Transcriptome evolution in black cottonwood (Populus trichocarpa)

by

Jasmine Ono

Hon.B.Sc., The University of Toronto, 2004

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate Studies

(Genetics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

December 2011

Abstract

In 1975, King and Wilson proposed that gene expression variation can play a role in the evolution of phenotypic variation since the variation in nature could not be explained by variation in protein coding sequences alone. When a mutation which causes a change in gene expression is introduced in to a population, either selection or neutral drift can act on it. When this mutation causes no change in fitness of the organism, it will be affected by neutral drift. The bounds for neutral drift are thought to be set by stabilizing selection. If the mutation is beneficial to the organism, it will be affected by positive selection. When different populations are located in different environmental conditions, different mutations can be beneficial in each population and divergent selection can result. We looked for these patterns of gene expression evolution among populations of *Populus trichocarpa*, black cottonwood, using a P_{ST} vs. F_{ST} approach.

P. trichocarpa is a model tree system that allows the study of an extended suite of tree biological processes. A suite of genomic tools have been developed for black cottonwood, including a genome sequence and a 15.5K microarray. It is broadly distributed in the far west of North America and shows an ecotypic mode of genetic differentiation, with populations divided into northern and southern groups.

In this study, we examined gene expression from 12 *P. trichocarpa* populations, 6 from the north and 6 from the south. We found evidence for divergent selection acting on the expression values of many genes, as well as stabilizing selection acting on a few. This supports the prevalence of natural selection acting on phenotypic traits, but we still found

an overwhelming majority of traits which seem to be drifting neutrally. We found no evidence for different selection acting on the northern and southern groups.

Preface

Chapter 3 resulted from a collaboration with my research supervisor, Dr. Kermit Ritland. K. Ritland first suggested looking at gene expression divergence among populations of *Populus*. I designed the experiment, with input from both K. Ritland and R. White, collected and analyzed the data, and wrote the manuscript. K. Ritland performed the estimations of F_{ST} and provided comments on the manuscript. Single nucleotide polymorphism data was provided by A. Geraldes, C. Douglas and Q. Cronk.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank Dr. Kermit Ritland for his guidance and supervision throughout my Master's. Without his help, this project would have never gotten started, let alone completed. I would also like to thank all of the other members of the Ritland lab and Treenomix group from 2008-2011 and Dr. Carol Ritland for all of their support and feedback. I would like to especially thank Dr. Kermit Ritland, Dr. Carol Ritland, Hesther Yueh, Agnes Yuen, Stuart Murray and Michelle Sun for all of their help in field collection and in the lab. I would also like to thank Nima Farzaneh for all of his help with the bioinformatics side of things and Rockney Albouyeh for sharing his experience.

I would like to thank Dr. Carl Douglas and Dr. Loren Rieseberg for their time as members of my Master's committee. Their advice and input helped guide this project in to its final form.

I would also like to thank Rick White for making the statistics more approachable. Thanks to the Bohlmann lab, for use of their microarray scanner when I desperately needed one. Thanks to Ryan Philippe and Mohamed Ismail for providing advice and information about the microarrays and trees that were used in this project. I would also like to thank Armando Geraldes, Dr. Quentin Cronk and Dr. Carl Douglas for use of a subset of their SNP data, which proved invaluable to the project. I would also like to thank them for meeting and talking with me about the project and sharing their own results, so that we could all try to make sense of the bigger picture.

I would like to thank everyone who organizes and attends DeltaTea, EDG, VEG and all

of the other reading groups available at UBC. These groups have helped me to develop my thinking as an evolutionary biologist and as a scientist in general.

I would also like to thank NSERC for funding this project through a grant awarded to Dr. Kermit Ritland.

Finally, I would like to thank many of the members of the evolution group at UBC for their continued support and willingness to help whenever possible. Of special note are: Dr. Sarah Otto, Dr. Michael Whitlock, Aleeza Gerstein, Kate Ostevik, Kieran Samuk, Milica Mandic, Rich Fitzjohn, Anne Dalziel, Gina Conte, Alana Schick, Michael Scott, Sam Yeaman, Florence Debarre, Karen Magnuson-Ford, Matthew Siegle, Greg Baute, Kathryn Turner, David Toews, Gwylim Blackburn, Brook Moyers, Laura Southcott, Jon Mee, Jean-Sebastien Moore, and Greg Owens. From help with R code and discussions of statistics to questions at practice talks and motivating me to work, completing this Master's without this group would have been near impossible.

Chapter 1

Introduction

Populus has recently been developed as a model system for long-lived organisms, specifically trees. The genome sequence of *Populus trichocarpa* has been elucidated (Tuskan et al., 2006) and large amounts of associated genomics resources (Jansson and Douglas, 2007) have been developed, specifically microarrays (Ralph et al., 2006) and thousands of single nucleotide polymorphism (SNP) markers (Geraldes et al., 2011), of relevance to this thesis. Trees offer the opportunity to study an extended suite of biological processes, many of which cannot be studied in other model plant systems such as *Arabidopsis* and rice. For example, trees in temperate climates need to be able to deal with seasonal changes and withstand winter conditions for many years running, while annual plants do not have these same pressures. Also, black cottonwood is a dioecious plant, meaning that there are two distinct sexes (DeBell, 1990), which is relatively rare among plants, and is not the case for other model plant species. It is also an important commercial plantation species, so insights into its biology can have commercial applications.

In 1975, King and Wilson found that the amount of phenotypic variation visible in nature could not be explained by the variation in proteins alone (King and Wilson, 1975). They explained this by suggesting that gene expression variation may also play a role in the phenotypic variation found in nature (King and Wilson, 1975). Since then, a central issue in evolutionary biology has been the relative roles of structural protein divergence (mutations that cause changes in amino acid sequence) and gene regulatory divergence (changes in the level of gene expression) (Hoekstra and Coyne, 2007). As a result of advances in

technology to measure gene expression, there has been an explosion of investigations that address the changes of gene expression observed among related species. These changes are due to putative evolutionary forces. Microarray technology was one of the key developments in recent years that propelled biological research into the post-genomic era (Shiu and Borevitz, 2008). The advent of microarrays allowed the ability to assay thousands of features at the same time, the most popular use of which was to profile messenger RNA (mRNA) levels (Shiu and Borevitz, 2008). The advent of microarrays has allowed the examination of the extent of variation in gene expression both within and among taxa, as well as allowed the formation of hypotheses about the evolutionary processes affecting this variation (Whitehead and Crawford, 2006b). For the purposes of this thesis, I have utilized the 15.5K poplar cDNA microarray developed by the Treenomix group (Ralph et al., 2006). This microarray, as well as cDNA libraries and ESTs, was developed as part of a genomics strategy to characterize inducible defences against insect herbivores in poplar (Ralph et al., 2006). This complemented previous genomic work in *Populus* by focusing on herbivore- and elicitor-treated tissues and incorporating normalization methods to capture rare transcripts (Ralph et al., 2006). We have also utilized a subset of the SNP resources developed by Geraldes et al. (2011).

As with all other phenotypes, gene expression can be affected by the evolutionary forces drift and selection. The methodologies to infer the relative roles of these evolutionary forces have also seen rapid development. Most are based on searching for departures from a neutral model (Fay and Wittkopp, 2008; Gilad et al., 2006a). Following Kimura (1983), the neutral model proposes that the greater the divergence is among taxa, the greater the divergence will be in their gene expression levels (Whitehead and Crawford, 2006b; Khaitovich et al., 2004). As our method of detecting evolutionary forces, we took a P_{ST} vs. F_{ST} approach. F_{ST} is a standardized measure of the degree of between population differentiation in alleles (Whitlock, 2011) and Q_{ST} is an analogous measure for the genetic

differentiation in a quantitative trait (Spitze, 1993). P_{ST} is an approximation of Q_{ST} that uses the phenotypic differentiation instead of the genetic differentiation in a trait (Brommer, 2011). Here, a departure from the neutral model would be a Q_{ST} (or P_{ST}) value which is significantly greater or less than the F_{ST} value. A greater Q_{ST} (or P_{ST}) value would be evidence for divergent selection and a smaller Q_{ST} (or P_{ST}) value would be evidence for stabilizing selection. In this thesis, I will apply this method to infer the patterns of evolution of gene expression in *Populus trichocarpa*.

1.1 Biology of black cottonwood

Black cottonwood, or *Populus trichocarpa*, is a member of the Salicaceae family of flowering plants (DeBell, 1990). It is among the fastest growing temperate trees and is the largest of the American poplars and the largest hardwood tree in western North America, able to exceed 60 m in height and up to 3 m in diameter (DeBell, 1990; Slavov and Zhelev, 2010). Black cottonwood is a long-lived tree species, growing for as long as 200 years (DeBell, 1990). It primarily grows on moist sites and preferably on alluvial soils (DeBell, 1990). The species is broadly distributed in a coastal range from Alaska to California. Inland, it is generally found on the west of the Rocky Mountains in British Columbia (BC), western Alberta, western Montana and northern Idaho (Fig. 1.1) (DeBell, 1990). A few scattered populations can also be found in southeast Alberta, eastern Montana, western North Dakota, western Wyoming, Utah and Nevada (DeBell, 1990). Observations by the BC Ministry of Forests confirmed previous reports that black cottonwood is absent from the central BC coast, referred to as the "no-cottonwood" belt, dividing the species' distribution into a northern region and a southern region (Xie et al., 2009). Small, isolated patches of cottonwood are found at small river plains along the belt (Xie et al., 2009). The northern and southern populations may have originated from different glacial refugia (Xie

Figure 1.1: Range of *Populus trichocarpa* in North America (Little, 1971).

et al., 2009; Soltis et al., 1997).

In its range, the annual precipitation varies between 10 inches to over 120 inches, with only about a third of that occurring during the growing season (DeBell, 1990). The frost-free period ranges from about 70 days to over 260 days, the maximum temperature can vary from 16°C to 47°C and the minimum temperature can vary from 0°C to -47°C (DeBell, 1990). Black cottonwood also grows over a range of elevations from sea level up to about 2100 m in British Columbia (DeBell, 1990).

Black cottonwood is normally dioecious, which means that male and female catkins are borne on separate trees (DeBell, 1990), although hermaphroditic trees have been reported (Slavov et al., 2009). Gender is genetically determined (Jansson and Douglas, 2007; Slavov and Zhelev, 2010), but there may be ecological determinants as well, as male clones are more frequent on drier sites (McLetchie and Tuskan, 1994). Under favourable conditions, *Populus* trees can reach maturity within four to eight years in intensively managed plantations and 10 to 15 years in natural populations (Slavov and Zhelev, 2010). Black cottonwood's flowering time can vary from early March to as late as mid-June across the range (DeBell, 1990). The relative timing of flowering follows a temperature-dependent progression, with populations at higher elevations, more northern latitudes and more continental climates flowering later (Slavov and Zhelev, 2010). Pollen is dispersed by wind and effective long-distance pollination can be extensive (Slavov and Zhelev, 2010). Large amounts of light and buoyant seeds can be produced and can be transported long distances by wind and water, although direct empirical data on dispersal distances is limited (Slavov and Zhelev, 2010; DeBell, 1990). Moist seed beds are essential for high germination, and seedling survival depends on continuously favourable conditions during the first month (DeBell, 1990). Young saplings are frequently injured and sometimes killed by unseasonably early or late frosts (DeBell, 1990). Frost cracks also decrease the quality of the wood and provide entrance for decay fungi (DeBell, 1990). Mortality in the first year is typically

high in *Populus* (up to 77-100%) (Slavov and Zhelev, 2010). Asexual reproduction is also common through root sprouting and rooting of shoots from broken branches or entire tree trunks (Slavov and Zhelev, 2010).

1.2 *Populus* as a model system for trees

Poplar (*Populus* spp.) is an established model system for genomic studies in angiosperm tree biology (Miranda et al., 2007; Tuskan et al., 2006) and includes species commonly known as aspens, cottonwoods and poplars. These trees are deciduous and mainly in the boreal, temperate and subtropical zones of the Northern Hemisphere (Slavov and Zhelev, 2010). *Populus* allows us to study many biological processes that better represent the breadth of plant biology (Jansson and Douglas, 2007). It will help us understand the evolution, function and adaptation of a genome of a long-lived, perennial, woody plant (Miranda et al., 2007). The ability of many species to be propagated by vegetative cuttings, a relatively short generation time, and susceptibility to *Agrobacterium*-mediated transformation are all useful traits in the development of a model system (Miranda et al., 2007). *Populus* is also a relatively close relative of *Arabidopsis* as a member of the Eurosid clade, which facilitates comparative genomics between the two species (Jansson and Douglas, 2007). It is a plantation forest tree with traditional uses as a species for wood and fibre (Jansson and Douglas, 2007; Miranda et al., 2007). *Populus* also has an unusual amount of natural variation that can allow us to explore many questions fundamental to tree biology, such as lignin and cellulose formation, perennial growth, dormancy and resistance against biotic and abiotic stress, many of which are now being addressed with genomic approaches in *Populus* (Jansson and Douglas, 2007; Miranda et al., 2007). These high levels of natural variation were supported by the *P. trichocarpa* genome, determined from a wild tree, which found levels of heterozygosity, or within individual genetic polymorphisms, at an overall

rate of approximately 2.6 polymorphisms per kilobase (Tuskan et al., 2006).

Trees are the opposite extreme to *Arabidopsis thaliana*, in that trees have a long life span and display woody growth forms (Jansson and Douglas, 2007). *Populus* is in the angiosperm Euroside I clade with *Arabidopsis* (Jansson and Douglas, 2007). A commonly used classification of *Populus* divides the genus into 29 species subdivided into 6 sections based on relative morphological similarities and crossability (Eckenwalder, 1996). The classification remains undecided, however, with the number of species varying from 22 to 85 (Slavov and Zhelev, 2010; Eckenwalder, 1996). *Arabidopsis* is more related to *Populus* than to most dicots, not to mention monocots like rice or gymnosperm trees like conifers (Jansson and Douglas, 2007). This facilitates comparative genomics approaches between the two model species, which is helpful since *Arabidopsis* has the most complete genome annotation of any plant (Jansson and Douglas, 2007).

Populus has been established as a system for genomic research of angiosperm tree biology (Ralph et al., 2006; Tuskan et al., 2006; Brunner et al., 2004). There are genomic and molecular biological resources available for *Populus*, including a genome sequence of *Populus trichocarpa*, or black cottonwood (Tuskan et al., 2006)(http://www.phytozome.net/poplar). The ~480 Mb genome is divided into 19 linkage groups and has been integrated with a detailed genetic map. The genome is only about 4.5-fold larger than the *Arabidopsis* genome and about 40-fold smaller than members of the pine family (Pinaceae) (Ralph et al., 2006). In version 2.2 of the *Populus* genome assembly and annotation, there were 45,000 promoted gene models, one of the largest for any completely sequenced plant genome to date (Jansson and Douglas, 2007)(http://www.phytozome.net/poplar). DNA microarrays have been developed in parallel with expressed sequence tag (EST) and genome sequencing (Jansson and Douglas, 2007), including the 15.5K element Treenomix cDNA microarray used for our study. These and other molecular and bioinformatic resources being developed for *Populus* make it an excellent system for studying tree genetics and genomics (Slavov

and Zhelev, 2010).

1.2.1 Biotic interactions of poplar

In their natural environment, poplars are often ecologically dominant trees and interact with a diverse array of mammals, insect pests, pathogens or symbionts over their relatively long lifespan (Miranda et al., 2007). *Populus* therefore needs to defend itself year after year and may also develop beneficial biotic interactions (Jansson and Douglas, 2007). Forest insect pests are a challenge to the sustainability of natural and planted forests because of the risk of forest insect pest epidemics (Ralph et al., 2006). This risk is increasing with global climate changes and the introduction of exotic pest species (Ralph et al., 2006). The larvae of several insect herbivores can cause extensive defoliation to stands of *Populus* species during outbreak periods (Ralph et al., 2006). The first lines of defence against insect herbivores are constitutive chemical and physical barriers (Ralph et al., 2006). Constitutive levels of phenolic products are likely involved in insect herbivore defence (Osier and Lindroth, 2006). Genetically determined variation in phenolic glycoside levels in aspen leaves have been shown to negatively impact growth and performance of forest tent caterpillars and other herbivores (Ralph et al., 2006; Osier and Lindroth, 2006). Interactions with a biotrophic fungus are not known for *Arabidopsis*, so *Populus* is now one of the best established genomic systems to study this biological interaction (Miranda et al., 2007). The 15.5K poplar cDNA microarray has been used to study both poplar's response to herbivory by forest tent caterpillars (*Malacosoma disstria*) (Ralph et al., 2006) and interactions with a biotrophic rust fungus *Melampsora medusa* (Miranda et al., 2007).

1.2.2 Silviculture of poplar

Populus is an important commercial plantation genus (Jansson and Douglas, 2007). Black cottonwood is planted as windbreaks and shelterbeds in conjunction with irrigated agriculture in the Columbia River basin (DeBell, 1990). It also has short, fine fibres and is used for pulp for high-grade book and magazine papers (DeBell, 1990). Its veneer is used in plywood, baskets and crates and it is also used to manufacture pellets and boxes (DeBell, 1990). More of the wood is used in concealed parts of furniture, fiberboard and flakeboard (DeBell, 1990).

These commercial uses offer application to research on the trees, such as research into the production of superior pulping trees as well as the use of woody plants as a source of ligno-cellulosic feedstock for biofuels (Pan et al., 2006; Miranda et al., 2007). The pulping characteristics of wood from field-tested lines showed the potential to make modified lignin trees with superior wood quality (Jansson and Douglas, 2007). The 15.5K Treenomix microarray was also used to identify a set of transcription factors common to *Populus* and *Arabidopsis* whose expression correlated to secondary wall formation in both, and sometimes spruce (Jansson and Douglas, 2007). This information can allow researchers to develop a better pulping tree as well as learn about the evolution of the wood-forming nature of trees.

1.3 Molecular genetic diversity of *Populus*

Trees usually have higher levels of genetic diversity within populations and lower genetic differentiation between populations than other plants (Hamrick et al., 1992). As a wind-pollinated obligate outbreeder with relatively large population sizes, *Populus* may have even higher variation than other trees (Slavov and Zhelev, 2010; Jansson and Douglas,

2007). A female *Populus* can produce tens of millions of seeds per year with potentially thousands of fathers (Jansson and Douglas, 2007). Studies of gene flow suggest that long-distance pollination can be extensive (Slavov and Zhelev, 2010). The seeds can then be dispersed many kilometres by wind (Slavov and Zhelev, 2010). Neutral molecular markers and adaptive traits reveal high levels of genetic variation within populations (Slavov and Zhelev, 2010; Jansson and Douglas, 2007). Deviations from Hardy-Weinberg equilibrium are not uncommon, but the magnitudes of the deviations are typically small to moderate (Slavov and Zhelev, 2010). The efficient mixing of alleles in outbreeding species ensures that those that give the highest fitness will accumulate in a population at a given site (Jansson and Douglas, 2007). In contrast to inbreeders like *Arabidopsis*, false positives from population structure are less of a problem (Jansson and Douglas, 2007). The differentiation among populations, as measured by F_{ST}, is typically weak (Slavov and Zhelev, 2010). The median F_{ST} for the genus is 0.047 as measured by allozymes and RFLPs and the microsatellite measures are comparable (Slavov and Zhelev, 2010). This is almost two times lower than the mean for long-lived woody species (0.084) and nearly five times lower than plants in general (0.228) (Slavov and Zhelev, 2010). The values for black cottonwood are 0.063, as measured by allozymes (Weber and Stettler, 1981), and 0.078/0.112 (F_{ST}/R_{ST}) using microsatellite markers (Ismail, 2010). These values are in agreement with long-distance pollination and seed dispersal.

1.4 Adaptive evolution in *Populus*

With a life span of decades, trees face challenges distinct from those of annuals (Jansson and Douglas, 2007). Patterns of geographic variation in forest trees are primarily shaped by three interactive evolutionary forces: natural selection, genetic drift and gene flow (Xie et al., 2009; Morgenstern, 1996). Continuous clinal variation is expected if environmen-

tal factors change gradually along geographic coordinates and gene flow between adjacent populations is not restricted (Xie et al., 2009). Either abrupt environmental change or geographically isolated populations can lead to discontinuous or ecotypic variation (Xie et al., 2009). This can be especially true if isolated populations are founded from different glacial refugia, even with gradual environmental change (Xie et al., 2009). Other tree species have been found to have clinal patterns of genetic variation along the Pacific Northwest coast because the environmental change is gradual and there are no barriers to gene flow between populations of those species (Xie et al., 2009). This may not be the case for black cottonwood, however, due to the "no-cottonwood" belt that may restrict gene flow (Xie et al., 2009).

There is considerable quantitative genetic variation in cottonwood throughout its range. Growth is considerably less in northerly and interior locations (DeBell, 1990). This could be partially because trees in temperate climates need to be able to adapt to seasonal changes that restrict growth and withstand winter conditions (Jansson and Douglas, 2007). *Populus* is a typical deciduous tree and its ability to anticipate winter conditions is highly adaptive (Jansson and Douglas, 2007). Temperate and boreal trees alternate between active growth in the summer and dormancy in the winter with tradeoffs existing between substantial cold hardiness and growth (Holliday et al., 2008). The timing of entry into and exit from dormancy is locally adaptive (Holliday et al., 2008), with the most important input for anticipation of winter conditions in *Populus* being the shorter days in autumn (Jansson and Douglas, 2007). For example, photoperiodic studies conducted on black cottonwood under uniform conditions in Massachusetts have found that northern provenances cease growth earlier than southern provenances (DeBell, 1990; Pauley and Perry, 1954). Also, the cessation of growth among clones from the same latitude was related to the length of the growing season (number of frost-free days) at places of origin (elevation) (DeBell, 1990; Pauley and Perry, 1954). These are evidence for genetic clines in cessation of growth.

Several aspects of shoot growth were found to be under genetic control in another study (DeBell, 1990); date of flushing, amount of early growth, growth rate in midseason, date of cessation and average length of internode. There is also a large range of variation in leaf, branch and phenology characters, many of which vary clinally with latitude, longitude or elevation (Weber et al., 1985). Southwest clones tended to have smaller leaves, more numerous and more erect branches and continued growth later in the fall than those from the northeast (Weber et al., 1985).

In conifers, there is evidence for significant among-population differential gene expression along a latitudinal cline that corresponds to the genetic cline in cold hardiness, bud phenology and growth (Holliday et al., 2008). This was found in Sitka spruce (*Picea sitchensis*) and was interpreted as evidence for adaptive variation in cold hardiness (Holliday et al., 2008).

There is also evidence for selection in the timing of bud flush after dormancy is broken, as it is under genetic control with a tree of a given genotype requiring a certain temperature run for bud flush (Jansson and Douglas, 2007). Variation in the timing of bud flush usually exists between populations from different latitudes (Jansson and Douglas, 2007). Genecological studies in *Populus* also revealed strong and repeatable correspondence between clinal genetic variation for adaptive traits and climatic and geographic factors believed to be important agents of natural selection (Slavov and Zhelev, 2010; Morgenstern, 1996). Climate change may make selection for traits related to local adaptation increasingly important in managed forests (Holliday et al., 2008).

A few studies of particular interest look for specific genes associated with growth cessation in the European aspen, *Populus tremula*. There is evidence for divergent selection on these genes, which are *phyB2*, a phytochrome photoreceptor (Ingvarsson et al., 2006), *PtCENL*-1 gene (*Centroradialis Like*-1), a *Populus* homolog of the Terminal Flowering Locus 1 (TFL1) in *Arabidopsis thaliana* (Hall et al., 2007), *LHY1* and *LHY2*, circadian clock-

associated genes (Ma et al., 2010). Clinal variation with latitude was observed for each of these genes (Ingvarsson et al., 2006; Hall et al., 2007; Ma et al., 2010). Phytochromes are thought to be the primary regulators of night length-mediated bud set and initiation of autumn cold acclimation in perennials (Holliday et al., 2008). In hybrid aspen (*Populus tremula x Populus tremuloides*), over expression of PHYA, another phytochrome photoreceptor, blocked growth cessation and cold acclimation under short day lengths (Olsen et al., 1997).

1.4.1 Gene flow in black cottonwood and its effect upon local adaptation

Gene flow is believed to be extensive in most forest trees (Slavov and Zhelev, 2010), but this may not be true for black cottonwood. Geologic and climatic information and genetic evidence from other species suggest that cottonwood in the north and the south may have originated from different glacial refugia (Soltis et al., 1997; Xie et al., 2009). Xie et al. (2009) performed a common-garden test of 180 provenances of 36 drainages from northern BC to Oregon and found an ecotypic mode of north-south regional differentiation, with these regions being divided by the "no-cottonwood" belt. Data on height, abnormal flushing and infection of *Valsa sordida* and *Melampsora occidentalis* were collected (Xie et al., 2009). *V. sordida* affects weakened or stressed trees and creates cankers while *M. occidentalis* causes leaf rust (Xie et al., 2009). Trees from the north showed higher mortality, grew more slowly, were more susceptible to both pathogens tested and had a higher frequency of abnormal bud flushing (Xie et al., 2009). Regional differentiation accounted for the highest amount of variation observed in all traits measured (Xie et al., 2009). It seems that northern trees are poorly adapted to the southern coastal environment in Surrey, BC (Xie et al., 2009). This provides compelling indirect evidence for local adaptation in

black cottonwood because genotypes from a given habitat tend to have higher fitness in that habitat (Slavov and Zhelev, 2010). Genotype by environment interactions are commonly detected and are also a condition for local adaptation (Slavov and Zhelev, 2010). The species' distribution biography, ecological characteristics and life history suggest that restricted gene flow is the main factor responsible for the observed geographic pattern of genetic differentiation (Xie et al., 2009).

If the populations from the north and south originated from two separate refugia, they subsequently have not been able to converge, possibly because of physical barriers and the species' biological limits to colonization (Xie et al., 2009). The northern and southern coasts used to be two separate crustal fragments, which converged around the present location of the "no-cottonwood" belt about 140 million years ago (Xie et al., 2009). This may have created the present land formation with uplifted mountains, deep narrow river channels and discontinuous riverine systems that has restricted the availability of favourable habitat for black cottonwood seeds in the region and therefore confined the species' expansion (Xie et al., 2009).

In general, the degree of local adaptation may be from reproductive isolation by distance or by barrier ("no-cottonwood belt", phenological asynchrony between populations growing under different climatic conditions), from very strong divergent selection acting on the trait (Slavov and Zhelev, 2010), or a combination of these factors. A similar pattern of adaptive genetic variation is seen in other tree species for this range, where northern provenances grow much slower and suffer more severe disease infection and mortality in the southern environment (Xie et al., 2009; Ying and Liang, 1994; Xie et al., 1996; Hamann et al., 1998), but they have continuous differentiation. These species include red alder (Hamann et al., 1998; Xie et al., 1996), Sitka spruce and Shore pine (Xie et al., 2009; Ying and Liang, 1994). The majority of climatic variables vary continuously across the two regions (Xie et al., 2009). It may be that small patches of cottonwood in the "no-cottonwood"

belt have failed to bridge the gene flow between the two regions and restricted gene flow is shaping and sustaining the geographic pattern of genetic differentiation (Xie et al., 2009). Differences may have developed during glaciations when there were few, small, scattered refugia (Xie et al., 2009). Separate refugia could have undergone local adaptation or been differently affected by drift while isolated, leading to differentiation between populations derived from them (Xie et al., 2009). Neutral microsatellite markers in 47 populations across the range were also found to have differentiated into northern and southern groups, similar to those of Xie et al. (2009)(Ismail, 2010). Unravelling the relative roles of gene flow and natural selection, and the molecular underpinnings of adaptive genetic variation will be critical for the basic understanding of the evolution of *Populus* and for designing conservation and commercial strategies (Slavov and Zhelev, 2010).

Chapter 2

The evolution of gene expression

2.1 Gene expression and its role in evolution

The underlying mechanism of evolution has traditionally been viewed as structural protein divergence, or mutations that lead to changes in amino acid sequence, but a central issue in evolutionary biology is the relative roles of structural protein divergence and gene regulatory divergence, or changes in the level of gene expression (Hoekstra and Coyne, 2007). To explain how species with highly similar and even identical genes can differ so substantially in anatomy, physiology, behaviour and ecology, it was suggested that evolutionary differences are often based on changes in expression rather than amino acid changes (King and Wilson, 1975). These changes in gene expression are expected to correlate with protein levels, and therefore biological functions (Khaitovich et al., 2004). Until recently, however, relatively little attention had been paid to this hypothesis (Whitehead and Crawford, 2006b). Supporting the importance of gene expression changes to evolution, substantial differences have been found to exist in gene expression between related species (Shiu and Borevitz, 2008; Fay and Wittkopp, 2008). Genome-wide measurements have revealed high rates of genetic variation in gene expression in humans, mice, fish, flies, yeast, plants and bacteria (for a list of references, see Fay and Wittkopp, 2008). If this variation in regulatory or coding regions is heritable, it can be the raw material for evolutionary processes (Whitehead and Crawford, 2006b). It is generally agreed that much of the variation in gene expression for a particular environmental condition is heritable (Stamatoyannopou-

los, 2004; Gibson and Weir, 2005). It is not known whether the majority of changes in gene expression fixed during evolution are caused by selection or drift, but it is likely that gene expression is affected by these processes (Khaitovich et al., 2004; Whitehead and Crawford, 2006b). The relative importance of changes in protein function versus regulatory changes is still a subject of debate (Fay and Wittkopp, 2008).

The advent of microarrays has allowed the examination of the extent of variation in gene expression both within and among taxa, as well as allowed the formation of hypotheses about the evolutionary processes affecting this variation (Whitehead and Crawford, 2006b). The ability to assay thousands of features at a time has fundamentally changed how biological questions are addressed (Shiu and Borevitz, 2008). Microarrays can be broadly defined as tools for massively parallel ligand binding assays, where features are placed at high density on a solid support, for recognizing a complex mixture of target molecules (Shiu and Borevitz, 2008). The features on a microarray can be a variety of things, but DNA microarrays are the most popular and well developed and the most well known use is the profiling of messenger RNA (mRNA) levels (Shiu and Borevitz, 2008). Differences in gene regulation are likely to have an important role in the phenotypic variation both within and between taxa (King and Wilson, 1975; Gilad et al., 2006a) because measures of gene expression are used as proxies for the active amount of protein present in the cells (Whitehead and Crawford, 2006b). When and where a gene is expressed, as well as how much is made, can be as important as the biochemical function of the encoded RNA or protein (Fay and Wittkopp, 2008). There have been an increasing number of studies in evolutionary biology that use microarray technology to look at the expression of thousands of genes at a time, instead of only looking at the usual candidate characters, traits and genes (Whitehead and Crawford, 2006b). This can lead to novel insights about links between certain genes and adaptations not previously thought to be related.

