, p. alba (white poplar) and p. tremula (european aspen): the role of ecology and life history in gene introgression. *Molecular Ecology*, 14(4):1045–1057.

Lexer, C., Joseph, J., van Loo, M., Prenner, G., Heinze, B., Chase, M. W., and Kirkup, D. (2009). The use of digital image-based morphometrics to study the phenotypic mosaic in taxa with porous genomes. *Taxon*, 58(2):349–364.

Lexer, C., Joseph, J. A., van Loo, M., Barbara, T., Heinze, B., Bartha, D., Castiglione, S., Fay, M. F., and Buerkle, C. A. (2010). Genomic admixture analysis in european populus spp. reveals unexpected patterns of reproductive isolation and mating. *Genetics*, 186(2):699–712.

Lindtke, D., Buerkle, C. A., Barbar, T., Heinze, B., Castiglione, S., Bartha, D., and Lexer, C. (2012). Recombinant hybrids retain heterozygosity at many loci: new insights into the genomics of reproductive isolation in populus. *Molecular Ecology*, 21(20):5042–5058.

Livingstone, K. and Rieseberg, L. (2004). Chromosomal evolution and speciation: a recombination-based approach. *New Phytologist*, 161(1):107–112.

Lowry, D. B., Modliszewski, J. L., Wright, K. M., Wu, C. A., and Willis, J. H. (2008). The strength and genetic basis of reproductive isolating barriers in flowering plants. *Philosophical Transactions of the Royal Society B-Biological Sciences*, 363(1506):3009–3021.

Lowry, D. B. and Willis, J. H. (2010). A widespread chromosomal inversion polymorphism

contributes to a major life-history transition, local adaptation, and reproductive isolation. *Plos Biology*, 8(9).

Luquez, V., Hall, D., Albrectsen, B. R., Karlsson, J., Ingvarsson, P., and Jansson, S. (2008). Natural phenological variation in aspen (populus tremula): the swasp collection. *Tree Genetics & Genomes*, 4(2):279–292.

Macaya-Sanz, D., Suter, L., Joseph, J., Barbara, T., Alba, N., Gonzalez-Martinez, S. C., Widmer, A., and Lexer, C. (2011). Genetic analysis of post-mating reproductive barriers in hybridizing european populus species. *Heredity*, 107(5):478–486.

Mallet, J., Meyer, A., Nosil, P., and Feder, J. L. (2009). Space, sympatry and speciation. *J Evol Biol*, 22(11):2332–41.

Moison, M., Roux, F., Quadrado, M., Duval, R., Ekovich, M., L, D.-H., Verzaux, M., and Budar, F. (2010). Cytoplasmic phylogeny and evidence of cyto-nuclear co-adaptation in arabidopsis thaliana. *The Plant Journal*, 63(5):728–738.

Morgenstern, E. K. (1996). *Geographic variation in forest trees: genetic basis and application of knowledge in silviculture*. UBC Press, Vancouver.

Moyle, L. C., Levine, M., Stanton, M. L., and Wright, J. W. (2012). Hybrid sterility over tens of meters between ecotypes adapted to serpentine and non-serpentine soils. *Evolutionary Biology*, 39(2):207–218.

Muller, H. (1940). *Bearing of the Drosophila work on systematics.* University Press,, Oxford.

Nadeau, N. J., Whibley, A., Jones, R. T., Davey, J. W., Dasmahapatra, K. K., Baxter, S. W., Quail, M. A., Joron, M., ffrench Constant, R. H., Blaxter, M. L., Mallet, J., and Jiggins, C. D. (2011). Genomic islands of divergence in hybridizing heliconius butterflies identified

by large-scale targeted sequencing. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 367(1587):343–353.

Nakazato, T., Jung, M. K., Housworth, E. A., Rieseberg, L. H., and Gastony, G. J. (2007). A genomewide study of reproductive barriers between allopatric populations of a homosporous fern, ceratopteris richardii. *Genetics*, 177(2):1141–1150.

Nathan, R., Katul, G. G., Horn, H. S., Thomas, S. M., Oren, R., Avissar, R., Pacala, S. W., and Levin, S. A. (2002). Mechanisms of long-distance dispersal of seeds by wind. *Nature*, 418(6896):409–13.

Navarro, A. and Barton, N. H. (2003). Chromosomal speciation and molecular divergence - accelerated evolution in rearranged chromosomes. *Science*, 300(5617):321–324.

Nei, M., Maruyama, T., and Wu, C. I. (1983). Models of evolution of reproductive isolation. *Genetics*, 103(3):557–579.

Noor, M. A. F. and Bennett, S. M. (2010). Islands of speciation or mirages in the desert? examining the role of restricted recombination in maintaining species (vol 103, pg 434, 2009). *Heredity*, 104(4):418–418.

Noor, M. A. F. and Feder, J. L. (2006). Speciation genetics: evolving approaches. *Nature Reviews Genetics*, 7(11):851–861.

Noor, M. A. F., Garfield, D. A., Schaeffer, S. W., and Machado, C. A. (2007). Divergence between the drosophila pseudoobscura and d. persimilis genome sequences in relation to chromosomal inversions. *Genetics*, 177(3):1417–1428.

Noor, M. A. F., Grams, K. L., Bertucci, L. A., and Reiland, J. (2001). Chromosomal inversions and the reproductive isolation of species. *Proceedings of the National Academy of Sciences of the United States of America*, 98(21):12084–12088.

Nosil, P. (2012). *Ecological speciation*. Oxford series in ecology and evolution. Oxford University Press, Oxford ; New York.

NOSIL, P., FUNK, D. J., and ORTIZ-BARRIENTOS, D. (2009). Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, 18(3):375–402.

Nosil, P. and Sandoval, C. P. (2008). Ecological niche dimensionality and the evolutionary diversification of stick insects. *Plos One*, 3(4).

Nosil, P. and Schluter, D. (2011). The genes underlying the process of speciation. *Trends in Ecology & Evolution*, 26(4):160–167.

Nosil, P., Vines, T. H., and Funk, D. J. (2005). Perspective: Reproductive isolation caused by natural selection against immigrants from divergent habitats. *Evolution*, 59(4):705–719.

Pellmyr, O. and Segraves, K. A. (2003). Pollinator divergence within an obligate mutualism: Two yucca moth species (lepidoptera; prodoxidae: Tegeticula) on the joshua tree (yucca brevifolia; agavaceae). *Annals of the Entomological Society of America*, 96(6):716.

Penaloza-Ramirez, J. M., Gonzalez-Rodriguez, A., Mendoza-Cuenca, L., Caron, H., Kremer, A., and Oyama, K. (2010). Interspecific gene flow in a multispecies oak hybrid zone in the sierra tarahumara of mexico. *Annals of Botany*, 105(3):389–399.

Petit, R. J. and Hampe, A. (2006). Some evolutionary consequences of being a tree. *Annual Review of Ecology, Evolution, and Systematics*, 37(1):187–214.

Przeworski, M. (2003). Estimating the time since the fixation of a beneficial allele. *Genetics*, 164(4):1667–1676.

Ramsey, J. and Schemske, D. W. (1998). Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annual Review of Ecology and Systematics*, 29:467–501.

Rice, W. R. and Holland, B. (1997). The enemies within: intergenomic conflict, interlocus contest evolution (ice), and the intraspecific red queen. *Behavioral Ecology and Sociobiology*, 41(1):1–10.

Rieseberg, L. H. (1997). Hybrid origins of plant species. *Annual Review of Ecology and Systematics*, 28:359–389.

Rieseberg, L. H., Desrochers, A. M., and Youn, S. J. (1995). Interspecific pollen competition as a reproductive barrier between sympatric species of helianthus (asteraceae). *American Journal of Botany*, 82(4):515–519.

Rieseberg, L. H. and Willis, J. H. (2007). Plant speciation. *Science*, 317(5840):910–4.

Roe, A. D., MacQuarrie, C. J. K., Gros-Louis, M. C., Simpson, J. D., Lamarche, J., Beardmore, T., Thompson, S. L., Tanguay, P., and Isabel, N. (2014). Fitness dynamics within a poplar hybrid zone: I. prezygotic and postzygotic barriers impacting a native poplar hybrid stand. *Ecology and Evolution*, 4(9):1629–1647.

Savolainen, O., Lascoux, M., and Merila, J. (2013). Ecological genomics of local adaptation. *Nature Reviews Genetics*, 14(11):807–820.

Savolainen, O., Pyhjrvi, T., and Knrr, T. (2007). Gene flow and local adaptation in trees. *Annual Review of Ecology, Evolution, and Systematics*, 38(1):595–619.

Schluter, D., on Evolution, W. B. P. C., and Genetics. (2000). *The ecology of adaptive radiation*. Oxford series in ecology and evolution. Oxford University Press, Oxford.

Scotti-Saintagne, C., Mariette, S., Porth, I., Goicoechea, P. G., Barreneche, T., Bodenes, K., Burg, K., and Kremer, A. (2004). Genome scanning for interspecific differentiation between two closely related oak species [quercus robur l. and q petraea (matt.) liebl.]. *Genetics*, 168(3):1615–1626.

Seehausen, O., Butlin, R. K., Keller, I., Wagner, C. E., Boughman, J. W., Hohenlohe, P. A., Peichel, C. L., Saetre, G. P., Bank, C., Brannstrom, A., Brelsford, A., Clarkson, C. S., Eroukhmanoff, F., Feder, J. L., Fischer, M. C., Foote, A. D., Franchini, P., Jiggins, C. D., Jones, F. C., Lindholm, A. K., Lucek, K., Maan, M. E., Marques, D. A., Martin, S. H., Matthews, B., Meier, J. I., Most, M., Nachman, M. W., Nonaka, E., Rennison, D. J., Schwarzer, J., Watson, E. T., Westram, A. M., and Widmer, A. (2014). Genomics and the origin of species. *Nature Reviews Genetics*, 15(3):176–192.

Segraves, K. A. and Thompson, J. N. (1999). Plant polyploidy and pollination: Floral traits and insect visits to diploid and tetraploid heuchera grossulariifolia. *Evolution*, 53(4):1114–1127.

Smouse, P. E. and Sork, V. L. (2004). Measuring pollen flow in forest trees: an exposition of alternative approaches. *Forest Ecology and Management*, 197(13):21 – 38. Dynamics and Conservation of Genetic Diversity in Forest Ecology.

Sobel, J. M. and Chen, G. F. (2014). Unification of methods for estimating the strength of reproductive isolation. *Evolution*, 68(5):1511–1522.

Stebbins, G. L., Dunn, L. C., on Evolution, W. B. P. C., and Genetics. (1950). *Variation and evolution in plants*. Columbia biological series,. Columbia University Press, New York,.

Stettler, R. F. (1996). *Biology of populus and its implications for management and conservation*. NRC Research Press, Ottawa.

Street, N. R., Skogstrm, O., Sjdin, A., Tucker, J., Rodrguez-Acosta, M., Nilsson, P., Jansson, S., and Taylor, G. (2006). The genetics and genomics of the drought response in populus. *The Plant Journal*, 48(3):321–341.

Thompson, J. D. and Lumaret, R. (1992). The evolutionary dynamics of polyploid plants - origins, establishment and persistence. *Trends in Ecology & Evolution*, 7(9):302–307.

Turner, T. L. and Hahn, M. W. (2010). Genomic islands of speciation or genomic islands and speciation? *Molecular Ecology*, 19(5):848–850.

Turner, T. L., Hahn, M. W., and Nuzhdin, S. V. (2005). Genomic islands of speciation in anopheles gambiae. *Plos Biology*, 3(9):1572–1578.

Tuskan, G. A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R. R., Bhalerao, R. P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., Busov, V., Campbell, M., Carlson, J., Chalot, M., Chapman, J., Chen, G. L., Cooper, D., Coutinho, P. M., Couturier, J., Covert, S., Cronk, Q., Cunningham, R., Davis, J., Degroeve, S., Dejardin, A., Depamphilis, C., Detter, J., Dirks, B., Dubchak, I., Duplessis, S., Ehlting, J., Ellis, B., Gendler, K., Goodstein, D., Gribskov, M., Grimwood, J., Groover, A., Gunter, L., Hamberger, B., Heinze, B., Helariutta, Y., Henrissat, B., Holligan, D., Holt, R., Huang, W., Islam-Faridi, N., Jones, S., Jones-Rhoades, M., Jorgensen, R., Joshi, C., Kangasjarvi, J., Karlsson, J., Kelleher, C., Kirkpatrick, R., Kirst, M., Kohler, A., Kalluri, U., Larimer, F., Leebens-Mack, J., Leple, J. C., Locascio, P., Lou, Y., Lucas, S., Martin, F., Montanini, B., Napoli, C., Nelson, D. R., Nelson, C., Nieminen, K., Nilsson, O., Pereda, V., Peter, G., Philippe, R., Pilate, G., Poliakov, A., Razumovskaya, J., Richardson, P., Rinaldi, C., Ritland, K., Rouze, P., Ryaboy, D., Schmutz, J., Schrader, J., Segerman, B., Shin, H., Siddiqui, A., Sterky, F., Terry, A., Tsai, C. J., Uberbacher, E., Unneberg, P., et al. (2006). The genome of black cottonwood, populus trichocarpa (torr. & gray). *Science*, 313(5793):1596–1604.

Varela, M. C., Bras, R., Barros, I. R., Oliveira, P., and Meierrose, C. (2008). Opportunity

for hybridization between two oak species in mixed stands as monitored by the timing and intensity of pollen production. *Forest Ecology and Management*, 256(8):1546–1551.

Via, S. (2011). Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 367(1587):451–460.

Wallace, A. R., on Evolution, W. B. P. C., and Genetics. (1912). *Darwinism; an exposition of the theory of natural selection, with some of its applications, with portrait of the author, map and illustrations*. Macmillan, London,, [3d edition.

Whittemore, A. T. and Schaal, B. A. (1991). Interspecific gene flow in sympatric oaks. *Proceedings of the National Academy of Sciences of the United States of America*, 88(6):2540–2544.

Williams, J. H., Boecklen, W. J., and Howard, D. J. (2001). Reproductive processes in two oak (quercus) contact zones with different levels of hybridization. *Heredity*, 87:680–690.

Wood, T. E., Takebayashi, N., Barker, M. S., Mayrose, I., Greenspoon, P. B., and Rieseberg, L. H. (2009). The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences of the United States of America*, 106(33):13875–13879.

Wright, J. W. (1976). *Polyploidy, aneuploidy and haploidy*, pages 399–437. Academic Press, New York.

Wright, K. M., Lloyd, D., Lowry, D. B., Macnair, M. R., and Willis, J. H. (2013). Indirect evolution of hybrid lethality due to linkage with selected locus in mimulus guttatus. *Plos Biology*, 11(2).

Yeaman, S. (2013). Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proceedings of the National Academy of Sciences*, 110(19):E1743–E1751.

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

Zeng, Y. F., Liao, W. J., Petit, R. J., and Zhang, D. Y. (2011). Geographic variation in the structure of oak hybrid zones provides insights into the dynamics of speciation. *Molecular Ecology*, 20(23):4995–5011.

Chapter 2

Genome–wide insights into the fine scale population structure and early demographic history of Western Poplar (*Populus trichocarpa*)

Abstract

Rangewide patterns of genomic variation provide insights into the geographical factors and neutral genetic forces that shape contemporary population structure and gene flow, as well as the early demographic history of populations. Here, we present and analyze population genomic dataset (with $>10^6$ SNPs) from 391 rangewide samples of the deciduous tree *Populus trichocarpa*, and provide a comprehensive assessment of genomic diversity within and among populations. We show that the current distribution of *P. trichocarpa* originated from ancestral refugia in the south dating to ∼100KYA. Population structure analysis show evidence of extensive gene-flow and assortative mating between the populations. We also found that current populations at the northern range limit are isolated from the rest of the samples, and show signatures of introgression with *P. balsamifera*. We hypothesize that some of this introgression is adaptive, and discuss the evolutionary implications of this kind of

introgression.

Introduction

The wide availability of whole genome re-sequencing data and recent developments in the application of coalescent theory to such data have brought significant advances in our understanding of genetic variation, and what the underlying patterns of genetic variation can tell us about organisms history (Bustamante et al., 2011; 100, 2012; Gazave et al., 2014). Coalescent analyses, which were previously limited to a relatively small number of independent neutral loci, can now be applied to whole genome sequences, providing unprecedented resolution of species' evolutionary history (Li and Durbin, 2011; Schiffels and Durbin, 2014; Liu and Fu, 2015). These genome-wide reconstructions of population history shed light on the evolutionary processes that shape the genetic composition of populations with complex demographic histories (Schraiber and Akey, 2015; Bustamante et al., 2011). This genetic reconstruction is especially relevant in temperate and boreal forest trees, as they have experienced dramatic migrations and demographic events during the late Pleistocene (Holliday et al., 2012; Savolainen et al., 2007; Austerlitz et al., 2000).

Range expansions and contractions leave genetic signatures in both plastid and nuclear DNA (Excoffier et al., 2009; Cavalli-Sforza et al., 1993; Barbujani et al., 1995). These signatures, coupled with demographic (founder effects) and selective factors, can increase genetic differentiation among population (Nielsen and Slatkin, 2013; Feder and Nosil, 2010). This differentiation can be opposed by gene flow, which is a significant factor in forest trees, extending from few to thousands of meters (Hewitt, 1999; Goto et al., 2006; Valbuena-Carabana et al., 2005). The interplay between neutral (gene-flow, random genetic drift, demographic factors) and selective forces lead to complicated genetic diversity and divergence

patterns (Schraiber and Akey, 2015; Nielsen, 2005), and disentangling these correlates of diversity, is therefore central to understanding patterns of extant genetic variation.

Populus trichocarpa is a widespread tree species of primarily riparian ecosystems in western North America, with a range extending from California to Alaska (Figure 2.1a). Past studies using ecological modeling in *P. trichocarpa* have indicated strong effects of Pleistocene glaciation, with refugial habitats restricted to coastal and southern parts of its current range (Levsen et al., 2012). Other studies focusing on introgression and hybridization dynamics have also shown admixture between *P. trichocarpa* and its sister species, *P. balsamifera* (Geraldes et al., 2014; Suarez-Gonzalez et al., 2016). Moreover, *P. trichocarpa* is one of the most widely studied tree speices, with a relatively small and fully sequenced genome (Tuskan et al., 2006), making it a promising candidate for genetic reconstruction of demographic history. Despite several population genomic studies conducted on *P. trichocarpa* in the past few years (Slavov et al., 2012; Zhou and Holliday, 2012; Evans et al., 2014), none have focused on detailed dissection of population structure and demographic history among populations. Here, we investigated the genome-wide effects of demographic change during range expansion using exome resequencing data from range-wide population of *P. trichocarpa*. We use contemporary coalescent methods to infer the demographic history and historical introgression between population of *P. trichocarpa* and discuss impact of demographic history and migration patterns on extant genetic variation.

