THE GENETIC REGULATION OF GENE EXPRESSION, TRANSCRIPTIONAL NETWORKS, AND LEAF DEVELOPMENT IN THE PERENNIAL MODEL PLANT *Populus*

By

DEREK R. DROST

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2009

To Mom, Dad, and Abigail

ACKNOWLEDGMENTS

I would like to gratefully acknowledge the contributions of my graduate committee supervisors, Dr. Matias Kirst and Dr. Gary Peter, for their guidance, support, and committment to the success of my graduate education. Furthermore, I am grateful to Dr. Alice Harmon and Dr. Marta Wayne for their advice and service as members of my supervisory committee. I would also recognize the contributions of several University of Florida post-doctoral scholars, research scientists, and graduate students to the success of this research: Evandro Novaes, Carolina Novaes, Chris Dervinis, Dr. Catherine Benedict, Brianna Miles, Dr. Alison Morse and Ryan Brown. Additionally, I thank the faculty, staff, and students of the PMCB program for both their friendship and for lively discussions about science. I also thank Dr. Donald J. Lee (Professor of Agronomy, University of Nebraska-Lincoln) and Dr. Jerry Tuskan (Distinguished Scientist, Oak Ridge National Laboratory) for their continuing mentorship of my academic and professional development. Finally, and most importantly, I am indescribably grateful to my parents, Dirk and Sue, and my sister Abigail, for their steadfast love and support throughout all aspects of my education and life.

TABLE OF CONTENTS

page

LIST OF TABLES

LIST OF FIGURES

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

THE GENETIC REGULATION OF GENE EXPRESSION, TRANSCRIPTIONAL NETWORKS, AND LEAF DEVELOPMENT IN THE PERENNIAL MODEL PLANT *Populus*

By

Derek R. Drost

December 2009

Chair: Matias Kirst
Co-chair: Gary Peter
Major: Plant Molecular and Cellular Biology

Most heritable phenotypic traits are affected by the concurrent inheritance of alleles at multiple loci. As opposed to simple patterns of single gene inheritance, polygenic traits result in complex patterns of segregation, which are frequently assessed through quantitative trait loci (QTL) mapping methods. While QTL detection is a commonplace task in modern biology, cloning specific genetic variants underlying QTL is not – the task remains a challenge often associated with laborious and time-consuming fine mapping.

Recently, one novel method has been proposed to speed the identification of candidate genes and polymorphisms from QTL. “Genetical genomics” relies on allying traditional quantiative trait analysis to higher-level genomic data, including whole-genome expression or proteomic data. By considering these genomic “phenotypes” together with morphological or molecular phenotypes of interest, the role of genetics in these traits' regulation can be compared. QTL coordinately affecting phenotypes and transcript or protein abundance indicate a putative causal relationship between the phenotype and gene(s) represented by those transcripts or proteins, which can be functionally tested. Genetical genomics has recently demonstrated

significant promise to speed QTL cloning, and has implicated co-expressed gene networks underlying disease and complex phenotypes in model species.

Here, I directly implement genetical genomics methods aimed to reduce the challenges associated with moving from QTL to causal polymorphisms in poplar. First, I demonstrate a novel genotyping and genetic mapping approach for outcrossing species, by leveraging RNA-based molecular markers for high-resolution genetic analysis in *Populus*. These results lay a solid foundation for QTL and expression QTL (eQTL) mapping by increasing marker coverage and reducing breadth of QTL confidence intervals. Secondly, I apply whole-genome eQTL data to identify co-expressed transcriptional networks in three diverse plant tissues. These results build upon the understanding of transcriptional regulation in all plant species. Finally, I utilize genetical genomics to discover a candidate gene for leaf morphological variation in *Populus*, a high value phenotype for photosynthetic productivity and a vital trait for evolutionary classification of genus members into evolutionary sections. Taken together, these results paint a favorable picture for additional forward-genetic, genome-level studies to characterize complex phenotypes in poplar and other forest trees.

CHAPTER 1
INTRODUCTION AND LITERATURE REVIEW

The importance of forest ecosystems in the health and sustainability of life on Earth are irrefutable. Forest trees cover 30% of earth's terrestrial surface and forest ecosystems are responsible for harboring the majority of terrestrial biodiversity [1]. In addition to essential ecological roles, tree species provide substantial socioeconomic benefits in the forms of timber, cellulosic bioenergy, atmospheric carbon fixation and sequestration, and pulp and paper production [1]. Among the most ecologically and economically important forest tree species in the northern hemisphere are poplars; genus *Populus sp*. Poplars inhabit nearly 80 million hectares worldwide, including nearly 70 million hectares of indigenous forests in North America and Russia [2]. Various species of genus *Populus* are managed for wood and biomass production worldwide, while others serve as ecological foundation species of riparian and deciduous forest ecosystems throughout North America and Europe [3].

The recent sequencing and annotation of the *Populus trichocarpa* genome [4] provides new opportunities to understand how genetic diversity and genetic mechanisms contribute to forested ecosystems by enabling the identification of genes, markers, and regulatory networks that affect traits with significant roles in adaptation, ecology, and production. Identifying these linkages between genetics and phenotypic traits facilitates a better understanding of the role of genetics in forest tree and ecosystem function. Concurrently, this knowledge can direct tree improvement and breeding efforts toward more productive or environmentally fit genotypes with improved yield, wood quality, disease resistance, or environmental adaptation. Finally, and perhaps most importantly, knowledge of the role of genetic mechanisms in forest health and productivity can promote improved science-based policies for forest resource management and forest conservation on a global scale.

Poplar as a Reference for Tree Biology and Genetics

In addition to their role in terrestrial ecosystems and silviculture, several attributes of the poplar species have contributed to their emergence as a model genus for tree genetics, genomics, and molecular biology [5-7]. First, poplars are among the most rapidly growing hardwood trees when cultivated under intensive management, with field- and greenhouse-grown genotypes frequently achieving as much as three meters of growth per year [8]. Secondly, flowering and reproduction are also rapid – these processes normally occur within 3-6 years under field conditions, but transgenic genotypes overexpressing FT or LFY transgenes will flower in less than one year [9-11]. Third, most poplar species are routinely propagated from vegetative cuttings, allowing specific genotypes to be maintained indefinitely and studied over long periods of time or in different environments [12]. Fourth, several poplar genotypes and species are also amenable to stable, agrobacterium-mediated genetic transformation and regeneration [13], a protocol routinely performed in a number of different laboratories [4]. The transformation process facilitates reverse-genetic approaches to analyze and verify function of specific genes or regulatory elements of interest.

In addition to favorable growth and developmental characteristics, the facile interspecific hybridization among species within genus *Populus* benefits both fundamental scientific discovery and breeding of superior hybrids for deployment in managed forests. *Populus* is composed of six evolutionary sections, which vary extensively in their morphology, adaptation, and natural distribution ([12], Table 1-1). A number of interspecific segregating populations have been produced within and between these sections, most notably between *P. trichocarpa* and *P. deltoides*, *P. nigra* and *P. deltoides*, *P. tremula* and *P. tremuloides*, and *P. angustifolia* and *P. fremontii* [12]. These interspecific populations present several benefits for genetic analysis. First, alleles for certain genes or gene combinations may commonly be fixed within a single species. In

this case, progeny populations derived from intraspecific hybridizations will not produce variation for those genes or the phenotypes they influence. In contrast, interspecific hybrid progeny produced from species fixed for alternative alleles will vary for phenotypes affected by those genes. Secondly, in cases where alleles for genes affecting production or adaptation are fixed in a given species, interspecific crosses with an alternative species could be used to introgress new genetic combinations, increasing fitness or yield in the species of interest. Finally, interspecific pedigrees are useful to discover naturally occurring, specifically "tuned" variants for differential gene expression or function. Superior variants are rarely produced through biotechnology, since functional study of genes through genetic engineering is generally only effective to constitutively initiate or repress gene expression. However, gene discovery through naturally occurring variation provides the opportunity to identify alleles that have been subjected to the forces of evolution in a genomic context, finely tuning expression or activity of gene products for specific ecological niches. These novel combinations are likely to be superior in function relative to simple constitutive constructs produced through genetic engineering in a laboratory setting.

Poplar's position as a model hardwood species was solidified with the release of the genome sequence of the *P. trichocarpa* genotype 'Nisqually-1'in 2006 [4, 6]. *P. trichocarpa* was only the third plant species, and the first woody perennial to be sequenced, following the release of the *Arabidopsis thaliana* genome in 2000, and the rice genome in 2002 [14-16]. The *P. trichocarpa* genome sequencing project produced an excellent draft sequence that accelerates genetic, molecular, biochemical and evolutionary studies in poplar. Whole-genome Sanger-based shotgun sequencing and reassembly resulted in over 470 million base pairs (Mbp, 95% of the estimated genome size) of sequence distributed in 22,136 scaffolds with a 7.5× average

sequencing depth. Approximately 385Mbp of the total 470Mbp of sequence was anchored to the 19 chromosomes of poplar on the basis of sequence-tagged sites converted from previously identified simple-sequence repeat (SSR) microsatellite markers. Initial gene identification efforts, based on a compendium of previously generated expressed sequence tags (ESTs) and computational gene modeling algorithms, tallied ~45,600 genes -- about two-fold more than Arabidopsis (~27,000) and comparable to the ~41,000 identified in rice [4].

Also contributing to its place as a model hardwood, a vast collection of gene expression data is available for a number of species in genus *Populus*. Early gene expression analyses were based on transcript counting in EST libraries, or spotted cDNA microarrays produced from EST resources. These analyses generally focused on tissue-specific expression cataloging [17-25] or classifying transcriptional response to specific biological [22, 26-29], abiotic and environmental [30-33], or mechanical stimuli [18, 24, 34]. Subsequent studies have incorporated advances in modern microarray technology using oligonucleotide probes generated by *in situ* synthesis to obtain refined estimates of gene expression and catalog comparative patterns of gene expression between different organ types to produce a transcriptional "atlas" for *P. trichocarpa* [35]. Since all poplar species are expected to contain genes very similar (or identical) to that of sequenced *P. trichocarpa* [36], differences in expression caused by regulatory and expression-level polymorphisms are proposed to be critical for inter- and intraspecific variation within the genus. Identifying these putative interspecific differences in gene expression, through whole-genome expression analysis and quantitative genetics, forms the central focus of the study described herein.

Quantitative Genetics in the *Populus* System

Currently Established Genetic Marker Resources

Reliable methods to genotype DNA sequence polymorphisms are essential for quantitative genetic studies, as they contribute to the production of genetic maps -- the genomic "groundwork" upon which trait data is analyzed via forward genetic analysis. In poplar, several types of genetic markers are available and have been utilized to characterize genetic variation. Early population genetic studies in *P. fremontii* and *P. angustifolia* utilized restriction fragment length polymorphism (RFLP) [37]. However, RFLPs were largely abandoned after the development of polymerase chain reaction (PCR) -based markers, due to the low multiplex ratio of RFLP assays and the high cost and labor required per genotypic data point. Development of PCR-based markers such as random amplified polymorphic DNA (RAPD) and amplified fragment length polymorphism (AFLP) facilitated the first high-coverage genomic surveys of polymorphisms in *Populus* [38-40]. Still, dominant inheritance and low transferability among genotypes and species made these markers less suitable for adoption by the scientific community. Anchoring these markers to the genome sequence can also be challenging [41], yet sequence-tagged markers were shown to be crucial for assembling the *P. trichocarpa* Nisqually-1 genome sequence [4]. Co-dominant markers such as microsatellites (or simple sequence repeats – SSRs) and single nucleotide polymorphism (SNPs) have become more widely used in animal and plant genetics. The main advantages of SSRs are their multi-allelic nature and high levels of heterozygosity; SNPs also bear these advantages, in addition to being the most abundant type of DNA sequence variant. Since probes or primers used to assay SSRs and SNPs are designed specifically for their targeted locus, they are frequently transferable within and (sometimes) across species.

The International *Populus* Genome Consortium (IPGC) provides a large repository (4166) of known poplar SSRs, with most (84%) being identified from the genome and physical map end sequences of Nisqually-1 [4, 42, 43]. The remaining SSRs were generated from a series of other studies [44-47]. One third (1,395) of the IPGC SSRs have been tested in at least one *Populus* segregating pedigree, where 553 of them (40%) were successfully mapped. In addition, a set of 92 of the SSR markers were tested for their transferability in 23 species, including five from the related *Salix* genus. The SSR resource is being extensively utilized for genetic mapping and analysis of QTLs. From a practical standpoint, SSRs have already been applied in breeding programs for fingerprinting and ancestry determination of cultivars [48, 49].