The proportion of expression divergence attributable to natural selection remains un-

clear but there is large inter-individual variation, composed of a minor non-genetic component and a large heritable component, as has been demonstrated with crosses between strains of inbred lines (Whitehead and Crawford, 2006b). Variation is expected to be minimal between genetically identical individuals and increase among more distantly related individuals (Whitehead and Crawford, 2006b). Variation among individuals within outbred populations has typically been measured in humans and fish and is consistently high (Whitehead and Crawford, 2006b). Variation among populations and species appears to be primarily affected by neutral drift (Whitehead and Crawford, 2006a,b; Khaitovich et al., 2004). For example, Khaitovich et al. (2004) found that expression differences between species of primate and mouse accumulated roughly linearly with time, supporting a neutral model of expression evolution. They also used expressed pseudogenes as a control. Since pseudogenes don't produce any functional gene products, it is reasonable to expect that they are not the direct targets of selection (Khaitovich et al., 2004). They found that the rate of expression divergence between species doesn't differ significantly between intact genes and expressed pseudogenes, supporting the hypothesis that the majority of expression differences between species are selectively neutral (Khaitovich et al., 2004). For the pseudogenes to have been used, however, they were required to be present and expressed in both species, which may suggest that they were not evolving neutrally (Fay and Wittkopp, 2008). Also, only 23 pseudogenes were suitable for this analysis, and it's not clear whether sample size affected the results (Fay and Wittkopp, 2008). This study seems to indicate that a null hypothesis assuming functional neutrality should be used to identify gene expression differences between species that are fixed by selection (Khaitovich et al., 2004). This is in agreement with *Drosophila* (Rifkin et al., 2003), where differences in gene expression are consistent with phylogenetic relationships among species, and fish (Oleksiak et al., 2002).

2.1.1 Examples of positive selection on gene expression

There also exists selection on gene expression. Experimental evolution and evolutionary comparisons of development provide strong evidence that adaptation in natural populations often occurs by changes in gene regulation (Fay and Wittkopp, 2008; Whitehead and Crawford, 2006a; Rifkin et al., 2003; Gilad et al., 2006a). Genetic and transgenic experiments have shown that changes in gene regulation often underlie morphological differences between species, for example: changes in the pelvic structure in threespine stickleback mediated by *Pitx1*, trichome patterns in *Drosophila* by Ubx, butterfly eyespots by Distal-less and beak size among Galapagos finches by BMP4 (Fay and Wittkopp, 2008). Experimental evolution in microorganisms and studies elucidating the molecular basis of adaptations in domesticated crops also indicate a role for regulatory evolution in phenotypic evolution (for examples, see Fay and Wittkopp 2008).

An example of a microarray study that found divergent selection in gene expression was that performed by Oleksiak et al. (2002) of natural populations of the teleost fish from the genus *Fundulus*. Much of the expression divergence was described as random drift because neutral theory states that variation between populations should be a positive function of the variation within populations, and this is what was observed (Oleksiak et al., 2002). They did find, however, that some genes showed an unexpected pattern of expression changes, unrelated to evolutionary distance (Oleksiak et al., 2002). Clustering among individuals showed that some differences in expression separated the northern *Fundulus heteroclitus* population from both a southern *F. heteroclitus* population and a southern *Fundulus grandis* population (Oleksiak et al., 2002). This is not supported by neutral theory since the gene expression of the northern population differs from the expression in both southern populations, despite the fact that one is of the same species and one is of another (Oleksiak et al., 2002). Under neutral drift, the pattern of expression should be most similar among

populations within a species (Whitehead and Crawford, 2006b). These patterns of expression may be the result of evolution in different environments: cold water for the northern population and warmer waters for the southern ones (Oleksiak et al., 2002). This suggests that the natural variation that exists in gene expression may be important for evolution by natural selection (Oleksiak et al., 2002).

Another study examined the covariation of gene expression between five populations of *Fundulus* and an ecologically important parameter: native habitat temperature (Whitehead and Crawford, 2006a). They measured the expression of metabolic genes in common-gardened populations of *Fundulus heteroclitus*, whose habitat is distributed along a steep thermal gradient (Whitehead and Crawford, 2006a). After correcting for phylogeny, they found that much of the variation in gene expression fits a null model of neutral drift, but that selection seemed to be acting on 44 out of 329 genes, 13 of which were under directional selection, 24 under stabilizing selection and 7 under balancing selection (Whitehead and Crawford, 2006a). (Gilad et al., 2006b) also found evidence for selection in expression of certain genes among humans and other primates, both stabilizing and lineage-specific selection. Lineage-specific selection was judged from significantly elevated or reduced expression in the human lineage compared to the other primate lineages (Gilad et al., 2006b).

2.1.2 Stabilizing selection and gene expression

While there are examples of divergent or directional selection acting on gene expression, many studies have found a dominant signature of stabilizing selection. Rifkin et al. (2003) studied the gene expression variation of four strains of *Drosophila melanogaster*, one of *D. simulans* and one of *D. yakuba* during *Drosophila* metamorphosis. They could not reject overall low variation in 44% of the genes studied, which was considered to be evidence for stabilizing selection (Rifkin et al., 2003). Directional selection and neutral evolution

seemed to play smaller roles (Rifkin et al., 2003). Another example is Lemos et al. (2005), who analyzed published inter-species gene expression data sets of mice, *Drosophila* and apes. They calculated minimal and maximal rates of gene expression diversification consistent with neutrality, or evolution without constraint, based on a neutral model (Lynch and Hill, 1986) and found that the vast majority of genes exhibited far less between species variation than expected, which was interpreted as stabilizing selection (Lemos et al., 2005). A minority of genes were found to be under neutral drift and a few genes were under diversifying selection (Lemos et al., 2005). These studies indicate that changes in gene expression are often deleterious and therefore under stabilizing selection (Gilad et al., 2006a).

2.2 Finding evidence for selection

Extensive differences in gene expression can be detected across demographically distinct groups, like populations or species, which can generally be covered by the term "taxa" (Whitehead and Crawford, 2006b). As for nucleotide changes and other characters that are variable and heritable, some expression changes have phenotypic consequences and should be affected by drift or fixed by selection (Khaitovich et al., 2004; Whitehead and Crawford, 2006b). Using standard quantitative genetic methods, gene expression has been shown to be a heritable, often polygenic trait (Fay and Wittkopp, 2008). Distinguishing adaptive changes driven by positive selection from those driven by neutral divergence, mutation and drift, is critical for understanding the evolution of gene expression (Fay and Wittkopp, 2008). Methods originally developed to detect signatures of selection on morphological characters and DNA sequences have now been applied to expression data (Fay and Wittkopp, 2008), but one must distinguish between expression diversity due to genetic differences from that caused by environmental factors (Khaitovich et al., 2004). We have decided to focus on an F_{ST} vs. P_{ST} approach, with support from correlations with environ-

mental variables. A, related, Q_{ST} approach for expression data has been previously taken by Kohn et al. (2008) and Roberge et al. (2007). Here, we will review other methods used to detect selection in gene expression.

To find evidence for selection, we search for departures from the neutral model Gilad et al. (2006a); Fay and Wittkopp (2008). Kimura's neutral model assumes that the level of polymorphism (differences within a population) and divergence (differences between populations) is a simple function of the mutation rate Gilad et al. (2006a). Following Kimura (1983), it has been proposed that under drift, we would expect that the greater the divergence is among taxa, the greater the divergence will be in gene expression level (Whitehead and Crawford, 2006b; Khaitovich et al., 2004). In other words, if the majority of evolutionary changes are caused by historical accidents and not selection, they should accumulate mainly as a function of time (Khaitovich et al., 2004). This is only in the case that changes in gene expression don't affect the fitness of the individual and are therefore only affected by stochastic processes, such as drift (Gilad et al., 2006a).

Under the "nearly neutral theory", a large proportion of mutations will be slightly deleterious (Kimura, 1983; Gilad et al., 2006a). These mutations will contribute to polymorphism within taxa, but at a low frequency, and will rarely reach fixation (Gilad et al., 2006a). The ratio of polymorphism to divergence is expected to be higher than in the neutral theory because the within population variance is higher but the mean between populations will remain similar (Gilad et al., 2006a). With quantitative phenotypes like gene expression level, the evolutionary constraint is likely to take the form of stabilizing selection, which maintains a constant mean and reduces the variance both within and between populations (Gilad et al., 2006a).

Further, if expression is under natural selection, we would expect that the divergence between taxa should increase or decrease depending on the native ecological conditions (Whitehead and Crawford, 2006b). If most mutations in a locus are beneficial (or under

positive selection), they are more likely to reach fixation than under the neutral or nearly neutral theories, and therefore the ratio of polymorphism to divergence should be lower than expected under those models (Gilad et al., 2006a). Also, with a beneficial change, there should be a difference in mean expression level between populations corresponding to the difference between those populations' native ecological conditions (Gilad et al., 2006a).

2.2.1 Tests for selection

Neutral models estimate the rate at which mutation and drift create variation within and divergence between taxa (Fay and Wittkopp, 2008). If there is less divergence than expected, it is evidence of stabilizing selection and greater divergence than expected is evidence of divergent selection (Gilad et al., 2006a). The simplest neutral model is that the variation among taxa should be a positive function of the variation within taxa (Whitehead and Crawford, 2006b). You would then do an F test to test whether the variance among taxa is actually significantly higher than the variance within, and if it is, that is evidence for divergent selection (Whitehead and Crawford, 2006b). The problem with this model is that the function that relates the neutral variances is unknown (Whitehead and Crawford, 2006b). It also varies between genes and comparison groups because there will be larger ratios for genes with fewer constraints, as well as when using more divergent taxa (Whitehead and Crawford, 2006b).

A second approach is to compare the observed variance within and among to the expected variance scaled by time since divergence and the effective population size, as used by Hsieh et al. (2003), Khaitovich et al. (2004) and Rifkin et al. (2003). A modified version of this approach sets upper and lower limits on the range of expected trait divergence among taxa due to drift (Whitehead and Crawford, 2006b). Lemos et al. (2005) used this

modified version, which is based on the neutral model of Lynch and Hill (1986). Gene expression divergence rates outside of the neutral interval were considered to be a signature of stabilizing selection, if they were lower, or directional selection, if they were greater (Gilad et al., 2006a).

A third approach is to examine the asymmetry in gene expression variation along branches of a phylogenetic tree to identify patterns that reject the neutral expectation (Whitehead and Crawford, 2006b). Changes in gene expression can be tested for rate heterogeneity across phylogenetic lineages (Fay and Wittkopp, 2008). Change in rate of expression divergence can be explained by positive selection or by change in a functional constraint (Fay and Wittkopp, 2008).

A fourth approach uses neutral genetic markers to quantify genetic distance and uses genetic distance matrices to correct among taxon trait variation for nonindependence due to phylogeny (see phylogenetic comparative approach – Felsenstein (1985)) (Whitehead and Crawford, 2006b). Residual variation in expression at a locus, after taking phylogenetics into consideration, is then tested for correlation with ecological parameters of hypothesized evolutionary importance (Whitehead and Crawford, 2006b). For more on this, see the section “Comparative method”.

2.3 Drift and stabilizing selection

Many studies find that drift tends to dominate among-taxon variation Oleksiak et al. (2002); Khaitovich et al. (2004); Yanai et al. (2004); Whitehead and Crawford (2006a) while others find the dominance of stabilizing selection (Rifkin et al., 2003; Lemos et al., 2005). This is because most tests assume that the phenotype can evolve without mutational constraints, so the distribution of mutational effects is independent of phenotype (Fay and Wittkopp, 2008). This may be valid over short periods for fold changes, but will be violated if the

absolute effect of a mutation is ever dependent on the current value of the phenotype (Fay and Wittkopp, 2008). We must consider that neutral drift and stabilizing selection may not be entirely exclusive forces on gene expression (Gilad et al., 2006a).

Drift and stabilizing selection interact to diverge or constrain variation and this interaction is more complex as phylogenetic distance increases (Whitehead and Crawford, 2006b). Drift randomly traverses character space over which fitness is unaffected, but the boundaries of this character space are defined by the biological constraints set by stabilizing selection (Whitehead and Crawford, 2006b). Constraints for gene expression can also be set by technical factors (Gilad et al., 2006a). At the low end, expression can't go below 0 and detection on a microarray is only significant above the background level (Gilad et al., 2006a). At the high end, the energetic costs and physical limitations might put a limit on gene expression levels and saturation of RNA binding limits the level of expression that can be detected on a microarray (Gilad et al., 2006a). The unconstrained limit in neutral models is probably not realistic (Gilad et al., 2006a). Boundaries reduce the range of possible differences, and this effect will be greater for more divergent taxa and will be gene-specific (Whitehead and Crawford, 2006b). Neutral evolutionary divergence in gene expression will become nonlinear with greater divergence times due to this constraint, as drift is more likely to have hit the boundaries set by stabilizing selection (Whitehead and Crawford, 2006b). Depending on the gene and its function, some genes will appear primarily affected by drift while others will appear to be affected by stabilizing selection (Whitehead and Crawford, 2006b). It may be more useful to think of a continuum with stabilizing selection predominating for traits that vary less than expected and drift predominating for traits that vary linearly with time, across taxa (Whitehead and Crawford, 2006b).

Empirical evidence for this comes from mutation accumulation lines in both *D. melanogaster* and *Caenorhabditis elegans*. Rifkin et al. (2005) measured the mutational variation for gene expression in mutation accumulation lines of *D. melanogaster* and concluded that

stabilizing selection places severe limits on gene expression divergence. In *C. elegans* mutation accumulation lines maintained for 280 generations, it was found that expression diverged for 9% of the 7014 genes studied but expression difference between natural isolates that had been separated for thousands of generations affected only about one fifth as many genes (Denver et al., 2005). This was evidence that new mutations are not limiting the expression divergence but that stabilizing selection is minimizing differences in wild populations (Fay and Wittkopp, 2008).

There may be some merit in using population comparisons over species comparisons, in order to avoid neutral divergence in expression that has become a nonlinear function of time because of biological and technical constraints (Whitehead and Crawford, 2006b). For shorter phylogenetic distances, drift should drive linear divergence over time and the influences of drift and directional selection may be more readily distinguished (Whitehead and Crawford, 2006b). Also, specifically for microarray studies, sequence divergence in the hybridized probes confounds differences in mRNA concentration when interpreting the differential spot signal intensities (Whitehead and Crawford, 2006b). The more similar the mRNA sequence is likely to be, as with more closely related taxa, the more clearly actual differences will be distinguished.

2.4 F_{ST}, Q_{ST}, and P_{ST}

2.4.1 F_{ST}

Genetic differentiation among populations is affected by mutation, migration, drift and selection (Whitlock, 2011). F_{ST} is a standardized measure of the degree of among population genetic differentiation and can be estimated as:

$$F_{ST} = \frac{V_b}{(V_b + V_w)}$$

where V_b is the between population variance and V_w is the within population variance, together adding to the total genetic variation in neutral markers (Merila and Crnokrak, 2001). F_{ST} is the expected degree of population differentiation as the result of drift and gene flow (Merila and Crnokrak, 2001) and has the same expectation for all neutral alleles with low mutation rates (Whitlock, 2011).

F_{ST} is used with the allele frequency of a locus to predict the distribution of allele frequencies across populations and therefore understand evolution in structured populations (Whitlock, 2011). It can be interpreted as the proportional loss in heterozygosity at a locus caused by spatial population structure, compared to what is expected for a panmictic population with the same allele frequency (Whitlock, 2011). F_{ST} can also be a description of the relative time to the most recent common ancestor for the alleles chosen within and between populations (Whitlock, 2011; Slatkin, 1995). This is a common description of the average evolutionary history of all neutral loci, and is referred to as the coalescent F_{ST} (Whitlock, 2011). We expect the coalescent F_{ST} to be roughly similar for all loci and it can be inferred from data if the mutation process of marker alleles leaves a traceable history (Whitlock, 2011). Coalescent F_{ST} increases monotonically with increasing isolation of the populations and gives a good measure of the evolutionary uniqueness of separate populations (Whitlock, 2011). If the genetic variation increases proportionally with the time of divergence of alleles, the coalescent F_{ST} allows the partitioning of the proportion of genetic variance that is between populations from that which is within (Whitlock, 2011).

Mutation and selection vary widely from locus to locus, while migration and drift are roughly equal at all autosomal loci (Whitlock, 2011). Loci only strongly affected by migration and drift are roughly similar in F_{ST} while loci with high mutation rates or those experiencing high selection may have a different F_{ST} than the rest of the genome (Whitlock, 2011; Merila and Crnokrak, 2001). Repeatability across loci for F_{ST} makes it possible to establish a neutral baseline from which to infer selection at some loci (Whitlock, 2011).

More reliable inference may be possible with markers with lower mutation rates, like single nucleotide polymorphisms (SNPs) (Ritland, 2000).

2.4.2 Q_{ST}

Local adaptations stem from spatial and temporal heterogeneity in selection pressures acting on heritable traits, which are thought to underlie most phenotypic diversity in the wild (Merila and Crnokrak, 2001). Testing for selection requires the partitioning of the observed variation in a quantitative trait into its genetic and non-genetic components (Gilad et al., 2006a). Minimizing the differences in environment between samples reduces the environmental variance (Gilad et al., 2006a) and it is generally agreed that much of the variation in gene expression for a particular environmental condition is heritable (Stamatoyannopoulos, 2004; Gibson and Weir, 2005). The quantitative measure of the genetic basis for phenotypic variation is h^2, the narrow sense heritability, which is the additive genetic variation in a trait divided by the phenotypic variation (Whitehead and Crawford, 2006b). Significant heritable variation in gene expression is common in yeast, mice and humans, where h^2 has been found to be over 30% (Whitehead and Crawford, 2006b). Heritability of gene expression has also been investigated in the terpenoid pathways of Interior spruce (*Picea glauca x engelmannii*) (Albouyeh and Ritland, 2011). In any given pathway segment, the median heritability was always found to be above 40% (Albouyeh and Ritland, 2011). Also, much of the genetic variation in gene expression is due to many loci (Whitehead and Crawford, 2006b). These data, along with measures of natural variation, suggest that polymorphism in mRNA expression should provide ample material for evolution (Whitehead and Crawford, 2006b).

When species are spread over a heterogeneous landscape, individuals in different parts experience different environments and different selective pressures (Whitlock, 2008). Habi-

tats capable of sustaining a population of a particular species may also be spatially separated, and therefore species are subdivided over space (Whitlock, 2008). Local adaptation is enhanced by selective differences between populations, which creates genetic differences, and is opposed by migration, which lowers genetic differences (Whitlock, 2008). This is complicated by the fact that genetic differentiation among populations can also occur due to neutral drift alone (Whitlock, 2008).

Q_{ST} vs. F_{ST}

Q_{ST} vs. F_{ST} comparisons provide insights into the relative importance of drift and selection as causes of population differentiation in quantitative traits (Merila and Crnokrak, 2001). Q_{ST} is a metric of the degree of genetic differentiation among populations displayed by quantitative traits, which was proposed by Spitze (1993) as a parallel measure for F_{ST}. Q_{ST} for diploids can be calculated as:

$$Q_{ST} = \frac{\sigma^2_{GB}}{(\sigma^2_{GB}+2\sigma^2_{GW})}$$

where σ^2_{GB} is the additive genetic variance in a trait between populations and σ^2_{GW} is the additive genetic variance in a trait within populations (Merila and Crnokrak, 2001; Whitlock, 2008). Usually, Q_{ST} is compared to F_{ST} calculated from neutral loci (Whitlock, 2008). For a trait with an additive genetic basis and in linkage equilibrium, in a diploid organism, Q_{ST} is expected to be the same as F_{ST} if estimated from the allele frequency at the quantitative trait loci (Merila and Crnokrak, 2001). If the trait is neutral and differentiation is due to genetic drift, the global Q_{ST} should be the same as the global F_{ST} of neutral loci, if the trait is controlled by purely additive genes that have no dominance or epistasis (Whitlock, 2008; Merila and Crnokrak, 2001). In principle, when Q_{ST} of a trait is compared to the F_{ST} of neutral loci from the same set of populations, if Q_{ST} is greater than F_{ST}, the trait has diversified more than expected by drift, and this is evidence for divergent selection (Whitlock, 2008; Merila and Crnokrak, 2001). If Q_{ST} is lower than F_{ST}, there is evidence for stabi-

lizing selection acting on the trait (Whitlock, 2008). This means that natural selection is favouring the same mean phenotype in different populations (Merila and Crnokrak, 2001). If Q_{ST} is approximately equal to F_{ST}, as is the expectation if the trait is neutral, there is little evidence for selection (Whitlock, 2008). This does not prove that the differentiation was caused by drift, but that the effects of drift and selection are indistinguishable (Merila and Crnokrak, 2001). This is assuming that the chosen genetic markers are actually behaving neutrally (Merila and Crnokrak, 2001).

Q_{ST} is typically used to address two types of questions (Whitlock, 2008). The first is whether particular traits are under spatially divergent or uniform selection or whether a particular trait has undergone local adaptation (Whitlock, 2008). The second is whether a series of populations relates to their environment in a way to produce local adaptation in general, or whether traits on average adapt to local conditions (Whitlock, 2008). These two kinds of questions require distinct statistical methods (Whitlock, 2008). The second type compares the mean Q_{ST} over all traits to the global F_{ST} (Whitlock, 2008). The mean Q_{ST} may be a measure of the overall importance of local adaptation in the species, but it may be biased to average values over traits due to a priori choice of traits or non independence of traits (Whitlock, 2008). That is why it is better to employ univariate Q_{ST} and evaluate the degree of differentiation against the null on a trait-by-trait basis (Merila and Crnokrak, 2001). The first type of question uses the Q_{ST} of a single trait, and asks whether it is greater or less than expected for a trait evolving neutrally (Whitlock, 2008). Here, one would compare the Q_{ST} to a distribution of F_{ST} values of putatively neutral markers (Whitlock, 2008).

The challenge is that the F_{ST} for neutral loci and Q_{ST} for neutral traits are expected to be extremely variable, even for a given mean (Whitlock, 2008). Any given locus or trait can be very different from the expectation. Estimates of F_{ST} are heterogeneous among loci because of direct selection, indirect effects of selection such as linkage to loci under strong

selection, sampling error and drift (Whitlock, 2008). One can get more robust estimates of the expected variance if the number of populations and number of loci used is increased, which increases the precision of the estimate of global F_{ST} (Whitlock, 2008). As the number of local populations increases, there is a larger sample of the possible range of evolutionary processes and estimated F_{ST} values are less heterogeneous (Whitlock, 2008). Q_{ST} is also difficult to measure precisely, but it is better if there are more demes, more families per deme and when the rearing conditions are controlled (Whitlock, 2008). For evidence of selection, it should be shown that the Q_{ST} value is in the tail of the predicted distribution of F_{ST} (Whitlock, 2008). The error in estimating Q_{ST} is usually relatively large, however, so using a method such as bootstrapping to get the tail probability is recommended (Whitlock, 2008).

For most studies, Q_{ST} is usually greater than F_{ST}, which suggests a prominent role for natural selection in different populations of the same species (Merila and Crnokrak, 2001). Differences between Q_{ST} and F_{ST} estimates are largely restricted to morphological traits while life history traits have a similar degree of differentiation as DNA markers (Merila and Crnokrak, 2001). Few Q_{ST} values are smaller than F_{ST}, suggesting that selection in different populations is unlikely to be similar (Merila and Crnokrak, 2001). However, previous literature may have a bias in favour of populations and traits known to be phenotypically divergent, so the conclusion about the ubiquity of natural selection could be premature (Merila and Crnokrak, 2001; Whitlock, 2008). All of these interpretations are subject to the assumptions of Q_{ST} and methods used to derive it (Merila and Crnokrak, 2001).

Assumptions and conditions of Q_{ST}

Q_{ST}'s critical assumption is that the estimates of variance represent purely additive genetic effects, free of maternal, environmental and non-additive genetic effects (Merila and Crnokrak, 2001). If ignored, the conclusions can be misleading. To measure the additive

genetic variance within populations, a breeding design is required that allows the phenotype to be correlated with relatedness, as well as a common environment (Whitlock, 2008). To measure the genetic variance among populations, one must include only genetic differences, which can be obtained by using a common garden (Whitlock, 2008). Uncontrolled maternal or common environment effects can lead to smaller estimates of Q_{ST} while unaccounted for cross-generational maternal and environmental effects that are population specific can inflate the estimate of Q_{ST} (Merila and Crnokrak, 2001). The variance components derived in a common garden are often assumed to be unaffected by the rearing environment, but this may not be the case (Merila and Crnokrak, 2001). The complete removal of geographic differences due to persistent environmental and maternal effects may require several generations (Merila and Crnokrak, 2001). Both within and among population variance is subject to biases from genotype by environment effects and a novel environment may influence the expression of genetic variance unpredictably, especially if there is plasticity for the trait (Merila and Crnokrak, 2001; Whitlock, 2008).

Variation in gene expression shows evidence of dominance and non-additive (epistatic) interactions among loci, but this is true of many other quantitative traits (Fay and Wittkopp, 2008). Whitlock (1999) showed that additive by additive epistasis will lead to smaller estimates of Q_{ST} and dominance will lead to smaller or equal Q_{ST} estimates for neutral traits, if it follows the island model (Whitlock, 2008). Q_{ST} can be greater than F_{ST} under pure drift with no migration, but it is unlikely if there are multiple loci involved (Whitlock, 2008). Sex chromosomes or cytoplasmic factors can also make Q_{ST} larger than F_{ST} if they underlie the trait (Whitlock, 2008). In general, overestimating the within population variance underestimates Q_{ST}, making the estimation conservative if looking for evidence of divergent selection, while overestimating the among population variance has the opposite effect (Merila and Crnokrak, 2001; Whitlock, 1999). The inflation of Q_{ST} will also vary between traits (Merila and Crnokrak, 2001).

Q_{ST} has the same problems as other techniques to estimate natural selection in the wild (Whitlock, 2008). For example, if a trait is correlated with a trait under selection, it will look like it is under selection (Whitlock, 2008). Q_{ST} can be used in an exploratory manner and if it is used only to generate ideas and not test a priori hypotheses, the difficult statistical properties of Q_{ST} are less important (Whitlock, 2008). Comparative and exploratory methods do not absolutely require much information about F_{ST} as candidate traits can be compared to other traits (Whitlock, 2008). Q_{ST} is a crude measure of the amount of genetic differentiation of a trait caused by local adaptation. The comparison of Q_{ST} and F_{ST} allows us to examine the null hypothesis of neutral divergence among populations (Whitlock, 2008). Other techniques, like the correlation of a trait with an environmental measure can be used to learn the pattern caused by selection and the nature of selection (Whitlock, 2008) (see the section "Comparative method").

2.4.3 P_{ST}

Calculation of Q_{ST} requires unbiased estimates of the additive genetic variance within populations and the genetic variance among populations (Whitlock, 2008). Sometimes the total phenotypic variance in a trait across populations is used instead, and this measure is called P_{ST} (Brommer, 2011; Saether et al., 2007). P_{ST} as a term was introduced by Leinonen et al. (2006). The critical aspect for how well P_{ST} approximates Q_{ST} depends on the extent that additive genetic effects determine the variation between populations relative to that within populations (Brommer, 2011). The quantification of P_{ST} is usually based on phenotypic measures of a trait in the wild in several individuals across a number of populations (Brommer, 2011). A way of denoting the scaling of phenotypic to additive genetic variances is to say:

$$P_{ST} = \frac{c\sigma_B^2}{(c\sigma_B^2 + 2h^2\sigma_W^2)}$$

where σ^2_B is the phenotypic variance component between populations, σ^2_W is the phenotypic variance component within populations and h^2 is the heritability (Brommer, 2011; Saether et al., 2007). Non-additive genetic variances or environmental factors and genotype-environment interactions may give a distorted picture of the additive genetic variance when only phenotypic variances are investigated (Pujol et al., 2008). We use the two parameters c and h^2 to determine the accuracy of the approximation of Q_{ST} by P_{ST} (Brommer, 2011). There are no set values for these parameters so it is best to consider the sensitivity of your conclusions to a variety of values of c and h^2 (Brommer, 2011). We can rewrite the above equation as:

$$P_{ST} = \frac{\frac{c}{h^2}\sigma^2_B}{(\frac{c}{h^2}\sigma^2_B + 2\sigma^2_W)}$$

where the unknown ratio c/h^2 is the critical aspect that describes how well P_{ST} approximates Q_{ST} (Brommer, 2011). To evaluate robustness, c/h^2 is varied for calculations of P_{ST} and its 95% confidence interval, and each is compared to the neutral expectation (Brommer, 2011). When testing a conclusion of divergent selection, the parameter space where $c < h^2$ will be the most important, and when testing for stabilizing selection, the parameter space where $c > h^2$ will be most important (Brommer, 2011). This is because P_{ST} is an increasing function of c/h^2. So, as long as a trait is heritable, a precise estimate of h^2 is not as important for comparing against the neutral expectation as is finding your conclusion robust to deviations in c/h^2 (Brommer, 2011).

Previous P_{ST} studies have qualified their assumptions about the likely magnitude of P_{ST}, providing sensitivity analyses and verbal arguments to suggest their conclusions are robust to deviations from the assumed values (Merila and Crnokrak, 2001), but they have failed to consider sensitivity in a systematic fashion (Brommer, 2011). They have also ignored that c/h^2 is the critical aspect and have failed to consider how c and h^2 will affect the confidence interval for P_{ST}, focusing only on point estimates (Brommer, 2011). Taking the confidence interval into account will allow more exact and conservative interpretations of

P_{ST} (Brommer, 2011). In general, P_{ST} is error-prone, has biases, and may be more suitable as an exploratory technique on the operation of selection when Q_{ST} studies are not possible (Whitlock, 2008). P_{ST} should always be interpreted very conservatively (Brommer, 2011).

2.5 Comparative method

Environmental conditions can be the basis of natural selection on organisms, and many environmental variables correlate with latitude across the Pacific Northwest (Xie et al., 2009). Adapting to environmental conditions is especially important in a sessile, long-lived organism such as *Populus trichocarpa* that cannot move into a more favourable environment. For this reason, correlation of a phenotypic character with an environmental variable can be an important signature of selection (Fay and Wittkopp, 2008). Difference in environmental conditions and genetic distance often increase with physical distance, both of which can affect the phenotypic character in question. For this reason, the neutral genetic distance must be taken in to account when looking for correlations between phenotype and environment (Felsenstein, 1985; Whitehead and Crawford, 2006a,b).

This topic was briefly touched on in the section "Tests for selection". We will refer to this approach as the "comparative method". The comparative method examines the evolution of a trait in relation to the evolution of other traits or environmental variation in a phylogenetic context (Fay and Wittkopp, 2008). It uses genetic markers to quantify genetic distance and uses genetic distance matrices to correct for the among taxon trait variation due to phylogeny (Felsenstein, 1985; Whitehead and Crawford, 2006b). Closely related populations tend to share more similar environments, so clinal variation in expression could be due to drift where genetically similar populations have similar patterns of expression, or it could be adaptive divergence (Whitehead and Crawford, 2006b). If the evolutionary history is known, it can be taken into account (Felsenstein, 1985), making the

gene expression independent of genetic relatedness and then using this to examine correlation with the environmental gradient (Whitehead and Crawford, 2006b). The maximum trait variation among taxa is allocated to genetic distance and residual variation is then tested for correlation with ecological parameters of hypothesized evolutionary importance (Whitehead and Crawford, 2006b). When phylogeny accounts for much of the variance in expression among populations, it is a sign of neutral divergence. Phylogenetically independent variance regressing significantly with an environmental variable is suggestive of adaptive differences (Whitehead and Crawford, 2006a). This approach is gene-specific so the covariation between gene expression and the genetic distance is determined for each locus separately (Whitehead and Crawford, 2006b). We expect that different loci will have different relationships with genetic distance because of differing constraints on gene expression levels (Whitehead and Crawford, 2006b).