Materials and Methods

Sample collection and sequencing

Young leaf tissue was collected from 391 *P. trichocarpa* individuals growing in a common garden located at the Reynolds Homestead Forestry Research Center in Patrick County, Virginia. These samples originated from collections spanning most of the latitudinal range of the species, ranging from 37° N to 61° N latitude (Fig. 2.1a and Supplementary File 3). Genotyping was accomplished using sequence capture, and has been previously described (Zhou and Holliday, 2012). Briefly, 173,040 RNA baits targeting most exons, upstream regulatory regions, and 600 intergenic control regions were synthesized by Agilent Technologies ((Santa Clara, CA). Library preparation followed the manufacturers protocol (Agilent SureSelectXT version 1.2), and sequencing was carried out with 12 samples per lane on an Illumina HiSeq 2500 instrument. Following quality control and demultiplexing, the data were aligned with the *P. trichocarpa* v3.0 reference genome using the Burrows-Wheeler aligner mem algorithm (Li and Durbin, 2009) and SNPs were called using the GenomeAnalysisToolKit (GTK) Unified Genotyper (DePristo et al., 2011). Prior to GATK SNP calling, duplicates were identified and removed using the Picard (https://broadinstitute.github.io/picard/command-line-overview.html) MarkDuplicates and GATK DuplicateReadFilter functions, respectively (DePristo et al., 2011), and indels realigned using the GATK IndelRealigner function. Candidate SNPs from GATK were considered where their quality score was >50 and missing data <25%, resulting in a final dataset consisting of 1,058,513 SNPs.

Population structure

We used ADMIXTURE software (Alexander et al., 2009) and Principal Component Analysis (PCA) to explore the patterns of population structure in our dataset. Prior to this, we removed SNPs in high linkage disequilibrium ($r^2 > 0.1$) using PLINK software (Purcell et al., 2007), which left 188,774 SNPs. We tested nine ancestral clusters from K=2 through K = 10 in admixture, and used the ten-fold cross-validation procedure to choose an optimal value of K for our dataset. For PCA, we used the SNPRelate (Zheng et al., 2012) package in R. Both of these analyses were performed with and without out-groups to investigate the change in clustering pattern and degree of hybridization in the *P. trichocarpa* samples with the out-groups.

Linkage Disequilibrium and Genome-wide heterozygosity

We estimated linkage disequilibrium (LD) for all our population using PLINK (Purcell et al., 2007). 1,014,751 SNPs, which could be placed on a genetic map (kindly provided by Dr. Steve DiFaizo, West Virginia University) were used to calculate correlation between each marker pair for each population. Two measures of LD were considered, the correlation (r_{LD}) and squared correlation ($r_{LD}{}^2$) in genotype frequencies. The genome-wide data were binned into 50 uniformly spaced recombination distance categories (0.005 0.25 cM) and the $r_{LD}{}^2$ was averaged for each category. This averaged squared correlation was compared among all populations for all recombination distance categories. We used Plink (Purcell et al., 2007) and custom scripts in R to calculate the observed heterozygosity rate for each individual in our dataset, and then used that to calculate average genome-wide heterozygosity for each population. Following that, we compared the average levels of heterozygosity among all groups of populations using the nonparametric Kruskal-Wallis test.

Introgression analysis and Gene-flow

We used two complementary approaches, TreeMix (Pickrell and Pritchard, 2012) and Estimated Effective Migration Surface (EEMS) (Petkova et al., 2015), to estimate introgression and gene-flow within and between populations and out-groups. Treemix uses a composite-likelihood approach to model the evolutionary history of populations, which may include population splits and gene-flow events (Pickrell and Pritchard, 2012). Balsam poplar (*Populus balsamifera*) samples were included in this analysis to assess introgression with *P. trichocarpa*. Parameters for the model were a window size of 1000 SNPs and between one and four migration events. The four-population test (Patterson et al., 2012) and three population test (Reich et al., 2009) were used to test robustness of gene-flow events predicted by Treemix.

EEMS was used to identify potential regions with high and low historic gene-flow. EEMS uses genetic data from geo-referenced samples and an "isolation by resistance" method to estimate the migration surface in a continuous landscape (Petkova et al., 2015). This method can decipher complex patterns of genetic dissimilarity, which otherwise cannot be identified using PCA or admixture. We used the downsampled data ($(r^2 > 0.1)$ with 188,774 SNPs, and divided the region into three parts (north, center and south), to examine subtler variation in each subgroup. All runs used a randomly generated seed for at least 8 million iterations in the MCMC sampler.

Effective population size estimation and Demographic Analysis

In order to explore the demographic and migratory history of our populations, we used three approaches. First, we estimated changes in effective population size (N_e) over time using the LD-based method of McEvoy et al. (2011), which exploits the relationship of LD with

recombination distance to estimate N_e. Since time (t) can be approximated to 1/2c (Hayes et al., 2003), where c is the recombination distance category, we used average ${r_{LD}}^2$ for each binned time interval (0.005 – 0.25cM) to calculate change in effective population size over time. The equations used to measure the effective population size are:

$$E({r_{\mathrm{LD}}}^2) = 1/(\alpha + 4N_e c) \tag{2.1}$$

$$N_e = 1/[(4c) * [(1/{r_{\mathrm{LD}}}^2) - 2]] \tag{2.2}$$

where $\alpha = 2$ and accounts for impact of mutation; c is the recombination distance category. The details of these methods can be found in McEvoy et al. (2011)

Second, we applied the Multiple Sequentially Markovian Coalescent (MSMC) approach (Schiffels and Durbin, 2014) to infer the demographic history of our populations. Phased genome sequence data for two individuals per group were used to estimate N_e. Input data were generated per chromosome from Variant Call Format (VCF) files, and the analysis was run using for 100 iterations with fixed recombination. A generation time of 15 years and mutation rate of 1×10^{-8} was used to convert coalescent units into chronological time.

Finally, we used Peter and Slatkin (2013) range expansion method to detect the origin and direction of population expansion in our data. This approach finds asymmetries in the site frequency spectrum to detect founder effects and model complicated scenarios of expansion. We used LD–based pruning to remove highly linked SNP variants ($r^2 > 0.5$, and ∼700,000 SNPs) and divided our data into smaller sub regions to test for multiple origins of range expansion.

Divergence time estimation

We used LD in two different ways to estimate the divergence times of our populations. The first method (T_F) exploits the relationship between F_{ST} and effective population size, and the second method uses the correlation of LD for two populations to estimate divergence times (McEvoy et al., 2011). All the estimates were done using $\leq$ 0.1cM distance categories, since it provided the better longer term estimate of Ne over time.

T_F – Assuming a neutral evolutionary scenario with no migration, genetic drift determines genetic differentiation (F_{ST}) (Holsinger and Weir, 2009), which depends on effective population size (N_e) and divergence time (T). According to McEvoy et al. (2011) and Nei et al. (1987), this relationship can be approximated as:

$$F_{ST} \approx T/2N_e \tag{2.3}$$

$$T_F \approx 2N_e F_{ST} \tag{2.4}$$

We calculated the SNP–wise F_{ST} for all our populations and estimated average F_{ST} between them. N_e was calculated as the harmonic mean for each population, which was averaged among populations to calculate divergence time.

T_{LD} – This method is based on the premise that LD between two populations after a recent split should be in perfect correlation (r_{pop}), and that this correlation should deteriorate as the function of time. Thus, the following relationship was used to estimate divergence times of our population:

$$ln(r_{\text{pop}}) = c(-2T_{\text{LD}}) \tag{2.5}$$

where, r_{pop} is the correlation in LD pattern between the two population, c is the recombination distance category and T_{LD} is the timing of divergence (McEvoy et al., 2011). The files were split by chromosome, and divergence times was calculated across each chromosome to calculate confidence intervals for both divergence time estimates.

Results

Genetic relationships between population

We used principal component analysis (PCA) to explore the structure present in our populations. The first two axes explained 5.2% and 3.3%, respectively, and were driven mainly by latitudinal variation among the populations. The Anchorage and Kodiak populations clustered separately, while rest of the populations were arrayed along a roughly north-south cline (Fig. 2.1b). This suggests geographical isolation of the northern populations (Anchorage and Kodiak), and high gene-flow (both direct and indirect) and allelic clines between the rest of the samples. In order to further understand population structure, we used the maximum likelihood based clustering algorithm ADMIXTURE (Alexander et al., 2009). This method considers the samples to have originated from some hypothetical number (K) of ancestral populations, and uses an unsupervised clustering algorithm to assign the proportion of genetic ancestry to each of these ancestral populations. We tested K = 2 through K = 10 ancestral populations, and used the consistency between the runs and ADMIXTURE's 10-fold cross-validation procedure to determine best K for our dataset. K = 8 had lowest cross–validation error, and showed a closer relationship of Anchorage and Kodiak popula-

tions to *P. balsamifera* (balsam poplar) (Fig. 2.1c, and Fig. S3.1). The other outgroups (*P. deltoides*, *P. fermontii*, *P. tremula* and *P. tremuloides*) did not show any admixture with the *P. trichocarpa* samples. This was not a surprising result because *P. trichocarpa* is known to hybridize with balsam poplar, while other outgroups are either reproductively isolated (*P. tremula* and *P. tremuloides*) or do not hybridize extensively with *P. trichocarpa* (*P. deltoides* and *P. fermontii*), at least in the areas we sampled. Moreover, as with PCA, ADMIXTURE assigned Anchorage and Kodiak populations into a separate cluster, again representing geographic isolation of these populations from rest of the samples. Populations from British Columbia, Canada (BC) and Washington state (WA) were assigned to a one large cluster suggesting extensive gene-flow between these samples. Unlike the populations from BC and WA, southern samples were assigned to three distinct clusters, which suggests assortative mating and some isolation between these populations and the rest of the samples. Admixture also showed signs of introgression between balsam poplar and the Coquihalla (CQ) populations. Even though the CQ population originates in southern BC, these samples reach into a pass that connects coastal and interior BC, where the balsam poplar is found, which may result in introgression. As demographic history in trees is predominantly affected by post-glacial migration events and long distance dispersal followed by isolation (Savolainen et al., 2007; Excoffier et al., 2009), we used both latitudinal and elevation information along with population structure results to split our rangewide sample into distinct groups. These groups were named based on their location and origin of the samples and the names are: Anchorage (AK), Kodiak (KD), Northern B.C. (NBC), Central B.C. (CBC), Coquihalla (CQ), Southern B.C. (SBC), Northern U.S. (NUS), Greenwood (GW), Oregon (OR), Sierras (CA). More information about the samples and population assignments can be found in File S1.

Patterns of Linkage Disequilibrium and Genome-wide heterozygosity

We examined genome-wide linkage disequilibrium (LD) decay patterns in our populations, and the results (Fig. 2.2a) showed that LD increases with increasing geographical distance from the southern range limit. This has been previously reported by Slavov et al. (2012), and likely reflects postglacial migration of the species from southern refugia to higher latitudes. The highest values of LD were found in Anchorage and Kodiak populations, which had higher LD over short genomic distances (0.005–0.075 cM; ($r_{Mean}^2 = 0.631$ for Kodiak and $r_{Mean}^2 = 0.617$ for Anchorage) as compared to rest of the populations (NBC: $r_{Mean}^2 = 0.57$; CBC: $r_{Mean}^2 = 0.55$; CQ: $r_{Mean}^2 = 0.56$; SBC: $r_{Mean}^2 = 0.57$; NUS: $r_{Mean}^2 = 0.58$; GW: $r_{Mean}^2 = 0.53$; OR: $r_{Mean}^2 = 0.53$; Sierras: $r_{Mean}^2 = 0.53$). High LD over long genomic distances (0.075–0.25 cM) was also seen in Anchorage ($r^2 = 0.591$) and Kodiak ($r^2 = 0.579$) populations as compared to rest of the groups (NBC: $r_{Mean}^2 = 0.52$; CBC: $r_{Mean}^2 = 0.48$; CQ: $r_{Mean}^2 = 0.50$; SBC: $r_{Mean}^2 = 0.50$; NUS: $r_{Mean}^2 = 0.53$; GW: $r_{Mean}^2 = 0.46$; OR: $r_{Mean}^2 = 0.45$; Sierras: $r_{Mean}^2 = 0.45$).

Since the northern populations clustered closely with balsam poplar, we estimated genome–wide heterozygosity in our samples to assess whether hybridization may be contributing to these LD patterns. The results revealed significantly higher heterozygosity in Anchorage ($H_{Mean} = 0.0913$) and Kodiak populations ($H_{Mean} = 0.0922$) compared with the rest of the groups (NBC: $H_{Mean} = 0.076$; CBC: $H_{Mean} = 0.07$; CQ: $H_{Mean} = 0.082$; SBC: $H_{Mean} = 0.069$; NUS: $H_{Mean} = 0.07$; GW: $H_{Mean} = 0.073$; OR: $H_{Mean} = 0.074$; Sierras: $H_{Mean} = 0.068$) (Fig. 2.2b and S2.2). This suggests that the Anchorage and Kodiak populations have hybridized with balsam poplar, leading to increase in heterozygosity. Interestingly, the CQ population had the highest H_{Mean} of all the southern populations, which is consistent with the low level

of introgression apparent from the ADMIXTURE results.

Introgression and Gene-flow

To further investigate hybridization in northern populations we used Treemix and EEMS to model introgression and gene–flow events. First, we used Treemix (Pickrell and Pritchard, 2012) to build a maximum likelihood tree with our populations, using balsam poplar as an outgroup, and added migration events until the best fit of the tree was achieved. The tree with no migration events explained 91.5% of the variation, but the residuals from the inferred tree were very high, suggesting the presence of gene flow events. We therefore sequentially added migration events to the tree. The new population tree allowing four migration edges explained 99.7% of the variation in relatedness among-population, with p-values for each edge less than 2.2 10_{-308} (refer to Fig. S2.3a 2.3b for the no migration graph and residuals). The new inferred graph showed clear signs of recent admixture between balsam poplar and Kodiak, and showed ancestral introgression between balsam poplar and Anchorage populations (Fig. 2.3a and 2.3b). We further examined these migration events using three–population and four–population tests for treeness (Reich et al., 2009; Patterson et al., 2012). The trees [KD; [CBC, BAL]] and [CQ; [NUS, BAL]] failed the three-population tests ($Z = -32.6$, $p = 1 \times 10^{-30}$; $Z = -31.87$, $p = 1 \times 10^{-30}$), which suggests gene–flow between these populations. Similarly, the four-population test [[GW, ANCHORAGE]; [CBC, BAL]] and [[KD, NBC]; [OR, BAL]] failed ($Z = 35.1$, $p = 1 \times 10^{-30}$; $Z = -37.3$, $p = 1 \times 10^{-30}$), providing additional evidence of introgression between balsam poplar and Kodiak, as well as ancestral Anchorage and Kodiak populations.

We also used EEMS method to identify possible barriers to gene-flow, to better characterize admixture and gene–flow events (Petkova et al., 2015). We identified three possible barriers

to gene–flow (Fig.2.4a). All of these barriers were geographical and caused by the heterogeneous landscape and presence of mountain ranges. The first barrier was between Anchorage, Kodiak, and rest of the populations, while the second barrier separated southern BC and WA population from the northern samples. The third barrier corresponded to the Sierra Nevada mountains and separated Sierra populations from rest of the samples. EEMS also illustrates the distribution of genetic diversity on the landscape, and showed that the Kodiak and Sierra populations had low genetic diversity, suggesting geographic isolation. The barrier between SBC and NUS populations did not show low genetic diversity, suggesting higher levels of gene-flow (both direct and indirect) between the populations, despite the barrier.

Effective population size and Divergence time estimation

We further investigated the demographic history of the species, and used two different approaches to infer ancestral population sizes. Simulation studies investigating the effect of introgression and gene–flow and introgression on estimation of effective population size have shown that, even a small amount of introgression from a genetically divergent population can affect the estimation of effective population size (Hawks, 2017). We, therefore excluded Anchorage and Kodiak populations from the demographic analysis. We first used the method of McEvoy et al. (2011) to estimate the harmonic mean N_e over all recombination distance categories. The average N_e ranged from 130 in OR to 30 in NBC populations (Fig. 2.5b) These estimates were considerably lower than the ones reported by MSMC method, which ranged from 1×10^6 to 4×10^6. This huge difference in estimation of effective population sizes could be attributed to the approach used by both of these methods. LD based method uses genetic distance as a measure of time (1/c), and uses its relationship with linkage disequilibrium to measure effective population size. The genetic distance can only extend to

so much, in this case 0.005 – 0.25 cM, limiting the extent of N_e by this method. Moreover, MSMC method has also been known to be affected by introgression and gene-flow events, which lead to the inflation in the estimates. So, it could be combination of both the differences in approaches as well as the affect of continuous gene–flow between population, that might have affected the estimates of effective population size. We also found a considerable amount of both temporal (genetic distance) and spatial variation (populations) in N_e (Fig. 2.5a). Ne estimates from 100-20 KYA period showed a much smaller effective population size for most populations, with OR populations originating first around 60 KYA ago, followed by remaining populations. All other populations originated between 40-20 KYA, with the Kodiak populations being the most recent. Beginning 17 KYA, all the populations stabilized over time, with the Oregon populations decreasing in growth around 7 KYA. As a complementary approach, we estimated population size changes with the Multiple Sequential Markovian Coalescent (MSMC) method, which uses hidden markov model with recombination to estimate N_e (Schiffels and Durbin, 2014). Similar to the results described above, MSMC analysis suggest the OR populations had the highest effective population size (2×10^6) around 100KYA years ago (Figure 2.6a). All other populations showed expansion much later, around 20KYA, which was consistent with results of the LD-based method.

Range Expansion

Following the estimation of effective population size, we wanted to look at the range expansion dynamics, and find different possible origins of range expansion. This was important because all the recent populations in the north are hypothesized to have expanded from the southern refugia, and if the above results of OR populations being the ancestral refugia are correct, we should be able to identify the origin of range expansion in the south. To do this, we used the method of Peter and Slatkin (2013). The results showed that the populations

in fact have migrated to higher latitudes from the south, with the origin of range expansion close to OR populations (Figure 2.6b). Overall, based all the effective population size and range expansion results, we identify a likely ice-age refugium close to the contemporary Oregon populations.

Divergence time estimation

We used linkage disequilibrium (LD) estimates to measure the divergence time between each pair of our populations. LD can be used in two different ways to estimate divergence time. The first method uses the relationship of N_e and F_{ST} and the second uses similarity in LD structure over time to estimate divergence time (refer to methods and McEvoy et al. (2011) for details). The results showed signatures of isolation by distance, with more geographically proximal populations having smaller divergence time than more distant ones. The results of both the T_F and T_{LD} estimates are summarized in the distance matrix and neighbor-joining tree (Fig. S2.4, S2.5; File S1). Both divergence time estimates also showed that the Anchorage population was more closely related to NBC than Kodiak populations, even though the Kodiak populations are more geographically proximal to Anchorage than NBC populations.

Discussion

Patterns of population structure, linkage disequilibrium and gene-flow

Our results revealed complex patterns of genetic ancestry in our populations and shed light on the factors that have influenced the relationship between them. Principal component analysis and admixture results suggest rampant gene-flow in the center and south of the range. Admixture showed a presence of large population cluster in the center, while PCA showed a northward to southward cline reflecting high gene–flow among these populations (Reich et al., 2008). Forest tree populations in general are known for high gene-flow, which ranges from few hundred to thousands of meters (Savolainen et al., 2007; Hewitt, 1999; Valbuena-Carabana et al., 2005; Goto et al., 2006). All these factors could have led to the overall low level of genetic structure in populations (Holliday et al., 2012; Petit et al., 2003).