While SSRs have been implemented frequently in poplar research, the use of SNPs is only in its infancy. The largest SNP discovery for the genus came as a by-product of the *Populus trichocarpa* genome sequence [4, 42]. Since poplars are obligate outcrossing species, the sequencing and assembly of the heterozygous diploid genotype Nisqually-1 identified over one million SNPs and 162,000 single nucleotide insertion/deletion (indel) polymorphisms, with 20% of all polymorphisms falling withing gene coding sequences (Table 1-2). As expected by the action of purifying selection, nucleotide diversity of SNPs was lower (1.4×) in genes when compared to the diversity found in the whole genome. Selection is apparently even stronger against indels within gene sequences, where they are 1.8-fold less abundant than in the remainder of the genome. SNP frequency is variable between scaffolds of Nisqually-1, with the highest abundance found on LG_V (average of 1 SNP every 330 bp) and the lowest on LG_XVI (average of 1 SNP every 484 bp). SNP frequency in gene sequences ranged from 591 bp per SNP on LG_XII to 433 bp per SNP on LG_XIII. However, variation in SNP diversity between scaffolds should be analyzed with caution, as differences in sequencing depth were not

considered in this estimate. In another study 1,635 and 610 SNPs were identified in *P. tremula* and *P. trichocarpa*, respectively, by aligning approximately 70,000 ESTs [50]. Aside from these two large scale studies, additional poplar SNPs have been discovered and characterized in only a few targeted genes. These genes were sequenced in a sample of genotypes in studies aimed at estimating population genetic parameters, searching for departures from neutral evolution, or contributing to candidate gene association studies [51-55]. The main limitation of SNPs identified in these studies lies in the fact that forest species generally harbor an excess of low frequency alleles [56, 57]. As a result, many SNPs identified in a given study may not be suitable as molecular markers in different genetic backgrounds. However, as sequencing technologies advance [58, 59], polymorphisms will be discovered at markedly increased rates and will reveal specific SNPs with higher allele frequencies and utility for marker-based forward genetic studies.

While SSRs and SNPs represent the most useful marker classes in poplar today, other classes of sequence-anchored molecular markers have been described that could be incorporated into the poplar genetic system. Principal among these are single-feature polymorphisms (SFPs) and expression-level polymorphisms (ELPs), also called gene expression markers (GEMs). SFPs were first described by researchers in the yeast and *Arabidopsis* genetic systems, who hybridized labeled genomic DNA to oligonucleotide microarrays and noted predictable signal variance at certain probes for different genotypes [60, 61]. Sequencing revealed that SFPs may correspond to SNPs, small insertion/deletion polymorphisms, or small genomic rearrangements. Subsequently, expression studies based on hybridization of RNA and cDNA also demonstrated the ability to yield an abundance of SFP markers, and coordinately identify GEMs based on Mendelian segregation of divergent gene expression values between two parents of a mapping population [62, 63]. While these markers appear to be abundant and highly robust in haploid

models and inbred plant species [62-64], it remains to be discovered whether these small polymorphisms detected by differential hybridization or expression will segregate discernibly in a highly heterozygous, obligate outcrossing genetic system such as *Populus*.

Linkage/QTL Population Structures and Associated Analysis Methods

The previously mentioned interspecific populations of *Populus* have proven to represent a strong framework for QTL mapping, since variation between individuals of different species is expected to be far greater than variation between individuals of the same species. Thus, individual species, while potentially heterozygous for specific alleles, can be treated as "inbred lines" relative to one another and traditional inbred line mapping approaches can be adapted for genetic mapping and QTL analysis [12]. A number of pedigrees have been established based on this assumption, using two primary and distinct strategies for population structuring and genetic map production.

Early efforts were directed towards development of an "inbred F_2"-type pedigree system, wherein an interspecific cross was produced between two grandparents, one *P. trichocarpa* and one *P. deltoides*, which were expected to be "inbred" relative to one another. Two resulting F1 full-sib progeny were selected and mated, producing a segregating progeny population wherein ¼ of the resulting progeny were expected to be homozygous for either *P. trichocarpa* or *P. deltoides* alleles, with the remaining ½ bearing a heterozygous F_1-type genotype [65]. The resulting population can be analyzed as an inbred F_2 in standard QTL detection analysis [66, 67]. QTL for a number of traits, including stem growth characteristics, crown architecture, leaf morphology, and disease resistance were subsequently detected on the basis of this procedure [68-72]. Later efforts refined these mapping strategies to account for the outbred population structure and improve the mapping methods [73, 74]. Yet, these initial studies provided proof

that genetic mapping and QTL detection were feasible in the highly heterozygous poplar genetic system.

More recent efforts have focused on F_1 pseudotestcross and F_2 pseudobackcross pedigree structures for genetic mapping and QTL detection in outcrossing forest species [73, 74]. Pseudotestcross F_1 populations are dependent on identifying markers where one parent is homozygous for a given allele at a marker, while the other is heterozygous. Such an arrangement of alleles will result in F_1 progeny segregation of 1:1, allowing the population to be analyzed as a testcross in mapping and QTL detection software [12]. However, the arrangement of alleles is known only *a posteriori* and can vary between parents from locus to locus, hence the utilization of the term "pseudo"-testcross. This fact dictates that pseudotestcross analysis will result in two single-tree genetic maps for the pedigree, both of which are tested for QTL. Alternatively, pseudobackcross F_2 populations can be generated, whereby an interspecific F_1 hybrid can be "backcrossed" to a genotype of one of the F_1 donor species (Figure 1-1). Generally this is a different genotype than the one utilized to create the F_1 (hence the "pseudo"-backcross), owing to the self-incompatibility barriers common within forest trees. Since the two species are expected to be "inbred" relative to one another, many markers will again segregate at a ratio of 1:1. Fully informative markers (i.e. those which are biallelic in both the F_1 hybrid and the recurrent pseudobackcross parent) can be utilized to create a separate genetic map for the recurrent parent background, promoting QTL detection both with respect to the interspecific differences fixed between the two species (based on the hybrid parent map) and intraspecific differences within the recurrent species background (based on the map derived from the recurrent species). This strategy has been extremely effective and popular for genetic mapping and QTL detection in forest tree species including *Populus*, where rare alleles are common

within species and alleles affecting important morphological differences between species are often fixed. Thus, this strategy has contributed to the detection of numerous QTL for plant and organ morphology, biomass, phenology, disease resistance, and even molecular metabolites [75-77].

Quantitative Genetic Analyses of Gene Expression: Genetical Genomics

Analogous to the process of analyzing phenotypic trait QTL, wherein variation in phenotype among members of a segregating population is statistically associated to prevalence of a DNA marker genotype at a specific locus, variation in the abundance of an mRNA can be associated with marker genotypes through the process of expression QTL (eQTL) mapping [78]. Classical insight into the role of genetic variation in gene expression regulation was made by King and Wilson [79], followed by the assessment of protein abundance by Damerval *et al.* [80]. However, without technology to concurrently analyze the abundance of large numbers of transcripts or proteins, these studies were limited in their scope to specific suites of well classified genes. DNA microarray technology [81, 82] facilitated the analysis of thousands of genes in parallel to reveal that transcript abundance is highly variable between individuals in yeast [62], mice [83], plants [84, 85], and humans [86, 87]. Furthermore, these differences behave as traditional quantitative phenotypes, segregating predictably under quantitative genetic models in progeny derived from crosses between divergent parents [88].

Global studies of gene expression diversity among segregating populations were first proposed in 2001, yet work toward these ends was clearly underway at this point as the seminal articles describing true genome-wide linkage analysis of expression traits appeared in 2002 [62, 89]. In agreement with the original propositions of Jansen and Nap in 2001 [78], these studies showed that variation in expression phenotypes is complex, with genes being regulated by combinations of local, *cis*-acting variants and distant, *trans*-acting variants (Figure 1-2). In the

majority of genetic systems considered, distant *trans*-acting variation seems to be more common, yet these loci tend to explain smaller portions of heritable variation than their *cis*-acting counterparts [62, 87, 89-91]. Furthermore, often only small portions of a transcript's heritable variation are captured in an eQTL experiment, presumably due to the large number of small effect loci affecting transcript abundance that even the largest experiments lack the statistical power to resolve into significant eQTLs [88]. Most transcripts, in addition to being subjected to multiple points of *cis*- and *trans*-acting regulation, demonstrate non-additivity and transgressive segregation, and pleiotropic loci that coordinately regulate the abundance of multiple transcripts are quite common [91-93].

As indicated by the initial proposition of the genetical genomics concept, eQTL data gathered from segregating populations has been exceedingly useful to understand phenotypic variance, both at the level of single gene:phenotype relationships, and more complex biological processes [94]. Initial efforts to associate individual, differentially expressed genes with QTL phenotypes were carried out primarily in the murine genetic system. Candidate genes for complex clinical diseases including diabetes [95], asthma [96], and obesity [83] were discovered by assessing the correlations between expression levels and clinical phenotypes for genes with eQTL co-locating to the position of phenotypic QTLs (Figure 1-3). Efforts in *Drosophila melanogaster*, *Saccharomyces cerevisiae*, and *Arabidopsis* validated these initial findings, associating candidate genes with variation in traits of interest [89, 97, 98]. Later studies began to address the task of linking known biochemical pathways to phenotypic variation through eQTL data. Since genes encoding proteins involved in a common biochemical pathway are functionally connected through the metabolite pools of the pathway to which they belong, they are frequently co-regulated at the level of gene expression [99]. Therefore, alterations in the level of expression

for a single gene in a pathway can feed back to other members through these co-regulatory mechanisms, leading to correlated expression for genes in that pathway (Figure 1-4). This fact was first exploited to demonstrate the role of differential pheromone response pathway regulation with a variation in a G-protein coupled receptor in yeast [98]. Subsequently, variation for growth and lignin traits in a *Eucalyptus* segregating population was shown to be associated with coordinate transcriptional regulation of the lignin biosynthetic pathway enzymes [100]. A number of similar findings have been described for several fundamental biochemical and regulatory pathways in *Arabidopsis* [93, 101].

These demonstrated cases of allele-dependent differential regulation of genes in known biochemical and signaling pathways has broadened to a 'global' approach to analysis of genome-wide eQTL data. Specifically, demonstrated co-expression of genes that are members of conserved pathways suggests that identifying the correlation structure among whole-genome expression data may identify regulatory connections between biochemically distinct pathways, as well as implicate novel pathway members and regulators that have not been previously annotated for their role in an established pathway. So-called 'whole genome co-expression network analysis' has begun to successfully accomplish these goals in a number of organisms [102]. Initial efforts to describe larger scale co-expressed networks were fundamentally based upon well characterized biological processes, such as the cell cycle [103], or from analyzing the transcriptomic response to directed genetic or environmental perturbations [104]. Nonetheless, these initial experiments clearly demonstrated the regulatory interplay between known pathways with seemingly distinct roles in cellular function.

Subsequent efforts sought to overcome the barriers of chemical and directed genetic perturbation by implementing genome-wide genetic "perturbation" – leveraging segregating

populations constructed from divergent progenitors, wherein evolution and natural selection have produced simultaneous genetic variation at multiple loci [105]. The aforementioned interconnectedness of the transcriptome, coupled with independent assortment of loci in a QTL population, enables the discovery of relationships between expressed genes without preconditioning biochemical or functional connections [106, 107]. These *a posteriori* generated co-expression networks have proven to be a powerful tool to elucidate signatures and pathways associated with complex phenotypes and disease traits in model and non-model organisms [101, 106, 108-113]. Post hoc incorporation of *a priori* knowledge about networked genes (i.e. transcription factor binding sites, gene annotations and ontology, protein-protein interactions, and/or protein cellular localization) has substantially aided in the interpretation of these co-expressed networks, by providing further connections between data points that could not be inferred simply from their co-expression patterns [114, 115]. The ability to predict whole-system response to genetic, environmental, or pharmacological perturbation is a key goal of modern biology, and these whole-genome-based network studies have shown initial promise towards realization of this goal.

Leaf Morphological Variation in *Populus*

Evolutionary Conservation and Previous Analyses of Leaf Morphology

Variation for leaf morphology, specifically leaf blade shape, is substantial among the members of genus *Populus*. Leaf shapes range from cordate to lanceolate and while shape can be somewhat variable within species, leaf morphology and shape are generally conserved across evolutionary sections [116, 117]. In fact, leaf morphology has often been utilized to diagnose evolutionary relationships between the species/sections from the fossil record [118]. Furthermore, leaf morphological characteristics in juvenile trees have been shown to be indicative of longer-term productivity and yield [119, 120]. Variation in leaf shape among

poplars is also heritable and relatively immune to the effects of environmental deviations [70, 121]. Taken together, this evidence makes leaf morphological variation an attractive candidate for quantitative trait analysis, leading to previous efforts to understand the genetic components controlling these traits through QTL-based methods [69, 70, 122-124]. Leaf shape has been generally measured by blade length and width, as well as their ratio [69, 70], since these measures differ predictably and consistently between a lanceolate leaf shape as in *P. trichocarpa* or *P. angustifolia*, versus a deltoid or cordate leaf shape common to *P. deltoides*, *P. fremontii*, or *P. nigra* (Figure 1-5). Several early QTL studies identified loci for these traits among interspecific pedigrees of *P. trichocarpa* and *P. deltoides* [70, 124]. However, due to changes in linkage group naming convention and the popularity of 'in-house' anonymous DNA markers (AFLP and RFLP), it is difficult to determine whether loci detected in different pedigrees represent conserved genetic architecture or variation due solely to a set of rare alleles restricted to one parent of the cross. More recent studies of sex-determination [125] and plant growth and biomass accumulation [75, 122, 126, 127], comprising multiple pedigrees constructed from several different interspecific combinations suggest that QTL detected in these pedigrees are often associated with evolutionarily conserved differences rather than rare alleles. Nonetheless, limited mapping resolution in QTL populations [128], coupled with the unavailability of genome sequence, has until recently hindered the pursuit of candidate genes for leaf morphological traits in *Populus*. Recently, novel genetic and transcriptomic approaches [123] have shown promise to implicate candidate genes for leaf morphological variation – a concept further explored within the work described herein.