This approach has been used in the Whitehead and Crawford (2006a) study, as mentioned in the section "Examples of positive selection on gene expression", where the expression values under directional selection were detected by finding the remaining variation associated with an ecological factor (temperature) after correction for genetic relatedness. Much of the variation was accounted for by phylogeny, but directional selection seemed to be acting on 13/329 genes (Whitehead and Crawford, 2006a). Other Q_{ST} studies have also taken the approach of correlating with environment somehow. For example, Saether et al. (2007) correlated both tail white and tarsus length with region in great snipe, while adjusting for neutral genetic divergence as represented by F_{ST}.

We must keep in mind, however, that observed changes in gene expression correlating with environmental variables does not prove that this is the molecular change responsible for adaptive divergence (Fay and Wittkopp, 2008). We expect that many changes in gene expression will correlate with each other or to other phenotypes (Fay and Wittkopp, 2008). Mutation accumulation studies indicate that groups of functionally related genes often ac-

quire regulatory changes together (Denver et al., 2005). Many changes in expression are not independent and observing a large number of genes correlated with a variable may be just as much evidence as a small number of genes (Fay and Wittkopp, 2008). It is difficult to separate cause and effect, which confounds evolutionary interpretation (Fay and Wittkopp, 2008). Correlations with environmental variables can also arise as the result of genetic, developmental or environmental constraint, unrelated to natural selection (Fay and Wittkopp, 2008).

Chapter 3

The prevalence of divergent selection on gene expression differences among populations of black cottonwood

3.1 Introduction

In 1975, King and Wilson (King and Wilson, 1975) found that levels of phenotypic variation visible in nature could not be explained by variation in protein coding sequences alone. They proposed that gene expression variation can play a role in the evolution of phenotypic variation in nature. Gene expression levels have been found to be heritable (Stamatoyannopoulos, 2004; Gibson and Weir, 2005) and since gene expression can affect macroscopic phenotype, and therefore fitness, it can underlie evolutionary change. Many examples of changes in gene expression have been found to underlie morphological differences between species (Fay and Wittkopp, 2008). Examples include changes in the pelvic structure of threespine sticklebacks mediated by *Pitx1*, trichome patterns in *Drosophila* by Ubx, butterfly eyespots by Distal-less and beak size among Galapagos finches by BMP4 (Fay and Wittkopp, 2008).

At the population level, gene expression studies can be informative about processes of gene expression evolution. When a mutation which causes a change in gene expression is introduced in to a population, either selection or neutral drift can act on that mutation.

When this change in gene expression causes no change in fitness of the organism, it will be affected by neutral drift. The bounds for neutral drift are thought to be set by stabilizing selection (Rifkin et al., 2005; Denver et al., 2005). Stabilizing selection prevents a deleterious mutation in gene expression from drifting to a high frequency within a population. Finally, if the change is beneficial to the organism in the environment that the population is located in, it will be affected by positive selection, allowing the mutation to increase in frequency in that population. When different populations are located in different environmental conditions, different changes will be beneficial to different populations, which can result in divergent selection. This is where the differences between populations are higher than would be expected through random drift because of the divergent selection pressures between environments. We looked for these patterns of gene expression evolution among populations of *Populus trichocarpa*, black cottonwood.

To detect patterns of evolution of gene expression, we took a P_{ST} vs. F_{ST} approach. F_{ST} is a standardized measure of the degree of between population differentiation in alleles (Whitlock, 2011) and Q_{ST} is an analogous measure for the genetic differentiation in a quantitative trait (Spitze, 1993). P_{ST} is an approximation of Q_{ST} that uses the phenotypic differentiation instead of the genetic differentiation in a trait (Brommer, 2011). It is used when Q_{ST} estimation is not possible, such as when genetic crosses were not performed. The results from a P_{ST} vs. F_{ST} comparison can be interpreted in the same way as the Q_{ST} vs. F_{ST} comparison, but should be done so conservatively (Brommer, 2011). In principle, if both are measured from neutral alleles and traits, F_{ST} and Q_{ST} should be equal (Merila and Crnokrak, 2001). This is the expectation based on neutral drift and divergence from this is evidence for selection. If Q_{ST} is larger than F_{ST}, it is suggestive of divergent selection acting on the trait, and if Q_{ST} is smaller than F_{ST}, it is suggestive of stabilizing selection acting on the trait (Whitlock, 2008; Merila and Crnokrak, 2001). A Q_{ST} approach for expression data has been previously taken by Kohn et al. (2008) and Roberge et al. (2007).

To investigate the divergent selection of gene expression in black cottonwood, we used the comparative method. When a trait evolves in the same direction repeatedly in different populations, it is a good indicator that it is an adaptive change, but similar patterns of change can occur in the absence of selection if there is genetic similarity between the populations due to shared ancestry (Felsenstein, 1985; Fay and Wittkopp, 2008). The comparative method looks for correlations between a phenotype of interest and some other condition, while controlling for genetic distance between populations. Whitehead and Crawford (2006a) used this method to find expressed genes likely adapted to water temperature differences in populations of *Fundulus*. When genetic distance is included, repeated evolution of a trait in the same direction that is correlated with the levels of an environmental variable is evidence for natural selection and adaptive forces.

Black cottonwood is a model tree system that allows the study of an extended suite of tree biological processes, not available in *Arabidopsis thaliana*, the existing plant model species (Jansson and Douglas, 2007). For example, trees in temperate climates need to be able to deal with seasonal changes and withstand winter conditions for many years running, while annual plants do not have these same pressures. Also, black cottonwood is dioecious (two distinct sexes), which is relatively rare (DeBell, 1990). It is also an important commercial plantation species, so insights into its biology can have commercial applications. A suite of genomic tools have been developed for black cottonwood. These include a genome sequence and a 15.5K microarray developed by the Treenomix group (Ralph et al., 2006). *Populus* is also at an advantage by being closely related taxonomically to *Arabidopsis*, more closely than some dicots, like the asterids, and much more closely than monocots like rice or gymnosperm trees like conifers (Jansson and Douglas, 2007). This is an advantage because the genes are much more likely to be conserved and, therefore, *Arabidopsis* annotations are more easily applied to *Populus* genes (Jansson and Douglas, 2007).

Black cottonwood is broadly distributed in western North America. Xie et al. (2009) found that it shows an ecotypic mode of genetic differentiation with the populations being divided into northern and southern groups. These groups are proposed to be separated by a "no-cottonwood" belt (Xie et al., 2009). Trees from the north suffered higher mortality, grew more slowly and were more susceptible to pathogens than those from the south when grown in the south (Xie et al., 2009), some of which has been noted before (DeBell, 1990). This indicates there may be adaptive expression differences between the northern and southern groups.

In this study, we examined gene expression from 12 black cottonwood populations, most from British Columbia, six from the north and six from the south (Fig. 3.1). The gene expression values seeming to be under divergent selection were then correlated with environmental variables, using the comparative method. A few candidate genes' expression values were also investigated. Previous literature has indicated that Q_{ST} is usually greater than F_{ST}, which has led to a conclusion about the ubiquity of natural selection, but there may be a bias in the literature in favour of populations known to be phenotypically divergent for the traits examined (Merila and Crnokrak, 2001). Many previous studies of gene expression, however, have found that drift or stabilizing selection dominates (Oleksiak et al., 2002; Khaitovich et al., 2004; Yanai et al., 2004; Whitehead and Crawford, 2006a; Rifkin et al., 2003; Lemos et al., 2005). By examining the expression of over 15,000 genes, we may be able to shed some light on this issue, as well as investigate how much gene expression levels seem to be affected by natural selection.

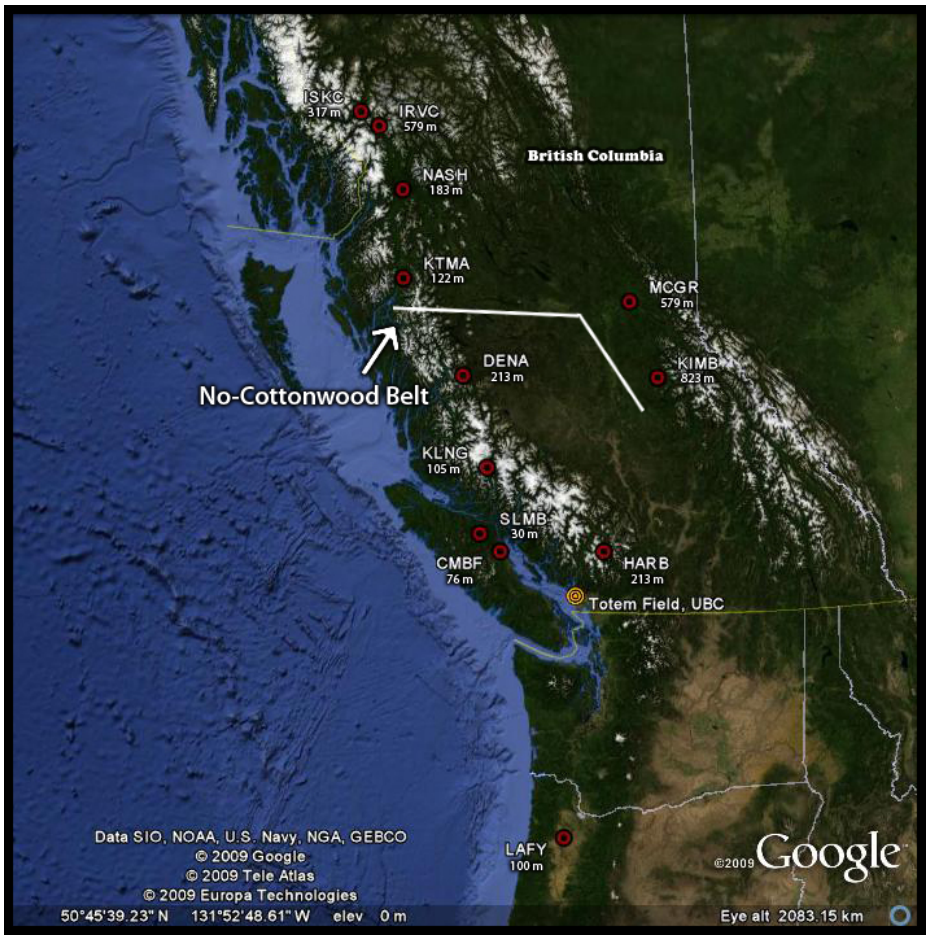


Figure 3.1: Locations of the 12 populations chosen for study. The numbers below the population names represent the elevation at which trees were sampled. The "no-cottonwood" belt separates northern and southern groups. Common field location located at Totem Field at the University of British Columbia.

3.2 Materials and methods

3.2.1 Plant material

In 1995, the BC Ministry of Forests completed a large-scale collection of scions from *Populus trichocarpa* clones (Xie et al., 2009). 854 clones from 188 provenances were collected along 36 river drainages (Xie et al., 2009). Scions were collected from trees with vigorous leader growth that were 10-25 years of age (Xie et al., 2009). Attempts were made to sample trees scattered within a stand but no minimal distance between trees was used (Xie et al., 2009). 835 trees from 180 provenances were propagated at the Ministry of Forests and Range Nursery in Surrey, British Columbia (BC) (Xie et al., 2009). Cuttings from these trees were taken and grown in a common field environment on Totem Field at the University of British Columbia in Vancouver, BC.

Up to one month after bud flush, gene expression in leaves is mainly determined by a developmental program while later in the summer, environmental factors are more important (Sjodin et al., 2008). Based on this, we collected early in the summer. On May 21 of 2009, leaves were collected from each tree growing in Totem Field. Tissue –specific expression patterns are a source of expression variation that can artificially inflate differences among individuals (Whitehead and Crawford, 2006b), so we tried to standardize our collection. We collected young, sink (net carbon importer) leaves from the leader of each tree. We used the leaf plastochron index (LPI) developed for *Populus* by Larson and Isebrands (1971) to standardize the collection of leaves. The youngest leaf with a length of 2 cm is designated LPI 0 (Larson and Isebrands, 1971), and the transition from sink to source status occurs between LPI 5 and LPI 7 (Philippe and Bohlmann, 2007). We collected four leaves of LPI 1 to LPI 4 and pooled them in a common tube. Leaves were flash frozen in liquid nitrogen and stored at -80°C prior to RNA isolation.

Table 3.1: Population locations for *Populus trichocarpa* used in this study.

Population	Location	Latitude (°N)	Longitude (°W)	Elevation (m)
CMBF	Campbell River, BC	49°57’	125°15’	76
DENA	Dean River, BC	52°46’	126°37’	213
HARB	Lillooet River, BC	50°02’	122°32’	213
IRVC	Bell Irving River, BC	56°44’	129°44’	579
ISKC	Iskut River, BC	56°56’	130°20’	317
KIMB	Kimball Creek, BC	52°56’	121°10’	823
KLNG	W. Klinaklini River, BC	51°18’	125°46’	105
KTMA	Kitimat River, BC	54°15’	128°31’	122
LAFY	Lafayette, OR	45°12’	123°05’	100
MCGR	McGregor River, BC	54°11’	122°00’	579
NASH	Nass River, BC	55°43’	128°49’	183
SLMB	Salmon River (Vancouver Island), BC	50°13’	125°49’	30

From the 180 provenances, we chose 12 populations from which we collected 3 individuals per population among which none of the leaves were visibly damaged. They were also chosen based on distance between the populations (we tried not to choose ones from the same drainages). These represent populations from both northern and southern groups (Fig. 3.1, Table 3.1).

3.2.2 RNA isolation, experimental design and microarray hybridization

The total RNA was isolated according to the protocol of Kolosova et al. (2004). RNA quantity and quality was assessed by measuring spectral absorbance between 200 and 350 nm and by visual assessment on a 1% agarose gel.

Reference and balanced are the two basic experimental designs for microarray experiments (Whitehead and Crawford, 2006b). In the reference design, all samples are labelled with one dye and cohybridized with a common reference sample labelled with a second dye (Whitehead and Crawford, 2006b). In balanced designs, like loops, experimental sam-

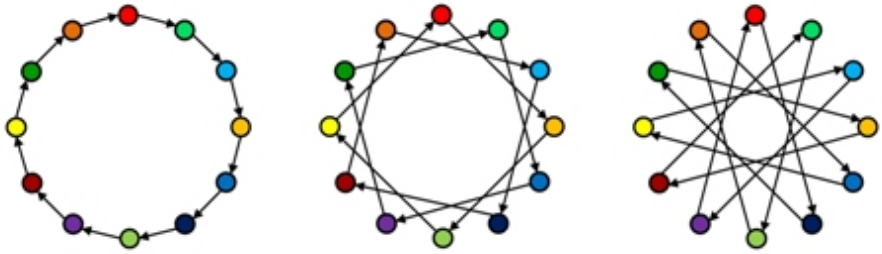


Figure 3.2: Design of the microarray experiment. Each population is represented by a differently coloured circle. Three loops were done, each using a different biological replicate from each population. The arrows are drawn between individuals that were hybridized, where the base of the arrow represents Cy3 and the arrow head represents Cy5.

ples are labelled with both dyes and hybridized to each other (Whitehead and Crawford, 2006b). For the same number of slides, twice the number of experimental samples can be included in balanced designs, which leads to improved precision and increased statistical power (Kerr and Churchill, 2001). The error due to technical variability is highest for reference designs when using the same number of arrays (Kerr, 2003). In light of this, we used a balanced loop design with dye swap. A pictorial representation of the design can be found in Fig. 3.2. Briefly, the microarrays were performed in three separate loops with a different individual from each population present in each loop. Each individual from each population was hybridized once with each dye. In total, for 12 populations of 3 individuals each, 36 hybridizations were performed.

The 15.5K poplar microarray, a cDNA microarray that contains 15 496 non-redundant ESTs and was developed by Ralph et al. (2006), was used for gene expression analysis. It contains elements from 14 cDNA libraries representing leaves, buds, phloem, xylem, bark and root tissues, as well as cultured cells (Ralph et al., 2006). The microarray is enriched for EST sequences from elicitor- or herbivore-treated libraries (Ralph et al., 2006). It was first applied in an initial study of the transcriptional response of poplar leaves to feeding by forest tent caterpillar larvae (Ralph et al., 2006). They also performed validation of the

microarray performance by doing self-self hybridizations and found a very low level of nonspecific hybridization and no genes reliably differentially expressed (lowest FDR was 48.9%) (Ralph et al., 2006). They also found that the results from real-time PCR were generally in agreement with the microarray results (Ralph et al., 2006).

Hybridizations were performed using the Genisphere Array350 kit (Genisphere). The microarray hybridization and conditions are described in detail by Ralph et al. (2006), with some modifications. Briefly, 40 ug of total RNA was reverse transcribed using M-MuLV reverse transcriptase (New England BioLabs) and oligo d(T) primers with a 5' unique sequence overhang specific to either Cy3 or Cy5 labeling reactions. After 1 hour of synthesis, the RNA strand of the cDNA:RNA hybrid was hydrolyzed in 0.5 M NaOH/0.05 M EDTA at 80°C for 10 min followed by neutralization in 1 M Tris-Cl (pH 7.5). After pooling of cDNAs, samples were precipitated with linear acrylamide and resuspended in a 45 µL hybridization solution of nuclease-free water, 2x SDS buffer, 4.0 µL LNA d(T) blocker, 0.3 µL Cy5-labeled GFP cDNA (Cy5-dUTP and Ready-To-Go labelling beads, Amersham Pharmacia Biotech) and 0.25 µL salmon sperm DNA. The slides were incubated at 60°C. The second hybridization included the Cy3 and Cy5 3DNA capture reagants (Genisphere) in a 45 µL volume consisting of 2x SDS buffer, nuclease-free water, 2.5 µL Cy3 capture reagent and 2.5 µL Cy5 capture reagent. The second hybridization was also incubated at 60°C.

3.2.3 Microarray analysis

All slides were scanned and images of hybridized arrays were acquired by using ScanArray Express (Perkin Elmer). The Cy3 fluor was excited at 543 nm and the Cy5 fluor at 633 nm. All scans were performed at the same laser power (90%) but the photomultipier tube (PMT) gain % was adjusted for each channel attempting to get the percentage of saturated array

elements to be about 1%, while minimizing background fluorescence. Fluorescent intensity data was quantified using the ImaGene 6.0.1 software (Biodiscovery). All spots were manually checked and, to correct for background intensities, auto segmentation was used. Auto segmentation uses an algorithm to define the foreground and background boundaries of a spot and uses this information to take the background buffer value into consideration when correcting for background intensity. Data were normalized to compensate for non-linearity of intensity distributions using variance stabilizing normalization (vsn) method (Huber et al., 2002). To get relative population estimates for gene expression, a mixed-effects model was fitted to the normalized intensities in the Cy3 and Cy5 channels of the 36 microarray slides. The model contained a fixed population effect, array effect and dye effect as well as random effect for biological replicate, each used twice in the data. Estimates for each population were obtained, as well as population variances and covariances between pairs of populations. All the above statistical analyses of gene expression data were carried out using the R statistical package (R Development Core Team, 2011). Functional annotation of the array elements was assigned according to the TAIR9 *Arabidopsis* protein set (Swarbreck et al., 2008) using BLASTX (Altschul et al., 1990). Only BLAST hits with expect values (E) < 10^{-10} were kept. Only the best BLASTX hit was kept for each cDNA on the microarray, as chosen based on E value. Another BLASTX against the NCBI non-redundant protein sequences (nr) was performed for the genes hypothesized to be under selection.

3.2.4 Individual F_{ST} calculations

Any given locus or trait can be very different from the expectation (Whitlock, 2008). Estimates of F_{ST} are heterogeneous among loci because of direct selection, indirect effects of selection such as linkage to loci under strong selection, sampling error and drift (Whit-

lock, 2008). More robust estimates of expected variance can be obtained by increasing the number of loci used, which increases the precision of the estimate of global F_{ST} (Whitlock, 2008). Rather than estimate F_{ST} based upon population gene frequencies, with the power of genomics we can estimate F_{ST} for individuals.

1910 single nucleotide polymorphisms (SNPs) were obtained for each of the trees, except for two individuals from one population. SNP genotypes were obtained based on date from a *P. trichocarpa* SNP database (Geraldes et al., 2011), together with other SNPs, to generate an Illumina SNP bead array (Carl Douglas, Gerald Tuskan et al., unpublished). Trees from the BC Ministry of Forests collection were genotyped for ~32,000 SNPs, among which 1910 were used in my study (data obtained from Carl Douglas, Armando Geraldes and Quentin Cronk). These SNPs all had GeneTrain scores (GenomeStudio, Illumina) over 0.85 and no missing data. We estimated individual F_{ST} divergence using the formula for individual inbreeding coefficients of Ritland (1996). For a diallelic locus, as all SNPs are, it is calculated as:

$f = \frac{S_i - p^2}{p(1-p)}$

where S_i is either 0, when the two alleles are different, or 1, when the two alleles are the same and p is the population allele frequency (Ritland, 1996). This value is then averaged over all loci for each individual, to give the individual F_{ST}. The individual F_{ST} values for each population are then averaged to obtain the population F_{ST} and the population F_{ST} values are averaged to obtain the global F_{ST} for the species.

3.2.5 Pairwise F_{ST} calculations

As our measure of pairwise F_{ST}, we used the relatedness between individuals, or the coefficient of kinship (r) (Ritland, 1996). This is the probability that two alleles, one sampled from each individual in the pair, are identical by descent (Ritland, 1996). Since we have

three individuals per population, there are 9 possible pairings for pairwise population measures. For each pair, the relatedness is measured as:

$r = \frac{S_i - p^2}{p(1-p)}$

where S_i is the number of alleles in common between the pair of individuals divided by 4 (the total number of pairs of alleles at a single locus) and p is the population allele frequency (Ritland, 1996). This is then averaged over all loci for each pair of individuals and averaged over each pair of individuals from a given pair of populations. This gives the population pairwise F_{ST} value. Some F_{ST} values were found to be below zero, but these were not adjusted to zero as this would create bias when the positive sampling variation is not similarly adjusted.

Isolation by distance was tested using a Mantel test (Mantel, 1967) between pairwise values of $F_{ST}/(1-F_{ST})$ and the natural logarithm of geographic distances (km) between all population pairs (Rousset, 1997) using 10000 permutations. A linear regression was also performed on these same values. Geographic distances were calculated from geographic coordinates using Passage 2 (Rosenberg and Anderson, 2011). The Mantel test was performed using the R statistical package (R Development Core Team, 2011). A neighbour-joining tree was built using the pairwise F_{ST} values as a distance matrix, using the program MEGA v5.05 (Tamura et al., 2011). For the purposes of this tree, the negative F_{ST} values were changed to 0 as negative distance values were not allowed.

3.2.6 P_{ST} calculations

The calculation of P_{ST} may be biased when averaging values over traits (Merila and Crnokrak, 2001), so we decided to use each gene's expression as a separate trait to calculate P_{ST} and evaluate against the null hypothesis. This gave us over 15,000 traits to evaluate. It is important that F_{ST} and Q_{ST} measurements are taken on the same collection of populations, so

we used the same individuals as those used for F_{ST} estimation (Whitlock, 2008), with an additional 2 individuals from one of the populations. P_{ST} was estimated as:

$$P_{ST} = \frac{c\sigma_B^2}{(c\sigma_B^2 + 2h^2\sigma_W^2)}$$

where σ^2_B is the phenotypic variance component between populations, σ^2_W is the phenotypic variance component within populations and h^2 is the heritability (the proportion of phenotypic variance that is because of additive genetic effects) (Brommer, 2011; Saether et al., 2007). The scalar c represents the proportion of the total variance that is presumed to be because of additive genetic effects across populations (Brommer, 2011). The equation above can be rewritten as:

$$P_{ST} = \frac{\frac{c}{h^2}\sigma_B^2}{(\frac{c}{h^2}\sigma_B^2 + 2\sigma_W^2)}$$

where the unknown ratio c/h^2 is the critical aspect that describes how well P_{ST} approximates Q_{ST} (Brommer, 2011).

Variance components were calculated for each gene for both global P_{ST} (including all populations) and pairwise P_{ST} (calculated for each pair of populations). The variance within populations was estimated as:

$$\sigma_W^2 = MS_{error}$$

$$MS_{error} = \frac{\Sigma s_i^2(n_i - 1)}{N - k}$$

where s^2_i is the variance for each population, n_i is the population sample size, N is the total number of data points and k is the total number of populations.

The variance between populations was estimated as:

$$\sigma_B^2 = \frac{MS_{groups} - MS_{error}}{n}$$

$$MS_{groups} = \frac{\Sigma n_i(\bar{Y}_i - \bar{Y})^2}{k - 1}$$

where n is the number of measurements within each population, $\overline{Y_i}$ is the population sample mean and $\overline{Y}$ is the grand mean of all measurements.

These values were then plugged in to the equation for P_{ST}, using c/h^2 of 0.25, 0.5, 1, 2 and 4. 95% confidence intervals for each measure of P_{ST} were then calculated using the

jackknife-1 method (Shao and Wu, 1989). Whenever we compare a measure of P_{ST} to a measure of F_{ST} based on neutral loci, there are three possible outcomes. The first is that P_{ST} is larger than F_{ST}, which means that divergence in the trait exceeds what is expected based on drift, and this is suggestive of divergent selection acting on the trait (Whitlock, 2008; Merila and Crnokrak, 2001). If the lower confidence interval was above the global F_{ST}, the gene's expression was considered to be under divergent selection. The second is that P_{ST} is roughly equal to F_{ST}, in which case there is no evidence for selection, and this is suggestive of neutral drift acting on the trait (Whitlock, 2008; Merila and Crnokrak, 2001). If the confidence interval encompassed the global F_{ST}, the gene's expression was considered most likely to be under neutral drift. The third outcome is that P_{ST} is smaller than F_{ST} and this is evidence for stabilizing selection acting on the trait (Whitlock, 2008; Merila and Crnokrak, 2001). If the upper confidence interval was below the global F_{ST}, the gene's expression was considered to be under stabilizing selection.

The global P_{ST} values were then used to look for patterns related to Gene Ontology (GO) terms (The Gene Ontology Consortium, 2000), as determined from the BLASTX against TAIR9 described above. Each annotation was pruned so that each GO category appears only once per gene, but any one gene can belong to multiple GO categories. For each GO category, the mean and standard error were calculated over all global P_{ST} values of all genes belonging to that category, but only for genes with E values $< 10^{-10}$. GO categories were also tested for under or overrepresentation in the group of genes whose P_{ST} values were deemed to be under selection by exact binomial tests, if their E value was $< 10^{-10}$. Binomial tests were performed using the R statistical package (R Development Core Team, 2011).

3.2.7 Environmental correlations

Elevation-corrected climate variables were determined for each population's home site using the program ClimateWNA v4.60 (Wang et al., 2006) for the years 1901-2009. Average values were then taken for each climate variable. The annual climate variables directly calculated were: mean annual temperature, mean warmest month temperature, mean coldest month temperature, continentality, mean annual precipitation, mean annual summer precipitation, annual heat:moisture index and summer heat:moisture index. The annual derived climate variables were: chilling degree-days, growing degree-days, heating degree-days, cooling degree-days, number of frost-free days, frost-free period, the Julian date on which the frost-free period begins, the Julian date on which the frost-free period ends, precipitation as snow between August in previous year and July in current year, extreme minimum temperature over 30 years, Hargreaves reference evaporation and Hargreaves climatic moisture deficit. The seasonal variables were all measured for winter (December of previous year to February), spring (March to May), summer (June to August) and autumn (September to November), and they were: mean temperature, mean maximum temperature, mean minimum temperature and precipitation. The monthly variables were taken separately for each month and were: mean temperature, maximum mean temperature, minimum mean temperature and precipitation.

The average value for each site was used to calculate a pairwise distance matrix between sites for each variable, and these were then used in Mantel tests with each other, undergoing 10000 permutations. The variables were then grouped based on the P values of the Mantel tests. Every member of each group had to be significantly correlated with a $P < 0.005$ with every other member of the group, which were originally based on finding variables correlated with each other with a $P < 0.0001$. Representative climate variables were then chosen from each group.

Using the genes which were found to be under divergent selection with $c/h^2 = 0.25$, the most conservative of estimates, we performed partial Mantel tests (Smouse et al., 1986) of the pairwise P_{ST} values with various variables, while controlling for pairwise F_{ST} or for natural logarithm of geographic distances (km). These were done using the pairwise P_{ST} values calculated with $c/h^2 = 2$ and 4. This was done because it is common to assume that $c = 1$ and that $h^2 = 0.25$ or 0.5 (Brommer, 2011). Also, while not correcting for many factors of phenotypic variance, the trees were raised in a common environment, which is the main requirement to measure genetic variance among populations (and assume $c = 1$) (Whitlock, 2008). The pairwise P_{ST} was correlated with whether the pair was within or between north-south regions, elevation, latitude, longitude and the representative variables from each climate variable group (Table 3.2). Mantel tests were also performed using pairwise P_{ST} and the natural logarithm of geographic distances (km). All Mantel and partial Mantel tests were performed with 10000 permutations using the R statistical package (R Development Core Team, 2011). Q values were calculated to adjust for the false discovery rate (FDR) using the program QVALUE (Storey and Tibshirani, 2003; Storey et al., 2004). The bootstrap method was used to calculate all Q values, but the general findings were verified using the smoother method.

Table 3.2: Groupings of climate variables based on P values in Mantel tests.