We also found strong signatures of isolation of northern populations from rest of the range. Both PCA and admixture clustered the samples from Anchorage and Kodiak separately. Similar results were also seen in the EEMS surface plot, which suggested the presence of a gene-flow barrier between northern populations (Anchorage and Kodiak) and rest of the samples. These populations also had very high values of linkage disequilibrium, suggesting high geographic isolation. Anchorage and Kodiak populations have a large geographic separation from rest of the samples in the form of the St. Elias mountains, and the ∼50km straight separating Kodiak Island from the mainland further contributes to its geographic isolation (Fig. 2.1a). That said, these populations are also characterized by very short growing seasons and flowering timing that likely differs from the rest of the samples (Viereck and Foote, 1970; Oubida et al., 2015). This phenological asynchrony may also contribute to

genetic isolation of samples (Rieseberg and Willis, 2007; Roe et al., 2014).

There was also evidence of restricted gene flow among the southern populations. Samples from GW, OR and Sierras populations, even though in close geographical proximity, were clustered into separate populations by the admixture analysis. EEMS results also indicated the same by showing high genetic diversity in the southern populations. These results suggest reproductive isolation among populations, and could be caused due to heterogeneous topography, resulting in limited gene-flow between the groups (Moreno-Estrada et al., 2013).

Introgression with balsam poplar

We found evidence of introgression between balsam poplar and samples from the northern populations. Admixture results clustered the Anchorage and Kodiak populations with *P. balsamifera* and showed high levels of introgression from balsam poplar into these populations. This signature was also corroborated by high levels of genome-wide heterozygosity in these samples. Hybridization can be a huge source of genetic variation, if the introgression happens with the genomically divergent population or a species. It can bring alleles which are different from the reference and segregating loci, and can therefore increase the estimate of genomic heterozygosity in the samples. Further genomic analyses using treemix gave us an in-depth picture of this admixture, showing evidence of both ancient and recent introgression between balsam poplar and northern populations (Anchorage, Kodiak, and NBC).

P. balsamifera and *P. trichocarpa* are sister species that diverged recently (∼75Ka) due to allopatry at the time of Pleistocene glaciation (Levsen et al., 2012). Due to their recent divergence, they have very similar reproductive morphology and growth times and are known to hybridize where their geographic ranges overlap (Geraldes et al., 2014; Viereck and Foote, 1970). The presence of a hybrid zone where the geographical ranges of two closely related

species overlap is common in forest tree species (Lexer et al., 2006; Lexer and Widmer, 2008; Suarez-Gonzalez et al., 2016), and allows introgression between balsam poplar and northern populations. There are several evolutionary and adaptive implications to the observed introgression pattern. Previous studies in trees in general, and *Populus* sp. in particular, have suggested that hybridization could be a source to adaptive genetic variation if that introgression allows movement of adaptive genomic regions between species (Rieseberg, 1997; Rieseberg and Willis, 2007). For example, Geraldes et al. (2014) studied hybrid zones between *P. balsamifera* and *P. trichocarpa* and found that there was selective introgression of adaptive genes from balsam poplar to *P. trichocarpa* populations. Balsam poplar is adapted to extreme cold temperatures, able to withstand winter temperatures as low as low as -62°C, whereas *P. trichocarpa* is adapted to mild, humid conditions (Isebrands and Richardson, 2014). Given the locations of the northern populations (Anchorage and Kodiak) are significantly colder than most of the *P. trichocarpa* range, and have concomitant very short growing season, introgression from balsam poplar could facilitate local adaptation of these population (Geraldes et al., 2014; Suarez-Gonzalez et al., 2016).

Treemix results also suggested ancient admixture between balsam poplar and the Anchorage population. While we did not include these samples in our analysis of demographic history due to increased level of introgression, Levsen et al. (2012) used ecological niche modeling and found that Pleistocene glaciation events correspond to *P. trichocarpa* and balsam poplars divergence time. They showed that *P. trichocarpa* and *P. balsamifera* were one species before glaciation, and that it was an ice age scenario that isolated the two populations, leaving *P. balsamifera* restricted to unglaciated regions of the intermountain west and *P. trichocarpa* other to coastal and southern refugia. Thus, it is possible that the current Anchorage population originates from old hybrids of southern and northern populations of *P. trichocarpa*, or that a small northern *P. trichocarpa* refuge remained unglaciated and

then became the source to Anchorage populations (Breen et al., 2012; Levsen et al., 2012). Future studies focusing on northern populations demographic history are needed to gain more insight into the genetic diversity and admixture patterns in them.

Demographic history and effective population size

Our estimates of ancestral population size from both the LD–based method and MSMC revealed the presence of ancestral refugia in the vicinity of the Oregon populations. Using MSMC, we estimated that the OR population originated around 100Kya, while the LD–based method showed the origin of OR around 60Kya. As the LD–based method is an approximate estimate based on genetic distance and LD–decay, it is susceptible to underestimation of effective population size. On the other hand, MSMC uses the SMC approximation approach, which models multiple recombination events between the markers to estimate the effective population size. Moreover, we also estimated the variability in MSMC estimates by running this method on the smaller portions of data, providing broader confidence intervals in N_e estimates. Regardless of the timing, the origin of the OR populations pre-dated those of the other populations, indicating of it to be the likely refugium. Similar results were obtained using Peter and Slatkin (2013) range expansion method, which showed a similar ancestral refugia.

The Wisconsin glaciation event culminated ∼76Kya, which was followed by retreat of glaciers over time (Levsen et al., 2012). The glaciation likely restricted *P. trichocarpa* populations to southern refugia close to the OR populations, but with the retreat of the glaciers the species migrating to higher latitudes, forming the current latitudinal extent of the species range (Levsen et al., 2012; Taberlet and Cheddadi, 2002). Locations identified as ancestral refugia are characterized by high haplotypic diversity, along with large effective population size

(Petit et al., 2003). Similar patterns were found in our EEMS diversity plot, and a previous study by Slavov et al. (2012), who found large effective population sizes and high haplotypic diversity near the Willamette and Columbia river (the location where OR populations were sampled), indicating of the presence of refugia.

This pattern of range expansion has special implications for the genetic makeup of the populations. Range expansion generally entail recurrent bottlenecks, which can lead to loss of genetic diversity in newly founded populations (Excoffier et al., 2009). Even though high gene-flow in trees can compensate for some of this loss of genetic diversity (Savolainen et al., 2007), these founding bottlenecks and frequent colonization events can increase the rate of genetic drift and decrease the efficacy of adaptive selection (Charlesworth and Wright, 2001). Range expansion can also lead to allele surfing along the migration front, which can increase the persistence of maladaptive phenotypes in the population (Excoffier and Ray, 2008; Burton and Travis, 2008; Travis and Dytham, 2002). These factors may lead to adaptive lag, preventing populations from reaching their adaptive optimum (Aitken et al., 2008). Indeed, a recent study suggests that deleterious alleles are more prevalent in peripheral populations of *P. trichocarpa* (Zhang et al., 2016).

In conclusion, our analysis of exome data from intraspecific populations of *P.trichocarpa* provide a comprehensive view of genetic structure and demographic history of the species. We show that both geography and demographic history have played major roles in shaping extant genetic variation within and among populations of *P. trichocarpa*, and that the current species range was likely colonized from glacial refugia in the south. One question that remains unanswered is that of the demographic history in the northern populations, where both Anchorage and Kodiak samples showed introgression with balsam popular. Future studies of the fine scale hybridization patterns in northern populations and their demographic history will provide a more complete picture of the early history of western poplar.

References

(2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65.

Aitken, S. N., Yeaman, S., Holliday, J. A., Wang, T., and Curtis-McLane, S. (2008). Adaptation, migration or extirpation: climate change outcomes for tree populations. *Evol Appl*, 1(1):95–111.

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*, 19(9):1655–64.

Austerlitz, F., Mariette, S., Machon, N., Gouyon, P.-H., and Godelle, B. (2000). Effects of colonization processes on genetic diversity: Differences between annual plants and tree species. *Genetics*, 154(3):1309–1321.

Barbujani, G., Sokal, R. R., and Oden, N. L. (1995). Indo-european origins: A computer-simulation test of five hypotheses. *American Journal of Physical Anthropology*, 96(2):109–132.

Breen, A. L., Murray, D. F., and Olson, M. S. (2012). Genetic consequences of glacial survival: the late quaternary history of balsam poplar (populus balsamifera l.) in north america. *Journal of Biogeography*, 39(5):918–928.

Burton, O. J. and Travis, J. M. J. (2008). The frequency of fitness peak shifts is increased at expanding range margins due to mutation surfing. *Genetics*, 179(2):941–950.

Bustamante, C. D., De La Vega, F. M., and Burchard, E. G. (2011). Genomics for the world. *Nature*, 475(7355):163–165.

Cavalli-Sforza, L., Menozzi, P., and Piazza, A. (1993). Demic expansions and human evolution. *Science*, 259(5095):639–646.

Charlesworth, D. and Wright, S. I. (2001). Breeding systems and genome evolution. *Current Opinion in Genetics & Development*, 11(6):685–690.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nat Genet*, 43(5):491–498.

Evans, L. M., Slavov, G. T., Rodgers-Melnick, E., Martin, J., Ranjan, P., Muchero, W., Brunner, A. M., Schackwitz, W., Gunter, L., Chen, J.-G., Tuskan, G. A., and DiFazio, S. P. (2014). Population genomics of populus trichocarpa identifies signatures of selection and adaptive trait associations. *Nat Genet*, 46(10):1089–1096. Article.

Excoffier, L., Foll, M., and Petit, R. J. (2009). Genetic consequences of range expansions. *Annual Review of Ecology Evolution and Systematics*, 40:481–501.

Excoffier, L. and Ray, N. (2008). Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecology & Evolution*, 23(7):347 – 351.

Feder, J. L. and Nosil, P. (2010). The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution*, 64(6):1729–1747.

Gazave, E., Ma, L., Chang, D., Coventry, A., Gao, F., Muzny, D., Boerwinkle, E., Gibbs, R. A., Sing, C. F., Clark, A. G., and Keinan, A. (2014). Neutral genomic regions refine models of recent rapid human population growth. *Proceedings of the National Academy of Sciences*, 111(2):757–762.

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

Goto, S., Shimatani, K., Yoshimaru, H., and Takahashi, Y. (2006). Fat-tailed gene flow in the dioecious canopy tree species fraxinus mandshurica var. japonica revealed by microsatellites. *Molecular Ecology*, 15(10):2985–2996.

Hawks, J. (2017). Introgression makes waves in inferred histories of effective population size. *ArXiv e-prints*.

Hayes, B. J., Visscher, P. M., McPartlan, H. C., and Goddard, M. E. (2003). Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res*, 13(4):635–43.

Hewitt, G. M. (1999). Post-glacial re-colonization of european biota. *Biological Journal of the Linnean Society*, 68(1-2):87–112.

Holliday, J. A., Suren, H., and Aitken, S. N. (2012). Divergent selection and heterogeneous migration rates across the range of sitka spruce (picea sitchensis). *Proc Biol Sci*, 279(1734):1675–83.

Holsinger, K. E. and Weir, B. S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting fst. *Nat Rev Genet*, 10(9):639–650.

Isebrands, J. G. and Richardson, J. (2014). *Poplars and willows : trees for society and the environment*. CABI ; FAO, Boston, MA Rome.

Levsen, N. D., Tiffin, P., and Olson, M. S. (2012). Pleistocene speciation in the genus populus (salicaceae). *Systematic Biology*, 61(3):401–412.

Lexer, C., Buerkle, C. A., Joseph, J. A., Heinze, B., and Fay, M. F. (2006). Admixture in european populus hybrid zones makes feasible the mapping of loci that contribute to reproductive isolation and trait differences. *Heredity*, 98(2):74–84.

Lexer, C. and Widmer, A. (2008). The genic view of plant speciation: recent progress and emerging questions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1506):3023–3036.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–60.

Li, H. and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–U84.

Liu, X. M. and Fu, Y. X. (2015). Exploring population size changes using snp frequency spectra. *Nature Genetics*, 47(5):555–U172.

McEvoy, B. P., Powell, J. E., Goddard, M. E., and Visscher, P. M. (2011). Human population dispersal "out of africa" estimated from linkage disequilibrium and allele frequencies of snps. *Genome Research*, 21(6):821–829.

Moreno-Estrada, A., Gravel, S., Zakharia, F., McCauley, J. L., Byrnes, J. K., Gignoux, C. R., Ortiz-Tello, P. A., Martinez, R. J., Hedges, D. J., Morris, R. W., Eng, C., Sandoval, K., Acevedo-Acevedo, S., Norman, P. J., Layrisse, Z., Parham, P., Martinez-Cruzado, J. C., Burchard, E. G., Cuccaro, M. L., Martin, E. R., and Bustamante, C. D. (2013). Reconstructing the population genetic history of the caribbean. *Plos Genetics*, 9(11).

Nei, M., Nei, M., on Evolution, W. B. P. C., and Genetics. (1987). *Molecular evolutionary genetics*. Columbia University Press, New York.

Nielsen, R. (2005). Molecular signatures of natural selection. *Annu Rev Genet*, 39:197–218.

Nielsen, R. and Slatkin, M. (2013). *An introduction to population genetics : theory and applications*. Sinauer Associates, Sunderland, Mass.

Oubida, R. W., Gantulga, D., Zhang, M., Zhou, L., Bawa, R., and Holliday, J. A. (2015). Partitioning of multivariate phenotypes using regression trees reveals complex patterns of adaptation to climate across the range of black cottonwood (populus trichocarpa). *Frontiers in Plant Science*, 6:181.

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient admixture in human history. *Genetics*, 192(3):1065–93.

Peter, B. M. and Slatkin, M. (2013). Detecting range expansions from genetic data. *Evolution*, 67(11):3274–3289.

Petit, R. J., Aguinagalde, I., de Beaulieu, J.-L., Bittkau, C., Brewer, S., Cheddadi, R., Ennos, R., Fineschi, S., Grivet, D., Lascoux, M., Mohanty, A., Mller-Starck, G., Demesure-Musch, B., Palm, A., Martn, J. P., Rendell, S., and Vendramin, G. G. (2003). Glacial refugia: Hotspots but not melting pots of genetic diversity. *Science*, 300(5625):1563–1565.

Petkova, D., Novembre, J., and Stephens, M. (2015). Visualizing spatial population structure with estimated effective migration surfaces. *bioRxiv*.

Pickrell, J. K. and Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet*, 8(11):e1002967.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., and Sham, P. C. (2007). Plink: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3):559–575.

Reich, D., Price, A. L., and Patterson, N. (2008). Principal component analysis of genetic data. *Nat Genet*, 40(5):491–2.

Reich, D., Thangaraj, K., Patterson, N., Price, A. L., and Singh, L. (2009). Reconstructing indian population history. *Nature*, 461(7263):489–94.

Rieseberg, L. H. (1997). Hybrid origins of plant species. *Annual Review of Ecology and Systematics*, 28:359–389.

Rieseberg, L. H. and Willis, J. H. (2007). Plant speciation. *Science*, 317(5840):910–4.

Roe, A. D., MacQuarrie, C. J. K., Gros-Louis, M. C., Simpson, J. D., Lamarche, J., Beardmore, T., Thompson, S. L., Tanguay, P., and Isabel, N. (2014). Fitness dynamics within a poplar hybrid zone: I. prezygotic and postzygotic barriers impacting a native poplar hybrid stand. *Ecology and Evolution*, 4(9):1629–1647.

Savolainen, O., Pyhjrvi, T., and Knrr, T. (2007). Gene flow and local adaptation in trees. *Annual Review of Ecology, Evolution, and Systematics*, 38(1):595–619.

Schiffels, S. and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nat Genet*, 46(8):919–25.

Schraiber, J. G. and Akey, J. M. (2015). Methods and models for unravelling human evolutionary history. *Nat Rev Genet*, 16(12):727–740. Review.

Slavov, G. T., DiFazio, S. P., Martin, J., Schackwitz, W., Muchero, W., Rodgers-Melnick, E., Lipphardt, M. F., Pennacchio, C. P., Hellsten, U., Pennacchio, L. A., Gunter, L. E., Ranjan, P., Vining, K., Pomraning, K. R., Wilhelm, L. J., Pellegrini, M., Mockler, T. C., Freitag, M., Geraldes, A., El-Kassaby, Y. A., Mansfield, S. D., Cronk, Q. C. B., Douglas, C. J., Strauss, S. H., Rokhsar, D., and Tuskan, G. A. (2012). Genome resequencing

reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree populus trichocarpa. *New Phytologist*, 196(3):713–725.

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

Taberlet, P. and Cheddadi, R. (2002). Quaternary refugia and persistence of biodiversity. *Science*, 297(5589):2009–2010.

Travis, J. M. J. and Dytham, C. (2002). Dispersal evolution during invasions. *Evolutionary Ecology Research*, 4(8):1119–1129.

Tuskan, G. A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R. R., Bhalerao, R. P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., Busov, V., Campbell, M., Carlson, J., Chalot, M., Chapman, J., Chen, G.-L., Cooper, D., Coutinho, P. M., Couturier, J., Covert, S., Cronk, Q., Cunningham, R., Davis, J., Degroeve, S., Déjardin, A., dePamphilis, C., Detter, J., Dirks, B., Dubchak, I., Duplessis, S., Ehlting, J., Ellis, B., Gendler, K., Goodstein, D., Gribskov, M., Grimwood, J., Groover, A., Gunter, L., Hamberger, B., Heinze, B., Helariutta, Y., Henrissat, B., Holligan, D., Holt, R., Huang, W., Islam-Faridi, N., Jones, S., Jones-Rhoades, M., Jorgensen, R., Joshi, C., Kangasjärvi, J., Karlsson, J., Kelleher, C., Kirkpatrick, R., Kirst, M., Kohler, A., Kalluri, U., Larimer, F., Leebens-Mack, J., Leplé, J.-C., Locascio, P., Lou, Y., Lucas, S., Martin, F., Montanini, B., Napoli, C., Nelson, D. R., Nelson, C., Nieminen, K., Nilsson, O., Pereda, V., Peter, G., Philippe, R., Pilate, G., Poliakov, A., Razumovskaya, J., Richardson, P., Rinaldi, C., Ritland, K., Rouzé, P., Ryaboy, D., Schmutz, J., Schrader, J., Segerman, B., Shin, H.,

Siddiqui, A., Sterky, F., Terry, A., Tsai, C.-J., Uberbacher, E., Unneberg, P., Vahala, J., Wall, K., Wessler, S., Yang, G., Yin, T., Douglas, C., Marra, M., Sandberg, G., Van de Peer, Y., and Rokhsar, D. (2006). The genome of black cottonwood, populus trichocarpa (torr. & gray). *Science*, 313(5793):1596–1604.

Valbuena-Carabana, M., Gonzalez-Martinez, S. C., Sork, V. L., Collada, C., Soto, A., Goicoechea, P. G., and Gil, L. (2005). Gene flow and hybridisation in a mixed oak forest (quercus pyrenaica willd. and quercus petraea (matts.) liebl.) in central spain. *Heredity*, 95(6):457–465.