Candidate Genes and Pathways Discovered in *Arabidopsis*

As a well-established model angiosperm plant, *Arabidopsis* lends several clues to genes, pathways, and gene networks that, if conserved, may underlie variation in leaf morphology

observed within *Populus*. In angiosperms generally and *Arabidopsis* specifically, it is thought that growth and division at the leaf margins results in leaf blade initiation and may be the result of the junction formed between adaxial and abaxial cell fates [129]. Subsequent blade expansion is a result of cell divisions and expansion in the plate meristem, which gives rise to leaf cell layers and two dimensional (longitudinal and lateral) 'plates' of cells [130, 131]. Abundant variation for both blade length and width (and subsequently, their ratio) indicates that plate meristem processes are specifically regulated both in the longitudinal (length) and lateral (width) planes. Previous studies have implicated genes involved in both polar cell division and polar cell expansion that play a role in leaf shape variation.

Among the first genes isolated that affect leaf blade expansion, specifically in the longitudinal plane, was *ROTUNDIFOLIA3* (*ROT3*). Aberrant cellular expansion in *rot3-1*, the null allele homozygote, resulted in an overall stunted plant with leaves stunted for growth specifically in the longitudinal plane, whereas overexpression of *ROT3* triggered hyperexpansion in the longitudinal direction and elongated leaves [132]. Subsequent biochemical and bioinformatic evidence indicated that *ROT3* is a cytochrome P450-type gene that likely functions in the brassinosteriod signaling pathway, possibly in a leaf specific manner. A similar phenotype was obtained for mutations in another *ROTUNDIFOLIA* gene, *ROT4* [133]. Despite the naming convention, *ROT4* and *ROT3* are not functionally or biochemically related. *ROT4* is an intronless 53 amino acid peptide that includes a 29 amino acid domain of unknown function. The original mutation was generated in an enhancer trap system, where overexpression of the *ROT4* gene reduced cell division frequency in the longitudinal direction, decreasing the number of cells deposited in this plane of the leaf [133]. At least 22 homologs of *ROT4* exist in *Arabidopsis*, and at least two of these show overexpression phenotypes similar to that of *ROT4*, suggesting similar

function during leaf development perhaps mediated by the highly conserved domain of unknown function.

In addition to genes that affect leaf longitudinal growth, at least two other genes have been identified that when mutated result in aberrant leaf lateral growth. Analogous to the *ROT3* and *ROT4* mutations, these genes function to alter either cellular expansion or cellular division in the lateral plane. The *ANGUSTIFOLIA* (*AN*) gene was identified as a mutant with decreased cell expansion in the lateral plane, resulting in a narrowed-leaf phenotype [134]. *AN* is a functional homolog of animal carboxyl-terminal binding proteins (CtBPs) or brefeldin-A-ADP ribosylated substrate proteins (BARS), which are indistinguishable in function at the primary amino acid sequence level but function in diverse molecular roles [135]. CtBPs generally function as nuclear transcriptional co-repressors [136], while BARS function in membrane and vesicle fission dynamics, particularly at the Golgi complex [135]. Recent evidence has shown that AN possesses no CtBPs or transcriptional co-repressor function, indicating that it likely functions either as a BARS (which is thought to be vertebrate specific) or in some function that has yet to be attributed to CtBP/BARS superfamily proteins [137]. It is known, however, that AN affects microtubule dynamics, yet the mechanism of this activity remains unknown [134]. In addition to *AN*, which controls cell expansion, the *ANGUSTIFOLIA3* (*AN3*) gene also affects leaf shape, with null mutants resulting in a *an*-like phenotype [138]. The *an3* mutant, however, functions to decrease cell proliferation in the leaf width direction, resulting in elongated, narrow leaves. *AN3* encodes a protein with sequence similarity to the human transcriptional co-activator and chromatin remodeling complex interactor *SYNOVIAL SARCOMA TRANSLOCATION*. AN3 interacts with GROWTH-REGULATING FACTOR5 (GRF5) *in vitro*, a gene for which altered

expression mimics a mild *an3* phenotype [138]. GRF5 functions in cell cycle control, suggesting that *an3* mutants may be compromised for cell cycle progression at the leaf plate meristem.

Project Objectives

Despite the recent boom in research interest surrounding the *Populus* genetic system, fueled primarily by the aforementioned release of the full genome sequence of *P. trichocarpa*, little remains known about the role of transcriptional variation in shaping diversity between species of genus *Populus*. Within this project, I begin to address this shortcoming by leveraging whole-genome microarray expression data derived from a pseudobackcross of *P. trichocarpa* and *P. deltoides* to characterize novel aspects of transcriptional regulation in the *Populus* system. In Chapter 2, I describe the utilization of microarray data to discover, genotype, and genetically map two novel classes of *Populus* molecular markers – GEMs and SFPs – and in conjunction with SSR data, show the unexpected utility of these markers to improve the map-based shotgun sequence assembly of the *P. trichocarpa* genome. In Chapter 3, this map data is utilized to characterize the genome-wide genetic diversity for expression traits (eQTL) in three distinct tissues (leaf, root, and xylem) and eQTL results are subsequently employed to construct and annotate *a posteriori* defined co-expression networks for each of the three tissues. Chapter 4 describes the integration of map and eQTL data from leaves, in a genetical genomics context, to implicate an ADP-ribosylation factor GTPase as a candidate gene for evolutionarily conserved variation in leaf shape between *P. trichocarpa* and *P. deltoides*. Finally, in Chapter 5, I summarize the findings of this research and suggest avenues for future studies to better understand the role of transcriptional diversity in shaping phenotypic variation in *Populus*.

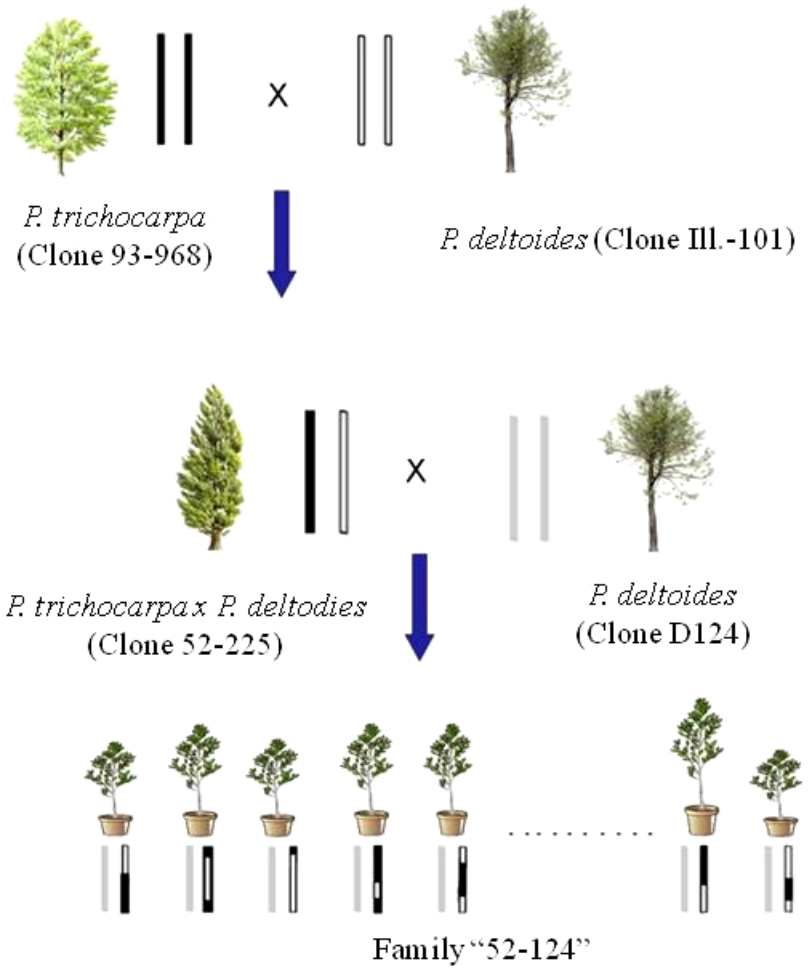


Figure 1-1. Overview of the pseudobackcross pedigree construction method for outcrossing forest trees. An interspecific hybrid F_1 population is constructed from a controlled cross of two elite genotypes. An individual is selected from this population and backcrossed to an individual of the same species, but different genotype, as one of the elite F1 progenitorsThe majority of the genetic variance segregating in the pedigree can be accounted for due to differences between *P. trichocarpa* and *P. deltoides*.

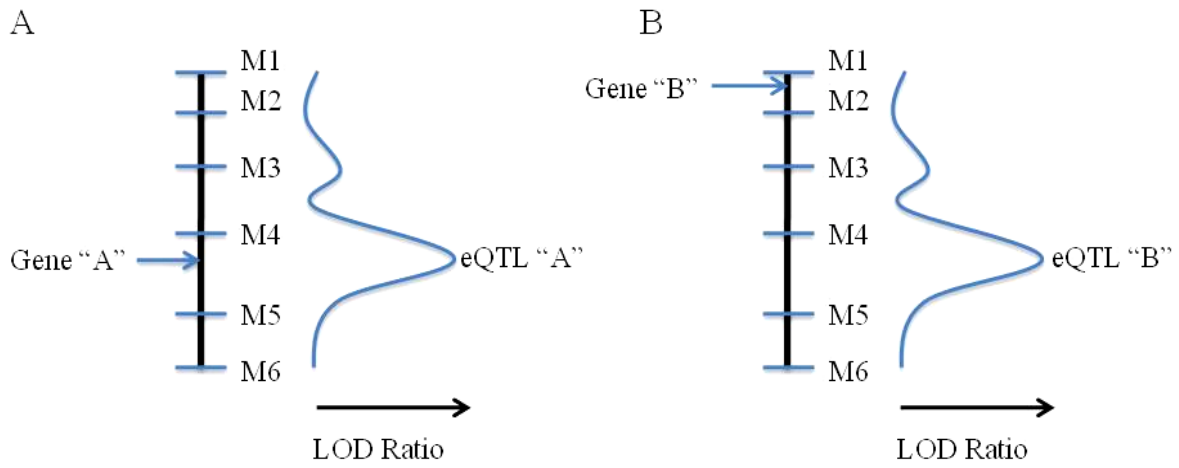


Figure 1-2. Hypothetical examples of *cis*- and *trans*-acting eQTL regulation. A) An eQTL for gene "A" is detected in *cis*-, since the position of the significance of the eQTL peak is located within the marker interval (M4-M5) known to physically contain gene "A" on the basis of a physical map or genome sequence assembly. *Cis*-acting eQTL may be caused by polymorphisms in a regulatory region for the gene in question, a mutation that leads to missense-mediated transcript decay, or a *trans*-acting regulatory factor tightly linked to the gene in question. B) An eQTL for gene "B" is detected in *trans*-, since the significance for the eQTL peak is located within a marker interval (M4-M5) different from the interval known to contain gene "B" (M1-M2). *Trans*-acting eQTL may be located on the same linkage group as the gene (depicted above) or on a different linkage group (not depicted), and may be the result of a sequence or regulatory polymorphism affecting a regulatory factor governing the expression of the gene in question.