Grp. 1	**Grp. 2**	**Grp. 3**	**Grp. 4**	**Grp. 5**	**Grp. 6**	**Grp. 7**	**Grp. 8**	**Grp. 9**	**Grp. 10**	**Grp. 11**	**Grp. 12**	**Grp. 13**	**Grp. 14**
Jan. MT	Apr. MT	Nov. MT	May Min. MT	Jan. MP	**May MP**	**Jun. MP**	Aug. MP	**Sep. MP**	Oct. MP	Sum. MP	Jan. MT	Aut. MP	**PAS**
Feb. MT	May MT	Dec. MT	Jun. Min. MT	Feb. MP			MSP		Aut. MP	**SHM**	Feb. MT	**AHM**	
Mar. MT	Jun. MT	Jan. Min. MT	Jul. Min. MT	Mar. MP			SHM		**MAP**		Dec. MT		
Apr. MT	Jul. MT	Feb. Min. MT	Aug. Min. MT	Apr. MP			**CMD**				Dec. Max. MT		
May MT	Aug. MT	Mar. Min. MT	**Sum. M Min. T**	Nov. MP							Jan. Min. MT		
Aug. MT	Apr. Max. MT	Apr. Min. MT		Dec. MP							Feb. Min. MT		
Sep. MT	May Max. MT	May Min. MT		Win. MP							Mar. Min. MT		
Oct. MT	Jun. Max. MT	Jun. Min. MT		**Spr. MP**							Nov. Min. MT		
Nov. MT	Jul. Max. MT	Aug. Min. MT		MAP							Dec. Min. MT		
Dec. MT	Aug. Max. MT	Sep. Min. MT									Win. M Min. T		
Jan. Max. MT	Sep. Max. MT	Oct. Min. MT									Win. MT		
Feb. Max. MT	Oct. Max. MT	Nov. Min. MT									MCMT		
Mar. Max. MT	Jul. MP	Dec. Min. MT									**TD**		

Grp. 1	Grp. 2	Grp. 3	Grp. 4	Grp. 5	Grp. 6	Grp. 7	Grp. 8	Grp. 9	Grp. 10	Grp. 11	Grp. 12	Grp. 13	Grp. 14
Apr. Max. MT	Spr. M Max. T	Mar. MP									DD<0		
Sep. Max. MT	Sum. M Max. T	Dec. MP									NFFD		
Oct. Max. MT	Sum. MT	Win. M Min. T									bFFP		
Nov. Max. MT	**MWMT**	Spr. M Min. T									EMT		
Dec. Max. MT	DD>5	Sum. M Min. T											
Feb. Min. MT	DD>18	Aut. M Min. T											
Mar. Min. MT	Eref	Aut. MT											
Apr. Min. MT		Win. MP											
May Min. MT		MAT											
Nov. Min. MT		**MCMT**											
Jul. MP		DD<0											
Win. M Max. T		**NFFD**											
Spr. M Max. T		bFFP											
Aut. M Max. T		eFFP											

Grp. 1	Grp. 2	Grp. 3	Grp. 4	Grp. 5	Grp. 6	Grp. 7	Grp. 8	Grp. 9	Grp. 10	Grp. 11	Grp. 12	Grp. 13	Grp. 14
Spr. M Min. T		FFP											
Win. MT													
Spr. MT													
Aut. MT													
MAT													
MCMT													
DD>5													
DD<18													
NFFD													
EMT													

Abbreviations: M, mean; T, temperature (°C); P, precipitation (mm); Win., winter; Spr., spring; Sum., summer; Aut., autumn; MAT, mean annual temperature (°C); MWMT, mean warmest month temperature (°C); MCMT, mean coldest month temperature (°C); TD, temperature difference between MWMT and MCMT (continentality) (°C); MAP, mean annual precipitation (mm); MSP, mean annual summer precipitation (mm); AHM, annual heat:moisture index; SHM, summer heat:moisture index; DD<0, degree-days below 0°C (chilling degree-days); DD>5, degree-days above 5°C (growing degree-days); DD<18, degree-days below 18°C (heating degree-days); DD>18, degree-days above 18°C (cooling degree-days); FFP, frost-free period; NFFD, number of frost-free days; bFFP, Julian date on which FFP begins; eFFP, Julian date on which FFP ends; PAS, precipitation as snow (mm) since August of previous year; EMT, extreme minimum temperature over 30 years; Eref, Hargreaves reference evaporation; CMD, Hargreaves climatic moisture deficit

3.2.8 Candidate genes

Another approach we took was to look at candidate genes. Some likely genes to be under selection are those involved in the control of growing season, including timing of bud set and growth cessation, because it plays a major role in the trade-off between growth and survival (Hall et al., 2007). Previous studies of traits involved in phenology have used the European aspen, *Populus tremula*, and have found evidence for divergent selection on *phyB2* (Ingvarsson et al., 2006), *PtCENL*-1 (a *Populus* homolog of TFL1 in *Arabidopsis thaliana*) (Hall et al., 2007), *LHY1* and *LHY2* (Ma et al., 2010). We searched our microarray for genes whose best BLASTX hit was PHYB, TFL, or LHY. Expression patterns of the resulting genes were then checked for their global P_{ST} vs. the global F_{ST}.

3.3 Results

3.3.1 Estimates of F_{ST}

The estimates of F_{ST} for each population were: CMBF F_{ST} = 0.1331, DENA F_{ST} = 0.0763, HARB F_{ST} = 0.1498, IRVC F_{ST} = 0.0596, ISKC F_{ST} = 0.1068, KIMB F_{ST} = 0.0573, KLNG F_{ST} = 0.0758, KTMA F_{ST} = 0.0521, LAFY F_{ST} = 0.1826, MCGR F_{ST} = 0.0456, NASH F_{ST} = 0.0468 and SLMB F_{ST} = 0.1856. The global F_{ST} over all populations was 0.0976. The pairwise values of F_{ST} can be found in Table 3.3. All standard errors of estimates of F_{ST} were 0.0000, except for that of SLMB, which was 0.005. This is because of the large number of loci used to perform the calculations.

Table 3.3: Population pairwise F_{ST} estimates.

	NORTHERN						SOUTHERN					
	ISKC	**IRVC**	**NASH**	**KTMA**	**MCGR**	**KIMB**	**CMBF**	**DENA**	**HARB**	**KLNG**	**LAFY**	**SLMB**
IRVC	0.0867											
NASH	0.0467	0.0399										
KTMA	-0.0003	0.0084	0.0227									
MCGR	0.006	0.0118	0.0231	0.0291								
KIMB	0.0007	0.0087	0.019	0.0307	0.0248							
CMBF	-0.0107	-0.0031	0.0091	0.0292	0.0282	0.0051						
DENA	0.0015	0.0078	0.014	0.0365	0.0086	0.0184	0.0111					
HARB	-0.0125	-0.0041	0.0032	0.0207	0.0145	0.0334	0.0125	0.0238				
KLNG	-0.005	-0.0024	0.0096	0.0194	0.0183	0.0147	0.034	0.0332	0.0158			
LAFY	0.003	0.0172	0.0126	0.012	0.0083	0.0057	0.006	0.0053	0.0085	0.0038		
SLMB	0.0217	0.0131	0.0152	0.0033	-0.0243	0.0259	0.04	0.0448	0.0172	0.0365	0.0209	

3.3.2 Isolation by distance

We found no evidence of linear isolation by distance when pairwise genetic and geographic distances were compared (Mantel r = -0.7034867, P = 1, Fig. 3.3). In fact, a linear regression line fit to the data gives a negative slope (-0.017509), but this may be partially due to the non-independence of the data. A neighbour-joining tree of the pairwise F_{ST} values shows almost no clustering between closely-situated populations, as divided into groups based on their approximate geographic location (Fig. 3.4, Fig. 3.1).

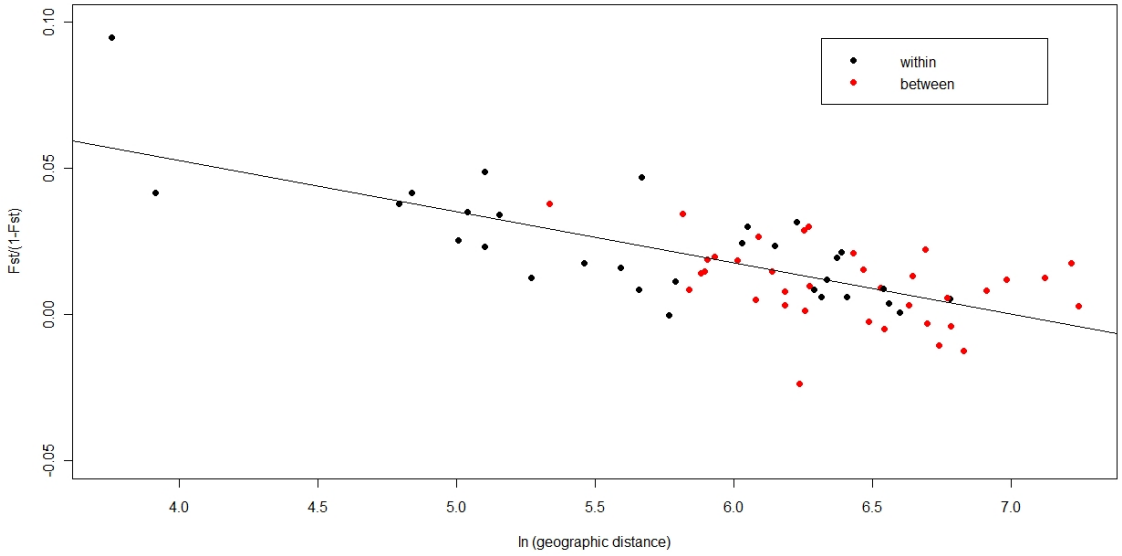


Figure 3.3: Isolation by distance plot. No evidence of linear isolation by distance was found (Mantel r=-0.7034867, P=1). Neutral genetic distance decreases slowly with geographic distance as shown by the linear regression line (slope = -0.017509). Comparisons of populations within a region are indicated by the black circles and comparisons of populations between North-South regions are indicated by the red circles.

3.3.3 Distribution of global P_{ST} values

Global P_{ST} values and jackknife confidence intervals were calculated with the parameter c/h^2 = 0.25, 0.5, 1, 2 and 4. To see the distribution of global P_{ST} values, see Fig. 3.5.

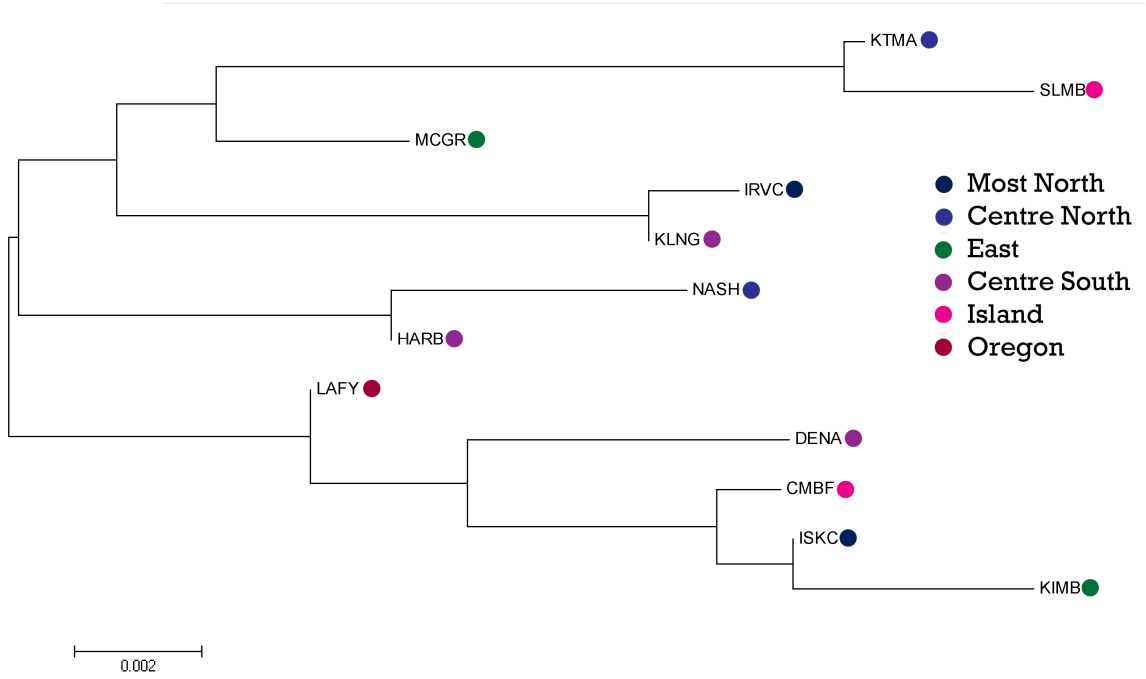


Figure 3.4: Neighbour-joining tree of the pairwise F_{ST} values. Groups are defined according to the populations' approximate geographic location (see Fig. 3.1). Populations from the North are in the groups "Most North", "Centre North" and "East", while those from the South are in the groups "Centre South", "Island" and "Oregon".

The distribution of P_{ST} values calculated when $c/h^2 = 2$ can be thought of as the proportion of phenotypic variance which is found between populations (it is the phenotypic variance component between populations divided by the sum of the variance betwen and the variance within). This distribution peaks at a value a little below 0.5 and is slightly left-skewed. For each value of c/h^2, confidence intervals for each trait were checked to see if they encompassed the global F_{ST} value. If they did, the gene was considered to be most likely drifting neutrally. If the upper confidence interval was below the global F_{ST}, the gene's expression was considered to be under stabilizing selection. If the lower confidence interval was above the global F_{ST}, the gene's expression was considered to be under divergent selection. For a summary of the number of genes under each mode at each value of c/h^2, see Table 3.4. The most conservative estimates for genes under divergent selection are

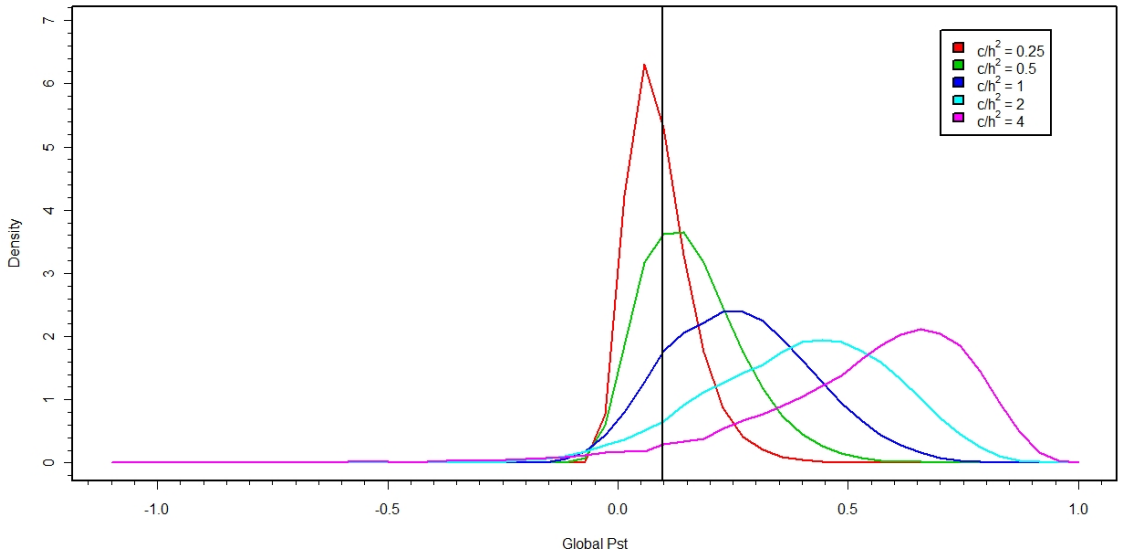


Figure 3.5: Distribution of global P_{ST} values. The global F_{ST} (0.0976) is shown as a solid black line. The density plots of all five tested values of c/h^2 are overlaid. As c/h^2 increases, the peak moves towards the right and the width of the peak increases. In most cases, many of the P_{ST} values lie above the global F_{ST}.

those when $c/h^2 = 0.25$ and the most conservative estimates for genes under stabilizing selection are those when $c/h^2 = 4$. Under these most conservative conditions, 368 (2.37%) gene expression values are under divergent selection (Table 3.5) and 27 (0.17%) are under stabilizing selection (Table 3.6).

Table 3.4: Number of genes whose expression values seem to be affected by divergent selection, drift and stabilizing selection as judged by global P_{ST} compared to global F_{ST}.

c/h^2	Divergent selection	Neutral Drift	Stabilizing Selection
0.25	368	12091	3037
0.5	1707	12920	869
1	4336	10903	257
2	7449	7962	85
4	9959	5510	27

Table 3.5: Genes under divergent selection when $c/h^2 = 0.25$.

Clone ID	BLASTX vs. Arabidopsis	AGI code	E-value	BLASTX vs. Non-redundant	Organism	E-value	P_{ST}± CI
No hit							
WS01127_H06	no hit	no hit		antimicrobial peptide 1	*Pinus pinaster*	1.7E-47	0.30±0.17
WS0151_L19	no hit	no hit		no hit			0.26±0.16
WS0151_F05	no hit	no hit		no hit			0.24±0.14
WS0224_A16	no hit	no hit		no hit			0.26±0.11
WS0223_D14	no hit	no hit		rhUS18	*Macacine herpesvirus 3*	5.2	0.24±0.10
WS0118_P04	no hit	no hit		rubredoxin-type Fe(Cys)4 protein	*Acidovorax citrulli*	4.9	0.34±0.13
WS0172_D14	no hit	no hit		hypothetical protein	*Plasmodium berghei*	5.6	0.20±0.09
WS0194_J15	no hit	no hit		predicted protein	*Populus trichocarpa*	5.5E-19	0.31±0.17
WS0173_G12	no hit	no hit		pleckstrin domain-containing protein	*Polysphondylium pallidum PN500*	1.0E-11	0.22±0.13
WS0201_K24	no hit	no hit		unknown protein	*Populus trichocarpa*	1.2	0.35±0.13
WS0213_G20	no hit	no hit		hypothetical protein	*Plasmodium chabaudi chabaudi*	3.3	0.45±0.18
WS01127_E01	no hit	no hit		hypothetical protein	*Plasmodium chabaudi chabaudi*	0.26	0.25±0.15
E-value > E-10							
WS0198_J09	unknown protein	AT3G61723.1	0.82	predicted protein	*Arabidopsis lyrata*	2.4	0.21±0.11
WS0192_L21	unknown protein	AT1G03106.1	0.34	predicted protein	*Populus trichocarpa*	3.1E-09	0.32±0.24
WS0161_C10	unknown protein	AT1G24575.1	3.7E-04	predicted protein	*Populus trichocarpa*	2.6E-35	0.24±0.12
WS0212_D21	unknown protein	AT1G27213.1	0.95	no hit			0.22±0.09
WS0198_P14	unknown protein	AT1G27850.1	1.3	maturase K	*Raphanus sativus*	4.1	0.31±0.17
WS0213_B21	unknown protein	AT1G30757.1	0.68	predicted protein	*Populus trichocarpa*	7.8E-20	0.33±0.25
WS01127_H22	unknown protein	AT1G36272.1	3	unknown protein	*Medicago truncatula*	0.0082	0.27±0.18
WS0196_P23	unknown protein	AT1G60783.1	5.4E-04	predicted protein	*Populus trichocarpa*	6.1E-59	0.24±0.11

Clone ID	BLASTX vs. Arabidopsis	AGI code	E-value	BLASTX vs. Non-redundant	Organism	E-value	P_{ST}± CI
WS01911_L04	unknown protein	AT2G15318.1	0.78	copper amine oxidase N-terminal domain superfamily	*Carboxydibrachium pacificum DSM 12653*	3.1	0.29±0.15
WS01127_G06	unknown protein	AT2G18938.1	1.5	unknown protein	*Medicago truncatula*	1.9	0.28±0.18
WS0175_L10	unknown protein	AT2G31090.1	7.8	predicted protein	*Populus trichocarpa*	8.2E-48	0.33±0.18
WS0182_M17	unknown protein	AT2G35736.1	7.7E-08	hypothetical protein	*Vitis vinifera*	1.9E-07	0.33±0.17
WS0173_I24	unknown protein	AT2G37195.1	2.3E-06	predicted protein	*Populus trichocarpa*	9.6E-22	0.32±0.22
WS0173_O16	unknown protein	AT2G43386.1	4.8	no hit			0.28±0.16
WS0209_A01	unknown protein	AT2G45860.1	8.2	Circumsporozoite protein precursor, putative	*Ricinus communis*	0.0015	0.31±0.19
WS0207_M20	unknown protein	AT3G03170.1	0.000029	predicted protein	*Populus trichocarpa*	2.7E-42	0.45±0.25
WS0134_H06	unknown protein	AT3G10020.1	3.8E-06	unknown protein	*Populus trichocarpa*	2.3E-34	0.26±0.16
WS02010_E03	unknown protein	AT3G19660.1	2	predicted protein	*Populus trichocarpa*	2.8E-15	0.34±0.18
WS0173_F01	unknown protein	AT4G11385.1	4.3E-07	hypothetical protein	*Mycobacterium tuberculosis 210*	1.8E-19	0.43±0.24
WS0178_M17	unknown protein	AT4G11385.1	3.2E-03	unknown protein	*Schistosoma japonicum*	3.1E-08	0.44±0.20
WS0175_O13	unknown protein	AT4G11385.1	0.19	NADH dehydrogenase subunit 2	*Ranodon sibiricus*	0.93	0.59±0.13
WS0158_F18	unknown protein	AT4G11385.1	2.4E-06	GTP-binding protein alpha subunit, gna, putative	*Ricinus communis*	1.1E-10	0.34±0.26
WS0124_A08	unknown protein	AT4G11385.1	0.016	predicted protein	*Populus trichocarpa*	7.1E-33	0.28±0.13
WS0157_D04	unknown protein	AT4G11385.1	1.6E-04	Kunitz-type protease inhibitor KPI-B7.2	*Populus trichocarpa x Populus nigra*	4.9E-57	0.49±0.22
WS0233_O11	unknown protein	AT4G24380.1	5.8E-03	predicted protein	*Populus trichocarpa*	2.4E-41	0.45±0.26
WS0152_K23	unknown protein	AT4G29905.1	1.4E-10	predicted protein	*Populus trichocarpa*	1.6E-28	0.49±0.17
WS0145_O07	unknown protein	AT4G30780.1	8.7E-09	predicted protein	*Populus trichocarpa*	8.0E-89	0.23±0.12
WS0166_A07	unknown protein	AT4G33660.1	2.2E-04	glycine-rich protein	*Gossypium hirsutum*	2.7E-10	0.71±0.14
WS01224_J06	unknown protein	AT5G06980.2	0.19	reverse transcriptase-like protein	*Amaranthus quitensis*	6.4	0.22±0.12
WS0191_J21	unknown protein	AT5G19480.1	0.000023	RNA binding protein, putative	*Ricinus communis*	9.7E-05	0.29±0.16
WS0115_O05	unknown protein	AT5G24570.1	0.000012	KPNA4 protein	*Homo sapiens*	0.47	0.37±0.23

Clone ID	BLASTX vs. Arabidopsis	AGI code	E-value	BLASTX vs. Non-redundant	Organism	E-value	P_{ST}± CI
WS0175_E10	unknown protein	AT5G46030.1	3.6E-10	Transcription elongation factor, putative	*Ricinus communis*	3.1E-09	0.56±0.17
WS0165_M23	unknown protein	AT5G47928.1	5.5	amino acid permease fragment 2	*Helicobacter acinonychis*	1.9	0.34±0.19
WS0224_H05	unknown protein	AT5G48610.1	0.05	PAP2 family protein	*Treponema phagedenis F0421*	4.9	0.30±0.15
WS0214_H11	unknown protein	AT5G57747.1	5.4	CBL-interacting serine/threonine-protein kinase, putative	*Ricinus communis*	1.4E-05	0.28±0.19
WS0161_J22	60S ribosomal protein L37 (RPL37B)	AT1G52300.1	7.7E-08	hypothetical protein	*Xanthomonas gardneri ATCC 19865*	4.4E-09	0.25±0.15
WS0122_E15	actin binding	AT1G31810.1	3.5E-10	beta C1 protein	*Cotton leaf curl virus-associated DNA beta*	0.17	0.32±0.23
WS01911_J04	AGP19 (arabinogalactan-protein 19)	AT1G68725.1	0.000047	chloride channel, putative	*Toxoplasma gondii GT1*	1.4E-06	0.30±0.15
WS0211_M06	ANAC083 (NAC domain containing 83)	AT5G13180.1	3.2E-09	NAC domain class transcription factor	*Malus x domestica*	9.7E-08	0.39±0.21
WS0133_J24	AP2 domain-containing transcription factor, putative	AT1G01250.1	2.9E-07	AP2/ERF domain-containing transcription factor	*Populus trichocarpa*	2.2E-39	0.26±0.15
WS0195_D03	aspartyl protease family protein	AT5G37540.1	4	estrogen receptor alpha splice variant	*Homo sapiens*	1.3	0.24±0.13
WS0174_L17	ATB BETA	AT1G17720.2	1.8E-07	protein phosphatase 2, regulatory subunit B (PR 52), alpha isoform, isoform CRA_b	*Rattus norvegicus*	3.8E-11	0.48±0.34
WS0198_G07	BRI1 (brassinosteroid insensitive 1)	AT4G39400.1	0.1	predicted protein	*Populus trichocarpa*	7.3E-24	0.38±0.18
WS0193_P02	C2 domain-containing protein	AT1G07310.1	1.1E-10	C2 domain-containing protein	*Arabidopsis thaliana*	4.4E-08	0.27±0.17
WS0171_D21	carbon-nitrogen hydrolase family protein	AT5G12040.2	0.012	tRNA pseudouridine synthase B	*Pelobacter carbinolicus DSM 2380*	0.91	0.20±0.10
WS01116_I12	cation/hydrogen exchanger, putative (CHX21)	AT2G37910.1	0.49	extracellular protein	*Granulicatella adiacens ATCC 49175*	2.5	0.30±0.20
WS01127_B08	CPK6 (calcium-dependent protein kinase 6)	AT2G17290.1	5.2E-10	calcium-dependent protein kinase 5	*Solanum tuberosum*	2.9E-08	0.34±0.26
WS01221_K01	CXE12; carboxylesterase	AT3G48690.1	2.6E-08	CXE carboxylesterase	*Paeonia suffruticosa*	1.3E-10	0.32±0.18
WS0116_C05	CYP704A1	AT2G44890.1	2.1E-06	cytochrome P450	*Populus trichocarpa*	2.7E-12	0.36±0.15

Clone ID	BLASTX vs. Arabidopsis	AGI code	E-value	BLASTX vs. Non-redundant	Organism	E-value	P_{ST}± CI
WS0208_J13	dentin sialophosphoprotein-related	AT5G64170.1	1.3E-10	dentin sialophosphoprotein-related	*Arabidopsis thaliana*	5.3E-08	0.39±0.16
WS0185_F08	disease resistance protein (NBS-LRR class), putative	AT4G27220.1	4E-07	nbs-lrr resistance protein	*Populus trichocarpa*	7.3E-48	0.34±0.20
WS0208_P13	DNAJ heat shock family protein	AT4G39960.1	1.4E-10	DNAJ heat shock family protein	*Arabidopsis lyrata*	1.7E-08	0.29±0.19
WS0168_J01	ELM1 (elongated mitochondria 1)	AT5G22350.1	0.12	DNA adenine methyltransferase	*Aster yellows phytoplasma*	0.56	0.29±0.19
WS02011_M04	defensin-like (DEFL) family protein	AT5G05598.1	7.4	hypothetical protein	*Zea mays*	1.5	0.38±0.28
WS01117_B04	defensin-like (DEFL) family protein	AT5G05598.1	5	Ribonuclease H	*Medicago truncatula*	0.17	0.41±0.21
WS01222_K03	defensin-like (DEFL) family protein	AT1G13607.1	0.96	envelope glycoprotein	*Human immunodeficiency virus 1*	0.18	0.28±0.18
WS0162_I05	defensin-like (DEFL) family protein	AT2G36255.1	0.46	predicted protein	*Arabidopsis lyrata*	1.9	0.23±0.13
WS0207_G07	ECA1 gametogenesis related family protein	AT1G44191.1	0.023	glycoside hydrolase family protein	*Acidothermus cellulolyticus 11B*	0.0092	0.31±0.19
WS0224_K14	eukaryotic translation initiation factor-related	AT1G73180.2	8.6E-04	eukaryotic translation initiation factor 3 subunit, putative	*Ricinus communis*	0.056	0.26±0.15
WS0198_K19	HAT2	AT5G47370.1	2.6	predicted protein	*Populus trichocarpa*	9.6E-06	0.26±0.16
WS01213_I08	hydroxyproline-rich glycoprotein family protein	AT2G22180.1	0.22	CCHC-type integrase	*Populus trichocarpa*	4.3E-08	0.27±0.12
WS0174_E17	hydroxyproline-rich glycoprotein family protein	AT3G26910.1	0.69	predicted protein	*Populus trichocarpa*	4.5E-56	0.23±0.07
WS0199_K16	invertase/pectin methylesterase inhibitor family protein	AT1G62760.1	7.6E-06	integrase	*Populus trichocarpa*	3.2E-11	0.27±0.17
WS0183_O24	involved in protein catabolic process	AT1G68660.1	1.2E-10	clp protease adaptor protein	*Chlamydomonas reinhardtii*	1.8E-13	0.28±0.17
WS0197_H08	JAZ3 (jasmonate-ZIM-domain protein 3)	AT3G17860.2	0.02	prepilin peptidase	*Clostridium botulinum H04402 065*	2	0.30±0.13
WS0176_P20	KEU (keule); protein transporter	AT1G12360.1	4E-09	plant sec1, putative	*Ricinus communis*	1.5E-08	0.30±0.17

Clone ID	BLASTX vs. Arabidopsis	AGI code	E-value	BLASTX vs. Non-redundant	Organism	E-value	P_{ST}± CI
WS0113_D08	leucine-rich repeat family / extensin family protein	AT4G13340.1	0.0045	proline-rich	*Ajellomyces dermatitidis SLH14081*	0.061	0.20±0.07
WS0158_O08	LSU2 (response to low sulfur 2)	AT5G24660.1	2E-07	LSU2 (response to low sulfur 2)	*Arabidopsis thaliana*	8.6E-05	0.24±0.14
WS0173_P12	MT2A (metallothionein 2A)	AT3G09390.1	3.6E-10	metallothionein 2b	*Populus trichocarpa x Populus deltoides*	4.1E-14	0.35±0.17
WS0155_D08	myb family transcription factor	AT5G52660.1	0.03	MYB transcription factor	*Camellia sinensis*	3.4E-07	0.57±0.17
WS0156_N19	nuclear associated protein-related	AT3G02710.1	0.82	trypomastigote small surface antigen	*Trypanosoma cruzi*	5.1	0.27±0.17
WS0168_I21	proline-rich family protein	AT3G09000.1	0.000096	glucose-repressible gene protein	*Verticillium albo-atrum VaMs.102*	1.6E-12	0.24±0.12
WS0203_K03	prolyl oligopeptidase family protein	AT1G69020.1	0.000054	prolyl oligopeptidase family protein	*Arabidopsis thaliana*	0.023	0.21±0.10
WS0183_B12	protein phosphatase 2C (PP2C), putative	AT2G30020.1	1.4E-10	protein phosphatase 2C	*Medicago sativa*	2.6E-09	0.30±0.20
WS0119_N20	protein phosphatase 2C (PP2C), putative	AT5G27930.1	2.6E-10	protein phosphatase 2c, putative	*Ricinus communis*	1.0E-08	0.25±0.12
WS0232_B12	PWWP domain-containing protein	AT5G40340.1	1.7E-10	hypothetical protein	*Harpegnathos saltator*	5.8E-13	0.26±0.15
WS0151_H09	RDR1 (RNA-dependent RNA polymerase 1)	AT1G14790.1	3.4	RNA-dependent RNA polymerase	*Populus trichocarpa*	2.3E-15	0.29±0.19
WS0188_L08	SIB1 (SIGMA factor binding protein 1)	AT3G56710.1	5.0E-08	sigma factor binding protein 1	*Citrullus lanatus*	1.2E-12	0.36±0.16
WS0153_H02	SP1L3 (SPIRAL 1-like 3)	AT3G02180.3	0.081	predicted protein	*Populus trichocarpa*	2.3E-30	0.26±0.16
WS0115_I16	stress protein-related	AT5G16020.1	0.0026	CCHC-type integrase	*Populus trichocarpa*	0.1	0.29±0.18
WS0132_F23	trypsin and protease inhibitor/ Kunitz family protein	AT1G17860.1	0.000088	Kunitz-type protease inhibitor KPI-A1	*Populus trichocarpa x Populus nigra*	4.4E-72	0.40±0.17
WS0133_I11	trypsin and protease inhibitor/ Kunitz family protein	AT1G17860.1	5.2E-06	Kunitz-type protease inhibitor KPI-B3	*Populus trichocarpa x Populus deltoides*	7.0E-90	0.53±0.29
WS0133_N23	trypsin and protease inhibitor/ Kunitz family protein	AT1G17860.1	3.6E-09	Kunitz trypsin inhibitor TI7	*Populus nigra*	2.2E-86	0.36±0.19
WS0132_D18	trypsin and protease inhibitor/ Kunitz family protein	AT1G17860.1	4.3E-07	Kunitz-type protease inhibitor KPI-A1.2	*Populus trichocarpa x Populus deltoides*	1.7E-77	0.29±0.13
WS0151_M13	trypsin and protease inhibitor/ Kunitz family protein	AT1G73325.1	5.0E-08	Kunitz-type protease inhibitor KPI-A2	*Populus trichocarpa x Populus nigra*	2.1E-104	0.40±0.20