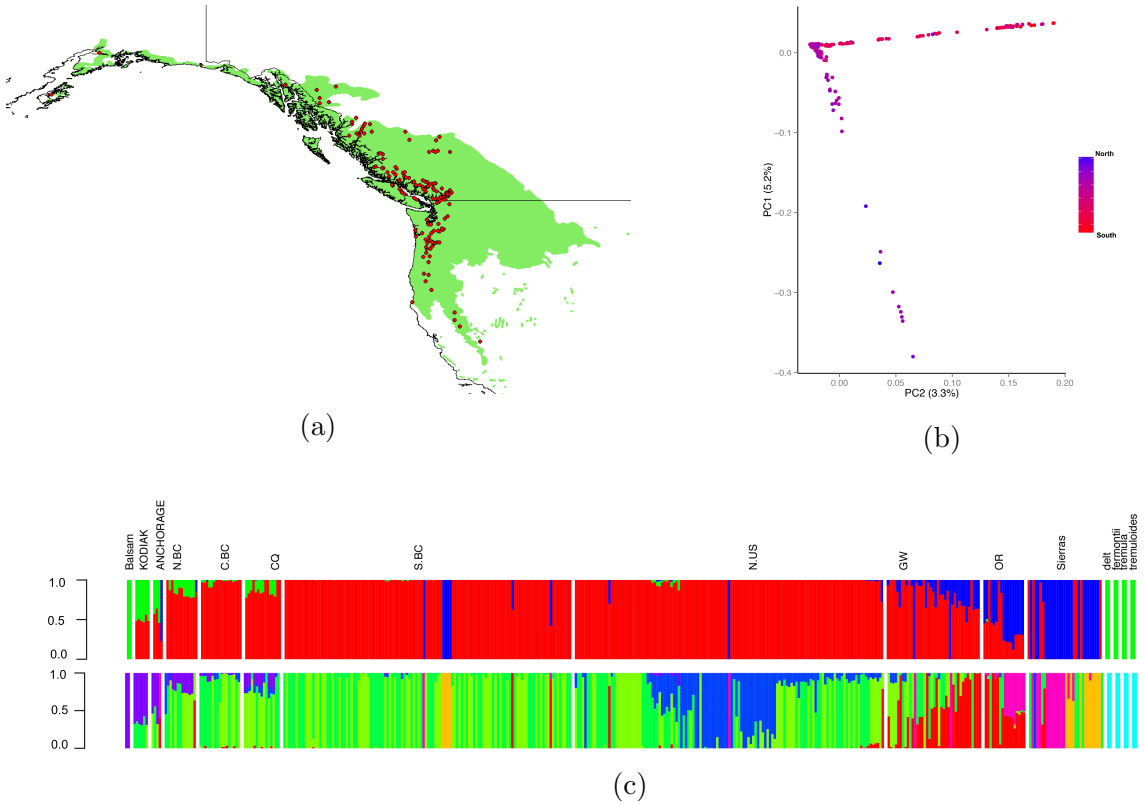
Viereck, L. A. and Foote, J. (1970). *The status of Populus balsamifera and P. trichocarpa in Alaska*. Canadian field-naturalist.

Zhang, M., Zhou, L., Bawa, R., Suren, H., and Holliday, J. (2016). Recombination rate variation, hitchhiking, and demographic history shape deleterious load in poplar. *Molecular Biology and Evolution*, 33(11):2899.

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., and Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of snp data. *Bioinformatics*, 28(24):3326–8.

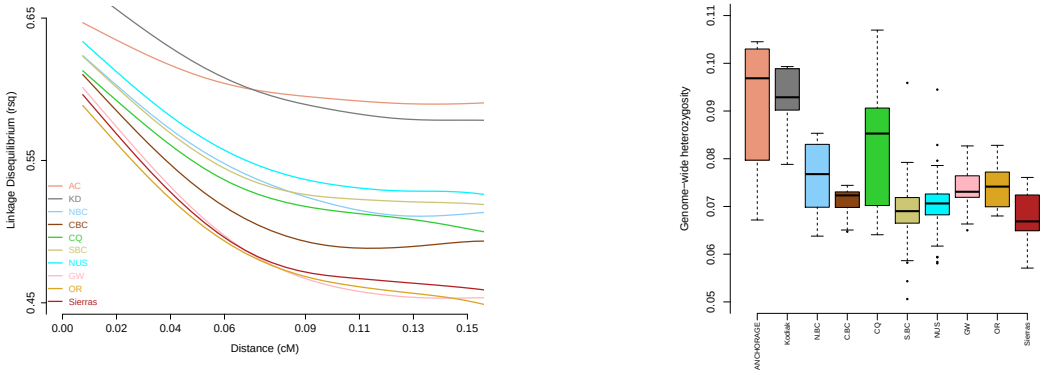
Zhou, L. C. and Holliday, J. A. (2012). Targeted enrichment of the black cottonwood (populus trichocarpa) gene space using sequence capture. *Bmc Genomics*, 13.

Figures and Tables



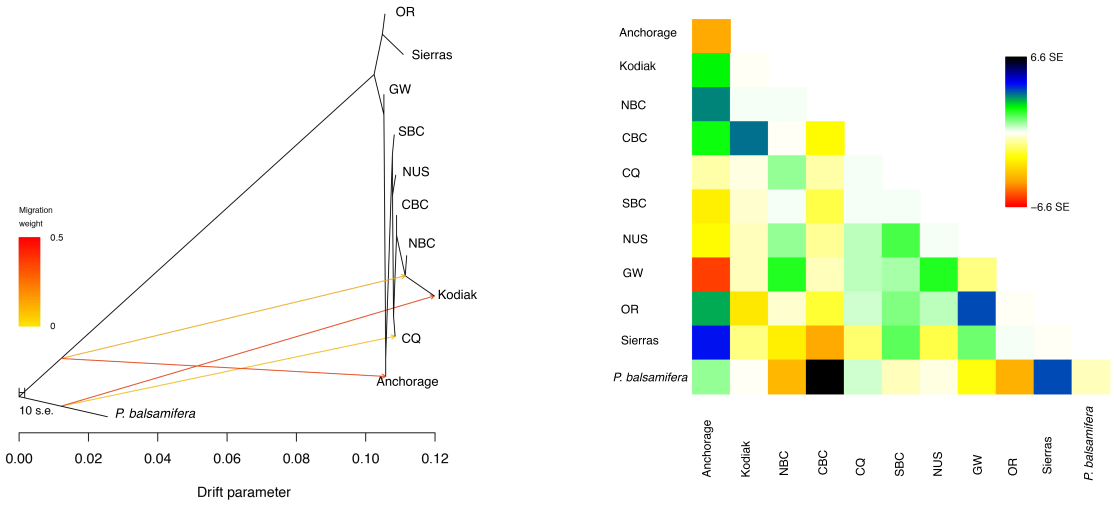
(a) (b) (c)

Figure 2.1: Genetic structure of *P.trichocarpa* samples.
A. Geographic distribution of sample locations. B. Principal component analysis of the data, excluding the outgroups. The transition in colors from blue to red represent change in latitude from north to south. C. Admixture results for K = 3, 8. Colors in the graph represent different population clusters.



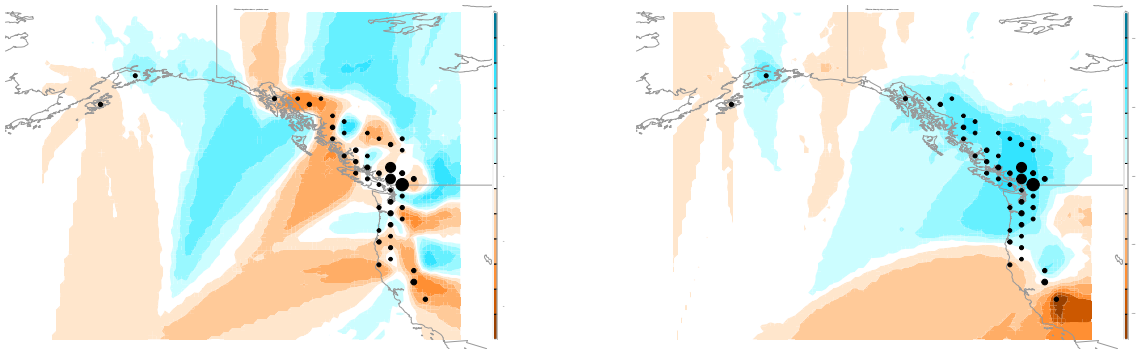
(a) (b)

Figure 2.2: Linkage disequilibrium patterns and genome-wide heterozygosity.
A. LD measured for each population and each pair of SNPs within the 50 evenly spaced recombination distance categories. B. Box-plots for average genome-wide heterozygosity for each population



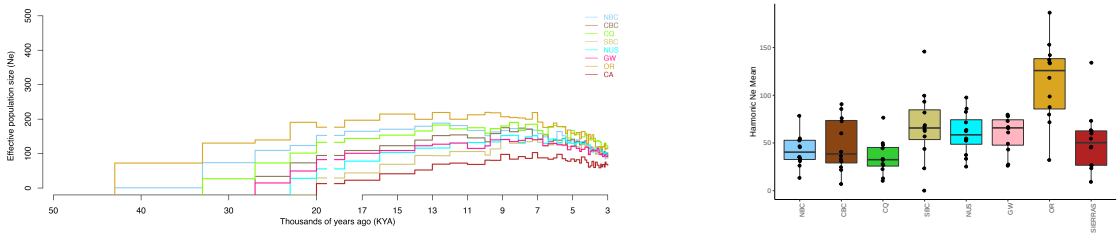
(a) (b)

Figure 2.3: Introgression between *P.balsamifera* and *P. trichocarpa.*
A. Inferred tree showing population tree and historical admixture between balsam poplar and trichocarpa populations. The color gradient on the arrows indicate different migration weights at the gene-flow event B. Residual fit from the phylogenetic tree in A.



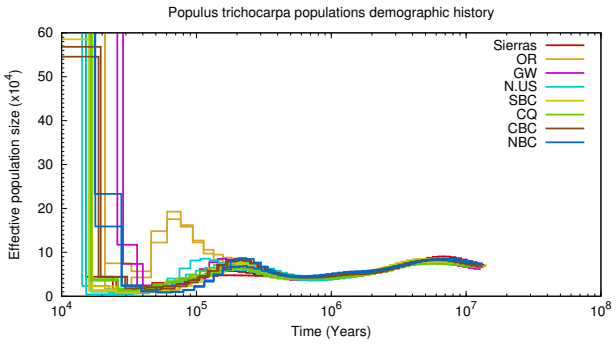
(a) (b)

Figure 2.4: EEMS analysis showing the migration and diversity rates between all populations.
A. Effective migration rates and gene-flow barriers between all populations. Color gradient from blue to orange represent change from high to low gene-flow regions B. Diversity surface plot showing distribution of genetic diversity in the populations. Color gradient from blue to orange represent high to low genetic diversity.

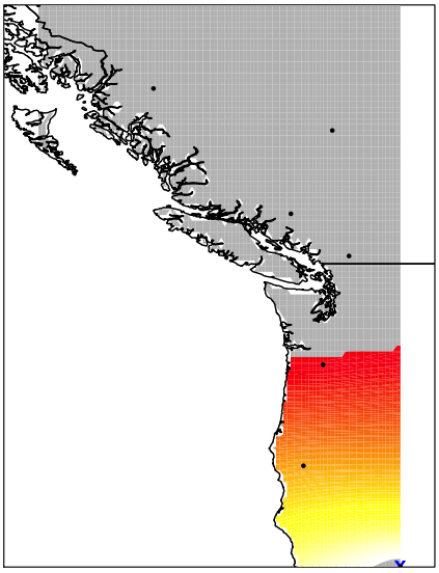


(a) (b)

Figure 2.5: Demographic history as inferred by Linkage disequilibrium method.
A. Spatial and temporal variation in N_e estimates. N_e was calculated using LD estimates from 50 evenly distributed recombination classes. B. Harmonic mean of the effective population size for each population over all recombination distance classes. N_e was also calculated separately for each chromosome, which was used to generate 95% confidence intervals.



(a)



(b)

Figure 2.6: Effective population size estimates and origin of range expansion. A. Historical changes in effective population size inferred using MSMC method. The time scale on the x-axis was calculated using 1×10^{-8} mutation rate B. Range expansion based estimates of origin of ancestral population using Peter and Slatkin (2013) method for all samples. Brighter yellow regions indicate more likely origin of expansion. Red and gray regions indicate unlikely origin.

Supplementary Figures

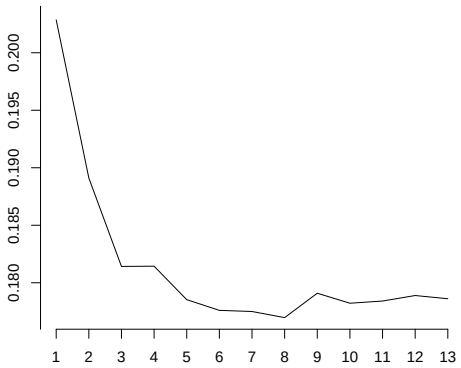


Figure S2.1: Cross-validation error from the admixture runs.

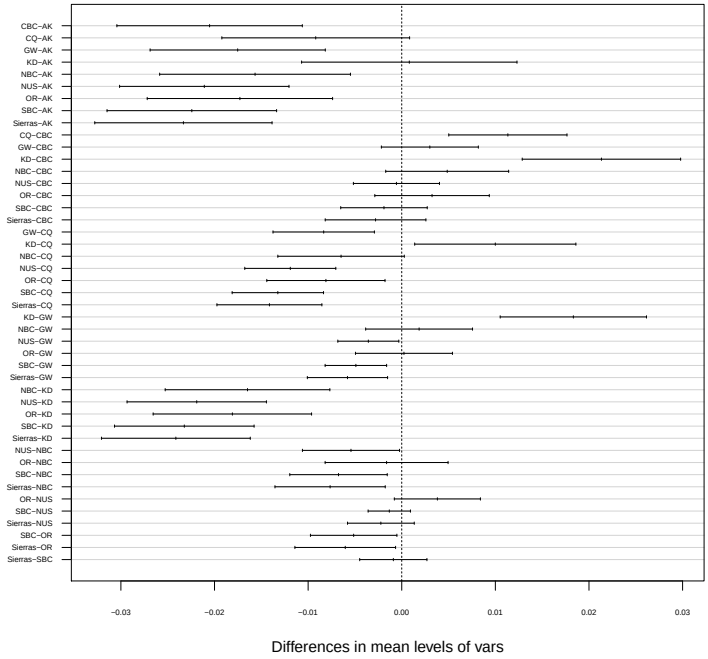
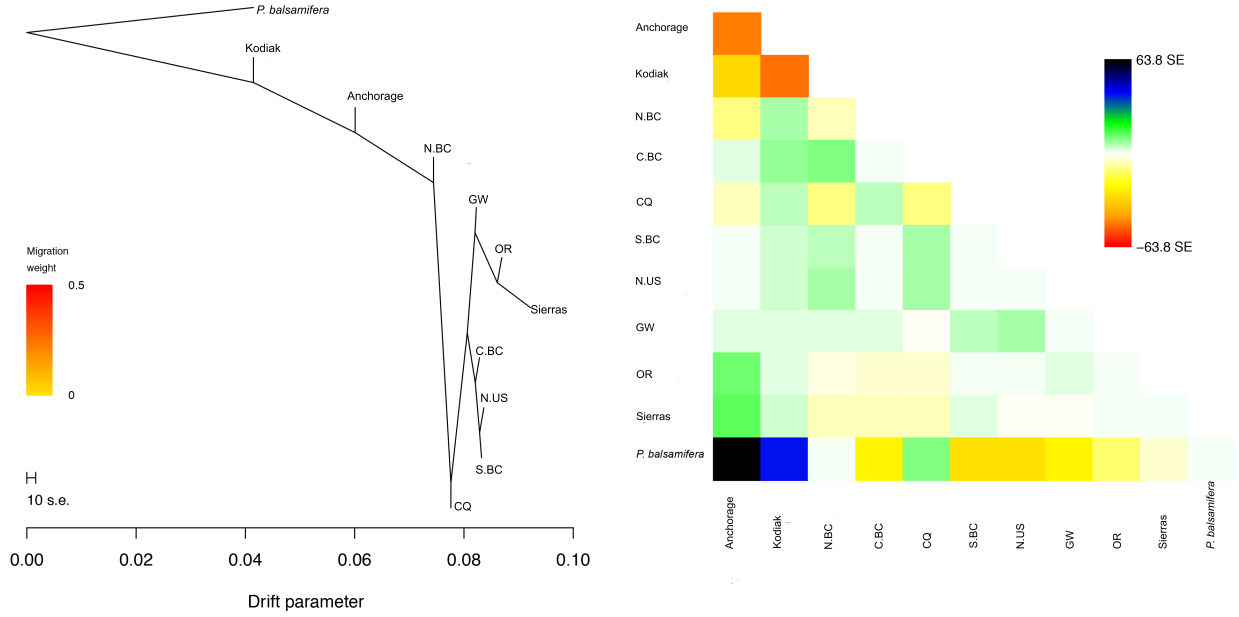


Figure S2.2: Pairwise comparisons of Genome-wide heterozygosity among all populations.



(a) (b)

Figure S2.3: Treemix results without adding the migration events.
A. Inferred graph. B. Residuals.

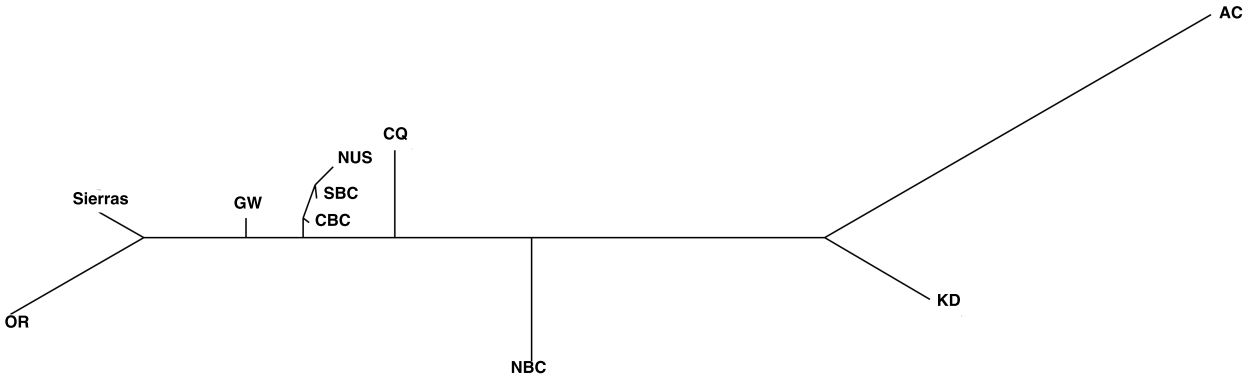


Figure S2.4: Neighbor joining tree showing relationship between populations using T_f divergence estimates.