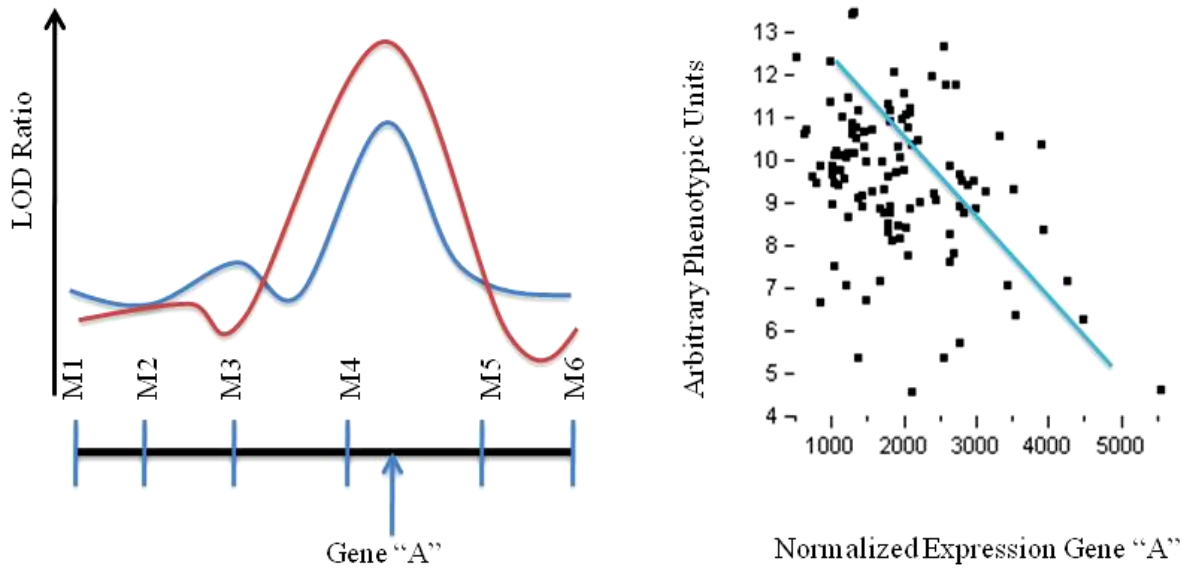
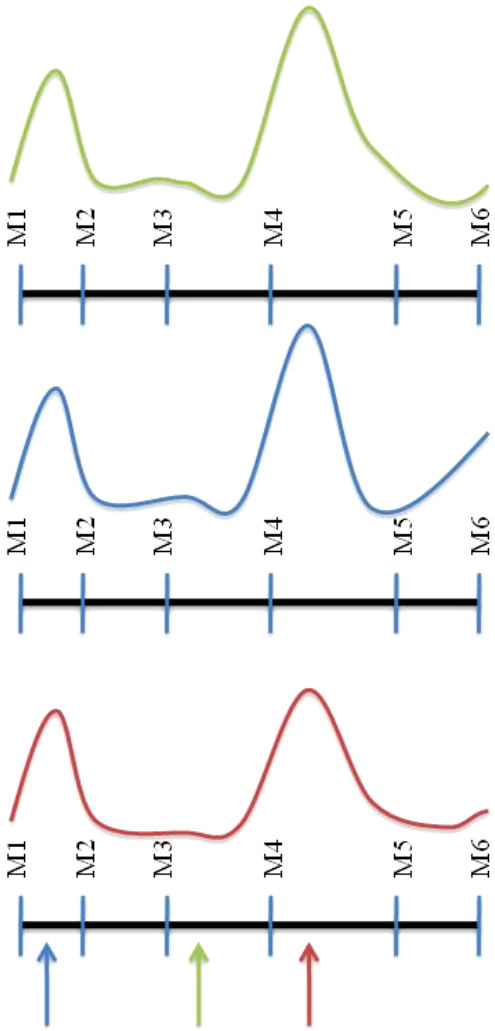


Figure 1-3. A hypothetical example of candidate gene identification by the genetical genomics method. A phenotypic trait of interest is measured and mapped as a quantitative trait using standard QTL mapping approaches. A significant QTL for this trait is identified near markers "M4" and "M5" (red LOD curve). Whole-transcriptome microarray analysis identifies a *cis*-acting eQTL for gene "A" (blue LOD curve) underlying the trait QTL. Statistical analysis indicates that the expression of gene "A" and the phenotype are negatively correlated at a statistically relevant threshold. The coincidence of QTL peaks, the physical position of the gene, and the correlation between trait and transcript abundance provide initial circumstantial evidence that differential transcriptional regulation of gene "A" may play a role in the phenotype in question.



	Trait (Black)	*Green*	*Blue*	*Red*
Trait (Black)	1	.75	.82	.68
Green		1	.47	.89
Blue			1	.76
Red				1

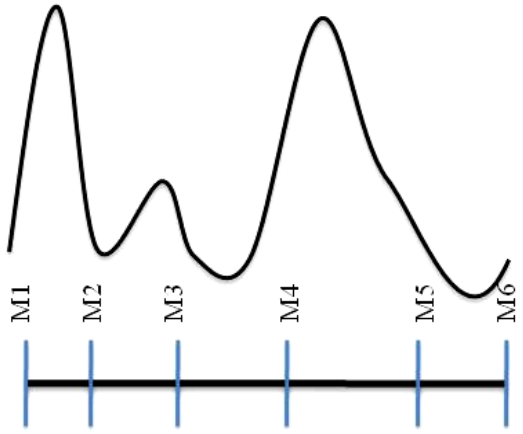


Figure 1-4. A hypothetical example of a mechanism of network construction and phenotypic association from genetical genomics eQTL data in a segregating population. eQTL detected for three genes ("green", "blue", and "red") share similar eQTL LOD curve profiles, suggesting their underlying regulation is shared. An arbitrary phenotypic trait LOD curve (black) also has a similar profile. Statistical correlations between the trait and the genes are highly significant, suggesting that these genes may act in a pathway that influences variance in the phenotype in question.

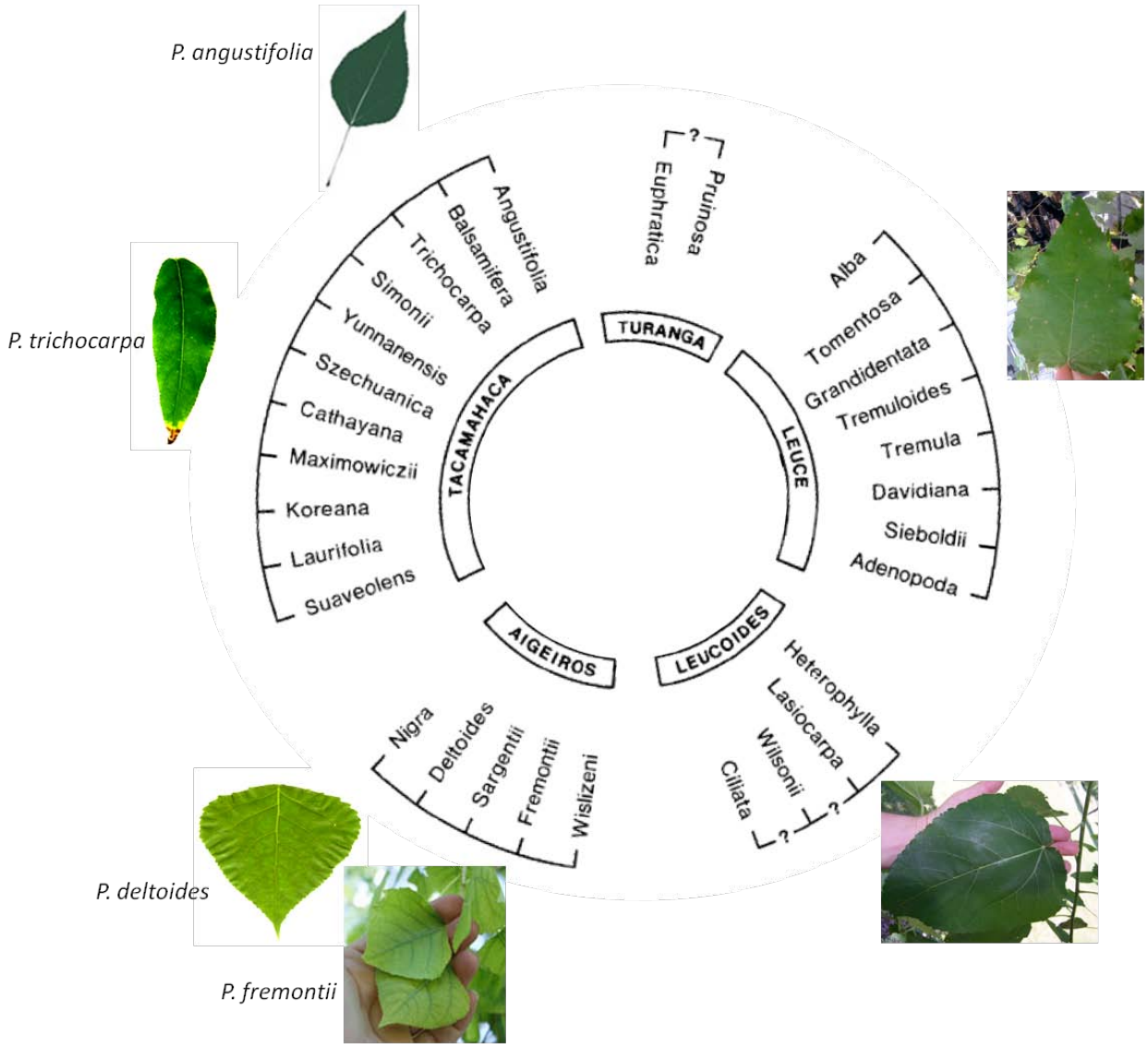


Figure 1-5. Leaf morphologies assoicated with the four most populated evolutionary sections of genus *Populus*. Narrow elliptical or lanceolate leaf shapes are observed among members of section Tacamahaca, whereas broader ovate, deltoid, or cordate shapes are observed among members of sections Aigeiros, Leucoides, and Leuce. Section Abaso is not diagrammed.

Table 1-1. Range, leaf morphology, and species count associated with sections of genus *Populus*.

Section[a]	Geographic (climatologic) range	Leaf morphology	Number of species[a]
Aigeiros	N. America, Asia, W. Europe (temperate)	Cordate/Deltoid	3
Leuce	Circumpolar (subarctic and cool temperate)	Oval/Deltoid	10
Leucoides	N. America, E. Asia (warm temperate)	Ovate	3
Tacamahaca	N. America, Asia (cool temperate)	Elliptical/Lanceolate	9
Turanga	S.W. Asia, E. Africa (subtropical to tropical)	Lanceolate/Cordate	3
Abaso	Mexico (tropical, subtropical, arid)	Deltoid/Lanceolate	1

[a.] Sections and species as recognized by Eckenwalder in [8].

Table 1-2. Frequency of polymorphism across chromosomes in the genome of *Populus trichocarpa*, clone Nisqually-1.

Scaffolds	Genome				Genes[a]			
	Number of polymorphisms		Diversity[b] (bp / polymorphism)		Number of polymorphisms		Diversity[b] (bp / polymorphism)	
	Indel	SNPs	Indel	SNPs	Indel	SNPs	Indel	SNPs
LG_I	13,344	85,410	2,410.01	376.53	2,132	16,772	4,409.58	560.53
LG_II	11,811	70,007	1,984.01	334.73	1,759	13,769	4,151.55	530.36
LG_III	7,767	46,293	2,246.28	376.88	1,340	9,721	3,990.46	550.07
LG_IV	5,798	39,659	2,600.94	380.25	953	7,180	4,244.67	563.39
LG_V	8,116	51,571	2,099.02	330.33	1,060	8,383	4,501.44	569.19
LG_VI	7,623	46,963	2,316.31	375.98	1,217	9,816	4,676.68	579.82
LG_VII	5,175	32,691	2,300.35	364.15	764	6,730	4,788.73	543.62
LG_VIII	6,940	42,758	2,223.37	360.87	1,184	10,362	4,678.92	534.63
LG_IX	5,589	33,874	2,220.73	366.41	1,005	7,980	4,321.07	544.19
LG_X	7,773	47,329	2,471.87	405.96	1,375	11,533	4,908.94	585.26
LG_XI	4,792	32,942	2,748.75	399.85	854	6,502	4,217.58	553.95
LG_XII	4,875	31,368	2,671.80	415.23	766	5,963	4,599.58	590.86
LG_XIII	4,979	32,977	2,308.85	348.60	921	8,179	3,844.72	432.94
LG_XIV	5,574	34,953	2,454.74	391.46	914	7,437	4,559.25	560.33
LG_XV	4,059	25,327	2,510.66	402.37	653	5,418	4,808.18	579.50
LG_XVI	4,458	26,454	2,875.34	484.55	842	6,465	4,219.58	549.56
LG_XVII	2,238	16,291	2,432.56	334.18	344	2,808	3,855.58	472.34
LG_XVIII	4,568	29,995	2,722.86	414.67	752	5,976	4,638.20	583.66
LG_XIX	3,644	24,744	2,806.48	413.30	636	5,832	4,378.72	477.51
Unmapped scaffolds	42,767	327,755	3,343.42	436.27	6,951	64,742	4,653.55	499.63
Total or Average	161,890	1,079,361	2,487.42	385.63	26,422	221,568	4,422.35	543.07

[a.] Genes are composed of 45,555 currently annotated gene models plus 10,288 additional less supported models. The whole gene sequence was considered, including its non-translated portions (introns, 3'UTR and 5'UTR). [b.] Sequence gaps of known size (symbolized in the genome sequence with 'Ns') were excluded from the calculation of genomic diversity.