Clone ID	BLASTX vs. Arabidopsis	AGI code	E-value	BLASTX vs. Non-redundant	Organism	E-value	P_{ST}± CI
WS0151_C13	trypsin and protease inhibitor/ Kunitz family protein	AT1G73325.1	5.2E-07	Kunitz-type protease inhibitor KPI-B7.2	*Populus trichocarpa x Populus nigra*	2.5E-108	0.37±0.20
WS0141_I19	trypsin and protease inhibitor/ Kunitz family protein	AT1G73325.1	1.3E-06	Kunitz trypsin inhibitor	*Populus balsamifera*	6.0E-84	0.37±0.19
WS0133_J21	trypsin and protease inhibitor/ Kunitz family protein	AT1G73325.1	3.3E-09	Kunitz-type protease inhibitor KPI-A1.2	*Populus trichocarpa x Populus deltoides*	3.4E-42	0.31±0.20
WS0134_G14	trypsin and protease inhibitor/ Kunitz family protein	AT1G73325.1	2.1E-06	Kunitz-type protease inhibitor KPI-B3	*Populus trichocarpa x Populus deltoides*	8.6E-56	0.48±0.25
WS0192_E02	U2AF splicing factor subunit, putative	AT3G44785.1	1.3	hypothetical protein	*Giardia lamblia ATCC 50803*	6.3	0.27±0.17
WS0207_G08	WBC11 (WHITE-BROWN complex protein 11)	AT1G17840.1	7.1	big map kinase/bmk, putative	*Ricinus communis*	9.7E-13	0.53±0.13
WS0202_A06	WRKY2; transcription factor	AT5G56270.1	0.035	WRKY transcription factor, putative	*Ricinus communis*	1.6E-52	0.22±0.10
WS0207_I22	zinc finger (B-box type) family protein	AT4G38960.1	4.6E-10	zinc finger (B-box type) family protein	*Arabidopsis thaliana*	2.5E-07	0.40±0.13
WS0152_O23	zinc finger (C3HC4-type RING finger) family protein	AT3G10815.1	2.9	no hit			0.34±0.22
WS0161_I11	zinc finger (C3HC4-type RING finger) family protein	AT5G01960.1	0.091	putative membrane protein	*Burkholderia multivorans CGD1*	5.6	0.38±0.28
WS0207_M07	zinc ion binding	AT1G70150.1	4.2E-04	similar to zinc finger, MYND-type containing 15	*Monodelphis domestica*	1.4E-06	0.22±0.13
Biological process unknown							
WS0126_L17	unknown protein	AT1G03250.1	2.3E-63	phenazine biosynthesis protein, putative	*Ricinus communis*	9.1E-65	0.29±0.14
PX0019_C13	unknown protein	AT1G08480.1	8.8E-55	predicted protein	*Populus trichocarpa*	8.2E-80	0.25±0.15
WS0116_B10	unknown protein	AT1G14020.1	2E-60	similar to auxin-independent growth promoter protein	*Arabidopsis thaliana*	8.6E-58	0.22±0.12
WS0168_G04	unknown protein	AT1G16080.1	2.0E-112	predicted protein	*Populus trichocarpa*	2.5E-124	0.25±0.12
WS0141_D11	unknown protein	AT1G65230.1	2.9E-63	predicted protein	*Populus trichocarpa*	2.3E-84	0.23±0.12

Clone ID	BLASTX vs. Arabidopsis	AGI code	E-value	BLASTX vs. Non-redundant	Organism	E-value	P_{ST}± CI
WS0173_G08	unknown protein	AT2G17710.1	1.9E-34	predicted protein	*Populus trichocarpa*	1.8E-75	0.24±0.13
WS0161_C24	unknown protein	AT2G28315.1	7.3E-67	UDP-glucuronic acid/UDP-N-acetylgalactosamine transporter, putative	*Ricinus communis*	3.8E-82	0.28±0.15
WS0153_N12	unknown protein	AT2G31710.1	7.8E-31	predicted protein	*Populus trichocarpa*	3.5E-47	0.29±0.19
WS0155_J22	unknown protein	AT2G35470.1	2.1E-14	predicted protein	*Populus trichocarpa*	2.2E-47	0.34±0.19
WS0209_O18	unknown protein	AT2G46630.1	3.7E-31	predicted protein	*Populus trichocarpa*	4.9E-96	0.26±0.16
WS0181_J06	unknown protein	AT3G02420.1	2.5E-64	predicted protein	*Populus trichocarpa*	1.7E-70	0.28±0.17
WS0132_I10	unknown protein	AT3G07090.1	3.7E-72	predicted protein	*Populus trichocarpa*	2.1E-72	0.35±0.17
WS0141_H14	unknown protein	AT3G07310.1	4.7E-24	predicted protein	*Populus trichocarpa*	2.2E-63	0.25±0.14
WS0187_I17	unknown protein	AT3G09860.1	3.7E-47	predicted protein	*Populus trichocarpa*	1.6E-53	0.29±0.18
WS0224_D17	unknown protein	AT3G19120.1	7E-46	predicted protein	*Populus trichocarpa*	4.7E-52	0.23±0.14
WS0178_K03	unknown protein	AT3G24100.1	1E-21	predicted protein	*Populus trichocarpa*	6.0E-29	0.28±0.15
WS0199_B07	unknown protein	AT3G25855.1	1.7E-13	predicted protein	*Populus trichocarpa*	1.8E-27	0.23±0.11
WS0195_N21	unknown protein	AT3G62370.1	2.4E-61	predicted protein	*Populus trichocarpa*	4.6E-86	0.35±0.18
WS0234_J21	unknown protein	AT4G02210.1	1.4E-21	conserved hypothetical protein	*Ricinus communis*	1.2E-44	0.22±0.11
WS0141_H10	unknown protein	AT4G04330.1	7.1E-53	predicted protein	*Populus trichocarpa*	7.2E-88	0.37±0.15
WS0148_G14	unknown protein	AT4G13500.1	1E-37	predicted protein	*Populus trichocarpa*	5.5E-58	0.25±0.15
WS0212_H22	unknown protein	AT4G16146.1	9.1E-21	unknown protein	*Populus trichocarpa*	9.4E-31	0.30±0.18
WS01213_C24	unknown protein	AT4G21740.1	2.6E-23	predicted protein	*Populus trichocarpa*	7.6E-77	0.33±0.16
WS0161_E14	unknown protein	AT4G24265.1	9.6E-17	predicted protein	*Populus trichocarpa*	1.9E-80	0.27±0.14
WS01121_E23	unknown protein	AT4G32020.1	4.7E-24	unknown protein	*Populus trichocarpa*	1.8E-21	0.31±0.18
WS0201_E12	unknown protein	AT4G32605.1	1.8E-70	predicted protein	*Populus trichocarpa*	5.6E-120	0.35±0.20
WS0212_O01	unknown protein	AT4G36980.1	1.8E-34	splicing factor, arginine/serine-rich 16	*Arabidopsis thaliana*	7.3E-32	0.42±0.24
PX0011_P21	unknown protein	AT4G36980.2	1.6E-22	splicing factor, arginine/serine-rich 16	*Arabidopsis thaliana*	6.9E-20	0.24±0.15
WS0232_C02	unknown protein	AT5G03670.1	4.7E-31	nuclease	*Aspergillus oryzae RIB40*	1.5	0.29±0.19

Clone ID	BLASTX vs. Arabidopsis	AGI code	E-value	BLASTX vs. Non-redundant	Organism	E-value	P_{ST}± CI
WS0112_J19	unknown protein	AT5G09960.1	2.8E-33	predicted protein	*Populus trichocarpa*	1.4E-52	0.29±0.16
WS01211_J20	unknown protein	AT5G10780.1	1E-12	predicted protein	*Populus trichocarpa*	2.9E-13	0.38±0.21
WS0155_H07	unknown protein	AT5G10780.1	1.8E-79	predicted protein	*Populus trichocarpa*	6.8E-92	0.31±0.18
WS0233_I18	unknown protein	AT5G40500.1	7.9E-38	predicted protein	*Populus trichocarpa*	3.1E-48	0.41±0.19
PX0011_F17	unknown protein	AT5G50011.1	2.2E-17	transcription factor, putative	*Ricinus communis*	5.4E-25	0.37±0.17
WS0181_L15	unknown protein	AT5G53650.1	7.7E-24	predicted protein	*Populus trichocarpa*	2.2E-29	0.21±0.10
WS0171_L20	unknown protein	AT5G65030.1	2.1E-37	predicted protein	*Populus trichocarpa*	1.1E-100	0.34±0.20
WS0151_K15	unknown protein	AT5G65250.1	5.2E-34	predicted protein	*Populus trichocarpa*	3.7E-107	0.44±0.35
WS01122_J17	unknown protein	AT5G65470.1	5.3E-97	predicted protein	*Populus trichocarpa*	8.1E-120	0.28±0.13
WS0206_N08	ATCSLC12 (cellulose-synthase like C12)	AT4G07960.1	2.6E-28	transferase, transferring glycosyl groups, putative	*Ricinus communis*	3.8E-34	0.27±0.18
WS0174_C06	ATP binding	AT3G52570.1	9.4E-90	ATP binding	*Arabidopsis thaliana*	4.0E-87	0.30±0.20
WS01911_B04	ATP-dependent Clp protease ClpB protein-related	AT1G07200.1	5.4E-21	predicted protein	*Populus trichocarpa*	4.0E-71	0.28±0.18
WS0163_A03	binding	AT3G13330.1	2.2E-18	binding	*Arabidopsis thaliana*	1.2E-15	0.47±0.18
WS0222_C05	calcium-binding EF hand family protein	AT3G10300.3	4.5E-58	EF-hand motif containing protein	*Juglans nigra*	2.1E-56	0.31±0.14
WS0163_E12	carbohydrate binding	AT2G25310.1	4.1E-73	carbohydrate binding	*Arabidopsis thaliana*	2.2E-70	0.29±0.19
WS0123_I11	CBS domain-containing/ octicosapeptide/Phox/Bemp1 (PB1) domain-containing protein	AT3G52950.1	4.4E-44	CBS domain-containing/ octicosapeptide/Phox/Bemp1 (PB1) domain-containing protein	*Arabidopsis thaliana*	0.014	0.24±0.13
WS0212_L21	cell cycle control protein-related	AT1G25682.1	5.9E-97	coiled-coil domain-containing protein 94	*Zea mays*	5.2E-101	0.29±0.15
WS0173_F22	CID9 (CTC-interacting domain 9)	AT3G14450.1	3.2E-105	CID9 (CTC-Interacting Domain 9)	*Arabidopsis thaliana*	1.3E-102	0.44±0.32

Clone ID	BLASTX vs. Arabidopsis	AGI code	E-value	BLASTX vs. Non-redundant	Organism	E-value	P_{ST}± CI
WS0145_J04	curculin-like (mannose-binding) lectin family protein	AT5G18470.1	9.5E-16	predicted protein	*Populus trichocarpa*	4.5E-95	0.38±0.19
WS0121_B18	CYP704A2	AT2G45510.1	1.3E-47	cytochrome P450	*Populus trichocarpa*	8.3E-57	0.35±0.17
WS0205_L04	CYP714A1	AT5G24910.1	2.2E-51	cytochrome P450	*Populus trichocarpa*	9.3E-95	0.33±0.11
WS0202_I12	dehydration-responsive family protein	AT1G26850.1	2E-38	ATP binding protein, putative	*Ricinus communis*	8.7E-44	0.25±0.12
WS01224_H11	dehydration-responsive family protein	AT4G18030.1	5.2E-105	dehydration-responsive family protein	*Arabidopsis lyrata*	3.5E-104	0.31±0.20
WS0163_D13	dehydration-responsive protein-related	AT5G64030.1	1.9E-23	ATP binding protein, putative	*Ricinus communis*	5.9E-24	0.24±0.14
PX0019_E22	ERG28 (Arabidopsis homolog of yeast ergosterol28)	AT1G10030.1	3.2E-57	ERG28 (Arabidopsis homolog of yeast ergosterol28)	*Arabidopsis thaliana*	1.4E-54	0.18±0.07
WS0118_N12	EXL3 (EXORDIUM like 3)	AT5G51550.1	1.7E-18	EXL3 (EXORDIUM like 3)	*Arabidopsis thaliana*	7.2E-16	0.28±0.12
WS01111_E17	FLA7 (FASCICLIN-like arabinoogalactan 7)	AT2G04780.1	1.2E-32	fasciclin-like arabinogalactan protein 12	*Populus tremula x Populus alba*	7.8E-59	0.27±0.17
WS01213_D05	glycine-rich protein	AT3G29075.1	7.9E-22	pro-resilin precursor	*Zea mays*	2.9E-27	0.25±0.15
WS0204_I09	glycine-rich RNA-binding protein, putative	AT1G60650.1	4.4E-28	glycine-rich RNA-binding protein, putative	*Ricinus communis*	2.4E-41	0.36±0.19
WS01119_P10	helicase domain-containing/ pentatricopeptide (PPR) repeat-containing	AT1G12700.1	1.5E-37	predicted protein	*Populus trichocarpa*	1.7E-88	0.29±0.16
WS01127_E07	hydrolase, acting on ester bonds	AT1G07230.1	6.7E-25	putative phospholipase	*Oryza sativa Japonica Group*	1.9E-25	0.27±0.17
WS0204_I18	hydrolase, alpha/beta fold family protein	AT2G39400.1	5E-52	esterase/lipase/thioesterase family protein	*Arabidopsis lyrata*	8.0E-50	0.30±0.16
WS01125_I17	hydroxyproline-rich glycoprotein family protein	AT3G45230.1	1.1E-13	structural constituent of cell wall, putative	*Ricinus communis*	1.4E-34	0.29±0.19
WS01125_I07	integral membrane Yip1 family protein	AT2G36300.1	1.6E-101	golgi membrane protein sb140	*Prunus armeniaca*	3.3E-44	0.27±0.16
WS01119_O24	iqd9 (IQ-domain 9); calmodulin binding	AT2G33990.1	2.6E-23	calmodulin binding protein, putative	*Ricinus communis*	3.0E-43	0.33±0.23
WS01210_A02	KH domain-containing protein	AT5G15270.2	2.1E-76	KH domain-containing protein	*Arabidopsis lyrata*	4.0E-78	0.33±0.18
WS01215_L10	located in chloroplast thylakoid lumen	AT1G14590.1	5E-75	putative Myb DNA binding protein	*Eutrema halophilum*	1.1E-75	0.17±0.07

Clone ID	BLASTX vs. Arabidopsis	AGI code	E-value	BLASTX vs. Non-redundant	Organism	E-value	P_{ST}± CI
WS0194_H22	located in extracellular region	AT5G17900.1	4E-34	microfibril-associated protein, putative	*Ricinus communis*	7.1E-33	0.34±0.23
WS0188_J11	located in nucleus, chloroplast	AT3G59780.1	2.8E-47	rhodanese/cell cycle control phosphatase-like protein	*Arabidopsis thaliana*	1.7E-46	0.25±0.16
WS0207_J19	NBP35 (nucleotide binding protein 35)	AT5G50960.1	3.7E-63	nuclear binding protein 35	*Zea mays*	7.3E-63	0.29±0.20
WS01125_E19	pectate lyase family protein	AT3G55140.1	4.7E-88	pectate lyase 2	*Hevea brasiliensis*	9.2E-72	0.26±0.16
WS0201_K19	pentatricopeptide (PPR) repeat-containing protein	AT1G12620.1	9.3E-15	predicted protein	*Populus trichocarpa*	4.6E-43	0.43±0.19
WS0142_O22	phospholipase/carboxylesterase family protein	AT5G20060.2	1.5E-59	phospholipase/carboxylesterase family protein	*Arabidopsis lyrata*	4.3E-58	0.26±0.16
WS02011_K12	photosystem II family protein	AT1G03600.1	2.8E-42	photosystem II family protein	*Arabidopsis thaliana*	1.5E-39	0.32±0.19
WS0143_K18	PSAE-1 (PSA E1 knockout); catalytic	AT4G28750.1	3.9E-36	photosystem I reaction center subunit IV A, chloroplast	*Ricinus communis*	2.8E-54	0.24±0.12
WS0156_A11	rhomboid protein-related	AT3G07950.1	1.5E-75	transmembrane protein, putative	*Ricinus communis*	1.5E-87	0.36±0.27
WS0191_H09	RNA recognition motif (RRM)-containing protein	AT3G52660.1	1.1E-19	nucleolar phosphoprotein, putative	*Ricinus communis*	2.1E-24	0.26±0.17
WS0172_K23	SAR DNA-binding protein, putative	AT3G05060.1	3.8E-23	matrix attachment region-binding protein	*Cucumis melo*	3.1E-30	0.24±0.10
WS0198_H04	serine/threonine protein phosphatase-related	AT1G56440.1	9E-23	serine/threonine protein phosphatase-related	*Arabidopsis thaliana*	3.8E-20	0.49±0.15
WS0161_G02	SKS6 (SKU5-similar 6); pectinesterase	AT1G41830.1	4.2E-31	multicopper oxidase	*Populus trichocarpa*	3.5E-32	0.26±0.16
WS0175_K03	small nuclear ribonucleoprotein-related	AT4G18372.1	1.4E-38	small nuclear ribonucleoprotein-related	*Arabidopsis thaliana*	6.0E-36	0.31±0.20
WS0142_H11	SOUL heme-binding family protein	AT1G17100.1	1.5E-73	soul heme-binding family protein	*Arabidopsis lyrata*	1.7E-72	0.36±0.16
WS0197_L15	tetracycline transporter	AT2G16990.1	2.5E-27	tetracycline transporter, putative	*Ricinus communis*	6.4E-32	0.32±0.18
WS0156_I17	zinc finger (C3HC4-type RING finger) family protein	AT3G06330.1	3.4E-30	zinc finger (C3HC4-type RING finger) family protein	*Arabidopsis thaliana*	1.4E-27	0.37±0.23
WS01119_O05	zinc finger (C3HC4-type RING finger) family protein	AT5G05830.1	3.8E-38	protein binding protein, putative	*Ricinus communis*	2.3E-59	0.24±0.14
WS01117_H13	zinc finger (Ran-binding) family protein	AT5G25490.1	2.8E-58	zinc finger (Ran-binding) family protein	*Arabidopsis thaliana*	1.2E-55	0.31±0.19

Clone ID	BLASTX vs. Arabidopsis	AGI code	E-value	BLASTX vs. Non-redundant	Organism	E-value	P_{ST}± CI
Transcription							
WS0233_M08	ATEBP (ethylene-responsive element binding protein)	AT3G16770.1	2.8E-26	AP2/ERF domain-containing transcription factor	*Populus trichocarpa*	3.0E-82	0.27±0.16
WS0125_L14	bZIP family transcription factor	AT1G58110.1	4.2E-39	bZIP transcription factor bZIP100	*Glycine max*	6.5E-55	0.27±0.17
WS0124_I04	COL1 (constans-like 1)	AT5G15850.1	1.4E-61	CONSTANS-like protein CO1	*Populus deltoides*	6.5E-118	0.46±0.22
WS0141_N18	COL2 (constans-like 2)	AT3G02380.1	4.7E-56	CONSTANS-like protein CO2	*Populus deltoides*	2.7E-111	0.36±0.18
WS0147_N16	ethylene-responsive family protein	AT4G29100.1	9.7E-47	transcription factor, putative	*Ricinus communis*	2.0E-52	0.29±0.17
WS0172_G19	HMGB3 (high mobility group B 3)	AT1G20696.1	1.4E-54	high mobility group family	*Populus trichocarpa*	2.9E-68	0.24±0.14
WS0188_G09	ILR3 (iaa-leucine resistant3)	AT5G54680.1	6.2E-77	BHLH domain class transcription factor	*Malus x domestica*	5.3E-76	0.31±0.18
WS0231_E15	myb family transcription factor	AT3G09600.1	1.2E-32	MYB transcription factor	*Camellia sinensis*	1.4E-34	0.53±0.15
WS0194_N14	nucleic acid binding / transcription factor/ zinc ion binding	AT2G01940.2	6.5E-16	C2H2L domain class transcription factor	*Malus x domestica*	3.5E-15	0.24±0.13
WS0117_J06	SPT42 (SPT4 homolog 2)	AT5G63670.1	1.7E-44	SPT42 (SPT4 HOMOLOG 2)	*Arabidopsis thaliana*	9.0E-42	0.29±0.18
WS0206_L05	WRKY21	AT2G30590.1	3E-14	WRKY transcription factor IId-3	*Solanum lycopersicum*	1.8E-13	0.20±0.10
WS01223_P17	zinc finger (B-box type) family protein	AT1G68520.1	1E-35	zinc finger (B-box type) family protein	*Cucumis melo*	5.4E-35	0.35±0.17
WS0132_F20	zinc finger (B-box type) family protein	AT1G68520.1	5.7E-35	zinc finger (B-box type) family protein	*Cucumis melo*	8.9E-35	0.32±0.18
WS01210_C06	zinc finger (B-box type) family protein	AT2G21320.1	3.6E-49	COL domain class transcription factor	*Malus x domestica*	4.7E-13	0.37±0.12
WS0113_E07	zinc finger (B-box type) family protein	AT2G21320.1	3.6E-43	COL domain class transcription factor	*Malus x domestica*	7.5E-38	0.37±0.12
WS0173_A15	zinc finger (CCCH-type) family protein	AT2G19810.1	1.5E-25	C3HL domain class transcription factor	*Malus x domestica*	9.9E-36	0.28±0.17
WS0234_H04	zinc finger (GATA type) family protein	AT5G25830.1	8E-29	Zinc finger, GATA-type	*Medicago truncatula*	2.8E-41	0.28±0.18

Clone ID	BLASTX vs. Arabidopsis	AGI code	E-value	BLASTX vs. Non-redundant	Organism	E-value	P_{ST}± CI
Response to stress							
WS0177_F13	4-coumarate-CoA ligase/ 4-coumaroyl-CoA synthase, putative	AT4G05160.1	1.1E-97	4-coumarate-coa ligase	*Populus trichocarpa*	1.1E-114	0.33±0.24
WS01223_B10	ACT7 (ACTIN 7); structural constituent of cytoskeleton	AT5G09810.1	5.7E-28	actin	*Cicer arietinum*	3.8E-27	0.31±0.12
WS0142_H18	ADC2 (arginine decarboxylase 2)	AT4G34710.1	5.6E-60	arginine decarboxylase	*Prunus persica*	1.7E-70	0.33±0.20
WS0127_N15	ADL1E (Arabidopsis dynamin-like 1E)	AT3G60190.1	1E-99	ADL1E (Arabidopsis dynamin-like 1E)	*Arabidopsis thaliana*	4.7E-53	0.34±0.25
WS0186_B22	AKR2 (ankyrin repeat-containing protein 2)	AT4G35450.3	4.8E-47	TGB12K interacting protein 3	*Nicotiana tabacum*	1.6E-51	0.30±0.15
WS01213_L02	APX3 (ascorbate peroxidase 3)	AT4G35000.1	1.1E-63	ascorbate peroxidase	*Populus tomentosa*	9.4E-47	0.23±0.14
WS0124_D01	ATGSR2; copper ion binding/ glutamate-ammonia ligase	AT1G66200.1	4.1E-14	glutamine synthetase	*Alnus glutinosa*	1.7E-12	0.27±0.13
WS0122_J04	ATHM2; enzyme activator	AT4G03520.1	7.9E-38	thioredoxin m	*Populus trichocarpa*	4.2E-54	0.27±0.18
WS0187_O24	ATPQ (ATP synthase D chain, mitochondrial)	AT3G52300.1	2.5E-50	mitochondrial F0 ATP synthase D chain	*Elaeis guineensis*	1.8E-50	0.24±0.12
WS0119_G23	ATRZ-1A; RNA binding / nucleotide binding	AT3G26420.1	4.1E-13	RNA-binding protein RZ-1	*Nicotiana sylvestris*	1.7E-17	0.35±0.21
WS01120_G07	CCR2 (cold, circadian rhythm, and RNA binding 2)	AT2G21660.1	1.3E-62	glycine-rich RNA-binding protein	*Ricinus communis*	2.7E-65	0.36±0.19
PX0015_M15	CIPK1 (CBL-interacting protein kinase 1)	AT3G17510.2	1.7E-99	CBL-interacting protein kinase 19	*Vitis vinifera*	2.1E-106	0.31±0.15
WS0234_B14	COR414-TM1	AT1G29395.1	1E-37	COR414-TM1	*Arabidopsis thaliana*	5.4E-35	0.35±0.14
WS0186_H21	CRB (chloroplast RNA binding)	AT1G09340.1	2.1E-115	CRB (chloroplast RNA binding)	*Arabidopsis thaliana*	9.0E-113	0.27±0.15
WS0112_E19	disease resistance-responsive protein-related	AT1G58170.1	1E-51	disease resistance response protein, putative	*Ricinus communis*	8.2E-64	0.40±0.23
WS0212_H10	GSH1 (glutamate-cysteine ligase)	AT4G23100.1	6.7E-64	glutamate-cysteine ligase, chloroplastic	*Solanum lycopersicum*	2.7E-65	0.26±0.16
WS0143_A15	HSP70 (heat shock protein 70)	AT3G12580.1	2.9E-79	heat shock protein 70	*Gossypium hirsutum*	3.5E-79	0.27±0.17

Clone ID	BLASTX vs. Arabidopsis	AGI code	E-value	BLASTX vs. Non-redundant	Organism	E-value	P_{ST}± CI
WS0184_H02	response to oxidative stress, high light intensity, hydrogen peroxide, heat	AT2G19310.1	2.7E-21	low molecular weight heat-shock protein	*Corylus avellana*	1.7E-31	0.28±0.19
WS0127_E10	JAR1 (jasmonate resistant 1)	AT2G46370.1	2.1E-44	GH3 family protein	*Populus trichocarpa*	1.4E-75	0.29±0.13
WS0173_F21	LOX1; lipoxygenase	AT1G55020.1	3.3E-23	lipoxygenase	*Prunus dulcis*	1.4E-26	0.31±0.22
WS0142_B07	MIPS2 (myo-inositol-1-phostpate synthase 2)	AT2G22240.1	1.4E-95	1L-myo-inositol 1-phosphate synthase	*Jatropha curcas*	3.5E-95	0.37±0.17
WS0157_N10	MLO1; calmodulin binding	AT4G02600.1	1.2E-21	MLO1; calmodulin binding	*Arabidopsis thaliana*	5.2E-19	0.26±0.15
WS0231_J06	NQR	AT1G49670.1	1.9E-25	putative NADPH oxidoreductase	*Capsicum chinense*	4.4E-25	0.21±0.10
WS0185_E12	PIP2A (plasma membrane intrinsic protein 2A)	AT3G53420.1	6.9E-126	aquaporin, MIP family, PIP subfamily	*Populus trichocarpa*	3.9E-144	0.17±0.07
WS01124_F22	PIP2B (plasma membrane intrinsic protein 2)	AT2G37170.1	7.8E-118	membrane protein	*Granulicatella adiacens ATCC 49175*	3	0.31±0.19
WS0195_N01	polcalcin/ calcium-binding pollen allergen, putative	AT1G24620.1	1.5E-52	calcium-binding pollen allergen	*Arachis hypogaea*	1.0E-49	0.23±0.13
WS0122_D03	PRK (phosphoribulokinase)	AT1G32060.1	3.2E-57	phosphoribulokinase	*Pisum sativum*	4.6E-10	0.25±0.16
WS0192_L06	RCI2A (rare-cold-inducible 2A)	AT3G05880.1	6.4E-20	stress-induced hydrophobic peptide	*Populus trichocarpa*	6.5E-23	0.36±0.27
WS0168_I02	RD21 (responsive to dehydration 21)	AT1G47128.1	5.4E-55	cysteine protease	*Hevea brasiliensis*	1.6E-58	0.38±0.25
WS0224_K12	RPN10 (regulatory particle non-ATPase 10)	AT4G38630.1	1.7E-81	26S proteasome non-ATPase regulatory subunit 4	*Zea mays*	6.2E-82	0.40±0.24
WS0182_A09	RSR4 (reduced sugar response 4)	AT5G01410.1	2.1E-21	pyridoxine biosynthesis protein	*Arachis diogoi*	3.4E-19	0.28±0.17
WS0165_O11	TUB6 (beta-6 tubulin)	AT5G12250.1	9.2E-20	tubulin beta-3 chain	*Gossypium hirsutum*	7.0E-18	0.25±0.15
WS0162_P11	VEP1 (vein patterning 1)	AT4G24220.1	1.7E-90	predicted protein	*Populus trichocarpa*	7.2E-136	0.25±0.15
WS02010_L17	wound-responsive family protein	AT4G10270.1	3.6E-26	unknown protein	*Populus trichocarpa x Populus deltoides*	3.0E-34	0.25±0.13
WS0188_D02	wound-responsive family protein	AT4G10270.1	5.7E-12	unknown protein	*Populus trichocarpa x Populus deltoides*	3.1E-25	0.41±0.26

Clone ID	BLASTX vs. Arabidopsis	AGI code	E-value	BLASTX vs. Non-redundant	Organism	E-value	P_{ST}± CI
Response to abiotic or biotic stimulus							
WS0133_H05	chlorophyll A-B binding protein CP29 (LHCB4)	AT5G01530.1	2E-116	light-harvesting complex II protein Lhcb4	*Populus trichocarpa*	1.7E-61	0.34±0.17
WS0131_B01	LHCA1; chlorophyll binding	AT3G54890.1	1.5E-112	light-harvesting complex I protein Lhca1	*Populus trichocarpa*	7.7E-36	0.22±0.13
WS0133_G18	LHCB4.2 (light harvesting complex PSII)	AT3G08940.2	3.6E-104	light-harvesting complex II protein Lhcb4	*Populus trichocarpa*	7.3E-95	0.27±0.18
WS0146_K10	UGT73B2 (UDP-glucosyltransferase 73B2)	AT4G34135.1	1.6E-56	UDP-glucosyltransferase, putative	*Ricinus communis*	2.5E-64	0.32±0.14
Developmental processes							
WS0184_F18	APO2 (accumulation of photosystem one 2)	AT5G57930.1	1.4E-109	APO2 (accumulation of photosystem one 2)	*Arabidopsis thaliana*	7.7E-107	0.27±0.13
WS0131_K20	CHC1	AT5G14170.1	6.3E-61	chromatin remodeling complex subunit	*Populus trichocarpa*	6.9E-67	0.35±0.21
WS0152_I16	CLE44 (CLAVATA3/ESR-related 44)	AT4G13195.1	5.1E-11	CLE44 (CLAVATA3/ESR-related 44)	*Arabidopsis thaliana*	2.2E-08	0.54±0.38
WS01214_K13	COP8 (constitutive photomorphogenic 8)	AT5G42970.1	2E-57	COP8 (constitutive photomorphogenic 8)	*Arabidopsis thaliana*	8.3E-55	0.27±0.12
WS0232_G19	ELF4 (early flowering 4)	AT2G40080.1	3.2E-25	ELF4 protein	*Manihot esculenta*	7.4E-31	0.32±0.24
WS0148_K02	emb2024 (embryo defective 2024)	AT5G24400.1	1E-53	6-phosphogluconolactonase	*Oryza brachyantha*	2.3E-52	0.27±0.14
WS0153_P05	FUS12 (FUSCA 12)	AT2G26990.1	9.8E-96	cop9 signalosome complex subunit, putative	*Ricinus communis*	1.6E-108	0.30±0.18
WS0158_P21	late embryogenesis abundant protein, putative	AT3G50790.1	1.9E-82	alpha/beta hydrolase domain containing 1,3, putative	*Ricinus communis*	3.4E-90	0.26±0.12
WS01123_M03	MEE14 (maternal effect embryo arrest 14)	AT2G15890.1	3.3E-55	MEE14 (maternal effect embryo arrest 14)	*Arabidopsis thaliana*	7.0E-35	0.27±0.15
WS0212_O23	MEE23 (maternal effect embryo arrest 23)	AT2G34790.1	5.1E-50	predicted protein	*Populus trichocarpa*	3.2E-101	0.23±0.11
WS0194_M08	MUM2 (mucilage-modified 2)	AT5G63800.1	1.5E-37	beta-galactosidase, putative	*Ricinus communis*	3.2E-36	0.38±0.27
WS0141_M24	PDX2 (pyridoxine biosynthesis 2)	AT5G60540.1	1.3E-14	PDX2 (pyridoxine biosynthesis 2)	*Arabidopsis thaliana*	6.9E-12	0.28±0.16
WS0151_J20	QQT1 (QUATRE-QUART 1)	AT5G22370.1	1.2E-66	ATP binding domain 1 family member B	*Zea mays*	1.1E-68	0.30±0.13