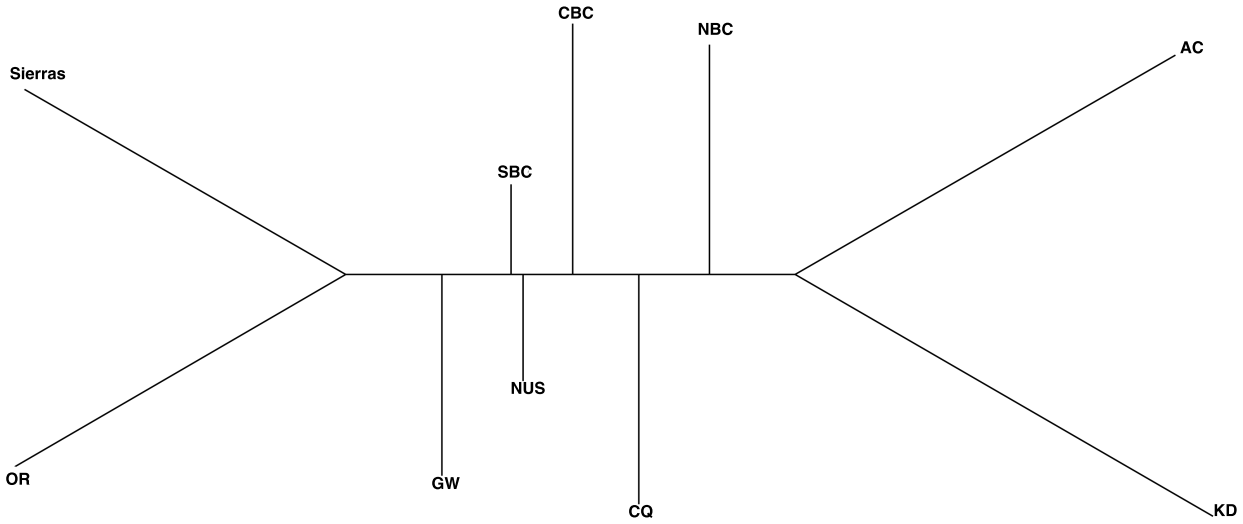


Figure S2.5: Neighbor joining tree showing relationship between populations using T_{LD} divergence estimates.

Chapter 3

Evolutionary Genomics of local adaptation in *Populus trichocarpa*

Abstract

Extant genetic variation in tree populations is a product of natural selection, as well as neutral processes. In order to identify the selective pressures shaping phenotypic diversification, we need to disentangle the effects of these neutral forces from selection. However, identifying selection signals in populations that have experienced dramatic migrations is not trivial. In this study, we examine patterns of genome-wide diversity in rangewide populations of the temperate tree Populus trichocarpa populations, and show that this diversity has been affected by both selective and demographic forces. To do this, we used two approaches. First, we scanned the genome for regions with unusually high and low levels of population differentiation, which are indicative of spatially diversifying and balancing selection, respectively. Second, scanned the genome for signatures of species-wide positive selection. We found significant overlap in the gene-ontology terms between highly divergent loci and targets of species-wide positive selection. Our results suggest genes involved in temperature sensing, light signaling pathway and plant immunity are major targets of selection. Overall, our study shows that P. trichocarpa populations experience both abiotic and biotic selection

pressures, which shape patterns of genetic variation among populations and in the species as a whole. It also indicates that environmental heterogeneity is a major driver of natural selection, and the genes important in taking cues from environment e.g. temperature and light are the potential targets of selection.

Introduction

Investigating the underlying causes of genetic differentiation is central to understanding the formation of species. Various population genetics studies have shown that contemporary genetic variation within and among populations is caused by a combination of neutral, demographic and selective forces (Nielsen, 2005; Excoffier et al., 2009; Savolainen et al., 2007). Various methods have been developed to identify different signatures of selection. However, it is difficult to determine the neutral background expectation distributions for diversity parameters in the face of demographic events that are common in temperate species (Nielsen, 2005; Hoban et al., 2016). The major the demographic processes that can distort allele frequencies include (i) population bottlenecks, which can dramatically decrease genetic diversity and lead to long haplotypes, mimicking the signatures of selective sweep (Simonsen et al., 1995), and (ii) range expansions, which can be accompanied by allele surfing, where genetic variants from a limited number of individuals can contribute disproportionately to new populations (Travis and Dytham, 2002; Excoffier and Ray, 2008). This can cause strong clinal variation in allele frequency in complete absence of selection. As a consequence, these confounding factors can increase the number of false positives, making it difficult to detect true signals of selection (Jensen et al., 2005; Thornton and Jensen, 2007).

Both bottlenecks and migratory history are especially relevant in the case of trees because the evolutionary history of trees has been marked by recurring range-expansions and associ-

ated bottlenecks (Savolainen et al., 2007; Bawa and Holliday, 2016; Zhou et al., 2014). This complex genetic history has left both neutral and selective genetic signatures on the populations. Many ecological studies have shown evidence of local adaptation among populations, the best studied of which involve phenological traits (bud set, bud flush) that vary among populations in response to environmental variation (Ducousso, A et al., 1996; Hnninen, 1995; Oubida et al., 2015).

Populus trichocarpa is a widespread tree species of primarily riparian ecosystems in western North America, with a natural range extending from California to Alaska (Fig.S2.1). Like most temperate tree species, P. trichocarpa has a complex demographic history involving recurrent range expansions and contraction due to post-glacial migration to higher latitudes (Zhou et al., 2014; Bawa and Holliday, 2017). It is also a model tree species, with relatively small and fully sequenced and annotated genome, making it an ideal candidate to study local adaptation (Tuskan et al., 2006).

Various studies of adaptation in *P. trichocarpa* have been carried out in recent years, focusing on various aspects of environmental heterogeneity and its possible effects on the genome (Evans et al., 2014; Zhou et al., 2014; Geraldes et al., 2014). While all of these studies have elucidated important aspects of local adaptation and evolutionary factors influencing it, most have been limited either by a relatively small number of genetic markers or through limited sampling of the species range. In this study, we seek to fill those gaps by focusing on populations encompassing most of the natural distribution of *P. trichocarpa*, and using markers that span the full gene space (the 'exome'). Using several complementary population genomic approaches, we identify numerous genomic intervals targeted by natural selection.

Material and Methods

Data collection, sequencing and population structure

All the sampling, genotyping, and population structure methods have been previously described (Zhou and Holliday, 2012; Bawa and Holliday, 2017). Briefly, we collected leaf tissue samples from 391 individuals of *P. trichocarpa* growing in common garden at Reynolds Homestead, located in Patrick county, Virginia. The original locations of these samples spanned a wide range along the west coast of North America, with their latitudes ranging from 37° N to 61° N (Fig. 1). All the samples were sequenced using the sequence capture method by Agilent Technologies ((Santa Clara, CA) on the Illumina Hi-seq platform. We utilized *P. trihcocarpa* v3.0 reference genome to align the paired end reads using burrows wheeler aligner (bwa) (Li, 2013), and called SNPs using the Genome Analysis Toolkit (GATK) Unified Genotyper (DePristo et al., 2011). The final dataset after quality control consisted of 1,058,513 SNPs.

We used Admixture (Alexander et al., 2009) and Principal Component Analysis (PCA) to determine number of populations in the samples. Prior to this, we sub-sampled the data to select SNPs in low linkage disequilibrium (r2 $<$ 0.1). We used SNPRelate package in R to for PCA analysis (Zheng et al., 2012). Admixture was run for K = 2 through K = 10 runs, and k-fold cross-validation procedure was used to estimate optimal number of populations.

Summary statistics

We began by estimating a variety of population genetics statistics, including pairwise nucleotide diversity (π), Tajimas D (D) (Tajima, 1989) and F_{ST} (Holsinger and Weir, 2009) for all the 391 individuals using VCFtools (Danecek et al., 2011) and PopGenome package in R

(Pfeifer et al., 2014). All the summary statistics were calculated separately for each chromosome using non-overlapping sliding window of 1kb. We also used snp-based sliding window for F_{ST} analysis, where 10 SNPs non-overlapping sliding window was used to calculate mean F_{ST}. Additionally, in order to differentiate positive selection from neutral and demographic signals, we simulated the demographic history both for all 10 populations and for three larger sub-populations (North, Center, and South) based on the demographic model estimated in Bawa and Holliday (2017). We simulated 100,000 100 kb windows using the cosi2 program (Shlyakhter et al., 2014) and estimated Tajimas D for those windows (Please refer to the supplementary file S2 for more details on the simulations). Custom scripts written in R and python were used to plot the results.

Signatures of selection

Genomic Islands of Divergence using Hidden Markov Model

We used the three-stage discrete Hidden Markov Model (HMM) (Rabiner, 1989; Felsenstein and Churchill, 1996) to identify genetically differentiated regions between different pairs of population. Traditionally, sliding window analysis of F_{ST} is used to identify the highly-diverged loci among the populations (Weir et al., 2005; Carlson et al., 2005), but the decision on an appropriate window size can be tricky (Schlamp et al., 2016). The HMM approach assumes that the genome is differentiated into contiguous states of divergence (low, intermediate and high), which are the hidden states of the HMM. A similar approach has been previously used by Hofer et al. (2012) and Soria-Carrasco et al. (2014) to summarize divergence between the populations. We modeled the Z-score of F_{ST} and assumed that its distribution varied among states. The Baum-Welch algorithm (Baum et al., 1970) was used to estimate the mean of this z-score and transition matrix among the different states. Follow-

ing that, we used Viterbi algorithm (Viterbi, 1967) to predict the most likely hidden states from the data. All these analyses were performed using hidden markov (Harte, 2009) and RHMM package (`http://rhmm.r-forge.r-project.org/`) in R. Following that, we used PLIS package in R to control the false discovery rate of identified high and low differentiated regions (Wei et al., 2009). We calculated the Local Index of Significance (LIS) under the genome-wide false discovery rate of 0.001 to identify the significant SNPs in high and low differentiated regions. Enrichment analysis was performed on the results of both high and low divergent states to identify the genes most and least differentiated among the population pairs.

Single locus genome scans

In order to test whether local adaptation and natural selection has left detectable signatures in the genome of *P. trichocarpa*, we used three different selection scan methods. We used two different population differentiation methods (hapFLK and OUTFLANK), which use differences in genotype and haplotype frequencies across populations to detect selection. Alongside these methods, we also included a simple haplotype statistics (H–statistics) that measure average length of pairwise haplotype homozygosity for each genetic locus. All the methods, except H–statistics calculate some sort of a null distribution or account of hierarchical population structure to remove confounding effects while detecting selection (Hoban et al., 2016; Whitlock and Lotterhos, 2015)

We used the hapFLK statistics calculated using the program hapFLK (Fariello et al., 2013) to identify selection outliers from the data. hapFLK is a population differentiation based selection scan method, that corrects for historical branching and population structure using kinship matrix and uses differentiation in haplotypes among populations as the basis to detect selection (Fariello et al., 2013; Schlamp et al., 2016). We used the following parameters: 10

population clusters (K = 10), 10 EM runs (nfit = 10) to fit the LD model and phased data (phased) option to run the program. Following that, we calculated the p-values and q-values associated with the hapFLK statistics by fitting the genome-wide standard normal distribution in R.

We also used OUTFLANK software by Whitlock and Lotterhos (2015) to detect selection based outliers in the data. Outflank calculates the null distribution of F_{ST} from large set of sampled loci and uses that as the reference neutral scenario representative of the species demographic history (Whitlock and Lotterhos, 2015). Any set of loci which fall outside this estimated null distribution are candidate of spatially diversifying selection. We performed this analysis on 188,774 SNPs (r2 < 0.1 and maf ≤ 0.05) after removing loci with heterozygosity < 0.1, and used 5% upper and lower tail trimming to calculate the null distribution of F_{ST}.

Following both of these tests, we used H-scan software (downloaded from: `https://messerlab.org/resources/`) (Garud et al., 2015) to estimate pairwise haplotype homozygosity tracts in each chromosome. H-scan calculates the length of homozygosity tracts ($h_{ij}(x)$) for pairs of samples, where x represents the genomic position between the first heterozygous site from left to right. These homozygosity tracts are averaged across all pairs to get H(x) statistics at each SNP position in data (Schlamp et al., 2016). H–scan statistics were measured in terms of physical distance between the SNPs (in base pairs) and calculated over 1 kb sliding window using all genome-wide SNP data.

Enrichment analyses

In order to gain insights into the genes involved in the evolutionary selective pressures related to local adaptation, we performed enrichment analysis on the selection scan based outliers.

The SNP position of the selection-scan outliers were matched with their gene annotations (downloaded from Phytozome V3.0) and the number of specific GO terms in the selection outliers list was compared against the specific GO terms in the whole annotation list. Following that, the hypergeometric distribution was used to calculate p-values for the over-represented terms. We also used Benjamani-Hochberg method (Benjamini and Hochberg, 1995) to control for false discovery rate, which was done using p.adjust function in R. We also estimated the fold-enrichment of GO terms in the data by calculating the odds ratio between the ratio of the genes of interest in the outliers list (m) and total number of genes in the list (n) and the ratio of genes belonging to a specific pathway (M) divided by all genes in the list (N) (Huang et al., 2008).

Results

Summary statistics

Genome-wide nucleotide diversity in Anchorage and Kodiak populations (Anchorage: π_{mean} = 0.0012 and Kodiak: π_{mean} = 0.0013) was nearly two times higher than the rest of the populations (NBC: π_{mean} = 0.00096; CQ: π_{mean} = 0.00096; CBC: π_{mean} = 0.00088; SBC: π_{mean} = 0.00079; NUS: π_{mean} = 0.00085; OR: π_{mean} = 0.00092; Sierras: π_{mean} = 0.00088) (Fig.S2.2). This increase in nucleotide diversity in Northern populations (Anchorage and Kodiak) could be attributed to introgression with *P. balsamifera*, which co-occur with our Anchorage and Kodiak populations (Geraldes et al., 2014; Suarez-Gonzalez et al., 2016). We also found the corresponding haplotypic homozygosity (Hstatistics) in Kodiak populations was significantly higher for Kodiak (H_{Mean} = 29840.43) than rest of the samples (Anchorage: H_{Mean} = 21250; NBC: H_{Mean} = 23634.07, CQ: H_{Mean} = 22770.7; CBC: H_{Mean} = 25112.37;

SBC: H_{Mean} = 24802.8; NUS: H_{Mean} = 25402.46; GW H_{Mean} = 22594.62; OR: H_{Mean} = 21446.6 and Sierras: H_{Mean} = 22615.82) (Fig. S2.2), suggesting the possibility of isolation and bottleneck in the Kodiak population, which might have caused large tracts of haplotypes with less genetic variation.

Mean values of Tajimas D were negative for all populations (AK: D_{Mean} = -0.13; NBC: D_{Mean} = -0.28; CQ: D_{Mean} = -0.43; CBC: D_{Mean} = -0.22; SBC: D_{Mean} = -0.33; NUS: D_{Mean} = -0.31; GW: D_{Mean} = -0.46; OR: D_{Mean} = -0.37 and Sierras: D_{Mean} = -0.28) except Kodiak (D_{Mean} = 0.13) (Fig. S3.1), suggesting an excess of rare genetic variants, which is a signature of population expansion (Nielsen, 2005; Simonsen et al., 1995). The Kodiak population, on the other hand, had an overall positive Tajimas D (D_{Mean} = 0.13), which could be caused by an excess of medium frequency SNPs as compared to neutral expectation (Fig. S3.1). Similar results have been reported in Sitka spruce (*Picea sitchensis*), and indicates of the recent bottleneck in the population (Holliday et al., 2010). We also used the coalescent simulation based on a previously estimated demographic model (Bawa and Holliday, 2017) to evaluate the selection signals relative to neutral expectation. The results showed that simulated Tajimas D to be negative in both Northern (AK: D_{Sim} = -1.1; KD: D_{Sim} =-1.04; NBC: D_{Sim} = -1.10; CQ: D_{Sim} = -1.03) and Central populations (CBC: D_{Sim} = -0.23; SBC: D_{Sim} = -0.215; NUS: D_{Sim} = -0.00049), where the magnitude of D-statistic was higher in Northern populations than Center. Southern populations on the other hand had positive Tajimas D statistics (GW: D_{Sim} = 0.18; OR: D_{Sim} = 0.66; Sierras: D_{Sim} = 0.50) (Fig. S2.3). This difference in the simulated and observed Tajimas D value suggest the possible role of both demography and selection in determining the current genetic variation patterns.

Signatures of selection

Genomic islands of divergence

Genomic divergence varied geographically among the larger subgroup of populations (North, Center and South). The mean F_{ST} was 0.045 for North vs. South population pairs, whereas North vs. Center and Center vs. South had the mean F_{ST} of 0.033 between them (Fig. S2.4). There was also significant variation in F_{ST} between the population pairs across the genome. We used a hidden markov model to identify highly divergent regions across the genome. Highly divergent regions in North vs South and North vs Center were enriched for SNPs with high F_{ST} values (NS: 00.316 and NC: 00.512), whereas the divergent SNPs between Center and South populations had medium to low F_{ST} values (CS: 00.294) (Table. 3.1). The number of highly differentiated genomic islands of divergence was significantly lower in Center vs South (13816) as compared to rest of the groups (NC: 21074; NS: 31347) (Table. 3.1).

We also estimated the overlap in diverged regions among different population pairs, and found large number of overlapping regions in different population comparisons. There were 3937 SNPs overlapping in Center vs South and North vs South population pairs, and 3829 SNPs similar in Center vs South and North vs Center population pairs (Fig. S3.2). North vs Center and North vs South population pairs had the largest number of overlapping SNPs (12040), which indicate of similarity in the selective forces in both of these cases (Soria-Carrasco et al., 2014; Holliday et al., 2016).

Among gene ontology terms that were over-represented in the highly divergent loci, there was significant enrichment for binding proteins (DNA binding, ATPase binding, Calcium dependent lipid binding, etc.), disease resistance proteins (LRR-kinase, NB-ARC domain, etc.), and proteins needed for photosynthetic and metabolic activity (Nuclear transporter

factor 2 (NFT-2), Phytochrome-B, inositol transporter, and FAR-1 related sequence) (Table S3.1 - S3.3). We also looked individually at each population pair comparison and the over-represented GO terms, and found that they belonged to similar functional categories: membrane receptor proteins, disease resistance proteins, calcium transport factors and proteins related to carbohydrate synthesis and metabolism (Table S3.2). This suggests that selective pressures are similar across the range, resulting in the same loci conferring local adaptation (Soria-Carrasco et al., 2014; Holliday et al., 2016).

Single locus scans

We next scanned the genomes of all 10 populations of *P. trichocarpa* to identify the genes targeted by selection. hapFLK revealed 3567 SNPs as outliers (p-value < 0.01), whereas OUTFLANK had 1500 outliers (OutlierFlag = True) (Fig. S2.5). There were 147 loci common between both of the tests (Table. 3.2). We also estimated H_{12} statistics to identify elevated levels of haplotype homozygosity between populations. There was substantial variation in H_{12} statistics across the genome, and signatures of selective sweeps (both soft and hard sweeps) were found in several locations, some of them coinciding with the selection outliers identified by OUTFLANK and hapFLK methods (Fig. S2.5).