CHAPTER 2

A MICROARRAY-BASED GENOTYPING AND GENETIC MAPPING APPROACH FOR HIGHLY HETEROZYGOUS OUTCROSSING SPECIES ENABLES LOCALIZATION OF A LARGE FRACTION OF THE UNASSEMBLED *Populus trichocarpa* GENOME SEQUENCE

This Chapter has been published in The Plant Journal 2009; 58(6):1054-1067 (PMID: 19220791)

Introduction

Microarrays revolutionized the study of gene expression and have recently been applied for high-throughput genotyping of sequence and expression-level polymorphisms. Single feature polymorphisms (SFPs) detected by differential hybridization of genomic DNA to whole-genome microarrays were reported initially in yeast [61, 62], *Arabidopsis* [60, 64], and later in rice [139]. Subsequent reports showed that hybridization of RNA could also identify SFP in haploid yeast [140] and several inbred plants [91, 141-145] while concurrently generating estimates of gene expression from segregants [91, 140]. Utilizing RNA to characterize SFPs also creates the opportunity to identify gene expression markers (GEMs) – genes that are differentially expressed between parents of mapping populations and show Mendelian segregation of expression values within progeny [145]. Generating genotypic and gene expression data in a common assay establishes a framework for powerful forward genetic approaches, including genetical genomics studies [78]. However, while microarray-based mapping has been successfully applied to haploid or homozygote lines, the approach has yet to be demonstrated in outcrossing plant species with high genetic diversity, where up to four alleles can segregate for each locus in a full-sib pedigree.

RNA-based SFP genotyping requires robust separation of microarray signal variance associated with differential hybridization kinetics between alleles from variance due to differences in mRNA abundance [140]. Previous studies in species with limited genetic diversity have relied on short ($\leq$25 -mer) oligonucleotide probes to detect genetic variants because a unique single nucleotide polymorphism (SNP) can result in differential hybridization and detection of

SFP [146]. Short oligonucleotide-based microarrays typically utilize multiple probes per gene (a probeset) to estimate gene expression. Thus, SFP-containing probes can be detected by comparing individual probe signal to the signal measured across the probeset. Probes for which the signal deviates significantly from the probeset mean in a subset of the segregating population suggest the presence of a segregating SFP, while the remainder of the probeset provides an estimate of gene expression [140, 143, 145]. However, in outcrossing species with extensive genetic diversity, abundant SNP variation and heterozygosity can result in significant bias for estimates of gene expression, since SFP may be present within many probes in a given probeset [146]. Such biases render platforms that utilize short probes less reliable for concurrent analysis of gene expression and genetic polymorphisms in these experimental settings. Utilizing long oligonucleotide probes may improve estimates of gene expression in these cases. However, approaches to select optimal long-oligonucleotide probes for gene expression analysis in highly diverse species or across multiple related species and their hybrids are lacking. Similarly, the capacity of longer probes to detect a useful quantity of segregating polymorphisms for genetic mapping has yet to be demonstrated.

Our objective in this study was to develop an approach to select optimal long-oligonucleotide probes for gene expression analysis and microarray-based genotyping in a highly heterozygous population. We utilized an interspecific pseudo-backcross of *P. trichocarpa* X *P. deltoides* and a long oligonucleotide (>50-mers) microarray platform to develop a two-step method to discover candidate SFP in parent lines, then genotype sequence- and expression-based polymorphic features in the progeny. We show that genotypic data generated by this method can contribute to the development of an accurate high-density, gene-based genetic map. Additionally, the value of these markers is demonstrated by the positioning of almost half of the previously

unassembled whole-genome shotgun (WGS) sequence scaffolds within the complex and highly heterozygous genome of *P. trichocarpa*. The results we describe provide a flavor for both the challenges and opportunities presented when undertaking a microarray-based genetic mapping study in a genetically diverse plant species. Therefore, we believe the techniques we present can provide a strong framework for future microarray-based genotyping in crops, forest tree species, and other complex plant genomes. Similarly, our approach for optimal probe selection for gene expression analysis within or between highly diverse species may prove useful for other agricultural and forest tree species with similar levels of genetic diversity.

Materials and Methods

Plant Growth Conditions and RNA Isolation

A pseudo-backcross population (Family 52-124) derived from the cross of a female *P. trichocarpa* × *P. deltoides* hybrid (genotype '52-225') and a male *P. deltoides* (genotype 'D124') was obtained from the Department of Forestry at the University of Minnesota-Duluth as hardwood cuttings. After rooting, bud break and shoot elongation, fresh softwood terminal cuttings were harvested and placed in rooting media pellets (Jiffy Forestry Products, Lorain, OH) for two weeks. Rooted cuttings were planted in two gallon pots and grown for six weeks in ebb-and-flow benches in a greenhouse under long day conditions (16h light:8hr dark) and standard nutrient regime (Hocking's Modified Complete Fertilizer [147]) supplemented with 25mM nitrogen (NH_4NO_3). Plants were distributed in the greenhouse in a partially balanced incomplete block design with three biological replications per genotype. At harvest, the main plant organs (stems, roots, leaves and sylleptic branches) were collected separately. Stems were further dissected into secondary xylem tissue and phloem/bark/immature xylem. Leaf, secondary xylem, and root tissue from two biological replicates of each genotype were selected for gene expression analysis. All tissue was flash-frozen in liquid nitrogen immediately after harvest and stored at -

80° Celsius (C) prior to lyophilization and subsequent RNA isolation [148]. RNA samples were treated with RQ1 DNase (Promega USA, Madison, WI), purified in RNeasy Plant Mini Kit columns (Qiagen USA, Valencia, CA), and integrity evaluated on 1% w/v agarose gels.

Microsatellite (SSR) Genotyping and Framework Map Construction

Parent trees and 418 progeny of Family 52-124 were genotyped for 167 framework SSR loci ([46, 47, 149]). DNA was isolated from leaf samples using the Qiagen DNeasy Mini Kit (Qiagen USA) following the manufacturer's protocol. PCR reagents and concentrations were as described [149], except that SSR loci were amplified from 7.5ng genomic DNA and amplified fragments were labeled by incorporation of fluorescein-12-dUTP (Roche Diagnostics, Germany) included at 8µM. Amplification conditions were 94°C initial denaturing for 5min; 9 cycles of touchdown: denaturing at 94°C for 15sec, annealing 59°C-50°C one cycle each in one degree increments for 15sec, extension at 72°C for 30sec; followed by 21 cycles of denaturing at 94°C for 15sec, annealing at 50°C for 15sec, and extension at 72°C for 30sec; with a final extension at 72°C for 3min. Fragments were detected as described [149] except in an Applied Biosystems (ABI; Foster City, CA) Prism 3730xl DNA analyzer. Alleles were identified and genotyped in GeneMapper 4.0 (ABI) and/or GeneMarker 1.5 (SoftGenetics LLC, State College, PA).

Single-tree framework maps were constructed in MapMaker v.3.0 [150] as described [73, 151] and were anchored to the *P. trichocarpa* genome assembly v.1.1 through a BLASTN analysis [152] of PCR primer sequences for each marker. Proper placement of markers was confirmed by comparison of sequence predicted and experimentally determined *P. trichocarpa* SSR amplicon lengths.

SSR used to confirm map position sequence scaffolds were identified using MsatFinder v.2.0 based on scaffold sequences from v.1.1 of the *P. trichocarpa* genome sequence. Primers were designed within the MsatFinder interface (Table 2-1) and SSR loci were amplified from 96

Family 52-124 progeny as described above. Thirteen of the sixteen loci segregated highly heterozygous alleles between the *P. trichocarpa* and *P. deltoides* backgrounds and were genotyped using agarose gel electrophoresis (1% w/v). The remaining three loci were scored using polyacrylamide gel electrophoresis as described [153].

Microarray Analysis of Parental Genotypes

RNA extracted from root, leaf, and secondary xylem of the parents of Family 52-124 was converted to double-stranded (ds-) cDNA (SuperScript Double Strand cDNA Synthesis Kit, Invitrogen USA, Carlsbad, CA) with oligo-dT primer (Promega USA) according to the manufacturer's protocol, except that synthesis of first and second strands were extended to 16h. Resultant ds-cDNA was labeled using cy3-tagged random 9mers and Klenow fragment for 2h at 37°C, denatured at 95°C for 5min and hybridized to custom *in situ* synthesized oligonucleotide microarrays (produced by NimbleGen, Madison, WI) at 42°C overnight (16-20h).

Microarray probe design

A total of 55,793 gene models derived from the annotation of the *P. trichocarpa* genome sequence were represented in the microarray used in the analysis of the two parents of Family 52-124. 60-mer oligonucleotide probes were designed based on NimbleGen standard procedures that optimize uniqueness of the targeted genomic region and GC content, while minimizing self-complementarities and homopolymer runs. The highest ranking 6-7 probes (probeset) were selected to represent each gene model, with optimal probe spacing leading to uniformly distributed, non-overlapping coverage. Twenty negative control probes utilized in previous studies [4] were also included to serve in background quantification.

Statistical analyses

Raw signal data from all hybridizations was background subtracted, $\log_2$ transformed, and quantile normalized [154]. Normalized signal detected for each probe was centered to zero and

analyzed in a gene-by-gene mixed ANOVA model (Equation 2-1) in SAS 9.1 (SAS Institute, Cary, NC) with genotype *i* (1 degree of freedom [df]), tissue *j* (2 df), tissue *i* by genotype *j* interaction (2 df), probe *k* (5-6 df) and genotype *i* by probe *k* interaction (5-6 df) as fixed effects.

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + e_{ijkl} \quad (2\text{-}1)$$

F-tests for all fixed effects, as well as least-square mean estimates were obtained, and correction for multiple tests was made using a modified false-discovery rate (FDR) threshold (FDR < .025, Table 2-2 [155]). Normalized $\log_2$ transformed signal values from microarrays derived from differentiating xylem tissue samples were analyzed separately using a similar model that excluded tissue effects. Pairwise *t*-tests were implemented to contrast least-square means estimates of the interaction detected between the two parents for each probe in a probeset. Resulting *p*-values were corrected for multiple testing as above (FDR<.1).

Microarray Analysis of Family 52-124

Based on the probes selected from the parent tree data, a modified microarray was designed for analysis of the progeny of Family 52-124. The modified microarray was comprised of 67,897 probes including the pre-selected 55,793 gene expression probes and 12,084 SFP genotyping probes, plus 20 controls [4]. Microarrays were synthesized in NimbleGen's four-plex platform and utilized for analysis in the progeny. RNA isolated from one biological replicate of secondary xylem in 154 progeny genotypes was converted to ds-cDNA, labeled, and hybridized as described above.

All 67,877 experimental probes were evaluated for Mendelian segregation in the progeny based on *k*-means clustering procedures modified from those described previously [143]. Briefly, quantile normalized, $\log_2$-transformed signal values detected for each probe in the progeny of Family 52-124 were separated into two clusters using "Proc Fastclus" in SAS 9.1. Cluster membership was tested for the expected 1:1 segregation using a χ^2-test. Probes for which cluster

frequencies deviated significantly ($\chi^2_{df=1} > 3.84$, $p < 0.05$) from the expected segregation were discarded.

Subsequently, the probability that an individual assigned to one cluster is not a member of the other cluster was evaluated by calculating the *p*-value (p_i) associated with the modified normal deviate (Equation 2-2):

$$z_i = |(x_i - m_j)/s_j| \quad (2\text{-}2)$$

Where x_i is the signal at a given probe for an individual assigned to cluster *i*, and m_j and s_j are the mean and standard deviation of signal at that probe for all individuals assigned to cluster *j* [143]. We used $z_i > 1.96$ ($p_i < 0.05$) as evidence that the two allelic classes were clearly distinguishable, and scored individuals below this threshold as missing data. Probes resulting in >10% missing data ($n \geq 15$) were not considered for mapping.

Grouping, Ordering, and Mapping of SSRs, GEMs, and SFPs to Linkage Groups

Selected GEM and SFP markers, in conjunction with SSR markers utilized for the framework mapping, were grouped and ordered using MadMapper V248 linkage mapping software essentially as described [145]. However, because MadMapper scripts were developed for marker grouping and ordering in advanced generation *Arabidopsis* recombinant inbred lines, estimates of pairwise recombination frequency provided differ from those experimentally observed in a first generation backcross pedigree structure [156]. In addition, only microarray-based markers grouping together with at least one SSR from the established framework map were included, as probes not linked to the framework are likely to have an excess of genotyping error.

Markers were re-genotyped after localization of recombination breakpoints using a structural change analysis method with the 'Strucchange' statistical module in R [157], in a

strategy initially described by Singer *et al*. [64]. Structural change analysis detects large pattern shifts in a dataset based on a Bayesian Information Criterion (BIC) statistical threshold, and can be used to detect overall change between phases of alleles that are characteristic of recombination breakpoints.