Clone ID	BLASTX vs. Arabidopsis	AGI code	E-value	BLASTX vs. Non-redundant	Organism	E-value	P_{ST}± CI
WS0162_E15	RIC4 (ROP-interactive CRIB motif-containing protein 4)	AT5G16490.1	9.9E-31	Wiscott-Aldrich syndrome, C-terminal	*Medicago truncatula*	2.3E-34	0.29±0.16
WS0148_G18	SCE1 (SUMO conjugation enzyme 1)	AT3G57870.1	7.3E-67	ubiquitin-conjugating enzyme E2 I	*Zea mays*	8.8E-67	0.26±0.17
WS0118_N18	SMT2 (sterol methyltransferase 2)	AT1G20330.1	1.7E-26	24-sterol C-methyltransferase	*Gossypium hirsutum*	1.7E-24	0.41±0.26
Cell organization and biogenesis							
WS0181_K17	60S ribosomal protein L35a (RPL35aA)	AT1G07070.1	1.3E-53	60S ribosomal protein L35a	*Vernicia fordii*	2.5E-53	0.30±0.16
WS0209_F04	ACT1 (actin 1)	AT2G37620.1	2.1E-30	actin 7	*Corchorus olitorius*	5.4E-28	0.33±0.26
WS01122_B07	ATEXLA2 (Arabidopsis thaliana expansin-like A2)	AT4G38400.1	8.3E-98	expansin-like protein	*Quercus robur*	3.2E-69	0.48±0.26
WS01221_K06	PDV1 (plastid division 1)	AT5G53280.1	2.7E-21	PDV1 (plastid division 1)	*Arabidopsis thaliana*	1.4E-18	0.28±0.17
WS0176_N05	proline-rich extensin-like family protein	AT2G43150.1	5.9E-17	extensin	*Solanum tuberosum*	2.1E-17	0.38±0.16
WS0188_O18	ribosomal protein L10 family protein	AT5G13510.1	1.8E-72	50S ribosomal protein L10, chloroplastic	*Nicotiana tabacum*	8.3E-71	0.26±0.17
Signal transduction							
WS0115_C09	ATRABA1D (Arabidopsis RAB GTPase homolog A1D)	AT4G18800.1	1.2E-89	GTP-binding protein	*Cucumis melo*	2.3E-52	0.21±0.10
WS01212_M04	AtRLP43 (receptor like protein 43)	AT3G28890.1	1.1E-23	verticillium wilt disease resistance protein	*Solanum torvum*	3.4E-26	0.26±0.15
WS0172_C03	BIN2 (brassinosteroid-insensitive 2)	AT4G18710.1	5.3E-80	shaggy-like kinase	*Ricinus communis*	1.4E-82	0.21±0.11
WS0231_F15	CIPK21 (CBL-interacting protein kinase 21)	AT5G57630.1	8.2E-14	CBL-interacting protein kinase 14	*Vitis vinifera*	2.0E-21	0.21±0.11
WS01222_P14	serine/threonine protein phosphatase 2A (PP2A) regulatory subunit B', putative	AT5G25510.1	1.9E-43	protein phosphatase 2A B'kappa subunit	*Oryza sativa Japonica Group*	1.2E-32	0.22±0.12
WS0196_I03	SPHK1 (sphingosine kinase 1)	AT4G21540.2	6.6E-34	SPHK1 (sphingosine kinase 1)	*Arabidopsis thaliana*	2.8E-31	0.30±0.20

Clone ID	BLASTX vs. Arabidopsis	AGI code	E-value	BLASTX vs. Non-redundant	Organism	E-value	P_{ST}± CI
Transport							
WS01213_K10	ADL6 (dynamin-like protein 6)	AT1G10290.1	7.2E-83	dynamin-2A, putative	*Ricinus communis*	5.2E-85	0.40±0.24
WS0212_O05	AHA5 (Arabidopsis H(+)-ATPase 5)	AT2G24520.1	1.7E-90	autoinhibited H+ ATPase	*Populus trichocarpa*	1.5E-96	0.44±0.24
WS0212_I21	APM1 (aminopeptidase M1)	AT4G33090.1	1.7E-63	APM1 (aminopeptidase M1)	*Arabidopsis thaliana*	7.0E-61	0.36±0.27
WS0156_O03	ATPT2 (Arabidopsis thaliana phosphate transporter 2)	AT2G38940.1	1.9E-86	high affinity inorganic phosphate transporter	*Populus trichocarpa*	9.2E-102	0.30±0.16
WS0163_M21	AVA-P4; ATPase	AT1G75630.1	3.8E-38	AVA-P2; ATPase	*Arabidopsis thaliana*	2.0E-35	0.26±0.11
WS0167_G13	AVA-P4; ATPase	AT1G75630.1	5.6E-53	AVA-P2; ATPase	*Arabidopsis thaliana*	3.0E-50	0.31±0.22
WS0141_A11	CWLP (cell wall-plasma membrane linker protein)	AT3G22120.1	2.4E-43	cell wall-plasma membrane linker protein	*Brassica napus*	3.8E-43	0.17±0.06
WS0114_M17	GLTP1 (glycolipid transfer protein 1)	AT2G33470.1	8.9E-62	glycolipid transfer protein, putative	*Ricinus communis*	4.0E-32	0.30±0.20
WS0152_B07	heavy-metal-associated domain-containing protein	AT1G01490.1	3.8E-13	metal ion binding protein, putative	*Ricinus communis*	7.3E-15	0.32±0.09
WS0123_F23	heavy-metal-associated domain-containing protein	AT2G37390.1	9.2E-28	heavy-metal-associated domain-containing protein	*Arabidopsis lyrata*	8.4E-27	0.26±0.13
WS0162_N21	ER to Golgi vesicle-mediated transport, located in intracellular	AT1G80500.1	3.7E-47	trafficking protein particle complex protein 2	*Zea mays*	2.8E-43	0.59±0.24
WS0214_E06	KUP10; potassium ion transmembrane transporter	AT1G31120.1	6.8E-87	KUP10; potassium ion transmembrane transporter	*Arabidopsis thaliana*	2.9E-84	0.24±0.14
WS0212_O13	LHT2 (lysine histidine transporter 2)	AT1G24400.1	8.2E-26	lysine/histidine transporter	*Populus trichocarpa*	2.9E-29	0.30±0.20
WS0144_J21	metal ion binding	AT5G50740.3	4.5E-35	predicted protein	*Populus trichocarpa*	7.2E-56	0.31±0.21
WS0205_K22	mitochondrial substrate carrier family protein	AT3G20240.1	4.8E-54	mitochondrial substrate carrier family protein	*Arabidopsis lyrata*	2.0E-51	0.35±0.19
WS01224_L02	nitrate transporter (NTP3)	AT3G21670.1	1.1E-65	nitrate transporter, H+/oligopeptide symporter POT family	*Populus trichocarpa*	2.7E-81	0.34±0.13
WS0157_I02	PIP2;5 (plasma membrane intrinsic protein 2;5)	AT3G54820.1	2.9E-40	aquaporin, MIP family, PIP subfamily	*Populus trichocarpa*	6.3E-41	0.25±0.15

Clone ID	BLASTX vs. Arabidopsis	AGI code	E-value	BLASTX vs. Non-redundant	Organism	E-value	P_{ST}± CI
WS0203_P14	proton-dependent oligopeptide transport (POT) family protein	AT5G19640.1	1.1E-102	TGF-beta receptor, type I/II extracellular region	*Medicago truncatula*	2.3E-114	0.34±0.17
WS0193_M19	SEC14 cytosolic factor, putative/ phosphoglyceride transfer protein, putative	AT4G39170.1	1.1E-22	phosphatidylinositol transporter, putative	*Ricinus communis*	1.1E-32	0.32±0.23
WS0113_F05	SEC22; transporter	AT1G11890.1	6.4E-59	SEC22; transporter	*Arabidopsis thaliana*	2.7E-56	0.25±0.12
WS01216_A06	SYP81 (syntaxin of plants 81)	AT1G51740.1	5.3E-48	syntaxin-81, putative	*Ricinus communis*	9.1E-49	0.26±0.12
WS0201_H04	VPS46.2 (vacuolar protein sorting)	AT1G73030.1	4.4E-92	SNF7 family protein	*Arabidopsis lyrata*	8.0E-89	0.62±0.15
Protein metabolism							
WS01116_L01	60S acidic ribosomal protein P1 (RPP1A)	AT1G01100.1	2.8E-33	60S acidic ribosomal protein P1	*Zea mays*	2.0E-37	0.47±0.31
WS0168_F18	60S acidic ribosomal protein P1 (RPP1A)	AT1G01100.1	2.8E-33	60S acidic ribosomal protein P1	*Zea mays*	2.0E-37	0.34±0.24
WS01118_N11	ARF3 (ADP-ribosylation factor 3)	AT2G24765.1	3.5E-90	ADP-ribosylation factor 1	*Brassica rapa*	2.2E-70	0.24±0.12
WS02012_L03	ATAAH (Arabidopsis thaliana allantoate amidohydrolase)	AT4G20070.1	2.2E-40	allantoate amidohydrolase	*Glycine max*	4.0E-39	0.25±0.14
WS0205_I19	ATP binding / nucleotide binding / phenylalanine-tRNA ligase	AT5G56075.1	2.7E-19	nucleic acid binding , related	*Medicago truncatula*	3.6E-41	0.30±0.20
WS0195_C17	CDPK19 (calcium-dependent protein kinase 19)	AT5G19450.1	8.6E-30	calcium dependent protein kinase 14	*Populus trichocarpa*	1.4E-34	0.39±0.31
WS0166_E04	Chloroplast encoded ribosomal protein S4	ATCG00380.1	1.2E-32	ribosomal protein S4	*chloroplast Populus alba*	1.5E-32	0.36±0.17
WS0212_M15	CK1 (casein kinase 1)	AT4G26100.1	5.2E-78	casein kinase I-like	*Oryza sativa Japonica Group*	3.5E-77	0.29±0.13
WS01110_G05	CYP5 (cyclophilin 5)	AT2G29960.1	8.4E-34	isomerase peptidyl-prolyl cis-trans isomerase	*Populus trichocarpa*	7.1E-33	0.29±0.16
WS02012_L15	DNAJ heat shock N-terminal domain-containing protein (J11)	AT4G36040.1	6.9E-23	Chaperone protein dnaJ 11, chloroplast precursor, putative	*Ricinus communis*	2.2E-45	0.24±0.15
WS0162_M14	eukaryotic translation initiation factor 2B family protein	AT2G05830.1	1.4E-79	eukaryotic translation initiation factor 2B family protein	*Arabidopsis lyrata*	1.1E-77	0.31±0.19

Clone ID	BLASTX vs. Arabidopsis	AGI code	E-value	BLASTX vs. Non-redundant	Organism	E-value	P_{ST}± CI
WS0143_J07	J8; heat shock protein binding / unfolded protein binding	AT1G80920.1	3.5E-44	heat shock protein binding protein	*Solanum lycopersicum*	3.1E-48	0.21±0.11
WS0148_E02	kinase	AT4G08850.1	8.3E-27	leucine-rich repeat receptor-like protein kinase 1	*Populus nigra*	1.2E-69	0.27±0.16
WS0116_F06	PAB1 (proteasome subunit PAB1)	AT1G16470.1	2.4E-84	PAB1 (proteasome subunit PAB1)	*Arabidopsis thaliana*	2.8E-54	0.29±0.14
WS0111_F10	protein kinase family protein	AT3G20530.1	2E-62	receptor serine-threonine protein kinase, putative	*Ricinus communis*	4.7E-63	0.35±0.15
WS0116_K22	protein kinase-related	AT5G59010.1	2.8E-33	protein kinase family protein	*Arabidopsis thaliana*	1.5E-30	0.26±0.17
WS0187_E01	ribosomal protein L3 family protein	AT2G43030.1	2.7E-101	50S ribosomal protein L3, chloroplastic	*Nicotiana tabacum*	1.1E-100	0.30±0.21
WS0192_N19	SCPL45 (serine carboxypeptidase-like 45 precursor)	AT1G28110.1	3.8E-54	SCPL45 (serine carboxypeptidase-like 45 precursor)	*Arabidopsis thaliana*	1.6E-51	0.30±0.20
WS0157_D11	SRF8 (STRUBBELIG-receptor family 8)	AT4G22130.2	1.3E-14	serine/threonine protein kinase like protein	*Arabidopsis thaliana*	1.8E-12	0.32±0.19
PX0015_K17	SWAP (Suppressor-of-White-APricot)/surp domain-containing protein	AT1G14650.1	9.2E-60	swap (Suppressor-of-White-APricot)/surp domain-containing protein	*Arabidopsis lyrata*	3.8E-59	0.25±0.13
WS0156_A01	UBC14 (ubiquitin-conjugating enzyme 14)	AT3G55380.1	9.6E-81	UBC14 (ubiquitin-conjugating enzyme 14)	*Arabidopsis thaliana*	5.2E-78	0.32±0.20
WS0173_C18	UBP22 (ubiquitin-specific protease 22)	AT5G10790.1	1.5E-36	ubiquitin carboxyl-terminal hydrolase, putative	*Ricinus communis*	2.3E-52	0.38±0.19
WS0174_N02	UBQ10 (polyubiquitin 10)	AT4G05320.2	1.2E-89	polyubiquitin containing 7 ubiquitin monomers	*Zea mays*	1.5E-87	0.27±0.13
Other metabolic processes							
WS0168_E19	ADP-glucose pyrophosphorylase family protein	AT1G74910.1	3E-29	mannose-1-phosphate guanyltransferase	*Zea mays*	7.6E-29	0.65±0.11
WS01127_N21	ADT1 (arogenate dehydratase 1)	AT1G11790.1	1.9E-27	arogenate/prephenate dehydratase	*Populus trichocarpa*	9.0E-26	0.27±0.16
WS0126_I11	amidase family protein	AT4G34880.1	3.2E-50	amidase	*Cucumis melo*	1.3E-42	0.51±0.22

Clone ID	BLASTX vs. Arabidopsis	AGI code	E-value	BLASTX vs. Non-redundant	Organism	E-value	P_{ST}± CI
WS01224_I21	AtCXE5 (Arabidopsis thaliana carboxyesterase 5)	AT1G49660.1	8.5E-46	CXE carboxylesterase	*Malus pumila*	3.3E-47	0.34±0.14
WS0171_N23	ATFD3 (ferredoxin 3)	AT2G27510.1	3.6E-42	non-photosynthetic ferredoxin	*Citrus sinensis*	1.3E-49	0.31±0.20
WS0211_N23	ATFD3 (ferredoxin 3)	AT2G27510.1	5.7E-35	ferredoxin-3	*Saccharum hybrid cultivar Funong 95-1702*	3.1E-39	0.31±0.22
WS01117_N24	carbon-sulfur lyase	AT5G16940.1	3.9E-52	carbon-sulfur lyase	*Arabidopsis thaliana*	1.7E-49	0.41±0.27
WS0208_C14	GCN5-related N-acetyltransferase (GNAT) family protein	AT1G72030.1	6.7E-46	GCN5-related N-acetyltransferase (GNAT) family protein	*Arabidopsis thaliana*	2.8E-43	0.28±0.14
WS0122_C21	GDSL-motif lipase/hydrolase family protein	AT2G04570.1	1.2E-52	GDSL-motif lipase/hydrolase family protein	*Arabidopsis thaliana*	1.6E-35	0.31±0.20
WS0123_C15	lecithin:cholesterol acyltransferase family protein	AT1G27480.1	1.3E-60	lecithin:cholesterol acyltransferase family protein	*Arabidopsis lyrata*	1.2E-32	0.41±0.24
WS0193_B14	PECT1 (phosphorylethanolamine cytidylyltransferase 1)	AT2G38670.1	4.8E-54	ethanolamine-phosphate cytidylyltransferase 1	*Gossypium hirsutum*	3.2E-62	0.30±0.17
WS0148_A08	phosphoglycerate/bisphosphoglycerate mutase family protein	AT5G62840.1	4E-43	phosphoglycerate/bisphosphoglycerate mutase family protein	*Arabidopsis lyrata*	2.4E-41	0.26±0.14
WS0194_F18	QUA1 (QUASIMODO 1); transferase	AT3G25140.1	8.8E-39	glycosyltransferase, CAZy family GT8	*Populus trichocarpa*	5.9E-38	0.20±0.11
WS01111_I23	RNA recognition motif (RRM)-containing protein	AT5G32450.1	7.6E-40	RNA recognition motif-containing protein	*Arabidopsis lyrata*	3.2E-37	0.29±0.18
WS0199_B22	thiF family protein	AT1G05350.1	4.4E-16	ubiquitin-like modifier-activating enzyme 5	*Arabidopsis thaliana*	2.0E-13	0.31±0.12
WS0153_L19	UDP-glucose 6-dehydrogenase, putative	AT5G15490.1	2.7E-92	UDP-glucose 6-dehydrogenase	*Zea mays*	2.5E-94	0.28±0.18
Other cellular processes							
WS01213_E24	ABC4 (aberrant chloroplast development 4)	AT1G60600.1	1.1E-38	ABC4 (aberrant chloroplast development 4)	*Arabidopsis thaliana*	4.7E-36	0.31±0.17
WS0122_E04	ACHT4 (atypical CYS HIS rich thioredoxin 4)	AT1G08570.1	1.9E-50	ACHT4 (atypical CYS HIS rich thioredoxin 4)	*Arabidopsis thaliana*	3.5E-17	0.23±0.14
WS01110_B06	ACP1 (acyl carrier protein 1)	AT3G05020.1	2.8E-33	acyl carrier protein	*Fragaria vesca*	4.3E-42	0.34±0.20

Clone ID	BLASTX vs. Arabidopsis	AGI code	E-value	BLASTX vs. Non-redundant	Organism	E-value	P_{ST}± CI
WS0147_P24	APR3 (APS reductase 3)	AT4G21990.1	6.9E-55	adenosine 5' phosphosulfate reductase	*Populus tremula x Populus alba*	4.6E-67	0.24±0.15
WS0204_J07	aspartate/glutamate/uridylate kinase family protein	AT1G26640.1	1.6E-39	aspartate/glutamate/uridylate kinase family protein	*Arabidopsis lyrata*	5.9E-38	0.32±0.15
WS0178_O11	ATCCS (copper chaperone for SOD1)	AT1G12520.1	3.4E-53	Cu/Zn-superoxide dismutase copper chaperone precursor	*chloroplast Glycine max*	1.7E-56	0.44±0.23
WS01213_N03	ATRSP35; binding	AT4G25500.1	5.7E-67	splicing factor-like protein	*Vitis riparia*	0.0002	0.23±0.13
WS0207_K12	DPE1 (disproportionating enzyme)	AT5G64860.1	1.6E-69	4-alpha-glucanotransferase, chloroplastic/amyloplastic	*Solanum tuberosum*	2.6E-67	0.31±0.16
WS0133_K03	FAD3 (fatty acid desaturase 3)	AT2G29980.1	4.2E-103	endoplasmic reticulum 18:2 desaturase	*Populus tomentosa*	1.3E-90	0.22±0.11
WS0203_J18	FTSZ2-2; GTP binding	AT3G52750.1	5.3E-32	cell division protein ftsZ, putative	*Ricinus communis*	3.3E-35	0.25±0.15
WS0153_B01	G6PD6 (glucose-6-phosphate dehydrogenase 6)	AT5G40760.1	3.4E-69	glucose-6-phosphate dehydrogenase	*Populus trichocarpa*	6.1E-75	0.18±0.09
WS0187_M12	glutaredoxin family protein	AT3G62930.1	1E-35	glutaredoxin	*Populus trichocarpa*	5.0E-48	0.27±0.18
WS0143_L03	histone H1.2	AT2G30620.1	2.3E-40	histone H1	*Populus trichocarpa*	1.5E-94	0.26±0.15
WS0234_D10	histone H3.2	AT4G40030.2	1.1E-65	histone H3.2	*Arabidopsis thaliana*	5.8E-63	0.23±0.12
WS0134_N08	HUA1 (enhancer of AG-4 1); RNA binding	AT3G12680.1	9.6E-49	HUA1 (enhancer of AG-4 1); RNA binding	*Arabidopsis thaliana*	4.1E-46	0.36±0.20
WS0127_N14	HYD1 (HYDRA1); C-8 sterol isomerase	AT1G20050.1	1.3E-44	HYD1 (HYDRA1); C-8 sterol isomerase	*Arabidopsis thaliana*	5.5E-42	0.32±0.23
WS0143_K21	LHCA4 (light-harvesting chlorophyll-protein complex I subunit A4)	AT3G47470.1	2.1E-99	light-harvesting complex I protein Lhca4	*Populus trichocarpa*	1.1E-107	0.29±0.15
WS0133_G24	MT2A (metallothionein 2A)	AT3G09390.1	3.9E-29	metallothionein 2b	*Populus trichocarpa x Populus deltoides*	4.8E-27	0.36±0.11
WS0178_L23	MT2A (metallothionein 2A)	AT3G09390.1	1.4E-24	metallothionein 2b	*Populus trichocarpa x Populus deltoides*	6.1E-36	0.27±0.16
WS0152_D14	MT2A (metallothionein 2A)	AT3G09390.1	2.3E-15	GRAS family transcription factor	*Populus trichocarpa*	2.1E-47	0.23±0.13

Clone ID	BLASTX vs. Arabidopsis	AGI code	E-value	BLASTX vs. Non-redundant	Organism	E-value	P_{ST}± CI
WS0213_A19	SHM3 (serine hydroxymethyltransferase 3)	AT4G32520.1	1.2E-39	serine hydroxymethyltransferase 8	*Populus trichocarpa*	1.5E-69	0.25±0.13
WS0124_N03	SMAP1 (small acidic protein 1)	AT4G13520.1	4.5E-19	SMAP1 (small acidic protein 1)	*Arabidopsis thaliana*	2.4E-16	0.20±0.09
WS0184_L09	sucrose-phosphatase 1 (SPP1)	AT2G35840.1	7.4E-58	sucrose phosphate phosphatase	*Ricinus communis*	7.0E-74	0.26±0.15
WS0164_E19	tRNA (adenine-N1-)-methyltransferase	AT5G14600.1	2.3E-54	tRNA (adenine-N1-)-methyltransferase	*Arabidopsis thaliana*	1.3E-51	0.23±0.10
WS01213_G24	unknown protein	AT4G24380.1	6.7E-48	unknown protein	*Populus trichocarpa*	3.6E-77	0.31±0.19
WS0122_B19	VIT1 (vacuolar iron transporter 1)	AT2G01770.1	1.4E-81	VIT1 (vacuolar iron transporter 1)	*Arabidopsis thaliana*	2.5E-62	0.25±0.12
Other biological processes							
WS0174_P20	AGP31 (arabinogalactan-protein 31)	AT1G28290.1	2.3E-40	arabinogalactan protein	*Daucus carota*	3.1E-48	0.25±0.15
WS0152_P20	APS1 (ATP sulfurylase 1)	AT3G22890.1	1.1E-119	ATP-sulfurylase	*Camellia sinensis*	1.0E-119	0.29±0.19
WS0187_D22	involved in aging; located in endomembrane system	AT2G17850.1	6.8E-32	oxysterol-binding protein	*Medicago truncatula*	5.8E-31	0.37±0.18
WS0155_D16	pfkB-type carbohydrate kinase family protein	AT5G51830.1	8.2E-22	fructokinase	*Dimocarpus longan*	9.9E-20	0.27±0.16

CI: 95% confidence interval

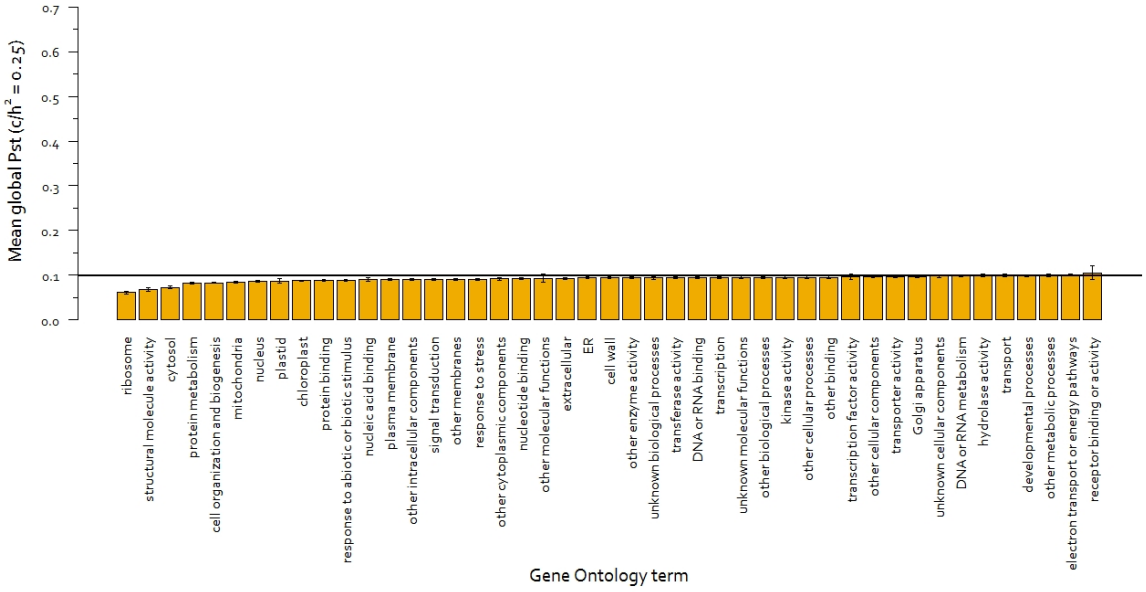
Table 3.6: Genes under stabilizing selection when $c/h^2 = 4$.

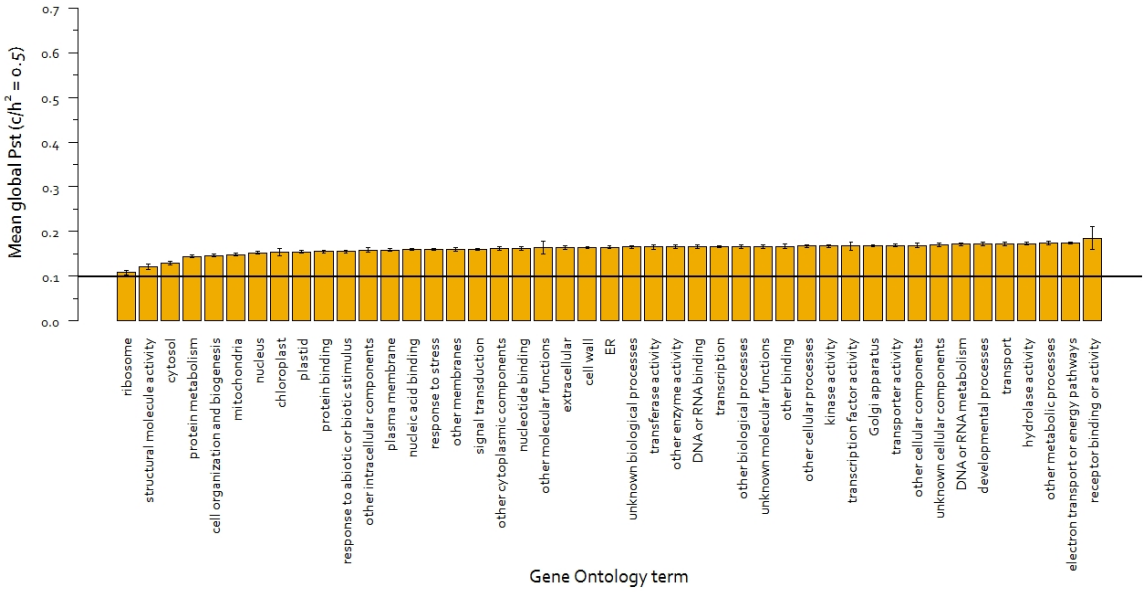
Clone ID	BLASTX vs. Arabidopsis	AGI code	E-value	BLASTX vs. Non-redundant	Organism	E-value	P_{ST}± CI
No hit							
WS0153_M17	no hit	no hit		predicted protein	*Populus trichocarpa*	1.7E-22	-0.58±0.57
E-value > E-10							
WS01125_O15	unknown protein	AT4G11385.1	5.9E-5	unknown	*Populus trichocarpa*	1.3E-19	-1.06±0.60
WS0113_O05	unknown protein	AT5G40855.1	1.4	serpentine receptor, class U family member (sru-2)	*Caenorhabditis elegans*	6.4	-0.75±0.46
WS0208_M03	unknown protein	AT5G57760.1	2.3E-8	transcription factor, putative	*Ricinus communis*	0.69	-0.96±0.97
WS0221_D15	unknown protein	AT4G11385.1	0.45	unknown	*Schistosoma japonicum*	5.1E-6	-0.49±0.51
WS02010_D10	zinc finger (CCCH-type) family protein	AT2G20280.1	3.9E-7	zinc finger (CCCH-type) family protein	*Arabidopsis lyrata*	5.3E-5	-0.97±0.30
Biological process unknown							
WS0185_G07	unknown protein	AT1G16840.4	3.3E-39	predicted protein	*Populus trichocarpa*	1.3E-81	-0.57±0.61
WS0202_G12	unknown protein	AT1G30880.1	6.2E-22	predicted protein	*Populus trichocarpa*	1.2E-53	-0.75±0.50
WS0113_L18	unknown protein	AT2G21870.1	2.0E-80	unknown	*Populus trichocarpa*	5.7E-79	-0.58±0.47
WS0112_O15	unknown protein	AT3G17300.1	2.8E-35	predicted protein	*Populus trichocarpa*	4.3E-42	-0.64±0.66
WS01111_A20	unknown protein	AT3G44960.1	2.0E-16	Shugoshin-1, putative	*Ricinus communis*	2.6E-35	-0.45±0.51
WS0151_C03	calmodulin-binding protein	AT5G57580.1	1.5E-16	calmodulin-binding protein	*Cicer arietinum*	2.2E-22	-0.69±0.72
WS0161_D17	cytochrome c oxidase subunit Vc family protein	AT2G47380.1	6.0E-24	cytochrome c oxidase polypeptide Vc	*Jatropha curcas*	1.7E-24	-0.45±0.46
WS0171_E15	ECA1 gametogenesis related family protein	AT1G44191.1	9.9E-29	salivary proline-rich protein	*Homo sapiens*	1.1E-27	-0.61±0.51
WS0174_J24	located in endomembrane system	AT3G28630.1	1.7E-76	predicted protein	*Populus trichocarpa*	2.4E-128	-0.76±0.71
WS0115_M04	located in vacuole	AT3G23760.1	4.2E-71	transferase, transferring glycosyl groups, putative	*Ricinus communis*	3.7E-68	-0.53±0.58
WS0195_M09	SPla/RYanodine receptor (SPRY) domain-containing protein	AT1G51450.1	2.0E-13	SPla/RYanodine receptor domain-containing protein	*Arabidopsis lyrata*	6.5E-11	-0.75±0.53