Following that, we matched the SNP locus positions of the outlier regions with annotation of their respective genes. Enrichment analysis were performed both separately for each selection test and for the common outliers between them. Outliers from the hapFLK test showed overrepresentation of premRNA processing proteins, nucleoside triphosphate hydrolases, Ankyrin repeat family protein, Cysteine proteinases, nucleotide binding proteins and chromatin remodeling factors (Table S3.4). OUTFLANK results showed overrepresentation for Homeodomain-family like proteins, RNAligase, basic helix-loop-helix (bHLH) DNA binding protein, Cytochrome P450 protein, DEX1 protein, and Protein kinase superfamily

protein (Table S3.5). We also looked for common outliers between both of these tests, which were enriched for following terms: Remorin family protein, basic helix-loop-helix (bHLH) DNA-binding proteins, Sugar isomerase (SIS) family protein, Homeodomain-like superfamily protein, Leucine rich repeats, Rho determination factor and Disease resistance protein (TIR-NBS-LRR class) family (Table. 3.2).

Discussion

Summary Statistics

Most tree populations have gone through dramatic migrations, followed by population bottlenecks and population growth (Savolainen et al., 2007; Excoffier et al., 2009; Aitken et al., 2008). These factors determine the current landscape of genetic variation (Nielsen and Slatkin, 2013; Hoban et al., 2016). We used several complementary population genetics statistics to summarize the patterns of genetic variation, and subsequently employed a simulation framework to control for confounding effects of bottleneck and expansion on the signatures of selection.

Overall nucleotide diversity in our populations was lower than reported by previous studies (0.002), which may reflect the enrichment of genes and surrounding genetic regions in our data, which brought down the overall estimates of nucleotide diversity (Ismail et al., 2012; Evans et al., 2014; Keller et al., 2011). Anchorage and Kodiak populations also had nearly double the nucleotide diversity than rest of the populations, which could be due to these northern populations experiencing high levels of gene-flow and admixture from *P. balsamifera*, which shares the same range as *P. trichocarpa* in higher latitudes (Geraldes et al., 2014; Suarez-Gonzalez et al., 2016). This introgression has also been shown by previous studies,

and may be adaptive in nature, for example, by transferring cold-tolerance alleles from *P. balsamifera* to *P. trichocarpa* populations (Geraldes et al., 2014; Lexer and Widmer, 2008). The Kodiak Island population had significantly higher haplotypic homozygosity, a surrogate measure for runs of homozygosity in haplotypes. This could be interpreted as a possible result of genetic isolation, as the Kodiak population is an island, and are separated by 50 km strait from the inland populations. Moreover, the St. Elias mountains south of this area may restrict geneflow from rest of the species range (Fig. S2.1).

9 out of 10 population groups also had negative Tajimas D (Fig. S2.3, S1), suggesting an excess of rare genetic variants in the population (Tajima, 1989; Nielsen, 2005). This is a common signature of population expansion, and has been previously reported in various *Populus* sp. studies (Evans et al., 2014; KELLER et al., 2010; Bawa, 2012). Kodiak populations had a positive value of Tajimas D (Fig. S1), which could be due to geographic isolation and a recent population bottleneck, which intensifies the effect of random genetic drift, leading to random fixation of alleles and decreasing overall genetic variation (Nielsen and Slatkin, 2013; Excoffier et al., 2009; Holliday et al., 2010). By comparison of the observed Tajimas D values with the simulated ones, we found incongruence between both of the measures. This mismatch may be caused by the interplay between demography and selection, but could also have been caused by variation in mutation rate and other evolutionary processes (Simonsen et al., 1995; Nielsen, 2005). That said, the observed values of Tajimas D for some of the Central and Southern populations still fell on to the outside the distribution of neutrally evolving regions, and might indicate the effect of positive selection (Fig. S2.3, S1).

Genomic islands of divergence

Hidden markov model based divergent SNPs identified from pairwise F_{ST} among the larger sub-group of populations showed that divergence varied significantly across the genome (Fig. S2.4). There was increase in the number of highly divergent islands between populations as geographical distance between increased (Table 3.1). This indicates the role of isolation by distance and possible effect of diversifying selection on the divergence patterns. We also found that the North vs South and North vs. Center populations were significantly enriched for high and medium F_{ST} values, with more even distribution across all chromosomes in the genome in North vs South population as compared to North vs Center (Fig. S2.4). On the other hand, the Center vs South comparison revealed SNPs with medium to low F_{ST}. This pattern may result from high gene-flow and counteracting the effects of diversifying selection between the populations (Aitken et al., 2008; Bawa and Holliday, 2017; Holliday et al., 2012). The North vs South comparison have very different climate and therefore selective pressures, whereas this difference is much less in other population pairs. Moreover, the rate of gene-flow also decreases with increase in geographical distance. There is likely very little direct/indirect gene-flow between North vs South, and very high between Center vs South population pairs, due to the contiguous nature of the range further south. High gene-flow leads to increase in allele sharing between populations and decreases divergence between them (Savolainen et al., 2007; Holliday et al., 2012).

There was also extensive parallelism between the genetic loci that were highly divergent in different pairs of populations, especially between North vs South and North vs Center population pairs (Fig. S2). This has been recently reported in the study of *P. trichocarpa* populations, where Holliday et al. (2016) found large number of same genetic loci involved in altitudinal as well as latitudinal adaptation. This also suggests the universal importance of some of the gene-complexes which are involved in the local adaptation and are targets of

diversifying selection. The highly divergent loci showed an enrichment for disease resistance related proteins (LRR repeats, NB-ARC domain, LRR-kinase protein), nucleic acid binding proteins (DNA binding proteins, ATP binding, Calcium dependent lipid binding), carbohydrate metabolism and photosynthetic activity proteins (Nuclear transportation factor-2, Phytochrome-B, inositol transporter, FAR-1 related sequence) and genes related to plant immunity (Homeodomain like protein, bHLH DNA binding protein). There was also significant enrichment for plasma membrane receptors (transmembrane receptors) and chromatin remodeling factors (Chromatin remodeling 1,4,5,24, DEK domain-containing chromatin associated protein, SPT2 chromatin protein). Previous studies have shown that chromatin modeling mechanisms and receptors in plasma membrane play an important role in temperature sensing in plants, and help in regulating the circadian clock (Penfield, 2008; Eriksson and Webb, 2011). Studies focusing on temperature sensing in plants have shown that chromatin remodeling complexes are required to incorporate H2A.Z, an alternative histone to H2A in the nucleosome. This helps in wrapping the DNA more tightly than the normal H2A nucleosome, and contributes to freezing tolerance in plants (Kumar and Wigge, 2010; Deal et al., 2007; Mizuguchi et al., 2003). We also found Phytochrome-B and FAR-1 (FAR-RED Impaired response1) genes enriched in highly divergent SNPs. Both of these genes are involved in light sensing (Hudson et al., 1999; Howe et al., 1996), and have been identified by previous studies for their role in local adaptation (Evans et al., 2014; Holliday et al., 2016; HALL et al., 2011). Light and temperatures are very important environmental cues used by trees to set their circadian clock, and since populations in different geographical areas have different day lengths and temperatures, the genes regulating those effects are expected targets of diversifying selection (Rohde and Bhalerao, 2007; Ibanez et al., 2010). There was also significant enrichment for plant immunity related genes and leucine rich repeats in the genomic islands of divergence (Table S3.1-S3.3). These genes have been shown to interact with phytohormomes, and provide immunity to various diseases (Xu et al., 2014; Carretero-

Paulet et al., 2010; Li et al., 2007; McKown et al., 2014). This suggests of the selective pressures due to biotic stresses in different geographical areas, which is leading to divergence among the populations.

Single locus scans

The loci identified as outliers in the single-locus selection scans were functionally similar to those reported in genomic islands of divergence results. There was significant enrichment for nucleotide binding proteins (ATP binding cassette protein 1, FAD-binding Berberine family protein, bHLH DNA-binding protein), carbohydrate metabolism related proteins (Sugar isomerase family protein, sugar transporter 14, Nucleotide-diphosphosugar transferases protein), genes conferring immunity (TIR-NBS-LRR, CC-NBS-LRR, Homeodomain like family protein, bHLH DNA-binding superfamily protein), and membrane receptors and chromatin modeling factors (Outer membrane OMP85 family protein, vesicle- associated membrane protein 726, Transmembrane amino acid transporter protein, Chromatin remodeling factor 4 and 8) (Table 3.2, S3.4 – S3.5). Various studies have shown fluctuations in carbohydrate metabolism with change in seasons, and its correlation to onset of dormancy (ANDERSON et al., 2005; Pagter et al., 2015). Since *P. trichocarpa* is a deciduous tree, and is known to go through a gradual metabolic shift before the start and close to end of dormancy, these genes (Sugar isomerase family protein, sugar transporter 14, Nucleotide-diphosphosugar transferases protein) could have an important role in reconfiguration of metabolism during growth cessation and break of dormancy (COOKE et al., 2012; Rohde et al., 2000). Carbohydrate metabolism genes are also known to be associated with abiotic stresses, like drought and lack of availability of nutrients and therefore could play important role in terms of optimizing the carbohydrate use given the resources (Cramer et al., 2007, 2011; Pinheiro and Chaves, 2011). We also found significant enrichment for genes which involved in plant immu-

nity, again showing the possible role of biotic stresses in causing selective pressures to cause divergence (Xu et al., 2014; Carretero-Paulet et al., 2010; Li et al., 2007). Similar study focusing on the population genomics of *P. trichocarpa* found variation in selective pressure caused from the leaf pathogens (*Melamsora* sp.) across the species range (McKown et al., 2014). They showed that this selective pressure was causing variation in size and number of stomata, to control the attack from these leaf pathogens (McKown et al., 2014). Chromatin remodeling factors and PPR-like genes are also known to play role in temperature sensing and DNA repair (Zoschke et al., 2013; Kumar and Wigge, 2010; Mizuguchi et al., 2003; Moreira and Philippe, 2017).

Taken together, these results showed that the *P. trichocarpa* populations in our study are challenged by both abiotic and biotic stress, and the selective pressure caused due to these factors lead to selection of various genes in *P. trichocarpa*. In particular, we found that genes involved in temperature sensing and light signaling are targets of local adaptation, which likely reflects geographic variation in the seasonal cycling between growth and dormancy that is governed by both temperature and daylength (Rohde and Bhalerao, 2007; Ibanez et al., 2010). These seasonal cycles enable freezing tolerance in winter and are correlated with reproductive timing (Bawa, 2012; Bawa and Holliday, 2016; Rohde and Bhalerao, 2007). Our study therefore provides novel insights into the role of diversifying selection and local adaptation processes ongoing in *P. trichocarpa* populations.

References

Aitken, S. N., Yeaman, S., Holliday, J. A., Wang, T., and Curtis-McLane, S. (2008). Adaptation, migration or extirpation: climate change outcomes for tree populations. *Evol Appl*, 1(1):95–111.

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*, 19(9):1655–64.

ANDERSON, J. V., GESCH, R. W., JIA, Y., CHAO, W. S., and HORVATH, D. P. (2005). Seasonal shifts in dormancy status, carbohydrate metabolism, and related gene expression in crown buds of leafy spurge. *Plant, Cell & Environment*, 28(12):1567–1578.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *Ann. Math. Statist.*, 41(1):164–171.

Bawa, R. and Holliday, J. (2017). Genome-wide insights into the fine scale population structure and demographic history of western poplar (*Populus trichocarpa*).

Bawa, R. and Holliday, J. A. (2016). *Genomics of Speciation in Temperate and Boreal Angiosperm Trees*. Springer New York, New York.

Bawa, R. K. (2012). *Signatures of natural selection and local adaptation in Populus trichocarpa and Populus deltoides along latitudinal clines.* PhD thesis, Virginia Tech.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological*, 57(1):289–300.

Carlson, C. S., Thomas, D. J., Eberle, M. A., Swanson, J. E., Livingston, R. J., Rieder, M. J., and Nickerson, D. A. (2005). Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Research*, 15(11):1553–1565.

Carretero-Paulet, L., Galstyan, A., Roig-Villanova, I., Martnez-Garca, J. F., Bilbao-Castro, J. R., and Robertson, D. L. (2010). Genome-wide classification and evolutionary analysis

of the bhlh family of transcription factors in arabidopsis, poplar, rice, moss, and algae. *Plant Physiology*, 153(3):1398–1412.

COOKE, J. E. K., ERIKSSON, M. E., and JUNTTILA, O. (2012). The dynamic nature of bud dormancy in trees: environmental control and molecular mechanisms. *Plant, Cell & Environment*, 35(10):1707–1728.

Cramer, G. R., Ergül, A., Grimplet, J., Tillett, R. L., Tattersall, E. A. R., Bohlman, M. C., Vincent, D., Sonderegger, J., Evans, J., Osborne, C., Quilici, D., Schlauch, K. A., Schooley, D. A., and Cushman, J. C. (2007). Water and salinity stress in grapevines: early and late changes in transcript and metabolite profiles. *Functional & Integrative Genomics*, 7(2):111–134.

Cramer, G. R., Urano, K., Delrot, S., Pezzotti, M., and Shinozaki, K. (2011). Effects of abiotic stress on plants: a systems biology perspective. *BMC Plant Biology*, 11(1):163.

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

Deal, R. B., Topp, C. N., McKinney, E. C., and Meagher, R. B. (2007). Repression of flowering in arabidopsis requires activation of flowering locus c expression by the histone variant h2a.z. *The Plant Cell*, 19(1):74–83.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nat Genet*, 43(5):491–498.

Ducousso, A, Guyon, JP, and Krmer, A (1996). Latitudinal and altitudinal variation of bud burst in western populations of sessile oak (quercus petraea (matt) liebl). *Ann. For. Sci.*, 53(2).

Eriksson, M. E. and Webb, A. A. (2011). Plant cell responses to cold are all about timing. *Current Opinion in Plant Biology*, 14(6):731 – 737. 14/6 Cell biology.

Evans, L. M., Slavov, G. T., Rodgers-Melnick, E., Martin, J., Ranjan, P., Muchero, W., Brunner, A. M., Schackwitz, W., Gunter, L., Chen, J.-G., Tuskan, G. A., and DiFazio, S. P. (2014). Population genomics of populus trichocarpa identifies signatures of selection and adaptive trait associations. *Nat Genet*, 46(10):1089–1096. Article.

Excoffier, L., Foll, M., and Petit, R. J. (2009). Genetic consequences of range expansions. *Annual Review of Ecology Evolution and Systematics*, 40:481–501.

Excoffier, L. and Ray, N. (2008). Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecology & Evolution*, 23(7):347 – 351.

Fariello, M. I., Boitard, S., Naya, H., SanCristobal, M., and Servin, B. (2013). Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics*, 193(3):929–941.

Felsenstein, J. and Churchill, G. A. (1996). A hidden markov model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution*, 13(1):93.

Garud, N. R., Messer, P. W., Buzbas, E. O., and Petrov, D. A. (2015). Recent selective sweeps in north american drosophila melanogaster show signatures of soft sweeps. *PLOS Genetics*, 11(2):1–32.

Geraldes, A., Farzaneh, N., Grassa, C. J., McKown, A. D., Guy, R. D., Mansfield, S. D., Douglas, C. J., and Cronk, Q. C. B. (2014). Landscape genomics of populus trichocarpa:

The role of hybridization, limited gene flow, and natural selection in shaping patterns of population structure. *Evolution*, 68(11):3260–3280.

HALL, D., MA, X.-F., and INGVARSSON, P. K. (2011). Adaptive evolution of the populus tremula photoperiod pathway. *Molecular Ecology*, 20(7):1463–1474.

Harte, D. (2009). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Statistics Research Associates, Wellington, New Zealand*.

Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., Poss, M. L., Reed, L. K., Storfer, A., and Whitlock, M. C. (2016). Finding the genomic basis of local adaptation: Pitfalls, practical solutions, and future directions. *The American Naturalist*, 188(4):379–397. PMID: 27622873.

Hofer, T., Foll, M., and Excoffier, L. (2012). Evolutionary forces shaping genomic islands of population differentiation in humans. *BMC Genomics*, 13(1):107.

Holliday, J. A., Suren, H., and Aitken, S. N. (2012). Divergent selection and heterogeneous migration rates across the range of sitka spruce (picea sitchensis). *Proc Biol Sci*, 279(1734):1675–83.

Holliday, J. A., Yuen, M., Ritland, K., and Aitken, S. N. (2010). Postglacial history of a widespread conifer produces inverse clines in selective neutrality tests. *Molecular Ecology*, 19(18):3857–3864.

Holliday, J. A., Zhou, L., Bawa, R., Zhang, M., and Oubida, R. W. (2016). Evidence for extensive parallelism but divergent genomic architecture of adaptation along altitudinal and latitudinal gradients in populus trichocarpa. *New Phytologist*, 209(3):1240–1251. 2015-19526.

Holsinger, K. E. and Weir, B. S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting fst. *Nat Rev Genet*, 10(9):639–650.

Howe, G. T., Gardner, G., Hackett, W. P., and Furnier, G. R. (1996). Phytochrome control of short-day-induced bud set in black cottonwood. *Physiologia Plantarum*, 97(1):95–103.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2008). Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nat. Protocols*, 4(1):44–57.

Hudson, M., Ringli, C., Boylan, M. T., and Quail, P. H. (1999). The far1 locus encodes a novel nuclear protein specific to phytochrome a signaling. *Genes & Development*, 13(15):2017–2027.

Hnninen, H. (1995). Effects of climatic change on trees from cool and temperate regions: an ecophysiological approach to modelling of bud burst phenology. *Canadian Journal of Botany*, 73(2):183–199.

Ibanez, C., Kozarewa, I., Johansson, M., Ogren, E., Rohde, A., and Eriksson, M. E. (2010). Circadian clock components regulate entry and affect exit of seasonal dormancy as well as winter hardiness in populus trees. *Plant Physiology*, 153(4):1823–1833.

Ismail, M., Soolanayakanahally, R. Y., Ingvarsson, P. K., Guy, R. D., Jansson, S., Silim, S. N., and El-Kassaby, Y. A. (2012). Comparative nucleotide diversity across north american and european populus species. *Journal of Molecular Evolution*, 74(5):257–272.

Jensen, J. D., Kim, Y., DuMont, V. B., Aquadro, C. F., and Bustamante, C. D. (2005). Distinguishing between selective sweeps and demography using dna polymorphism data. *Genetics*, 170(3):1401–1410.

Keller, S. R., Levsen, N., Ingvarsson, P. K., Olson, M. S., and Tiffin, P. (2011). Local selection across a latitudinal gradient shapes nucleotide diversity in balsam poplar, populus balsamifera l. *Genetics*, 188(4):941–U318.

KELLER, S. R., OLSON, M. S., SILIM, S., SCHROEDER, W., and TIFFIN, P. (2010). Genomic diversity, population structure, and migration following rapid range expansion in the balsam poplar, populus balsamifera. *Molecular Ecology*, 19(6):1212–1226.

Kumar, S. V. and Wigge, P. A. (2010). H2a.z-containing nucleosomes mediate the thermosensory response in arabidopsis. *Cell*, 140(1):136 – 147.