To contribute to the 'Strucchange' analysis of breakpoint positioning, the *p*-value (p_s) associated with the standard normal distribution (Equation 2-3) for the cluster of assignment was determined:

$$z_s = |(x_i - m_i)/s_i| \quad (2\text{-}3)$$

The *p*-values for each distribution (z_i and z_s, Equations 2-2 & 2-3) were compared by the calculating the ratio *R* (Equation 2-4), which has a range from zero to one, analogous to the procedure described previously [64]:

$$R = p_i /(p_i + p_s) \quad (2\text{-}4)$$

If the alleles are highly distinct (i.e., clearly form separate distributions), individuals from the population return values of *R* very close to zero or one, depending on their allele. However, markers with little allelic distinction accumulate individuals at intermediate levels of *R*. Utilizing a continuously distributed allele score such as *R* also provides a direct assessment of confidence associated with an assigned genotype on an individual-by-individual basis and thereby contributes to more concretely defined breakpoints in the 'Strucchange' analysis.

To verify proper placement of recombination breakpoints, agreement between 'Strucchange' genotypic results and raw SSR genotypes were inspected. Additional breakpoints supported by the 'Strucchange' minimum BIC statistic, but not present in the SSR data, were accepted if they included at least 3 microarray-based markers. Subsequently, genetic distances for the corrected genotypes were estimated with MapMaker v.3.0 [150].

Sequence-Level Characterization of SFP Alleles

A subset of mapped SFPs was arbitrarily selected for sequence-level characterization in each parent of Family 52-124. PCR primers were designed from genome sequence surrounding five mapped SFP (Table 2-3). Alleles were amplified from each parent tree using ~50ng xylem ds-cDNA, 200µM dNTPs, 2µL 10X Advantage 2 PCR buffer, and 0.4µL Advantage 2 polymerase mix (Clontech Laboratories, Inc., Mountain View, CA) in a total volume of 20µL. PCR was conducted in a two-step procedure with identical amplification conditions for each step: 95°C initial denaturing for 5min, 30 cycles of 95°C denaturing for 30sec, 58.5°C annealing for 30sec, 72°C extension for 1min 45sec, with a final extension of 72°C for 7min. Secondary PCR was conducted using identical reagent concentrations, except that a 1:25 dilution of the primary PCR was substituted as template. Amplicons from the secondary reaction were gel purified in 1% w/v agarose and cloned into pGEM-T (Promega USA) following the manufacturer's protocol. Eight to ten independent clones per construct were isolated using the QIAprep miniprep kit (Qiagen USA) and sequenced bidirectionally from the SP6 and T7 promoters in an ABI Prism 3730xl. Resulting sequences were aligned and analyzed in Sequencher v.4.6 (Gene Codes Corporation, Ann Arbor, MI) and ClustalW v.2.0 [158].

Results

SSR Framework Map of Genotype 52-225

We constructed a single-tree framework microsatellite map (Figure 2-1) for the maternal *P. trichocarpa* X *P. deltoides* hybrid parent (genotype '52-225') of Family 52-124 based on 167 SSR markers, using a pseudo-testcross strategy [73, 151]. The framework map represented the 19 consensus linkage groups (LG) of poplar [159], though an unresolved gap remained in LG_VI due to a lack of informative markers in this region. Markers shared with the genetic map of genotype '52-225' produced for a different population (Family 13, for which the genotype also

serves as the maternal parent; [40]) were largely collinear. Framework SSR loci represented a subset of the sequence-tagged sites used to assemble the *P. trichocarpa* WGS contigs and scaffolds into chromosome-scale linkage groups [4]. Based on that information, we anchored and oriented the framework map relative to the genome assembly. The framework map spanned 2970cM with average marker intervals comprising 17.8cM and served as the basis for subsequent grouping of SFP and GEM markers into linkage groups.

Identification of Probes for Genotyping Family 52-124

A microarray analysis was initially conducted in each parent line to (1) identify candidate SFP probes segregating in the pedigree and (2) identify a single optimal probe for gene expression analysis in the progeny. To develop a microarray platform that could be used for concurrent genotyping and transcript profiling the progeny of Family 52-124, we began by testing 6-7 probes per gene in the two parents. The custom platform was comprised of 384,287 60-mer features representing 55,793 annotated gene models (probesets) from the sequenced genome of *P. trichocarpa*. This gene set included 45,555 predicted gene models reported previously plus 10,238 ESTs and less supported gene models with transcriptional evidence [4].

For the probe selection study, RNA extracted from root, leaf, and secondary xylem of each parent of Family 52-124 was converted to double-stranded (ds-) cDNA, labeled, and hybridized to the microarrays. After normalization, the data was assessed in an analysis of variance (ANOVA) with genotype, tissue-by-genotype interaction, probe and genotype-by-probe interaction effects. Genotype effect accounts for overall differences in signal in a probeset between the two parental genotypes, and primarily reflects a difference in gene expression level between them (Figure 2-2 C, D). The tissue effect accounts for differences in expression detected by a probeset between different tissues, regardless of the genotype being profiled. Probe effect detects specific properties of probe that distinguish it from others in a probeset, independent of

parent genotype (Figure 2-2 A, D). Finally, the genotype-by-probe interaction accounts for specific properties of a probe that distinguish it from the rest of the probeset, dependent on the genotype being analyzed. Dependence on genotype suggests that these probes contain SFP between the parental genotypes that may segregate in the progeny (Figure 2-2 B, D).

To identify candidate probes for SFP genotyping, two separate analyses were conducted. In the first, a *t*-test was used to contrast least-square mean estimates of the interaction between the two parental genotypes at each probe, across all tissues. A probe within a probeset may be biased towards one or the other parent due to differential hybridization (i.e. an SFP) and therefore is a candidate to be tested for segregation in the progeny. Furthermore, only probes for which the difference in least-square means between the parental lines exceeded an arbitrary four-fold threshold were selected. We identified 2875 probes meeting these criteria (FDR < .1; p < .0085). When more than one probe from a probeset was identified, we selected the most significantly interacting probe. In total, candidate SFP probes were identified for 912 genes. Among these, 770 exhibited hybridization bias favoring the '52-225' hybrid parent, while 142 demonstrated stronger hybridization in the 'D124' *P. deltoides* parent. These results are expected because microarray probes were designed based on the genome sequence of *P. trichocarpa* [4], one of the species contributing to the hybrid parent. Therefore we hypothesized that the majority of candidate SFP may be explained by species-level polymorphism between *P. trichocarpa* and *P. deltoides* alleles. Based on this hypothesis and the interspecific pseudo-backcross pedigree structure, comprised of one *P. trichocarpa* and three *P. deltoides* alleles, we expected that most SFP and GEM alleles showing simple Mendelian inheritance should segregate at a ratio of 1:1.

To identify additional candidate SFP probes for genotyping and mapping in the progeny, we reanalyzed the parental expression data derived from secondary xylem in a separate ANOVA.

Similar to the previous analysis, we contrasted each parent's interaction with individual probes within a probeset, and selected those that were significant (FDR <.1, p <.0051) with at least a three-fold difference in least-square means estimates. The separate analysis focusing on xylem tissue was conceived after previous work showed this tissue to be among the most transcriptionally diverse in *Populus* [4]. From this dataset, we initially identified 13,191 additional candidate SFP probes, including 8986 with hybridization bias favoring the hybrid parent and 4205 showing hybridization bias favoring the *P. deltoides* parent. By again selecting only the most significantly interacting probe in each probeset, we identified an additional 11,172 genes harboring candidate SFP. In total, our two analyses identified single, specific probes from 12,084 genes containing candidate SFP, which were subsequently carried forward for analysis of the progeny.

Identification of Probes for Transcript Profiling Family 52-124

A second objective of the microarray analysis of parental genotypes was to identify a single, optimal probe for expression analysis of the 55,793 gene models in the 52-124 progeny. To identify probes unbiased for gene expression analysis in both parental species backgrounds, we assumed that the probeset mean best represents the true expression value in each parent. Therefore, contrary to the previous analysis, the goal was to select the probe that performs most consistently within the probeset in both parents.

To select the optimal probe for gene expression analysis, an iterative selection process was implemented. (1) For each gene, probes were ranked based on the deviation of the least-square mean estimate of each probe effect, relative to the probeset mean. Probes that do not deviate significantly from the probeset mean suggest that inherent properties of the probe do not contribute bias to the signal detected at that probe. (2) Next, the highest ranking probe was analyzed for its sequence alignment uniqueness scores assigned during probe design. Only

probes with no more than one unique match to the *Populus* genome sequence were further considered. (3) Finally, probes were evaluated for significant (FDR < .1) genotype-by-probe interaction. In cases where the probe was not unique or showed a significant genotype-by-probe interaction, the next highest ranked probe was evaluated (i.e., next step of the iteration). After seven iterative rounds of selection, all probes had been considered by these criteria, and probes to measure gene expression were selected for 46,001 genes.

The selection for the remaining 9792 genes was based on a rank variable provided by NimbleGen. The rank variable concurrently accounts for probe chemical properties and probe uniqueness characteristics. The highest ranked probe for each gene exhibiting non-significant probe effect and genotype-by-probe interaction effect was selected. For 149 genes, all probes in the probeset exhibited a significant probe effect or genotype-by-probe interaction. Single probes were chosen for these genes solely on the basis of the NimbleGen rank variable.

Genotyping SFP and GEM Probes in the Progeny of Family 52-124

To evaluate the candidate SFP probes identified in the parent genotypes, we assayed RNA abundance in xylem tissue from 154 progeny of Family 52-124. A modified microarray was designed, comprising the single selected expression probe per gene for each of 55,793 gene models and the 12,084 candidate SFP probes. Loci were genotyped using a *k*-means clustering allele-calling procedure (Materials and Methods). Normalized data for each of the 67,877 experimental probes was grouped into two separate clusters, and frequency of cluster membership was tested for 1:1 segregation ($\chi^2_{df=1} < 3.84$, $p > .05$). A total of 12,680 features followed the expected Mendelian segregation pattern, including 9782 probes selected for gene expression analysis (17.5%), and 2898 of the candidate SFP probes (24%). Gene expression

probes that segregate in the mapping population may be utilized as gene expression markers (GEMs) and therefore were considered in further analyses.

Next, signal separation between allelic classes was evaluated through a modified normal deviate and probes resulting in >10% ambiguous allele assignments were removed. Reliable genotypes in >90% of the progeny were obtained for 1733 probes, including 1014 GEMs and 719 SFPs (1.8% and 6% of total, respectively). The 1733 segregating features correspond to 1610 independent gene models – segregating probes corresponding to both GEM and SFP were identified for 123 gene models.

Genetic Mapping of Genotype '52-225'

The 1733 candidate SFP and GEM probes were utilized to generate a genetic map of genotype '52-225'. Marker grouping, ordering and mapping was carried out as described previously [145], with slight modifications (Materials and Methods). To correct for genotypic errors and ambiguities in the resulting linkage groups, markers were re-genotyped after localization of recombination breakpoints using a structural change analysis (Materials and Methods). In addition to the 167 framework SSRs, we unambiguously placed 324 SFP and 117 GEM loci in the map of '52-225' (Table 2-4, Figure 2-3). For most linkage groups, and the genome as a whole, average marker intervals were <5cM. The total genome length was 2798.5cM, in good agreement with recently published genetic maps for interspecific crosses of *Populus* [40]. Overall rate of marker placement error was low: for genes known to be physically located on specific chromosomes in the *P. trichocarpa* WGS sequence assembly, ten were not placed in their predicted linkage group – an error rate of 3.52% (10/284). Of the misplaced markers, seven corresponded to SFPs and three to GEMs. These ten markers were subsequently excluded from the map.

Physical Orientation of the '52-225' Genetic Map

We oriented and aligned the '52-225' genetic map to the chromosome-level WGS assembly of *P. trichocarpa* 'Nisqually-1' based on physical positions of genes interrogated by SFP and GEM probes [4] and our previously anchored SSR loci. In large measure, the predicted genetic orientation and physical orientation were collinear – several small inversions were detected that could be the result of error in map ordering or represent true differences in gene order between different *P. trichocarpa* clones or between *P. trichocarpa* and *P. deltoides* (data not shown). Slight variations in map order between 'Nisqually-1' and '52-225' have been reported elsewhere [125]. On average, predicted physical intervals between ordered markers contain 84.4 genes; however, the range is wide from one to 624 genes. Average physical distance spanned by marker intervals is 725 thousand base pairs (kb) and ranges from 146 base pairs (bp) to 5.31 million bp (Mbp).