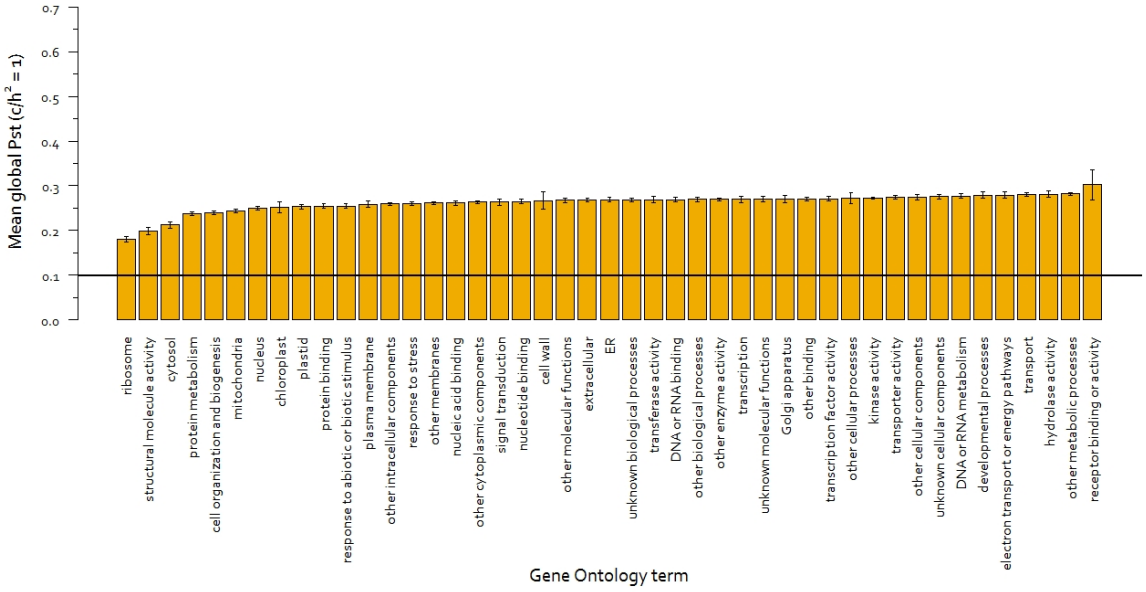
Clone ID	BLASTX vs. Arabidopsis	AGI code	E-value	BLASTX vs. Non-redundant	Organism	E-value	P_{ST}± CI
WS0156_F08	transferase family protein	AT5G17540.1	3.9E-29	benzoyl coenzyme A: benzyl alcohol benzoyl transferase	*Petunia x hybrida*	1.6E-42	-0.42±0.48
Transcription							
PX0011_D15	ZAP1 (zinc-dependent activator protein-1)	AT2G04880.1	1.8E-38	WRKY transcription factor 2	*(Populus tomentosa x P. bolleana) x P. tomentosa*	4.9E-66	-0.57±0.60
Response to stress							
WS0145_M16	TAPX (thylakoidal ascorbate peroxidase)	AT1G77490.1	3.2E-86	chloroplast thylakoid-bound ascorbate peroxidase	*Vigna unguiculata*	3.3E-90	-0.87±0.64
Transport							
PX0019_F18	ATP binding / microtubule motor	AT2G47500.1	6.1E-19	kinesin-related protein	*Gossypium hirsutum*	6.4E-38	-0.42±0.47
Protein metabolism							
WS0194_N02	60S ribosomal protein L13A (RPL13aD)	AT5G48760.1	4.9E-38	60S ribosomal protein L13A	*Arabidopsis lyrata*	7.7E-36	-0.68±0.50
WS0163_N15	mov34 family protein	AT1G48790.1	2.3E-47	mov34 family protein	*Arabidopsis thaliana*	1.2E-44	-0.84±0.51
WS0201_J17	RPS15A (ribosomal protein s15a)	AT1G07770.1	1.3E-12	40S ribosomal protein S15a	*Brassica napus*	7.1E-10	-0.77±0.67
PX0011_C19	TIF3H1; translation initiation factor	AT1G10840.1	4.5E-74	Mov34-1	*Medicago truncatula*	1.5E-73	-0.51±0.55
Other metabolic processes							
WS0185_P11	sugar isomerase (SIS) domain-containing protein/ CBS domain-containing protein	AT3G54690.1	5.2E-105	sugar isomerase (SIS) domain-containing protein / CBS domain-containing protein	*Arabidopsis thaliana*	2.8E-102	-0.68±0.63
Other cellular processes							
WS0168_E07	PFK3 (phosphofructokinase 3)	AT4G26270.1	4.5E-21	phosphofructokinase	*Elaeis oleifera*	3.3E-19	-0.55±0.43

CI: 95% confidence interval

Amongst the genes with BLAST hits in *Arabidopsis* whose E values were < 10^{-10}, almost all GO terms were found to have a mean global P_{ST} value above the global F_{ST} for most values of c/h^2 (Fig. 3.6). The only time any GO terms had a global P_{ST} value below the global F_{ST} was when $c/h^2 = 0.25$, which is the least robust condition. In the group of genes whose expression values were found to be under divergent selection, with hits whose E values were < 10^{-10}, no GO categories were found to be significantly underrepresented (Table 3.5). Three GO categories were found to be significantly overrepresented in the genes under divergent selection ($\alpha = 0.05$); Golgi apparatus, other binding and developmental processes. These results did not hold up when Bonferroni-corrected.




Mean global Pst (c/h^2 = 0.5)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
ribosome
structural molecule activity
cytosol
protein metabolism
cell organization and biogenesis
mitochondria
nucleus
chloroplast
plastid
protein binding
response to abiotic or biotic stimulus
other intracellular components
plasma membrane
nucleic acid binding
response to stress
other membranes
signal transduction
other cytoplasmic components
nucleotide binding
other molecular functions
extracellular
cell wall
ER
unknown biological processes
transferase activity
other enzyme activity
DNA or RNA binding
transcription
other biological processes
unknown molecular functions
other binding
other cellular processes
kinase activity
transcription factor activity
Golgi apparatus
transporter activity
other cellular components
unknown cellular components
DNA or RNA metabolism
developmental processes
transport
hydrolase activity
other metabolic processes
electron transport or energy pathways
receptor binding or activity
Gene Ontology term


Mean global Pst (c/h^2 = 1)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
ribosome
structural molecule activity
cytosol
protein metabolism
cell organization and biogenesis
mitochondria
nucleus
chloroplast
plastid
protein binding
response to abiotic or biotic stimulus
plasma membrane
other intracellular components
response to stress
other membranes
nucleic acid binding
other cytoplasmic components
signal transduction
nucleotide binding
cell wall
other molecular functions
extracellular
ER
unknown biological processes
transferase activity
DNA or RNA binding
other biological processes
other enzyme activity
transcription
unknown molecular functions
Golgi apparatus
other binding
transcription factor activity
other cellular processes
kinase activity
transporter activity
other cellular components
unknown cellular components
DNA or RNA metabolism
developmental processes
electron transport or energy pathways
transport
hydrolase activity
other metabolic processes
receptor binding or activity
Gene Ontology term

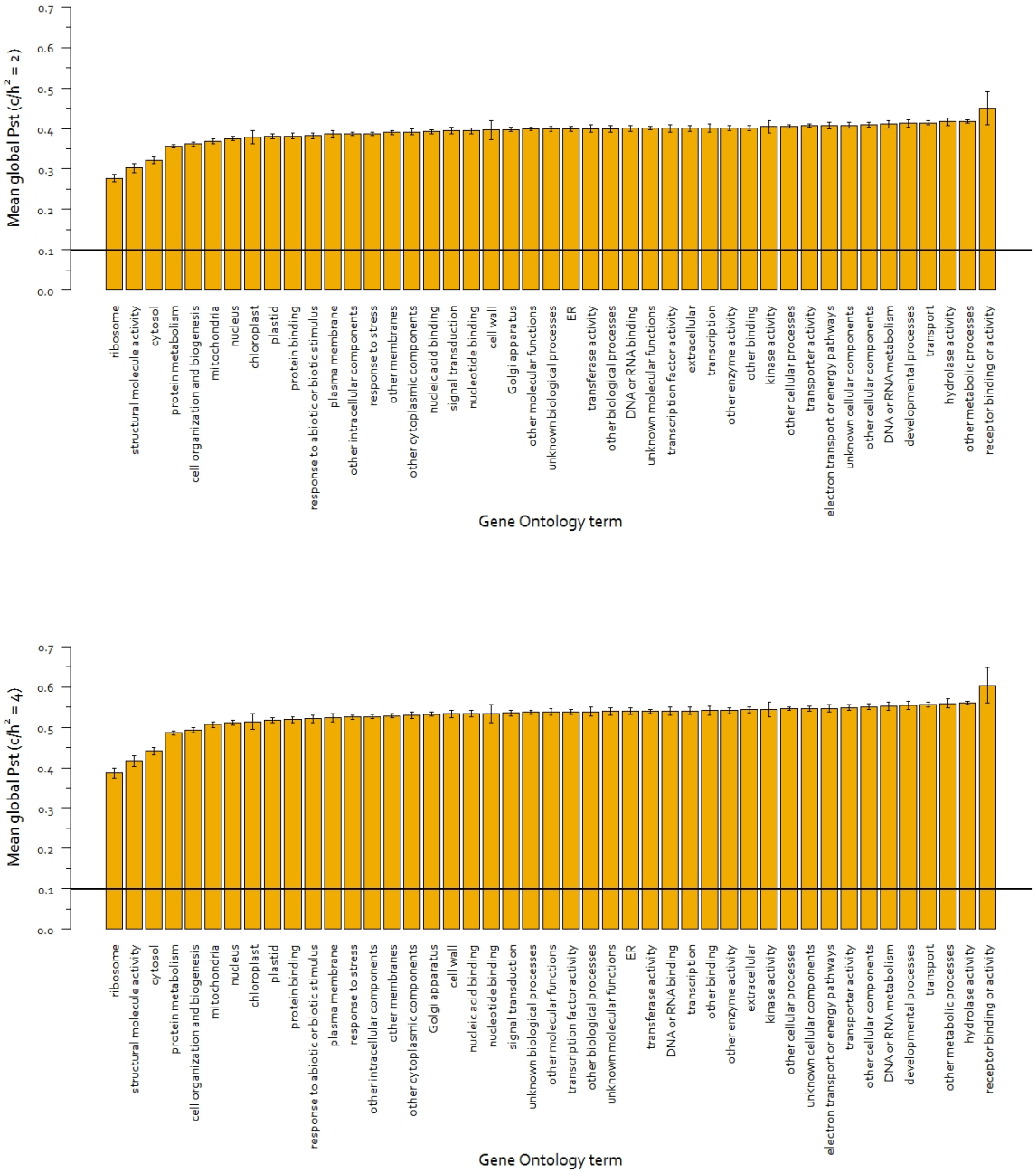


Figure 3.6: Mean P_{ST} for each Gene Ontology category. From top to bottom, P_{ST} is calculated when c/h^2 = 0.25, 0.5, 1, 2 and 4. The error bars are the standard error of the mean. The solid horizontal line represents the global F_{ST} (0.0976). Most GO categories seem to be under divergent selection for all values of c/h^2 excluding 0.25.

3.3.4 Environmental correlations

For the average values of each climate variable for each population location, see Table 3.7. For the groupings of the climate variables and those picked to be the representative variable for further analysis, see Table 3.2. The climate variables were grouped into 14 groups, with two representative variables chosen for one of the groups. Some variables belong to more than one group. Only 4 groups consist of a single variable. These are: May precipitation (mm), June precipitation (mm), September precipitation (mm) and precipitation as snow (mm) between August in the previous year and July in the current year. The representative climate variables were chosen with a preference for choosing annual climate variables.

The genes whose expression values were used for Mantel tests can be found in Table 3.5 and a summary of the results can be found in Table 3.8. Using a false discovery rate (FDR) cutoff of 10%, only a total of three tests were significant, all performed with $c/h^2 = 4$. A false discovery rate cutoff of 10% just means that we would only allow, on average, 10% of the tests deemed significant to be false positives. Two of these were using the same gene's expression data, correlated with May mean precipitation, while correcting for F_{ST} or geographic distance between populations. This gene was found to be an unknown protein in the *Arabidopsis* TAIR database (AT4G11385), although with a very large E-value (0.19). This gene may code for NADH dehydrogenase subunit 2, as detected by the BLAST search against the non-redundant protein sequences, but the E-value was very high (0.93). The other significant test was a Mantel test between another unknown protein (AT5G03670) and the geographic distance between populations. This gene may be a nuclease, but again the E-value was very high (1.5). For both of these genes, it is unlikely that their function can be predicted based on the BLAST results. They may represent proteins that have not yet been characterized in the NCBI database.

Table 3.7: Average values of climate variables at each population location.

Population:	CMBF	DENA	HARB	IRVC	ISKC	KIMB	KLNG	KTMA	LAFY	MCGR	NASH	SLMB
Monthly Mean Temperature (°C)												
January	1.845871	-3.872477	-3.828440	-11.09357	-9.811009	-8.533027	-1.637614	-4.144036	3.788990	-12.04770	-5.399082	2.080733
February	2.894495	-2.527522	-1.309174	-7.601834	-6.394495	-6.065137	-0.151376	-1.834862	5.725688	-7.31559	-3.558715	3.165137
March	4.941284	1.505504	2.98440	-3.290825	-2.300917	-1.613761	2.683486	1.457798	7.724770	-2.416513	0.636697	5.357798
April	7.944036	5.976146	7.656880	2.285321	3.348623	3.799082	6.032110	5.668807	10.12844	4.171559	5.149541	8.199082
May	11.40733	9.659633	11.86146	7.602752	8.302752	8.375229	9.780733	9.722018	13.35963	9.474311	9.497247	11.48899
June	14.49357	12.62110	15.95045	11.57614	12.01009	11.85963	12.70550	13.15137	16.31009	13.40091	12.70183	14.24954
July	16.98256	14.79082	18.63302	13.41743	13.80917	14.5266	14.91926	15.45963	18.94770	15.63119	14.66605	16.57706
August	16.97155	14.77981	18.16146	12.67706	13.23577	13.67706	15.08715	15.33302	18.98990	14.81284	14.55045	16.8146
September	13.60275	11.65045	13.46146	8.780733	9.222018	9.543119	11.99908	11.59541	16.28348	10.12844	10.77706	14.19449
October	8.98440	7.137614	8.135779	2.875229	3.789908	4.325688	7.496330	6.31559	11.76788	4.709174	5.366972	9.655963
November	5.050458	1.200917	1.551376	-4.440366	-3.372477	-2.24587	2.198165	0.922018	7.244036	-2.690825	-0.527522	5.165137
December	2.72293	-1.93853	-3.048623	-9.055963	-8.19266	-6.76146	-1.157798	-2.474311	4.324770	-8.385321	-4.405504	3.090825
Monthly Maximum Mean Temperature (°C)												
January	4.873394	-0.535779	-0.783486	-7.230275	-6.296330	-3.989908	1.311926	-1.536697	7.410091	-6.477981	-2.381651	4.553211
February	6.639449	1.781651	3.294495	-2.860550	-1.956880	-0.604587	3.59266	1.116513	10.17706	-1.609174	0.39266	6.385321
March	9.349541	6.178899	8.428440	1.818348	3.005504	4.217431	7.301834	5.393577	13.22477	3.691743	5.368807	9.578899
April	12.94495	11.31100	14.31926	7.855963	9.067889	10.60091	11.39816	10.6853	16.63394	11.09724	10.78807	13.27798
May	16.78165	15.55321	19.06146	13.87431	14.78807	15.70366	15.48348	15.45137	20.73302	17.09724	16.4733	16.69266
June	20.01559	18.61376	22.56330	17.81743	18.57155	19.01009	18.61376	18.93761	24.29449	20.69908	19.64770	19.39816
July	22.89816	20.95963	26.20275	19.33944	20.06605	22.2559	21.31834	21.06605	28.3266	23.30733	21.50366	22.07706
August	22.82752	20.89541	25.59266	18.5440	19.34678	21.42568	21.46055	20.79174	28.47798	22.58715	21.01100	22.33394
September	18.98256	17.37614	20.28348	13.67431	14.38807	16.73669	17.9293	16.22477	24.76422	17.26605	15.98256	19.42844
October	13.18348	11.5266	13.58348	6.297247	7.278899	9.720183	12.36972	9.23853	18.21743	10.12477	8.828440	13.71743
November	8.186238	4.099082	5.36146	-1.069724	-0.212844	1.860550	5.337614	3.104587	11.78073	1.251376	2.09266	7.866055
December	5.585321	0.962385	0.595412	-5.56146	-4.978899	-2.703669	1.773394	-0.233027	7.870642	-3.903669	-1.804587	5.466972

Population:	CMBF	DENA	HARB	IRVC	ISKC	KIMB	KLNG	KTMA	LAFY	MCGR	NASH	SLMB
Monthly Minimum Mean Temperature (°C)												
January	-1.19266	-7.214678	-6.866972	-14.95321	-13.33302	-13.08073	-4.583486	-6.74587	0.160550	-17.60366	-8.422018	-0.382568
February	-0.83853	-6.836697	-5.901834	-12.33577	-10.83669	-11.52385	-3.896330	-4.778899	1.268807	-13.01559	-7.51559	-0.052293
March	0.526605	-3.164220	-2.44587	-8.404587	-7.606422	-7.449541	-1.936697	-2.499082	2.223853	-8.530275	-4.095412	1.135779
April	2.947706	0.626605	0.982568	-3.280733	-2.380733	-2.996330	0.663302	0.663302	3.627522	-2.766055	-0.48440	3.12293
May	6.05412	3.764220	4.660550	1.257798	1.810091	1.053211	4.066972	4.002752	5.994495	1.83853	2.523853	6.290825
June	8.97706	6.623853	9.342201	5.355963	5.452293	4.699082	6.8	7.372477	8.317431	6.111009	5.741284	9.089908
July	11.0587	8.623853	11.06513	7.505504	7.552293	6.806422	8.510091	9.863302	9.579816	7.958715	7.842201	11.06972
August	11.1266	8.655963	10.7293	6.846788	7.125688	5.925688	8.70733	9.872477	9.503669	7.033027	8.08440	11.29266
September	8.214678	5.922018	6.641284	3.885321	4.058715	2.347706	6.07706	6.991743	7.796330	2.997247	5.559633	8.964220
October	4.781651	2.735779	2.685321	-0.562385	0.294495	-1.066055	2.634862	3.40733	5.328440	-0.704587	1.9	5.603669
November	1.923853	-1.69266	-2.263302	-7.841284	-6.529357	-6.343119	-0.937614	-1.248623	2.723853	-6.633944	-3.139449	2.472477
December	-0.15412	-4.826605	-6.691743	-12.53944	-11.40366	-10.8266	-4.094495	-4.723853	0.778899	-12.85504	-7.00733	0.716513
Monthly Mean Precipitation (mm)												
January	178.0275	207.8899	190.4036	88	53.70642	63.81651	168.1284	246.9816	170.6238	87.88990	96.49541	195.4770
February	125.6146	150.412	148.4770	62.77981	34.8440	44.88073	120.2477	161.0091	123.3486	59.60550	48.69724	175.7247
March	128.8990	131.3394	113.7614	42.4587	27.18348	45.88990	121.6880	146.2752	114.3577	52.33944	34.3853	147.1192
April	71.67889	92.18348	75.88073	35.87155	26.14678	39.71559	74.19266	109.7155	64.68807	37.63302	30.71559	89.53211
May	58.01834	68.05504	53.80733	33.19266	24.98165	55.09174	47.71559	58.98165	48.13761	53.90825	41.37614	78.60550
June	50.11009	61.50458	57.12844	45.41284	31.00917	83.46788	46.57798	53.30275	26.68807	73.5412	49.93577	74.63302
July	39.94495	58.26605	38.59633	76.44954	58.21100	72.58715	51.44954	57.73394	10.56880	59.04587	52.60550	45.66055
August	45.06422	67.25688	43.86238	74.43119	49.99082	70.16513	48.87155	63.33944	15.57798	66.46788	53.00917	65.69724
September	59.34862	130.8073	81.49541	98.62385	81.33027	61.99082	80.24770	116.0642	38.28440	78.86238	80.80733	77.00917
October	146.9724	238.4311	167.706	118.4311	93.11926	64.63302	197.4311	231.853	80.37614	94.32110	114.9633	164.5504
November	214.8990	247.4862	192.4495	119.8440	67.86238	65.75229	216.5321	259.3761	156.8440	88.49541	88.31192	272.0733
December	224.3853	234.559	199.2568	86.66972	65.98165	73.9266	201.3944	259.6605	187.293	88.58715	86.28440	305.0275

Population:	CMBF	DENA	HARB	IRVC	ISKC	KIMB	KLNG	KTMA	LAFY	MCGR	NASH	SLMB
Seasonal Mean Maximum Temperature (°C)												
Winter	5.705504	0.733944	1.037614	-5.201834	-4.411926	-2.433944	2.225688	-0.208256	8.488990	-4	-1.259633	5.472477
Spring	13.02844	11.01743	13.93761	7.871559	8.955045	10.17247	11.39816	10.51009	16.85963	10.63394	10.87522	13.17889
Summer	21.91926	20.15412	24.78440	18.54311	19.33302	20.90091	20.45963	20.26238	27.02844	22.20458	20.7266	21.26880
Autumn	13.44587	11.00091	13.07247	6.321100	7.146788	9.434862	11.87247	9.516513	18.25137	9.544954	8.968807	13.67522
Seasonal Mean Minimum Temperature (°C)												
Winter	-0.725688	-6.291743	-6.488990	-13.27522	-11.85229	-11.81009	-4.190825	-5.429357	0.733944	-14.48807	-7.644954	0.094495
Spring	3.175229	0.408256	1.062385	-3.477981	-2.721100	-3.130275	0.937614	0.723853	3.946788	-3.152293	-0.688990	3.512844
Summer	10.39449	7.967889	10.37247	6.573394	6.711926	5.813761	8.006422	9.036697	9.131192	7.033944	7.224770	10.49082
Autumn	4.972477	2.319266	2.359633	-1.496330	-0.723853	-1.688073	2.583486	3.048623	5.283486	-1.447706	1.440366	5.675229
Seasonal Mean Temperature (°C)												
Winter	2.489908	-2.779816	-2.727522	-9.236697	-8.136697	-7.121100	-0.982568	-2.81559	4.614678	-9.240366	-4.455045	2.77706
Spring	8.101834	5.711926	7.500917	2.195412	3.11559	3.519266	6.162385	5.625688	10.40366	3.740366	5.094495	8.35412
Summer	16.15137	14.06422	17.58715	12.56330	13.02568	13.35321	14.23761	14.65137	18.08256	14.61743	13.97155	15.87706
Autumn	9.213761	6.669724	7.716513	2.406422	3.209174	3.877981	7.234862	6.286238	11.76605	4.049541	5.20733	9.671559
Seasonal Mean Precipitation (mm)												
Winter	528.0091	592.8256	538.0642	237.5412	154.5137	182.5504	489.7522	667.5137	481.3027	236.1376	231.4862	676.2018
Spring	258.5779	291.559	243.3853	111.4954	78.32110	140.7247	243.6513	315.0183	227.293	143.8899	106.4587	315.1834
Summer	135.0642	187.0366	139.5963	196.293	139.1834	226.2018	146.8807	174.3577	52.79816	198.9633	155.559	186.0366
Autumn	421.2293	616.7706	441.6605	336.8623	242.2935	192.4862	494.2477	607.2201	275.4954	261.6513	284.0917	513.6513

Population:	CMBF	DENA	HARB	IRVC	ISKC	KIMB	KLNG	KTMA	LAFY	MCGR	NASH	SLMB
Annual variables												
MAT	8.990825	5.916513	7.521100	1.980733	2.803669	3.404587	6.658715	5.935779	11.2146	3.288990	4.955045	9.170642
MWMT	17.36422	15.35688	19.00091	13.6559	14.07614	14.81651	15.48073	15.95504	19.42018	15.87981	15.13577	17.07981
MCMT	1.404587	-4.756880	-4.308256	-11.61651	-10.3559	-9.33853	-2.188990	-4.728440	3.580733	-12.48073	-6.326605	1.630275
TD	15.95963	20.11376	23.30917	25.27247	24.43211	24.15504	17.66972	20.68348	15.83944	28.36055	21.46238	15.44954
MAP	1342.917	1688.220	1362.733	882.0825	614.3669	741.9541	1374.513	1764.220	1036.761	840.6788	777.6330	1691.183
MSP	252.440	385.8899	274.8899	328.0917	245.4954	343.3486	274.853	349.4678	139.1559	331.8073	277.9174	341.6146
AHM	14.43211	9.647706	13.10917	13.95504	21.37981	18.38440	12.43119	9.235779	21.10091	16.14311	19.67522	11.55229
SHM	73.23394	41.66055	73.39816	44.21926	61.06605	44.97981	59.09633	48.5559	156.6577	49.92201	56.91559	52.83027
DD<0	102.3577	439.9724	427.4036	1218.834	1066.385	930.0733	273.4954	443.9449	50.53211	1152.018	610.1559	94.5412
DD>5	1839.458	1393.155	1909.706	1010.825	1091.880	1145.541	1435.880	1416.825	2459.284	1313.87	1294.119	1872.302
DD<18	3345.284	4444.963	3909.954	5861.981	5565.990	5346.724	4178.394	4440.385	2605.440	5388.908	4790.963	3277.477
DD>18	78.18348	11.88073	143.146	0.467889	1.788990	3.642201	14.09174	24.78899	189.4770	19.65137	11.90825	71.13761
FFP	187.2201	143.9816	142.7706	103.0091	109.2385	87.79816	147.5779	151.6055	196.7614	96.24770	128.5504	199.4954
NFFD	273.6146	208.8715	213.412	154.3302	163.1651	149.1376	219.7614	218.5963	292.3486	156.8990	191.8256	285.412
bFFP	110.2568	132.5963	133.3669	156.7247	152.1284	160.6422	130.7614	131.3761	106.559	155.4954	145.1743	105.9174
eFFP	297.4770	276.5779	276.1376	259.733	261.3669	248.440	278.3394	282.9816	303.3211	251.7431	273.7247	305.412
PAS	91.23853	435.9724	352.559	395.2752	229.4954	255.5963	255.3119	489.3302	33.41284	334.5779	216.706	101.0183
EMT	-19.53302	-29.31100	-29.53853	-39.67889	-38.20183	-37.44587	-25.9853	-28.63211	-16.80458	-39.15321	-31.5146	-17.9733
Eref	667.5137	573.1192	727.266	458.7155	482.5963	578.3669	613.3944	532.0917	953.7522	587.7522	545.6422	647.9357
CMD	256.1192	163.559	323.7614	157.8715	232.1834	192.0733	231.1100	163.1192	574.1284	221.4220	234.2110	178.1926

Abbreviations: MAT, mean annual temperature (°C); MWMT, mean warmest month temperature (°C); MCMT, mean coldest month temperature (°C); TD, temperature difference between MWMT and MCMT (continentality) (°C); MAP, mean annual precipitation (mm); MSP, mean annual summer precipitation (mm); AHM, annual heat:moisture index; SHM, summer heat:moisture index; DD<0, degree-days below 0°C (chilling degree-days); DD>5, degree-days above 5°C (growing degree-days); DD<18, degree-days below 18°C (heating degree-days); DD>18, degree-days above 18°C (cooling degree-days); FFP, frost-free period; NFFD, number of frost-free days; bFFP, Julian date on which FFP begins; eFFP, Julian date on which FFP ends; PAS, precipitation as snow (mm) since August of previous year; EMT, extreme minimum temperature over 30 years; Eref, Hargreaves reference evaporation; CMD, Hargreaves climatic moisture deficit

Table 3.8: Summary of Mantel test results for the environmental correlations.

Correlated with	**Controlling for**	**c/h²**	**Number significant FDR < 10%**	**Number significant FDR < 30%**
geographic distance		2	0	10
		4	1	1
annual heat:moisture index	F_{ST}	2	0	0
		4	0	0
	geographic distance	2	0	0
		4	0	0
Hargreaves climatic moisture deficit	F_{ST}	2	0	0
		4	0	0
	geographic distance	2	0	0
		4	0	0
elevation	F_{ST}	2	0	0
		4	0	0
	geographic distance	2	0	3
		4	0	0
latitude	F_{ST}	2	0	0
		4	0	0
	geographic distance	2	0	0
		4	0	0
longitude	F_{ST}	2	0	0
		4	0	368
	geographic distance	2	0	368
		4	0	368
mean annual precipitation	F_{ST}	2	0	3
		4	0	3
	geographic distance	2	0	4
		4	0	4
mean annual temperature	F_{ST}	2	0	0
		4	0	0
	geographic distance	2	0	0
		4	0	0
mean coldest month temperature	F_{ST}	2	0	0
		4	0	0
	geographic distance	2	0	0
		4	0	0
mean warmest month temperature	F_{ST}	2	0	0
		4	0	0
	geographic distance	2	0	0
		4	0	0

Correlated with	Controlling for	c/h²	Number significant FDR < 10%	Number significant FDR < 30%
number of frost-free days	F_{ST}	2	0	0
		4	0	0
	geographic distance	2	0	0
		4	0	0
precipitation as snow	F_{ST}	2	0	0
		4	0	0
	geographic distance	2	0	0
		4	0	0
May mean precipitation	F_{ST}	2	0	12
		4	1	7
	geographic distance	2	0	19
		4	1	12
June mean precipitation	F_{ST}	2	0	0
		4	0	0
	geographic distance	2	0	0
		4	0	0
September mean precipitation	F_{ST}	2	0	0
		4	0	0
	geographic distance	2	0	0
		4	0	0
spring mean precipitation	F_{ST}	2	0	89
		4	0	0
	geographic distance	2	0	282
		4	0	75
region	F_{ST}	2	0	0
		4	0	0
	geographic distance	2	0	0
		4	0	0
summer heat:moisture index	F_{ST}	2	0	0
		4	0	0
	geographic distance	2	0	0
		4	0	0
continentality	F_{ST}	2	0	0
		4	0	0
	geographic distance	2	0	0
		4	0	0
summer mean minimum temperature	F_{ST}	2	0	0
		4	0	0
	geographic distance	2	0	1
		4	0	0

By increasing the allowed FDR to 30%, many more tests became significant. All of them seem significant for three out of four of the correlations with longitude. The one correlation out of the four with none significant is because all of the Q values were slightly over 0.3. This is simply an artefact of the cutoff value. In general, genes' expression that were found to correlate with a given environmental variable for one set of c/h^2 and accounting for either F_{ST} or geographic distance, tended to correlate with the other c/h^2 and accounting for the other factor.

For the Mantel tests with geographic distance, 10 tests were significant with an FDR of 30%. Out of these, only six were genes with matches in *Arabidopsis* with E-values $< 10^{-10}$. These were two unknown proteins (AT5G03670, AT5G10780), a pectate lyase family protein, CDPK19 (calcium-dependent protein kinase 19), ATAAH (an allantoate deiminase) and LHT2 (lysine histidine transporter 2).