Lexer, C. and Widmer, A. (2008). The genic view of plant speciation: recent progress and emerging questions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1506):3023–3036.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv e-prints*.

Li, H., Sun, J., Xu, Y., Jiang, H., Wu, X., and Li, C. (2007). The bhlh-type transcription factor ataib positively regulates aba response in arabidopsis. *Plant Molecular Biology*, 65(5):655–665.

McKown, A. D., Guy, R. D., Quamme, L., Klpt, J., La Mantia, J., Constabel, C. P., El-Kassaby, Y. A., Hamelin, R. C., Zifkin, M., and Azam, M. S. (2014). Association genetics, geography and ecophysiology link stomatal patterning in populus trichocarpa with carbon gain and disease resistance trade-offs. *Molecular Ecology*, 23(23):5771–5790.

Mizuguchi, G., Shen, X., Landry, J., Wu, W.-H., Sen, S., and Wu, C. (2003). Atp-driven exchange of histone h2az variant catalyzed by swr1 chromatin remodeling complex. *Science*.

Moreira, D. and Philippe, H. (2017). Smr: a bacterial and eukaryotic homologue of the c-terminal region of the muts2 family. *Trends in Biochemical Sciences*, 24(8):298–300.

Nielsen, R. (2005). Molecular signatures of natural selection. *Annu Rev Genet*, 39:197–218.

Nielsen, R. and Slatkin, M. (2013). *An introduction to population genetics : theory and applications*. Sinauer Associates, Sunderland, Mass.

Oubida, R. W., Gantulga, D., Zhang, M., Zhou, L., Bawa, R., and Holliday, J. A. (2015). Partitioning of multivariate phenotypes using regression trees reveals complex patterns of adaptation to climate across the range of black cottonwood (populus trichocarpa). *Frontiers in Plant Science*, 6:181.

Pagter, M., Andersen, U. B., and Andersen, L. (2015). Winter warming delays dormancy release, advances budburst, alters carbohydrate metabolism and reduces yield in a temperate shrub. *AoB PLANTS*, 7:plv024.

Penfield, S. (2008). Temperature perception and signal transduction in plants. *New Phytologist*, 179(3):615–628.

Pfeifer, B., Wittelsbuerger, U., Ramos-Onsins, S. E., and Lercher, M. J. (2014). Popgenome: An efficient swiss army knife for population genomic analyses in r. *Molecular Biology and Evolution*, 31:1929–1936.

Pinheiro, C. and Chaves, M. M. (2011). Photosynthesis and drought: can we make metabolic connections from available data? *Journal of Experimental Botany*, 62(3):869.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Rohde, A. and Bhalerao, R. P. (2007). Plant dormancy in the perennial context. *Trends in Plant Science*, 12(5):217 – 223.

Rohde, A., Howe, G. T., Olsen, J. E., Moritz, T., Van Montagu, M., Junttila, O., and Boerjan, W. (2000). *Molecular Aspects of Bud Dormancy in Trees*. Springer Netherlands, Dordrecht.

Savolainen, O., Pyhjrvi, T., and Knrr, T. (2007). Gene flow and local adaptation in trees. *Annual Review of Ecology, Evolution, and Systematics*, 38(1):595–619.

Schlamp, F., van der Made, J., Stambler, R., Chesebrough, L., Boyko, A. R., and Messer, P. W. (2016). Evaluating the performance of selection scans to detect selective sweeps in domestic dogs. *Molecular Ecology*, 25(1):342–356.

Shlyakhter, I., Sabeti, P. C., and Schaffner, S. F. (2014). Cosi2: an efficient simulator of exact and approximate coalescent with selection. *Bioinformatics*, 30(23):3427.

Simonsen, K. L., Churchill, G. A., and Aquadro, C. F. (1995). Properties of statistical tests of neutrality for dna polymorphism data. *Genetics*, 141(1):413–429.

Soria-Carrasco, V., Gompert, Z., Comeault, A. A., Farkas, T. E., Parchman, T. L., Johnston, J. S., Buerkle, C. A., Feder, J. L., Bast, J., Schwander, T., Egan, S. P., Crespi, B. J., and Nosil, P. (2014). Stick insect genomes reveal natural selection's role in parallel speciation. *Science*, 344(6185):738–742.

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

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics*, 123(3):585–595.

Thornton, K. R. and Jensen, J. D. (2007). Controlling the false-positive rate in multilocus genome scans for selection. *Genetics*, 175(2):737–750.

Travis, J. M. J. and Dytham, C. (2002). Dispersal evolution during invasions. *Evolutionary Ecology Research*, 4(8):1119–1129.

Tuskan, G. A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R. R., Bhalerao, R. P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., Busov, V., Campbell, M., Carlson, J., Chalot, M., Chapman, J., Chen, G.-L., Cooper, D., Coutinho, P. M., Couturier, J., Covert, S., Cronk, Q., Cunningham, R., Davis, J., Degroeve, S., Déjardin, A., dePamphilis, C., Detter, J., Dirks, B., Dubchak, I., Duplessis, S., Ehlting, J., Ellis, B., Gendler, K., Goodstein, D., Gribskov, M., Grimwood, J., Groover, A., Gunter, L., Hamberger, B., Heinze, B., Helariutta, Y., Henrissat, B., Holligan, D., Holt, R., Huang, W., Islam-Faridi, N., Jones, S., Jones-Rhoades, M., Jorgensen, R., Joshi, C., Kangasjärvi, J., Karlsson, J., Kelleher, C., Kirkpatrick, R., Kirst, M., Kohler, A., Kalluri, U., Larimer, F., Leebens-Mack, J., Leplé, J.-C., Locascio, P., Lou, Y., Lucas, S., Martin, F., Montanini, B., Napoli, C., Nelson, D. R., Nelson, C., Nieminen, K., Nilsson, O., Pereda, V., Peter, G., Philippe, R., Pilate, G., Poliakov, A., Razumovskaya, J., Richardson, P., Rinaldi, C., Ritland, K., Rouzé, P., Ryaboy, D., Schmutz, J., Schrader, J., Segerman, B., Shin, H., Siddiqui, A., Sterky, F., Terry, A., Tsai, C.-J., Uberbacher, E., Unneberg, P., Vahala, J., Wall, K., Wessler, S., Yang, G., Yin, T., Douglas, C., Marra, M., Sandberg, G., Van de Peer, Y., and Rokhsar, D. (2006). The genome of black cottonwood, populus trichocarpa (torr. & gray). *Science*, 313(5793):1596–1604.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269.

Wei, Z., Sun, W., Wang, K., and Hakonarson, H. (2009). Multiple testing in genome-wide association studies via hidden markov models. *Bioinformatics*, 25(21):2802.

Weir, B. S., Cardon, L. R., Anderson, A. D., Nielsen, D. M., and Hill, W. G. (2005). Measures of human population structure show heterogeneity among genomic regions. *Genome Research*, 15(11):1468–1476.

Whitlock, M. C. and Lotterhos, K. E. (2015). Reliable detection of loci responsible for local adaptation: Inference of a null model through trimming the distribution of f-st. *American Naturalist*, 186:S24–S36.

Xu, F., Kapos, P., Cheng, Y. T., Li, M., Zhang, Y., and Li, X. (2014). Nlr-associating transcription factor bhlh84 and its paralogs function redundantly in plant immunity. *PLOS Pathogens*, 10(8):1–14.

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., and Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of snp data. *Bioinformatics*, 28(24):3326–8.

Zhou, L., Bawa, R., and Holliday, J. A. (2014). Exome resequencing reveals signatures of demographic and adaptive processes across the genome and range of black cottonwood populus trichocarpa. *Molecular Ecology*, 23(10):2486–2499.

Zhou, L. C. and Holliday, J. A. (2012). Targeted enrichment of the black cottonwood (populus trichocarpa) gene space using sequence capture. *Bmc Genomics*, 13.

Zoschke, R., Qu, Y., Zubo, Y. O., Börner, T., and Schmitz-Linneweber, C. (2013). Mutation of the pentatricopeptide repeat-smr protein svr7 impairs accumulation and translation of chloroplast atp synthase subunits in arabidopsis thaliana. *Journal of Plant Research*, 126(3):403–414.

Figures and Tables

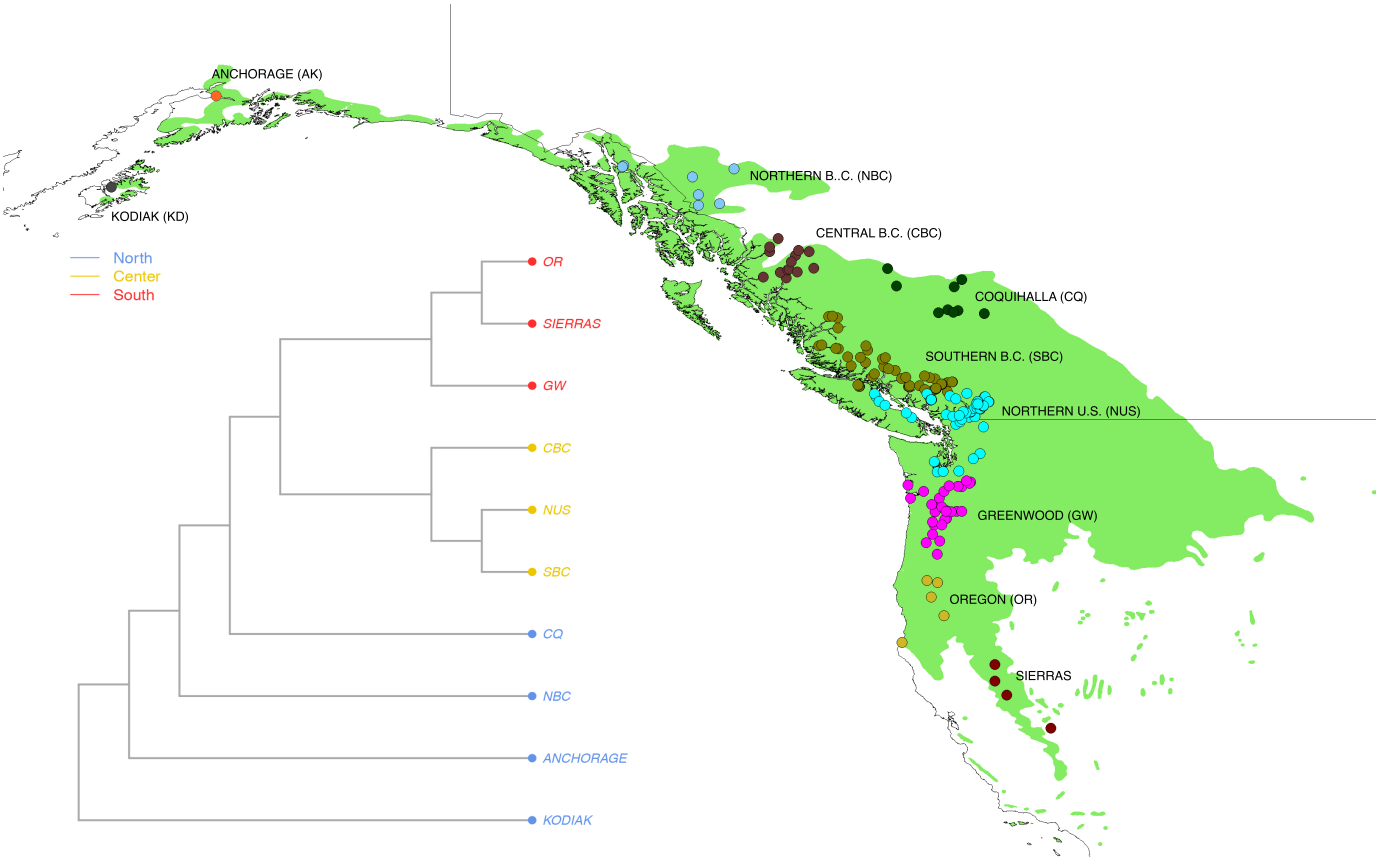


Figure S2.1: Geographical distribution of the samples used in the study, with color codings corresponding to different populations determined in the study Bawa and Holliday (2017). The neighbor joining tree on the left represents relationship between all 10 populations, and clustering of them to south, center and north for some of the analysis.

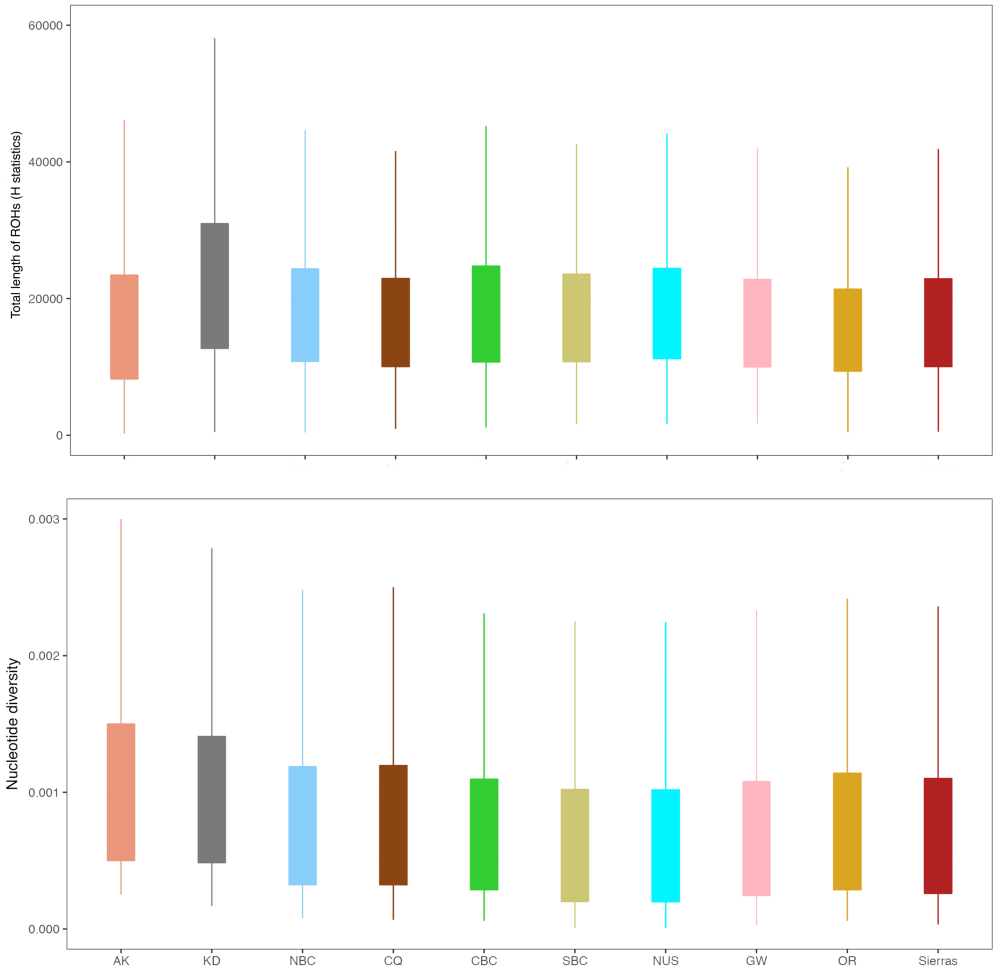


Figure S2.2: Differences in nucleotide diversity and haplotype diversity among different population of *P. trichocarpa*.

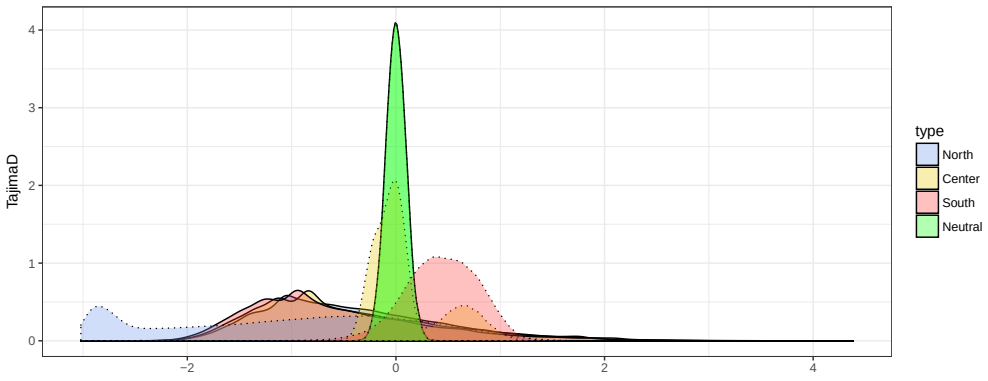


Figure S2.3: Distribution of Tajima's D statistic for 10 kb sliding window in three bigger sub groups (North, Center and South). The dotted line shows the simulated distribution of D expected under different demographic models, and the green density plot shows distribution of D under neutral expectation.

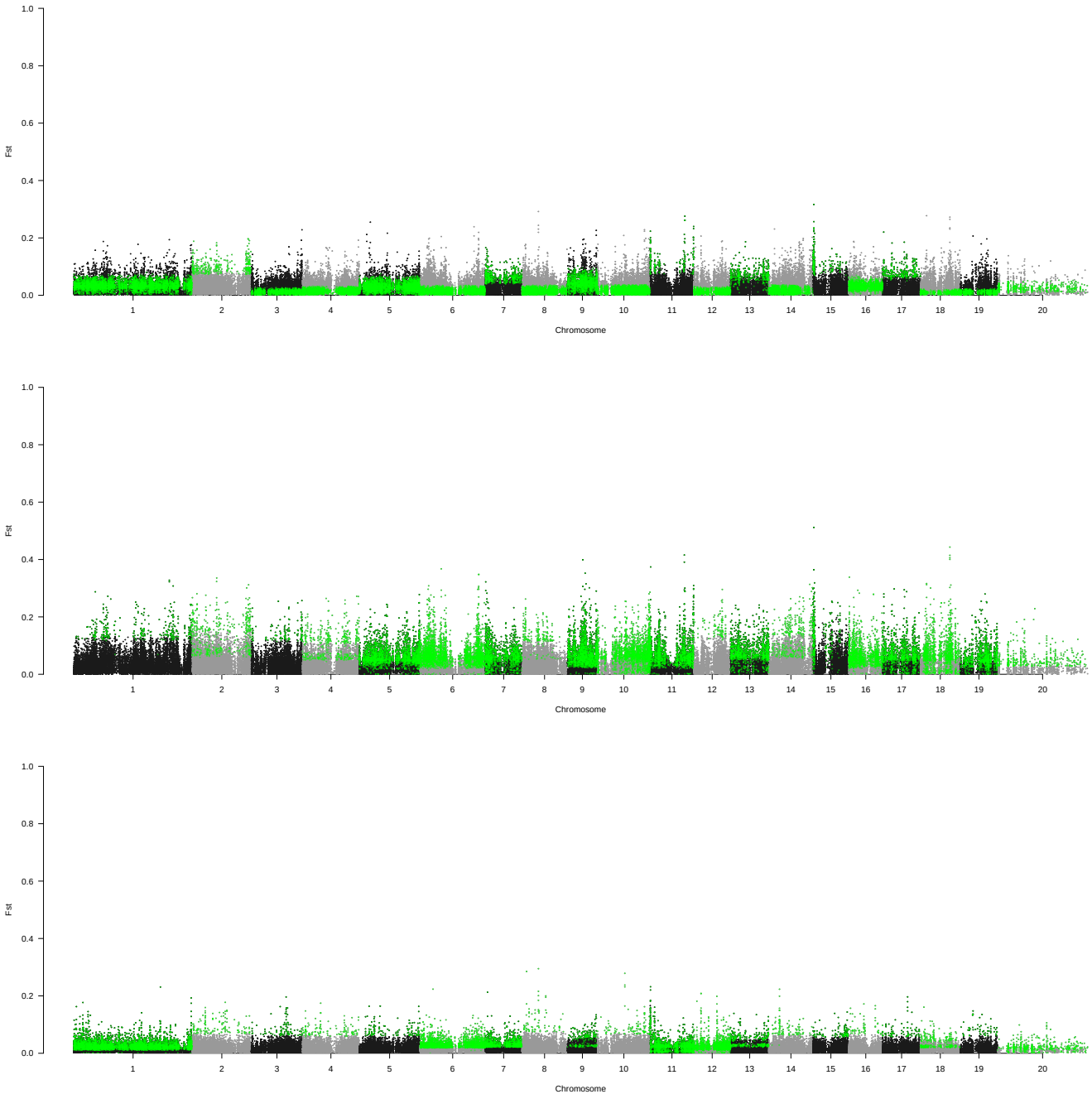


Figure S2.4: Patterns of genomic divergence between three populations (North, South and Center) of *P. trichocarpa*. Green color denote SNPs that occur in the highly divergent regions. The top manhattan plot is North vs. South, central manhattan plot is North vs. Center and last one is Center vs. South populations.