Genetic Mapping of the Unassembled *Populus* Genome

Approximately 7700 sequence scaffolds from the WGS assembly are not assigned to specific linkage groups in v.1.1 of the *P. trichocarpa* genome sequence. These scaffolds vary in size from <100bp to >3.5Mbp (mean ≈ 16.8kb) and represent 75Mbp of unplaced sequence [4]. Much of this sequence was postulated to be heterochromatic or derived from substantially divergent haplotypes in the sequenced clone [4, 42]. Our microarray-based mapping results provided an unprecedented opportunity to anchor a large amount of this unplaced sequence to potential genomic locations in *P. trichocarpa* based on the genes physically localized within these sequence scaffolds. Among our 1733 candidate GEM and SFP markers, 783 were contained in genes residing in 492 sequence scaffolds. We successfully mapped 167 of these 783 loci, thereby placing 116 sequence scaffolds to unique genetic locations in linkage groups (Table 2-4). Five remaining scaffolds demonstrated linkage to other markers in the map, but could not

be unambiguously placed within a single linkage group (data not shown). This error rate associated with scaffold mapping (4.13%; 5/121) is congruent with the mapping error rate observed for markers with known position in the linkage group WGS assembly (see above). The 116 sequence scaffolds placed to the genetic map correspond to 35.7Mbp of WGS sequence assembly, or nearly 50% of the unlinked sequence (Table 2-5). Among these mapped scaffolds, 34 (representing 23.3Mbp) could be linked by two or more markers, providing an orientation of sequence strands comprising the scaffolds.

Verification of Map Position for Unassembled Sequence Scaffolds

To confirm that our assembly of genomic scaffolds using SFP and GEM markers was reliable, we verified the position of a subset of mapped scaffolds using SSRs. From the *P. trichocarpa* v.1.1 sequence scaffolds [4], we identified SSR loci within ten distinct scaffolds mapped using GEM and SFP markers and designed PCR primers in their flanking sequences. Nine of the ten SSR segregated discernably, and after amplification and genotyping, we mapped these loci on the basis of only the original framework SSR map, to eliminate any bias that could be introduced due to genotyping error in linkage group anchored SFP and GEM alleles. For eight of the nine scaffolds, we successfully verified the putative map location of the scaffold sequence with respect to the framework SSRs (Table 2-6, Figure 2-1). Relative genetic distance between scaffold-anchored markers in both the SFP/GEM-based map and SSR framework map were also in agreement (Table 2-6).

The lone scaffold (scaffold_121) for which we could not verify map position using this technique was placed on the basis of a single GEM to LG_XVIII, whereas data from two SSR consistently positioned it within LG_VI. Therefore, we speculated that this result was attributable to strong *trans*-acting regulation from LG_XVIII acting on the gene characterized as a GEM. Since GEM may be the result of either *cis*- or *trans*-acting variation, we were interested

to determine if scaffolds mapped based on single GEM loci were less reliable for their genetic positioning. We considered the remaining 13 scaffolds that were placed on the basis of single GEMs in our map, and identified informative SSR in six of these 13 scaffolds. Using the framework SSR map, we successfully verified the predicted genetic placement for five of these six GEM-anchored scaffolds (Figure 1, Table 2-7). The single unverified scaffold (Scaffold_250) localized to the same linkage group, but a different SSR bin, than predicted by the GEM locus (Table 2-7).

Characterization of Sequence-Level Allelic Variation Represented By Mapped SFPs

SFP detected by short (≤25 -mer) oligonucleotide probes often correspond to one or few SNPs or small indels [143, 146, 160]. However, the implication of sequence mismatches on signal detected from long oligonucleotide probes has only recently been described [161]. Thus, we characterized the allelic variations present in a sample of mapped SFP probes from the microarray platform. Using ds-cDNA produced from xylem for each of the parent trees, we amplified, cloned, and sequenced regions corresponding to five mapped SFP loci and assayed polymorphisms between the alleles. We identified sequence-level variation ranging from single SNP in the 60-mer region to large in-del polymorphisms affecting >10bp (Figure 2-4 B-E). Of the five SFP we characterized, one exhibited no variation between alleles within the sequence interrogated by the genotyping probe, though sequence variation between the alleles and probe was observed (Figure 2-4 A). Therefore, this probe may correspond to an actual GEM that was mischaracterized as an SFP, as previously described [143]. As we hypothesized, SFP we detected are primarily due to species-level sequence polymorphisms between *P. trichocarpa* and *P. deltoides*, though multiple haplotypes were identified at two of the five probes (Figure 2-4 B,E).

Discussion

Parallel genotyping and expression quantification using mRNA microarray hybridization data require accurate classification of differences in signal intensity arising from DNA sequence variants versus transcript level abundance [140]. To separate genetic polymorphism from differences in transcript abundance, candidate genotyping probes can subsequently be detected by identifying individual probes that deviate significantly from the probe set mean signal (which provides a balanced measure of expression), and that segregate in the progeny. Although first demonstrated in populations with simple genetic segregation patterns (i.e. haploid, recombinant inbred line and doubled-haploid) and species with limited genetic diversity, we have extended mRNA-based microarray genotyping to a highly heterozygous, outcrossing plant species for which low resolution at the genotype level has often hampered forward-genetic gene discovery methods.

Contrary to previous studies, which relied on microarray platforms comprising multiple (11–30) short probes (≤25 -mer) per gene [63, 140, 143], we adopted a long-oligonucleotide microarray platform for use in our study. Furthermore, our analysis relied on single optimal genotyping and gene expression probes selected by analyzing the parental individuals before characterizing the segregating population. A set of six or seven probes per gene was first screened in the parental genotypes, and an analysis of variance was applied to identify probes interrogating potential polymorphisms and optimal probes for measuring transcript levels [142, 144]. Next, the microarray platform was re-designed to comprise a single optimal gene expression probe for each transcriptional unit and 12,084 candidate SFP probes for analysis of 154 segregating progeny. From this analysis, we identified 1733 segregating features with reasonably low levels of ambiguous data (<10%). After applying a statistically based genotyping correction described previously [64], we successfully mapped 441 of these segregating features

(25.4%). Our mapped features include probes that were pre-selected for gene expression analysis and those pre-selected for SFP genotyping, corresponding to 117 GEM and 324 SFP markers. The sample of sequenced SFP regions indicates that our data analysis approach robustly detected sequence variants from RNA-based microarray data.

Together with 167 framework SSR markers, our map represents one of the highest-resolution genetic maps derived from a single pedigree in the *Populus* genus. Markers from the framework SSR map represent an important tool to delineate true versus spurious linkage of GEM and SFP to linkage groups in the genome, analogous to the situation described when mapping largely homozygous barley RILs [143]. Nonetheless, we have demonstrated that GEM and SFP mapping in highly heterozygous species is both beneficial and feasible, and may serve as a supplement to traditional DNA-based markers. Our study focused on an inter-specific cross, in which sequence and gene expression variation may be extraordinary. However, estimates of genetic variation and nucleotide diversity within individual species of the *Populus* genus [54] and other economically significant outcrossing plants [59, 162-164] suggest that our analysis approach could also be adapted to identify genetically informative variants from diverse intra-specific accessions. However, it is expected that variables including probe length and statistical thresholds associated with allele calling may require optimization, and that the abundance of SFP and GEM detected may be lower.

Establishing a high-density, gene-based genetic map also provided an opportunity to position previously unlinked sequence scaffolds from the WGS sequence assembly of *P. trichocarpa* to putative genomic locations. The existing genome assembly comprises 410 Mbp of a total estimated genome size of 485 Mbp [4], but there is substantial variation in estimated chromosome sequence coverage, from 56% (chromosome XVII) and 65% (chromosome XIX) to

estimated completion (chromosomes X, XII and XIII). Of the 492 unplaced scaffolds in which we identified a segregating GEM or SFP marker, we unambiguously positioned 116 on our genetic map (23.6%). Scaffold sequences mapped using our GEM and SFP markers represent over 35 Mbp of previously unanchored sequence from the WGS assembly of *P. trichocarpa*, including more than 23 Mbp localized by at least two independent markers in the same scaffold.

Of a sample of 15 putatively placed genomic scaffolds, the placement of 13 could be verified using independent SSR markers, lending a good degree of confidence to our map-based re-assembly of nearly 50% of the *P. trichocarpa* scaffold sequence. In addition, 18 scaffolds that we mapped using SFPs or GEMs have been previously mapped using SSRs and amplified fragment length polymorphisms by other research groups (A. Rohde, Institute for Agriculture and Fisheries Research, University of Ghent; personal communication). Our microarray-based markers verified the genetic position for 17 of these 18 scaffolds. Misplacement of sampled scaffolds based on microarray marker data was generally due to mapping based on single GEM loci. Because GEMs can result from segregating *cis*- or *trans*-acting regulatory variation, scaffolds mapped based only on GEMs should be verified for their position using SSRs where possible. Despite this fact, localization of a large proportion of the previously unplaced genome sequence is a high-impact result for the *Populus* genomics community, even given the small degree of potential error in placements.

Interestingly, the newly mapped scaffolds are predominantly located in chromosomes with low sequence coverage, where larger gaps exist in the current assembly. It is unclear why there is bias towards mapping scaffolds in chromosomes with poor assembly. There may be a higher probability of mapping unassembled scaffolds to them simply because of their higher expected abundance there. Alternatively, smaller unmapped scaffolds could be more prevalent in

chromosomes that are populated by large numbers of hypervariable regions, as high levels of polymorphism are not favorable to long-range WGS assembly of a consensus haplotype [42]. Such an observation was recently made in the sex-determining telomeric region of *Populus* chromosome XIX [125]. Furthermore, chromosome XVII, which has the lowest estimated percentage of sequence fully assembled (56%, [4]), has the fourth highest rate of sequence polymorphism (unpublished data), and has the highest number of scaffolds mapped and total sequence added in our study. Although we can only speculate as to the basis for this phenomenon, our study provides a significant improvement to the WGS assembly of the *P. trichocarpa* sequence. Additional mapping studies using SFP and GEM markers that we have identified, and focusing on variation in the sequenced clone Nisqually-1, could shed light on the structural genomic nature of these scaffold sequences and their proper designation in the genome assembly as alternative haplotypes or bona fide unplaced WGS sequence segments. *De novo* sequencing and assembly of other *P. trichocarpa* and *P. deltoides* genotypes will also provide better indications of whether specific regions exist that are hypervariable within and between species haplotypes, and their genome location.

Perhaps most importantly, our effort demonstrates the power that microarray-based mapping may bring to future map-based WGS reassemblies. We have shown that mapping based on physically positioned genes can rapidly localize and orient large amounts of WGS-derived sequence within the context of a physical assembly, even when the sequence is scattered amongst a number of smaller scaffolds whose assembly is not supported by traditional WGS computerized assembly techniques or anonymous sequence marker anchoring methods. Thus, further application of microarray-based mapping in genetically diverse species will not only increase resolution at the level of genotype for forward-genetic analyses, but may drastically

improve the initial quality of draft WGS assemblies to the community as a whole. In addition, providing a putative location for an unplaced sequence can identify candidate genes affecting quantitative phenotypes that would otherwise go unconsidered if relying only upon the chromosome-level sequence assembly for characterization of a genomic interval.