From the partial Mantel tests with elevation, while controlling for geographic distance, three tests were significant with an FDR of 30%. These were QQT1 (quatre-quart 1) - an ATP and nucleotide binding protein, ATRZ-1A - another RNA and nucleotide binding protein, and the 60S ribosomal protein L35a. From those with mean annual precipitation, three were significant while controlling for pairwise F_{ST}, and four were significant while controlling for geographic distance. Out of the three genes, only two had E-values of 10^{-10} or less, which were ACHT4 (atypical cys his rich thioredoxin 4) and QQT1. The only additional gene that was significant while controlling for geographic distance had a large E-value. For the partial Mantel tests with May mean precipitation, with an FDR of 30%, there were 9, 12 or 19 tests that were significant, depending on the value of c/h^2 and what was controlled for. All of the genes that were significant for the 9 and 12 were significant in the 19. Out of these, only 11 had E-values of 10^{-10} or less. These were ATFD3 (ferre-

doxin 3), an unknown protein, another unknown protein located in the extracellular region, a cell cycle control related-protein, a ribosomal protein L3 family protein, a pfkB-type carbohydrate kinase family protein, APO2 (accumulation of photosystem one 2), LHT2 (lysine histidine transporter 2), CLE44 (clavata3/ ESR-related 44), CYP5 (cyclophilin 5), a peptidyl-prolyl cis-trans isomerase, and a myb family transcription factor. The partial Mantel tests with spring mean precipitation had a varying number of significant tests with an FDR of 30%, where most genes were significant in some cases. One test with summer mean minimum temperature was significant with $c/h^2 = 2$ while controlling for geographic distance, which was a histone H3.2.

3.3.5 Candidate genes

Out of the candidate genes, only *PHYB* and *LHY* had orthologs on the microarray. There was one *PHYB* gene and two *LHY* genes. Expression of *PHYB*, which encodes a phytochrome photoreceptor, appears to be under divergent selection with $c/h^2 = 2$ and 4. Out of the two *LHY* genes, circadian clock-associated genes, expression of one of the *LHY* genes appears to be under stabilizing selection when $c/h^2 = 0.5$, 0.25 and the other LHY gene seems to be diverging neutrally.

3.4 Discussion

Black cottonwood is broadly distributed along the coast of western North America with a "no-cottonwood" belt proposed to separate a northern group from a southern group in BC. We expected to find some divergent selection between these two groups, as well as many genes under stabilizing selection. From our genetic data, we have found a relatively low level of population differentiation in alleles, as measured by F_{ST}. Our global F_{ST} over all

populations was estimated as 0.0976. This generally agrees with previous studies of black cottonwood which have found F_{ST} for allozymes as 0.063 (Weber and Stettler, 1981) and F_{ST} = 0.078 and R_{ST} = 0.112 with microsatellite markers (Ismail, 2010). This is also in agreement with the differentiation among populations of *Populus* in general, which is typically weak. The median F_{ST} for the genus is 0.047, as measured by allozymes and RFLPs (Slavov and Zhelev, 2010). Studies of gene flow suggest that long-distance pollination can be extensive in *Populus* (Slavov and Zhelev, 2010). This and long-distance seed dispersal may account for the low population differentiation.

The lack of isolation by distance may indicate that gene flow is not restricted across the range and that the "no-cottonwood" belt may not be separating the populations into two genetic groups. This is supported by the neighbour-joining tree constructed using the pairwise F_{ST} data (Fig. 3.4), as there are no discernable patterns to the clustering. These results could also be an artefact of the sampling, but this is unlikely due to the strong support for no isolation by distance (Mantel test P = 1). More likely is that long-distance pollination or seed dispersal is common. This could also be the result of the populations having gone through recent bottlenecks. These bottlenecks could be due to the strong selection that takes place at the seedling stage, as mortality in the first year can be from 77-100% in *Populus* (Slavov and Zhelev, 2010).

The use of P_{ST} as an approximation of Q_{ST} is suited for exploratory studies of quantitative traits. The estimation of P_{ST} is usually based on phenotypic measures of a trait in the wild in several individuals across a number of populations (Brommer, 2011). The trees sampled were grown in a common environment, but were grown from cuttings from natural individuals and thus could have had residual effects from their natural environments (e.g. epigenetic effects, Raj et al. 2011). One challenge from Q_{ST} and F_{ST} comparisons is that the F_{ST} for neutral loci and Q_{ST} for neutral traits are expected to be extremely variable (Whitlock, 2008). Whitlock (2008) suggests that one should show that the Q_{ST} value is

greater than the global F_{ST} and that it is in the tail of the distribution to have evidence for divergent selection. By taking very conservative estimates of which genes' expression values were under selection, we also attempted to choose only those expression profiles that were in the tail of the distribution. We did this by using the most conservative values of c/h^2, 0.25 for divergent selection and 4 for stabilizing selection, as well as considering the 95% jackknife confidence intervals.

In general, we found that the distribution of P_{ST} values peaks around the global F_{ST} value for the lower values of c/h^2, but as c/h^2increased, the distribution widened and shifted to the right (Fig. 3.5). We also found that, even with lower values of c/h^2, divergent selection may be more prevalent than stabilizing selection on gene expression among these populations. More expression patterns show evidence for divergent selection over a broader range of c/h^2 (Table 3.4), with 368 (2.37%) genes always having evidence for divergent selection vs. 27 (0.17%) for stabilizing selection. This may have something to do with the fact that we are using population comparisons instead of species comparisons, as there has been less time since divergence and, therefore, they are more likely to be primarily affected by drift rather than stabilizing selection. This is because drift and stabilizing selection interact to diversify or constrain variation and this interaction is more complex as phylogenetic distance increases (Whitehead and Crawford, 2006b). For shorter phylogenetic distances, drift should drive linear divergence over time before it reaches the bounds set by stabilizing selection (Whitehead and Crawford, 2006b).

When looking at the GO categories of all genes under consideration, we found that all GO categories had a mean global P_{ST} value above the global F_{ST} for most values of c/h^2. This may support the ubiquity of natural selection on traits, but there are a few caveats. First, these are only mean values and don't take the 95% confidence interval into account. Second, as mentioned above, each gene can belong to multiple GO categories, and many do. Due to this, the means for each GO category are not independent of each other and

cannot be directly compared to one another. Unfortunately, many of the proteins which appear to be under selection are unknown, and therefore do not have GO annotation. Out of those that have a good match in *Arabidopsis*, no GO categories were significantly underrepresented when compared to the microarray as a whole, and only three were found to be overrepresented. These were Golgi apparatus, other binding and developmental processes, and the results did not hold up to Bonferroni correction. Still, this may indicate a stronger selection on genes involved in this component, function and process.

As mentioned in the methods section, it is common to assume that c = 1 and that h^2 = 0.25 or 0.5 (Brommer, 2011). Also, since the trees were raised in a common environment, c = 1 may be a valid assumption as this should eliminate much of the variance due to environmental factors. We cannot say this with certainty, however, since the trees grown were taken from cuttings from natural populations, which may affect their individual environmental history (Raj et al., 2011). Also, other effects may play a role, like the response to a novel environment, which may be idiosyncratic (Whitlock, 2008). If we are to say that c = 1, we can consider our range of c/h^2 to really be a range of h^2 where h^2 = 4, 2, 1, 0.5 and 0.25. Theoretically, heritability should not exceed 1, and values closer to 0.5 and 0.25 are more commonly assumed. Significant heritable variation in gene expression has been found in yeast, mice and humans, with h^2 being around or over 0.3 (Whitehead and Crawford, 2006b). In trees, the heritability of gene expression has also been investigated in the terpenoid pathways of Interior spruce (*Picea glauca x engelmannii*) (Albouyeh and Ritland, 2011). In any given pathway segment, the median heritability was always found to be above 0.4 (Albouyeh and Ritland, 2011). These agree most closely with h^2 = 0.25 and 0.5, so values of P_{ST} calculated with c/h^2 = 2 and 4 may be the most realistic.

For this reason, we chose to run our environmental correlations with the P_{ST} values calculated when c/h^2 = 2 and 4. We found that only three tests were significant at a FDR cutoff of 10%. These were two partial Mantel tests using the same gene's expression data

correlated with May mean precipitation while correcting for F_{ST} or the geographic distance between populations, respectively. This gene's best match in *Arabidopsis* was an unknown protein (AT4G11385), but this match was poorly supported. The best match for this gene in the non-redundant protein database was an NADH dehydrogenase subunit 2, but this match was also poorly supported. The only other significant test with a FDR of 10% was a Mantel test between the expression of another gene and the geographic distance between populations. This gene was found to have a good match in *Arabidopsis*, an unknown protein, but no good matches in the non-redundant protein database with informative biological functions. The best was a nuclease, but this match was not well supported (E = 1.5). These genes may warrant further investigation, as they are the ones most likely to be under selection.

By increasing the allowed FDR to 30%, many more tests become significant. Generally, when one gene's expression is found to correlate with a given environmental variable under one set of test conditions, it is likely to correlate under the other conditions. All of the tests seem significant for three out of four of the types of partial correlations with longitude. The other type has none significant, but this is an artefact of the cutoff value. All of the Q values for this set of tests were slightly over 0.3, but not by much. The partial Mantel tests with spring mean precipitation also had a varying number of significant tests, but with most genes being significant in some cases. Partial Mantel tests with May mean precipitation had from 9 to 19 significant tests and Mantel tests with geographic distance had at most 10. Partial Mantel tests with mean annual precipitation had up to four, with elevation had three and with summer mean minimum temperature had one. This may indicate that longitude, spring mean precipitation, May mean precipitation and geographic distance have the largest effect on selection on gene expression in *P. trichocarpa*, with mean annual precipitation, elevation and summer mean minimum temperature playing a smaller, but important, role.

We also investigated three candidate genes found to be under selection in previous

studies of the European aspen, *Populus tremula*. These were *PHYB*, a phytochrome photoreceptor, and two *LHY* genes, circadian clock-associated genes (Ingvarsson et al., 2006; Ma et al., 2010). We found that *PHYB* may be under divergent selection, as its P_{ST} value is greater than the global F_{ST} value when $c/h^2 = 2$ and 4. One of the *LHY* genes seems to be diverging neutrally while the other has some evidence for stabilizing selection, when $c/h^2 = 0.5$ and 0.25. This evidence for selection is weak, as it all comes from the least robust values of c/h^2. However, if we consider that $c/h^2 = 2$ or 4 may be the true value of this parameter, the evidence for divergent selection acting on *PHYB* may be valid.

Large-scale gene expression experiments like these are only a jumping-off point into the investigation of selection acting on these trees. We have found evidence for divergent selection acting on the expression values of many genes in black cottonwood, as well as stabilizing selection acting on a few. This supports the prevalence of natural selection acting on phenotypic traits, but we still find that an overwhelming majority of expression values cannot be conclusively shown to be under selection. Also, based on our results, it would appear that the dominance of stabilizing selection detected in other studies of gene expression evolution may be the result of drift having reached the bounds set by selection. This is more likely to occur when making comparisons between more phylogenetically distant taxa. We hope that this study will generate hypotheses that can now be investigated further by other techniques, complementing those that we have used.

Chapter 4

Conclusions

There are an increasing number of studies into the role that gene expression plays in evolution. These studies follow the observation by King and Wilson (1975) that the phenotypic variation seen in nature cannot be explained by the variation in protein coding sequences alone. Gene expression is not what usually comes to mind when the word 'phenotype' is used in evolutionary studies, but it is one that can be investigated as any other. Evolutionary biologists are interested in what kinds of changes lead to better adapted organisms. Whether these changes are typically structural in nature or lead to changes in the level of gene expression is one of the major questions in evolutionary biology. Another is whether changes in general, produced by mutation, are typically beneficial (increasing the organism's fitness), deleterious (decreasing fitness) or neutral in nature.

Our study of the evolutionary processes affecting gene expression in the tree *Populus trichocarpa* (black cottonwood) tries to address an aspect of these questions. We found that an overwhelming majority of genes on our microarray had expression levels that could not be distinguished from neutral divergence in the populations investigated. Out of those under selection, we found that many more seemed to be under divergent selection than stabilizing selection. Divergent selection is evidence of there being a beneficial change in at least one of the populations studied while stabilizing selection is evidence of a previous deleterious change. In our populations, there seems to be evidence for beneficial mutations in the expression level of more genes than there is evidence for deleterious mutations.

Part of the reason for the overwhelming majority of neutral expression values may

be our very stringent conditions for evaluating selection. These were used along with a neutral model, assumed to be the null hypothesis. This assumption has been supported by many studies which have found evidence for the mostly neutral accumulation of expression differences with time, using a few different methods (Whitehead and Crawford, 2006b,a; Yanai et al., 2004; Khaitovich et al., 2004; Rifkin et al., 2003; Oleksiak et al., 2002). It is always more difficult to reject the null hypothesis than accept it, and our stringent conditions make this even more difficult. This may make our study better suited to comparing the relative effects of stabilizing and divergent selection because the conditions to accept either of them were difficult. Even between these two categories, the acceptance rate compared to biological reality may not have been equal, however, because we do not know the true value of c/h^2. If we assume that the true value of this parameter is either 2 or 4, as is commonly assumed (Brommer, 2011), the conditions for accepting divergent selection were even more stringent than those for accepting stabilizing selection. This may mean that divergent selection plays an even larger role in the evolution of gene expression than was originally stated.

The conclusions of studies of gene expression evolution may be highly dependent on the phylogenetic distance between the taxa used for comparison. By using population instead of species comparisons, we wish to fill in some of the gaps in the literature. This is important because, for example, stabilizing selection may be more likely to be detected when comparing more distantly related taxa, such as species, than when comparing populations as the assumptions of neutral divergence due to drift are more likely to have broken down. In other words, as the phylogenetic distance increases, it is more likely that a trait will have diverged enough to hit the boundaries set by stabilizing selection, and gene expression is a trait that is very unlikely to be able to drift infinitely in any direction. At the same time, using taxa that are too closely related, such as lab strains, may not accurately represent questions about natural selection due to environmental pressures.

We also hope that our study will contribute to the literature on methods. There is no real standardized method of detecting evolutionary forces (drift and selection) in gene expression levels, but we believe that the Q_{ST} vs. F_{ST} approach may be a useful one. It has been used for many other quantitative traits, and can be applied to expression data. This has been done before to a limited extent (Kohn et al., 2008; Roberge et al., 2007), but has not been widely accepted in the field.

Ultimately, the conclusions of our study may be most readily applied to future studies of black cottonwood. *P. trichocarpa* is an important commercial species, and further investigation in to many of the genes for which we have found evidence for selection may help develop a better plantation tree, adapted and able to thrive in local conditions. Future investigation of these genes can be performed using other techniques, such as association with phenotypes of interest and investigation of biochemical function, in order to determine their role in the adaptation of black cottonwood to its range in western North America.

Bibliography

Albouyeh, R. and Ritland, K. (2011). Heritability and species divergence for gene expression of spruce terpenoids are highly correlated, indicating adaptive divergence of a key gene family involved in insect defense. Submitted to Journal of Heredity.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool, *J. Mol. Biol.* **215**: 403–410.

Brommer, J. E. (2011). Whither Pst? The approximation of Qst by Pst in evolutionary and conservation biology, *J. Evol. Biol.* **24**: 1160–1168.

Brunner, A. M., Busov, V. B. and Strauss, S. H. (2004). Poplar genome sequence: functional genomics in an ecologically dominant plant species, *Trends Plant Sci.* **9**: 49–56.

DeBell, D. S. (1990). Populus trichocarpa Torr. & Gray.; Black Cottonwood, *in* R. M. Burns and B. H. Honkala (eds), *Silvics of North America: Volume 2. Hardwoods, Agriculture Handbook 654*, United States Department of Agriculture (USDA), Forest Service, pp. 570–576.

Denver, D. R., Morris, K., Streelman, J. T., Kim, S. K., Lynch, M. and Thomas, W. K. (2005). The transcriptional consequences of mutation and natural selection in Caenorhabditis elegans, *Nat. Genet.* **37**: 544–548.

Eckenwalder, J. E. (1996). Systematics and evolution of Populus, *in* R. F. Stettler, H. D. Jr. Bradshaw, P. E. Heilman and T. M. Hinckley (eds), *Biology of Populus and its im-*

plications for management and conservation, NRC Research press, National Research Council of Canada, Ottawa, Ontario, Canada, chapter 1, pp. 7–32.

Fay, J. C. and Wittkopp, P. J. (2008). Evaluating the role of natural selection in the evolution of gene regulation, *Heredity* **100**: 191–199.

Felsenstein, J. (1985). Phylogenies and the comparative method, *The American Naturalist* **125**: 1–15.

Geraldes, A., Pang, J., Thiessen, N., Cezard, T., Moore, R., Zhao, Y., Tam, A., Wang, S., Friedmann, M., Birol, I., Jones, S. J., Cronk, Q. C. and Douglas, C. J. (2011). SNP discovery in black cottonwood (Populus trichocarpa) by population transcriptome resequencing, *Mol Ecol Resour* **11 Suppl 1**: 81–92.

Gibson, G. and Weir, B. (2005). The quantitative genetics of transcription, *Trends in Genetics* **21**(11): 616–623.

Gilad, Y., Oshlack, A. and Rifkin, S. A. (2006a). Natural selection on gene expression, *Trends Genet.* **22**: 456–461.

Gilad, Y., Oshlack, A., Smyth, G. K., Speed, T. P. and White, K. P. (2006b). Expression profiling in primates reveals a rapid evolution of human transcription factors, *Nature* **440**: 242–245.

Hall, D., Luquez, V., Garcia, V. M., St Onge, K. R., Jansson, S. and Ingvarsson, P. K. (2007). Adaptive population differentiation in phenology across a latitudinal gradient in European aspen (Populus tremula, L.): a comparison of neutral markers, candidate genes and phenotypic traits, *Evolution* **61**: 2849–2860.

Hamann, A., El-Kassaby, Y. A., Koshy, M. and Namkoong, G. (1998). Multivariate analy-

sis of allozymic and quantitative trait variation in Alnus rubra: geographic patterns and evolutionary implications, *Canadian Journal of Forest Research* **28**: 1557–1565.

Hamrick, J. L., Godt, M. J. W. and Sherman-Broyles, S. L. (1992). Factors influencing levels of genetic diversity in woody plant species, *New Forests* **6**(1-4): 95–124.

Hoekstra, H. E. and Coyne, J. A. (2007). The locus of evolution: evo devo and the genetics of adaptation, *Evolution* **61**: 995–1016.

Holliday, J. A., Ralph, S. G., White, R., Bohlmann, J. and Aitken, S. N. (2008). Global monitoring of autumn gene expression within and among phenotypically divergent populations of Sitka spruce (Picea sitchensis), *New Phytol.* **178**: 103–122.

Hsieh, W. P., Chu, T. M., Wolfinger, R. D. and Gibson, G. (2003). Mixed-model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles, *Genetics* **165**: 747–757.

Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A. and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression, *Bioinformatics* **18 Suppl 1**: 96–104.

Ingvarsson, P. K., Garcia, M. V., Hall, D., Luquez, V. and Jansson, S. (2006). Clinal variation in phyB2, a candidate gene for day-length-induced growth cessation and bud set, across a latitudinal gradient in European aspen (Populus tremula), *Genetics* **172**: 1845–1853.

Ismail, M. (2010). *Molecular genetic diversity among natural populations of Populus*, PhD thesis, University of British Columbia.

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

Kerr, M. K. (2003). Design considerations for efficient and effective microarray studies, *Biometrics* **59**: 822–828.

Kerr, M. K. and Churchill, G. A. (2001). Statistical design and the analysis of gene expression microarray data, *Genet. Res.* **77**: 123–128.

Khaitovich, P., Weiss, G., Lachmann, M., Hellmann, I., Enard, W., Muetzel, B., Wirkner, U., Ansorge, W. and Paabo, S. (2004). A neutral model of transcriptome evolution, *PLoS Biol.* **2**: E132.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*, Cambridge University Press.

King, M. C. and Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees, *Science* **188**: 107–116.

Kohn, M. H., Shapiro, J. and Wu, C. I. (2008). Decoupled differentiation of gene expression and coding sequence among Drosophila populations, *Genes Genet. Syst.* **83**: 265–273.

Kolosova, N., Miller, B., Ralph, S., Ellis, B. E., Douglas, C., Ritland, K. and Bohlmann, J. (2004). Isolation of high-quality RNA from gymnosperm and angiosperm trees, *BioTechniques* **36**: 821–824.

Larson, P. R. and Isebrands, J. G. (1971). The plastochron index as applied to developmental studies of cottonwood, *Canadian Journal of Forest Research* **1**: 1–11.

Leinonen, T., Cano, J. M., Makinen, H. and Merila, J. (2006). Contrasting patterns of body shape and neutral genetic divergence in marine and lake populations of threespine sticklebacks, *J. Evol. Biol.* **19**: 1803–1812.

Lemos, B., Meiklejohn, C. D., Caceres, M. and Hartl, D. L. (2005). Rates of divergence in gene expression profiles of primates, mice, and flies: stabilizing selection and variability among functional categories, *Evolution* **59**: 126–137.

Little, E. L. (1971). Atlas of united states trees, volume 1, conifers and important hardwoods, U.S. Department of Agriculture Miscellaneous Publication 1146. 9 p., 200 maps.

Lynch, M. and Hill, W. G. (1986). Phenotypic evolution by neutral mutation, *Evolution* **40**: 915–935.

Ma, X. F., Hall, D., Onge, K. R., Jansson, S. and Ingvarsson, P. K. (2010). Genetic differentiation, clinal variation and phenotypic associations with growth cessation across the Populus tremula photoperiodic pathway, *Genetics* **186**: 1033–1044.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach, *Cancer Res.* **27**: 209–220.

McLetchie, D. N. and Tuskan, G. A. (1994). Gender determination in Populus, *Norw. J. Agric. Sci.* **18**: 57–66.

Merila, J. and Crnokrak, P. (2001). Comparison of genetic differentiation at marker loci and quantitative traits, *Journal of Evolutionary Biology* **14**: 892–903.

Miranda, M., Ralph, S. G., Mellway, R., White, R., Heath, M. C., Bohlmann, J. and Constabel, C. P. (2007). The transcriptional response of hybrid poplar (Populus trichocarpa x P. deltoides) to infection by Melampsora medusae leaf rust involves induction of flavonoid pathway genes leading to the accumulation of proanthocyanidins, *Mol. Plant Microbe Interact.* **20**: 816–831.

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

Oleksiak, M. F., Churchill, G. A. and Crawford, D. L. (2002). Variation in gene expression within and among natural populations, *Nat. Genet.* **32**: 261–266.

Olsen, J. E., Junttila, O., Nilsen, J., Eriksson, M. E., Martinussen, I., Olsson, O., Sandberg, G. and Moritz, T. (1997). Ectopic expression of oat phytochrome A in hybrid aspen changes critical daylength for growth and prevents cold acclimatization, *Plant Journal* **12**: 1339–1350.

Osier, T. L. and Lindroth, R. L. (2006). Genotype and environment determine allocation to and costs of resistance in quaking aspen, *Oecologia* **148**: 293–303.

Pan, X., Gilkes, N., Kadla, J., Pye, K., Saka, S., Gregg, D., Ehara, K., Xie, D., Lam, D. and Saddler, J. (2006). Bioconversion of hybrid poplar to ethanol and co-products using an organosolv fractionation process: optimization of process yields, *Biotechnol. Bioeng.* **94**: 851–861.

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

Philippe, R. N. and Bohlmann, J. (2007). Poplar defense against insect herbivores, *Canadian Journal of Botany* **85**: 1111–1126.

Pujol, B., Wilson, A. J., Ross, R. I. and Pannell, J. R. (2008). Are Q(ST)-F(ST) comparisons for natural populations meaningful?, *Mol. Ecol.* **17**: 4782–4785.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. **URL:** *http://www.R-project.org*

Raj, S., Brautigam, K., Hamanishi, E. T., Wilkins, O., Thomas, B. R., Schroeder, W.,

Mansfield, S. D., Plant, A. L. and Campbell, M. M. (2011). Clone history shapes Populus drought responses, *Proc. Natl. Acad. Sci. U.S.A.* **108**: 12521–12526.

Ralph, S., Oddy, C., Cooper, D., Yueh, H., Jancsik, S., Kolosova, N., Philippe, R. N., Aeschliman, D., White, R., Huber, D., Ritland, C. E., Benoit, F., Rigby, T., Nantel, A., Butterfield, Y. S., Kirkpatrick, R., Chun, E., Liu, J., Palmquist, D., Wynhoven, B., Stott, J., Yang, G., Barber, S., Holt, R. A., Siddiqui, A., Jones, S. J., Marra, M. A., Ellis, B. E., Douglas, C. J., Ritland, K. and Bohlmann, J. (2006). Genomics of hybrid poplar (Populus trichocarpa x deltoides) interacting with forest tent caterpillars (Malacosoma disstria): normalized and full-length cDNA libraries, expressed sequence tags, and a cDNA microarray for the study of insect-induced defences in poplar, *Mol. Ecol.* **15**: 1275–1297.

Rifkin, S. A., Houle, D., Kim, J. and White, K. P. (2005). A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression, *Nature* **438**: 220–223.

Rifkin, S. A., Kim, J. and White, K. P. (2003). Evolution of gene expression in the Drosophila melanogaster subgroup, *Nat. Genet.* **33**: 138–144.

Ritland, K. (1996). Estimators for pairwise relatedness and individual inbreeding coefficients, *Genetical Research* **67**: 175–185.

Ritland, K. (2000). Marker-inferred relatedness as a tool for detecting heritability in nature, *Molecular Ecology* **9**: 1195–1204.

Roberge, C., Guderley, H. and Bernatchez, L. (2007). Genomewide identification of genes under directional selection: gene transcription Q(ST) scan in diverging Atlantic salmon subpopulations, *Genetics* **177**: 1011–1022.

Rosenberg, M. and Anderson, C. D. (2011). Passage: Pattern analysis, spatial statistics and geographic exegesis. version 2., *Methods in Ecology and Evolution* **2**: 229–232.

Rousset, F. (1997). Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance, *Genetics* **145**: 1219–1228.

Saether, S. A., Fiske, P., Kalas, J. A., Kuresoo, A., Luigujoe, L., Piertney, S. B., Sahlman, T. and Hoglund, J. (2007). Inferring local adaptation from QST-FST comparisons: neutral genetic and quantitative trait variation in European populations of great snipe, *J. Evol. Biol.* **20**: 1563–1576.

Shao, J. and Wu, C. F. J. (1989). A general theory for jackknife variance estimation, *The Annals of Statistics* **17**: 1176–1197.

Shiu, S. H. and Borevitz, J. O. (2008). The next generation of microarray research: applications in evolutionary and ecological genomics, *Heredity* **100**: 141–149.

Sjodin, A., Wissel, K., Bylesjo, M., Trygg, J. and Jansson, S. (2008). Global expression profiling in leaves of free-growing aspen, *BMC Plant Biology* **8**: 61.

Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies, *Genetics* **139**: 457–462.

Slavov, G. T., Leonardi, S., Burczyk, J., Adams, W. T., Strauss, S. H. and DiFazio, S. P. (2009). Extensive pollen flow in two ecologically contrasting populations of Populus trichocarpa, *Molecular Ecology* **18**: 357–373.

Slavov, G. T. and Zhelev, P. (2010). Salient biological features, systematics, and genetic variation of Populus, *in* S. Jansson, R. Bhalerao and A. T. Groover (eds), *Genetics and Genomics of Populus*, Springer NY, pp. 15–38.

Smouse, P. E., Long, J. C. and Sokal, R. R. (1986). Multiple regression and correlation extensions of the Mantel test of matrix correspondence, *Systematic Zoology* **35**: 627–632.

Soltis, D. E., Gitzendanner, M. A., Strenge, D. D. and Soltis, P. S. (1997). Chloroplast DNA intraspecific phylogeography of plants from the Pacific Northwest of North America, *Plant Systematics and Evolution* **206**: 353–373.

Spitze, K. (1993). Population structure in Daphnia obtusa: quantitative genetic and allozymic variation, *Genetics* **135**: 367–374.

Stamatoyannopoulos, J. A. (2004). The genomics of gene expression, *Genomics* **84**(3): 449–457.

Storey, J. D., Taylor, J. E. and Siegmund, D. (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach, *Journal of the Royal Statistical Society, Series B* **66**: 187–205.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genome-wide studies, *Proceedings of the National Academy of Sciences* **100**: 9440–9445.

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., Radenbaugh, A., Singh, S., Swing, V., Tissier, C., Zhang, P. and Huala, E. (2008). The Arabidopsis Information Resource (TAIR): gene structure and function annotation, *Nucleic Acids Res.* **36**: D1009–1014.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. and Kumar, S. (2011). Mega5: Molecular evolutionary genetics analysis using likelihood, distance, and parsimony methods. Molecular Biology and Evolution. (to be submitted).

The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology, *Nature Genetics* **25**: 25–29.

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., 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.

Wang, T., Hamann, A., Spittlehouse, D. and Aitken, S. N. (2006). Development of scale-free climate data for western Canada for use in resource management, *International Journal of Climatology* **26**: 383–397.

Weber, J. C. and Stettler, R. F. (1981). Isoenzyme variation among ten populations of Populus trichocarpa Torr. et Gray in the Pacific Northwest, *Silvae Genetica* **30**: 2–3.

Weber, J. C., Stettler, R. F. and Heilman., P. E. (1985). Genetic variation and productivity of Populus trichocarpa and its hybrids. I. Morphology and phenology of 50 native clones, *Canadian Journal of Forest Research* **15**: 376–383.

Whitehead, A. and Crawford, D. L. (2006a). Neutral and adaptive variation in gene expression, *Proc. Natl. Acad. Sci. U.S.A.* **103**: 5425–5430.

Whitehead, A. and Crawford, D. L. (2006b). Variation within and among species in gene expression: raw material for evolution, *Mol. Ecol.* **15**: 1197–1211.

Whitlock, M. C. (1999). Neutral additive genetic variance in a metapopulation, *Genetical Research* **74**: 215–221.

Whitlock, M. C. (2008). Evolutionary inference from QST, *Mol. Ecol.* **17**: 1885–1896.

Whitlock, M. C. (2011). G'ST and D do not replace FST, *Mol. Ecol.* **20**: 1083–1091.

Xie, C.-Y., Ying, C. C., Yanchuk, A. D. and Holowachuk, D. L. (2009). Ecotypic mode of regional differentiation caused by restricted gene migration: a case in black cottonwood (Populus trichocarpa) along the Pacific Northwest coast, *Canadian Journal of Forest Research* **39**(3): 519–525.

Xie, C.-Y., Ying, C. and Courtin, P. (1996). Genetic variability and performance of red alder (Alnus rubra), *in* P. G. Comeau, G. J. Harper, M. E. Blache, J. O. Boateng and K. D. Thomas (eds), *Proceedings of the Workshop on Ecology and Management of B.C. Hardwoods. 1-2 December 1993 Richmond, B.C.*, FRDA Report No. 155. British Columbia Ministry of Forests, Victoria, B.C., pp. 147–156.

Yanai, I., Graur, D. and Ophir, R. (2004). Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control, *OMICS* **8**: 15–24.

Ying, C. C. and Liang, Q. (1994). Geographic pattern of adaptive variation of lodgepole pine (Pinus contorta Dougl.) within the species' coastal range: field performance at age 20 years, *Forest Ecology and Management* **67**: 281 – 298.