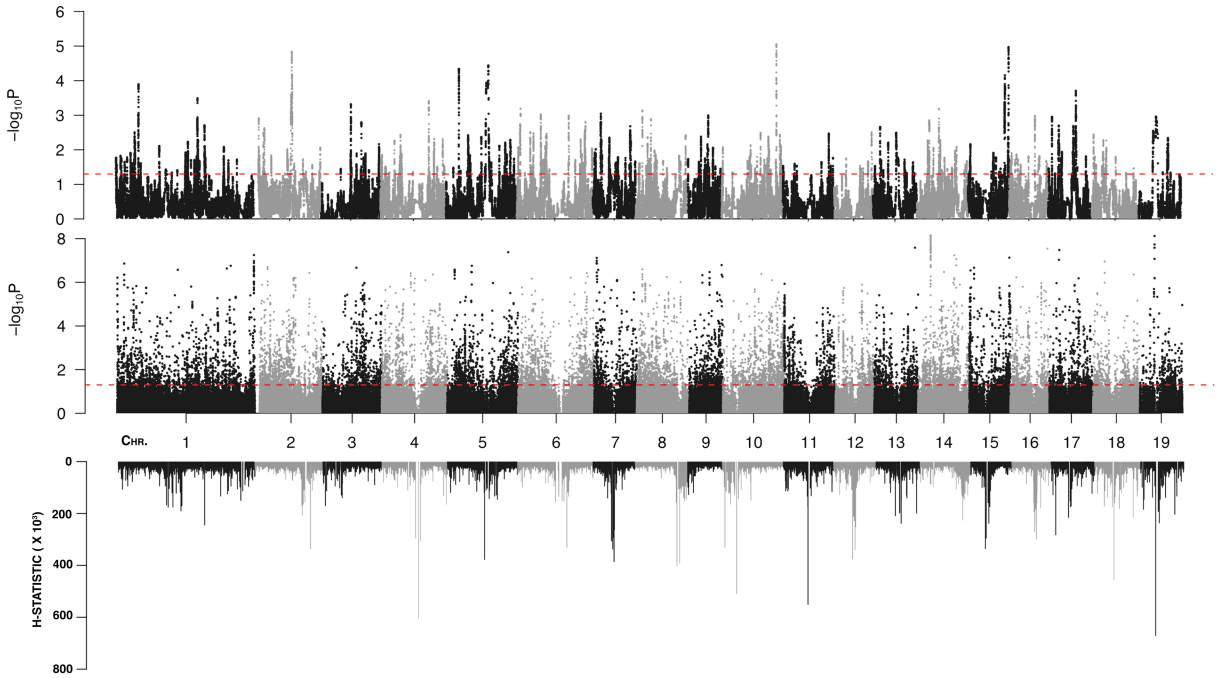


Figure S2.5: Selection scan results. The figure shows the results from HapFLK scan, OutFlank method and H-statistics respectively performed on all 10 demographic populations. The genome-wide thresholds correspond to $P < 0.01$ (red dashed lines).

Table 3.1: Properties of high and low genomic islands of differentiation identified using HMM approach.

North X Center	**High-differentiation islands (HDIs)**		**Low-differentiation islands (LDIs)**	
	Mean	**Min-max**	**Mean**	**Min-max**
No. of Islands	21074		26077	
Mean F_{ST}	0.0564	0-0.512	0.0581	0-0.0317
No. of SNPs	9.0	1 - 128	9	1 - 128
Mean Nuc. diversity	0.00093	0.00000376-0.0128	0.000942	0.00000376-0.0128

North X South	**High-differentiation islands (HDIs)**		**Low-differentiation islands (LDIs)**	
	Mean	**Min-max**	**Mean**	**Min-max**
No. of Islands	31347		15750	
Mean F_{ST}	0.045	0 - 0.316	0.055	0 - 0.292
No. of SNPs	7.87	1 - 124	7.54	1 - 119
Mean Nuc. diversity	0.00085	0.0000115 - 0.0128	0.000833	0.0000115 - 0.0128

Center X South	**High-differentiation islands (HDIs)**		**Low-differentiation islands (LDIs)**	
	Mean	**Min-max**	**Mean**	**Min-max**
No. of Islands	13816		25580	
Mean F_{ST}	0.0089	0 - 0.294	0.00425	0 - 0.016
No. of SNPs	7.2	1 - 133	7.04	1 - 133
Mean Nuc. diversity	0.000612	0.00000376 - 0.011	0.00000582	0.00000376 - 0.0122

Table 3.2: Gene Ontology enrichments for selection scan outliers (common between hapFLK and OUTFLANK).

GO biological process	**Fold Enrichment**	**q-value**
Potri.014G058900 (Remorin family protein); Chr14:4582906	1852.22	0.00
Potri.003G128000 (basic helix-loop-helix (bHLH) DNA-binding superfamily protein); Chr3:14833419	1002.17	0.00
Potri.009G085500 (function unknown); Chr9:8009162	987.85	0.00
Potri.015G140600 (Sugar isomerase (SIS) family protein); Chr15:14811342	987.85	0.00
Potri.016G098000 (function unknown); Chr16:9477486	823.20	0.00
Potri.002G180900 (function unknown); Chr2:14115077	823.20	0.00
Potri.002G180800 (Homeodomain-like superfamily protein); Chr2:14113573	811.77	0.00
Potri.005G225200 (Rho termination factor); Chr5:23536110	548.80	3.63e-14
Potri.006G244100 (Protein of unknown function (DUF630 and DUF632)); Chr6:25219510	514.50	2.61e-16
Potri.017G078800 (sec7 domain-containing protein); Chr17:9188398	470.40	1.13e-13
Potri.019G046000 (Disease resistance protein (TIR-NBS-LRR class) family); Chr19:5556316	470.40	1.13e-13
Potri.018G045500 (Leucine-rich receptor-like protein kinase family protein); Chr18:4181301	411.60	6.28e-11
Potri.009G048500 (sugar transporter 14); Chr9:5379229	365.87	6.46e-13
Potri.004G156500 (function unknown); Chr4:17827128	343.00	6.86e-15
Potri.002G176900 (Basic helix-loop-helix (bHLH) DNA-binding family protein); Chr2:13645399	329.28	2.52e-08
Potri.006G114000 (function unknown); Chr6:8882010	329.28	2.52e-08
Potri.014G059100 (RNI-like superfamily protein); Chr14:4592248	329.28	2.52e-08
Potri.019G075400 (Pentatricopeptide repeat (PPR-like) superfamily protein); Chr19:10969573	329.28	2.52e-08

Supplementary Figures and Tables

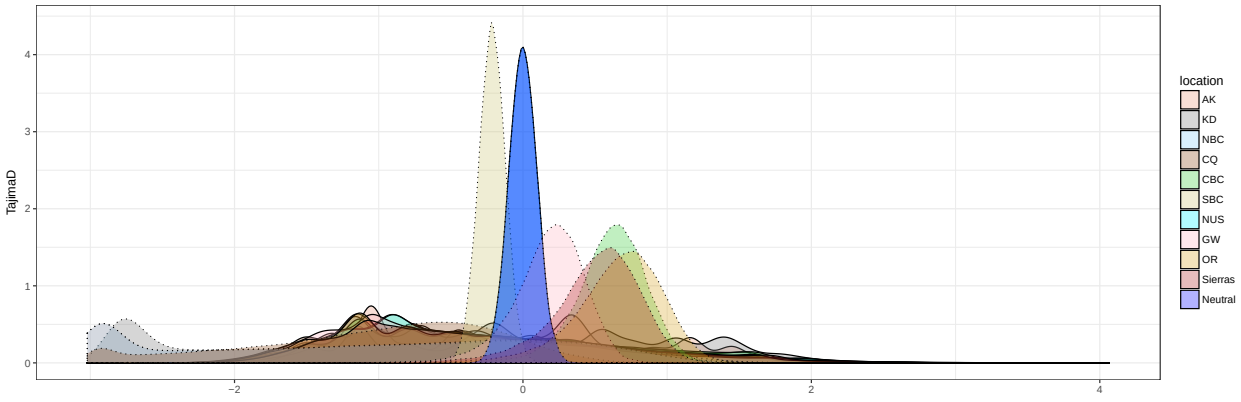
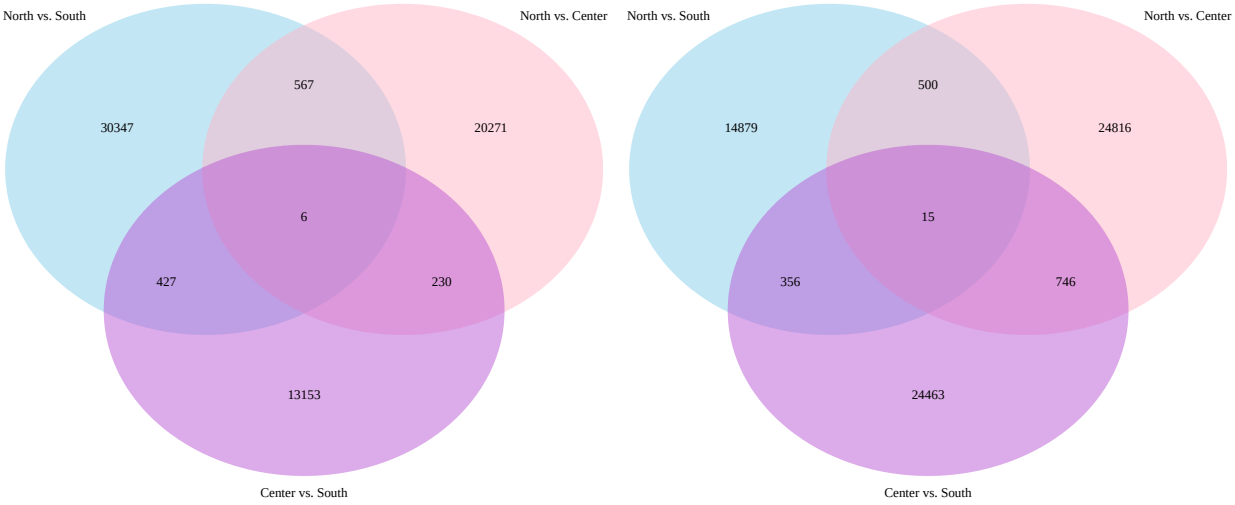


Figure S3.1: The distribution of observed and simulated values of Tajima's D in all 10 demographic populations.



(a) (b)

Figure S3.2: **Overlap in the genetic regions in the genomic islands of divergence** A. Overlap in the genetic regions between different pairs of populations for highly differentiated regions B. Overlap in the genetic regions between different pairs of populations for low differentiated regions

Table S3.1: Gene-Ontology enrichments for variants in highly differentiated regions in North vs. South populations.

GO biological process	**Fold Enrichment**	**q-value**
Potri.003G029400 (Receptor like protein 1)	11.9	0.00
Potri.003G013200 (receptor like protein 15)	10.5	0.00
Potri.006G275800 (NB-ARC domain-containing disease resistance protein)	7.65	0.00
Potri.010G121000 (PR5-like receptor kinase)	7.65	0.00
Potri.008G203700 (Replication factor-A protein 1-related)	7.27	0.00
Potri.008G203700 (PIF1 helicase)	7.27	0.00
Potri.001G420000 (NB-ARC domain-containing disease resistance protein)	6.88	0.00
Potri.006G272100 (NB-ARC domain-containing disease resistance protein)	6.25	0.00
Potri.012G108000 (Pentatricopeptide repeat (PPR-like) superfamily protein)	5.36	0.00
Potri.012G082900 (Tetratricopeptide repeat (TPR)-like superfamily protein)	4.97	0.00
Potri.004G191800 (wall-associated kinase 2)	4.89	0.00
Potri.016G114400 (Leucine-rich repeat protein kinase family protein)	4.08	0.00
Potri.004G000200 (zinc ion binding;nucleic acid binding)	3.83	0.00

Table S3.2: Gene-Ontology enrichments for variants in highly differentiated regions in North vs. Center populations.

GO biological process	**Fold Enrichment**	**q-value**
Potri.013G069500 (Unknown function)	6.22	0.0000000
Potri.005G145700 (Unknown function)	4.19	0.0000000
Potri.016G114400 (Leucine-rich repeat protein kinase family protein)	3.22	0.0000000
Potri.005G158200 (F-box protein-related)	3.04	0.0000000
Potri.006G185800 (Photosystem II reaction center PsbP family protein)	2.76	0.0000000
Potri.006G242500 (ATP binding;nucleic acid binding;helicases)	2.76	0.0000000
Potri.019G019300 (Pentatricopeptide repeat (PPR-like) superfamily protein)	2.21	0.0000000
Potri.006G058300 (cytochrome P450, family 93, subfamily D, polypeptide 1)	2.07	0.0000000

Table S3.3: Gene-Ontology enrichments for variants in highly differentiated regions in Center vs South populations.

GO biological process	**Fold Enrichment**	**q-value**
Potri.001G108500 (Tetratricopeptide repeat (TPR)-like superfamily protein)	5.01	0
Potri.001G066500 (Disease resistance protein (TIR-NBS-LRR class) family)	3.21	0
Potri.001G379700 (NB-ARC domain-containing disease resistance protein)	2.72	0
Potri.007G037500 (F-box associated ubiquitination effector family protein)	2.50	0
Potri.011G161600 (FAD-binding Berberine family protein)	2.50	0
Potri.011G091400 (damaged DNA binding protein 1A)	2.36	0
Potri.011G019600 (Ankyrin repeat family protein)	2.25	0
Potri.011G064700 (TRICHOME BIREFRINGENCE-LIKE 38)	2.17	0
Potri.011G001500 (cytochrome P450, family 716, subfamily A, polypeptide)	2.14	0
Potri.001G031600 (Auxin-responsive family protein)	2.08	0
Potri.001G217700 (Leucine-rich repeat receptor-like protein kinase family protein)	2.01	0
Potri.011G145800 (FAR1-related sequence 5)	1.42	0
Potri.001G419800 (NB-ARC domain-containing disease resistance protein)	1.40	0

Table S3.4: Gene-Ontology enrichments for variants from the hapFLK selection test

GO biological process	**Fold Enrichment**	**q-value**
Potri.006G240700 (Matrixin family protein)	583.72	0
Potri.013G087500 (hAT dimerisation domain-containing protein)	551.65	0
Potri.015G119800 (Ankyrin repeat family protein)	550.05	0
Potri.015G118200 (Bromodomain transcription factor)	502.90	0
Potri.017G047000 (hydroxyproline-rich glycoprotein family protein)	502.90	0
Potri.015G118600 (Ankyrin repeat family protein)	468.54	0
Potri.001G294600 (Plant protein of unknown function (DUF936))	449.02	0
Potri.008G044000 (Pentatricopeptide repeat (PPR) superfamily protein)	399.13	0
Potri.007G037500 (F-box associated ubiquitination effector family protein)	389.15	0
Potri.015G118800 (phloem protein 2-A9)	389.15	0
Potri.019G047100 (zinc ion binding;nucleic acid binding)	389.15	0
Potri.019G046200 (terpene synthase-like sequence-1,8-cineole)	381.66	0
Potri.001G379700 (NB-ARC domain-containing disease resistance protein)	367.38	0
Potri.001G293800 (voltage dependent anion channel 2)	347.99	0
Potri.019G046000 (Disease resistance protein (TIR-NBS-LRR class) family)	346.38	0
Potri.006G019300 (Heavy metal transport/detoxification superfamily protein)	333.55	0
Potri.001G254300 (DNA binding;zinc ion binding;DNA binding)	303.95	0
Potri.017G016000 (DNAJ heat shock family protein)	295.07	0

Table S3.5: Gene-Ontology enrichments for variants from OUTFLANK selection test.

GO biological process	**Fold Enrichment**	**q-value**
Potri.011G001500 (cytochrome P450, family 716, subfamily A, polypeptide 1)	349.99	0.00
Potri.017G089000 (UDP-glucosyl transferase 88A1)	282.69	0.00
Potri.008G150900 (sugar transporter 6)	240.81	0.00
Potri.014G059400 (hAT transposon superfamily)	175.89	0.00
Potri.011G001200 (cytochrome P450, family 716, subfamily A, polypeptide 1)	157.05	0.00
Potri.014G058900 (Remorin family protein)	141.34	0.00
Potri.001G410400 (Auxin-responsive GH3 family protein)	125.64	0.00
Potri.001G393000 (UDP-Glycosyltransferase superfamily protein)	113.07	0.00
Potri.012G074200 (plastid transcriptionally active 2)	107.69	0.00
Potri.011G001400 (cytochrome P450, family 716, subfamily A, polypeptide 1)	104.70	0.00
Potri.001G468300 (chloroplast sulfur E)	100.51	0.00
Potri.012G034700 (FAD-binding Berberine family protein)	100.51	0.00
Potri.002G165200 (RING/U-box superfamily protein)	94.23	0.00
Potri.015G004100 (NAC domain containing protein 62)	94.23	0.00
Potri.014G166600 (HXXXD-type acyl-transferase family protein)	86.98	0.00
Potri.016G110800 (defective in exine formation protein (DEX1))	85.06	0.00
Potri.011G001300 (cytochrome P450, family 716, subfamily A, polypeptide 1)	83.76	0.00
Potri.011G115500 (Tetratricopeptide repeat (TPR)-like superfamily protein)	83.76	0.00
Potri.017G088900 (Plant protein of unknown function (DUF863))	83.76	0.00
Potri.011G002000 (ATP binding cassette protein 1)	78.52	0.00