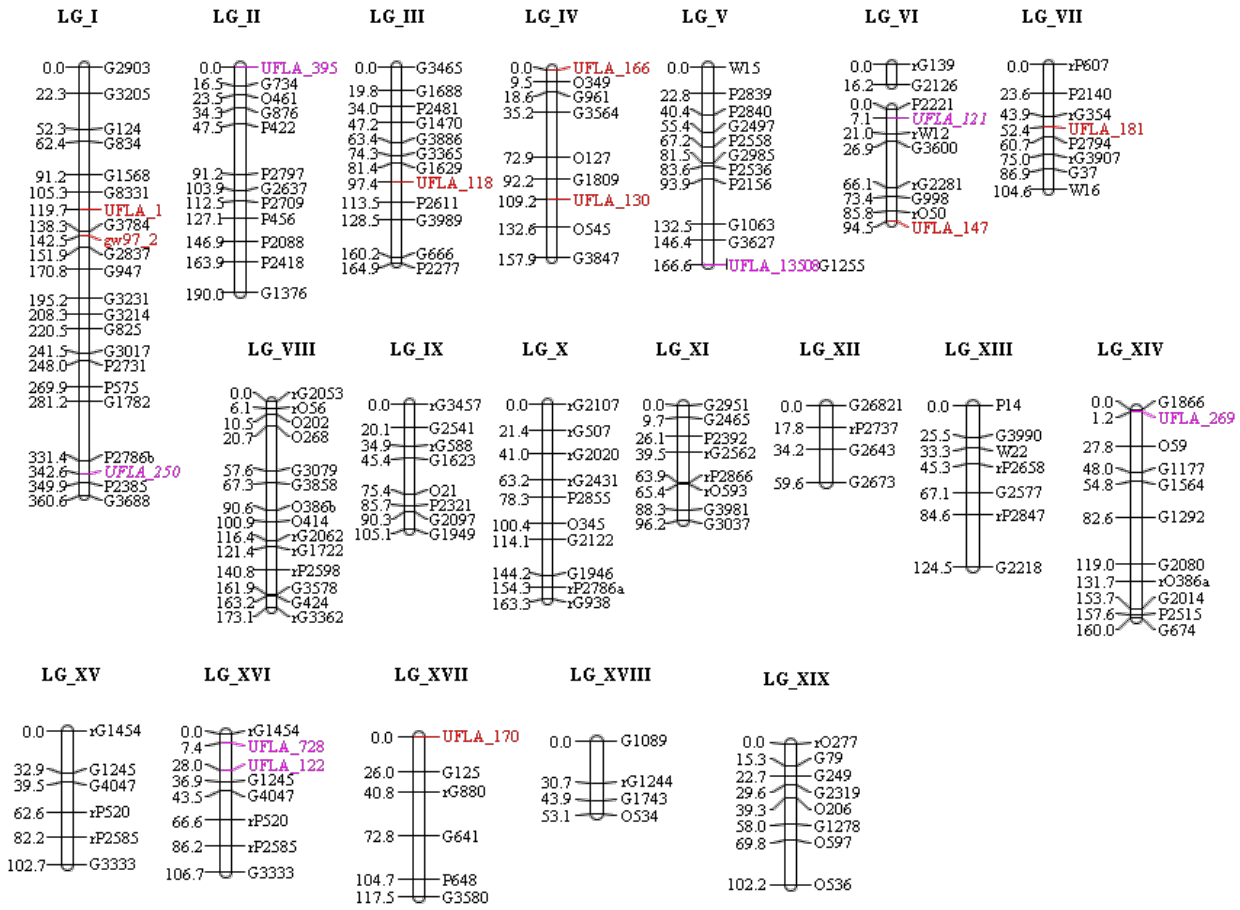


Figure 2-1. Framework SSR map of *P. trichocarpa* X *P. deltoides* clone '52-225'. Black loci: 167 framework markers genotyped in 418 segregating progeny. Red loci: SSR markers that verify positioning of scaffold sequences that were placed on the basis of genotyping SFP/GEM alleles in the mapping population. Pink loci, regular font: SSR loci verifying positioning of scaffold sequences placed on the basis of single GEM loci (Table 2-7). Pink loci, italicized font: SSR loci that fail to verify predicted positioning for scaffolds 121 and 250 based on GEM loci.

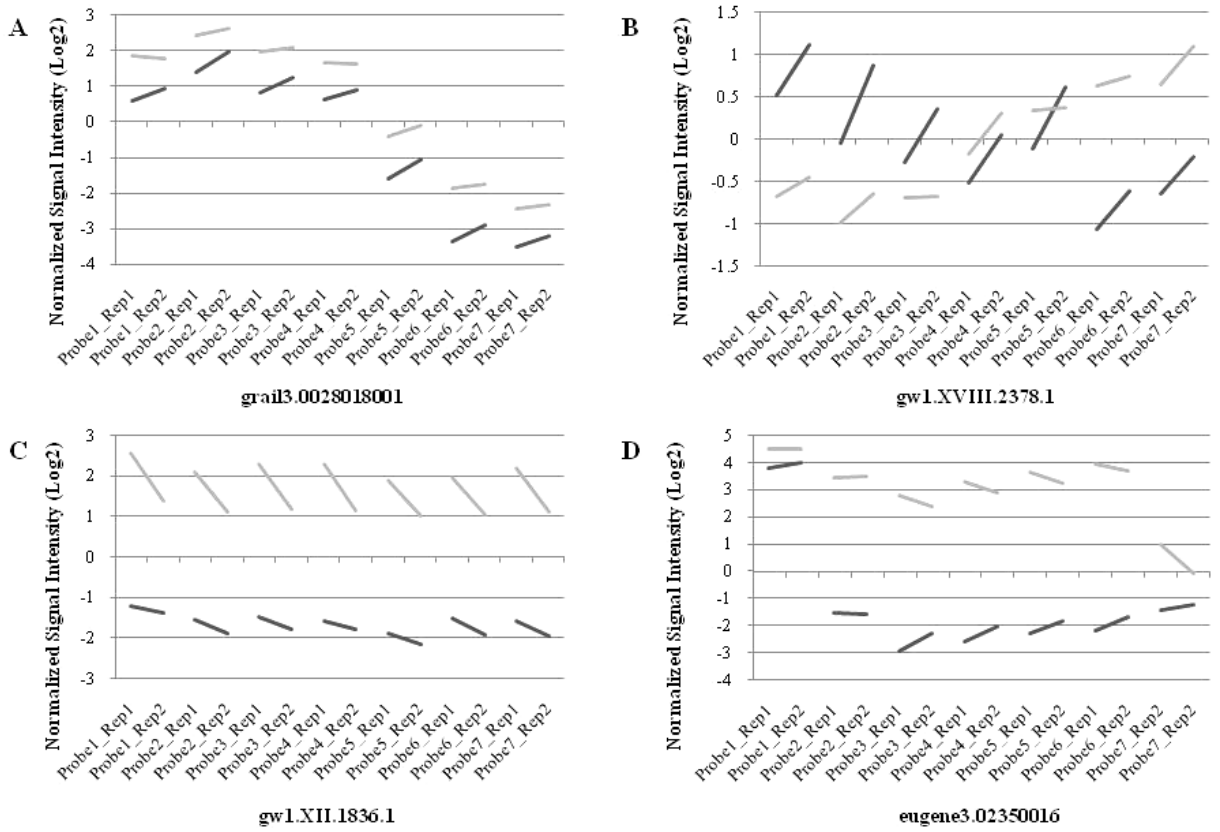


Figure 2-2. Examples of significant fixed effects detected by analysis of variance of microarray data from the parents of Family 52-124. Normalized, zero-centered signal measured in seven probes for each parent (*P. deltoides* D124 = black line, *P. deltoides* X *P. trichocarpa* 52-225 = grey line) in two biological replicates are displayed. A) Significant probe effect (gene ID grail3.0028018001) reflected by wide variation in measured signal intensity around the probeset mean (probes 2,6,7). B) Significant genotype*probe effect (gene ID gw1.XVIII.2378.1) revealed by difference in signal intensity across a probeset within one genotype (probes 1,2 for genotype 52-225; probes 6,7 for genotype D124). C) Significant genotype effect (gene ID gw1.XII.1836.1) represents a property of the probeset as a whole, and is reflected by relatively constant signal variance between genotypes at each probe across the probeset. D) Significant genotype (probes 2-6), genotype*probe (probe 7), and probe effects (probe 1) within a single gene (gene ID eugene3.02350016).

Figure 2-3. Microarray and SSR-based genetic map of *P. trichocarpa* X *P. deltoides* '52-225'. Colors and font styles represent marker types and genomic sequence locations: Black – Framework SSR markers; Green – SFP markers; Blue – GEM markers; Italicized – GEM/SFP markers contained within unplaced WGS scaffold sequences in v.1.1 of the *P. trichocarpa* genome; Plain font – GEM/SFP markers with known linkage group anchored genomic coordinates. Maps were generated with publicly available MapChart software, v.2.1 [165].

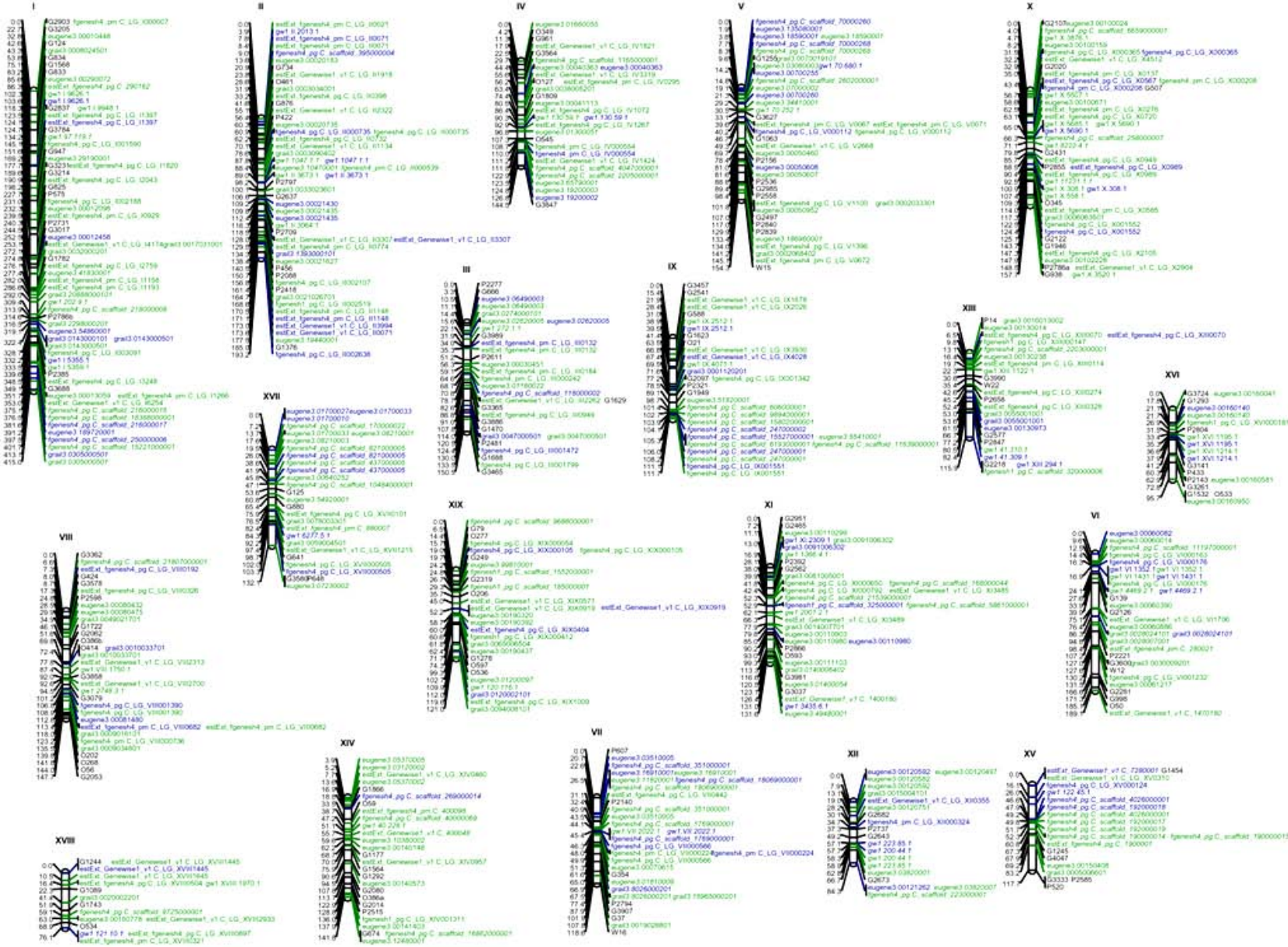


Figure 2-4. Allelic variations characterized by sequencing genomic DNA regions corresponding to mapped SFP probes. Among sequenced clones, haplotypes are shown as detected for *P. trichocarpa* X *P. deltoides* clone '52-225' and *P. deltoides* clone 'D124'. Variations between alleles or between detected sequence and probe sequence are depicted in red. A) No variation detected between parent trees for estExt_Genewise1_v1.C_LG_III2262. B) Extensive SNP and in-del polymorphism between haplotypes in grail3.0016013002. C) A 12bp deletion polymorphism in *P. deltoides* estExt_Genewise1_v1.C_LG_XVII1215. D) A single SNP distinguishes alleles of grail3.0005006601. E) Multiple SNP detected for extExt_Genewise1_v1.C_LG_XVIII1445.

Table 2-1. Primer sequences for amplification of scaffold anchored SSR loci.

UFLA locus ID	Primer name	v.1.1 scaffold	*P. trichocarpa* product size[a]	SSR motif	Motif repeats	Primer sequence
UFLA_29	Scaffold_29_ac_18_F	Scaffold_29	246	ac	18	GGATCCGTGACCAAGTTCAG
	Scaffold_29_ac_18_R					GAAGGTGCTCTTTTCCATGC
UFLA_166	Scaffold_166_ac_9_F	Scaffold_166	257	ac	9	ACACACCATGTGGCTCGTAG
	Scaffold_166_ac_9_R					AGCATCGCATCACATCACAG
UFLA_181	Scaffold_181_ag_8_F	Scaffold_181	250	ag	8	TCGGTTGGTCTGTATCATCC
	Scaffold_181_ag_8_R					ACGCAATGAGAGGTTTCTGG
UFLA_118	Scaffold_118_ag_12_F	Scaffold_118	249	ag	12	TCACGCCAGTAACCTTGTTG
	Scaffold_118_ag_12_R					TTCTCGAGGTAAGGTGTCAGG
UFLA_130	Scaffold_130_ac_5_F	Scaffold_130	246	ac	5	GGTTGGCAATCTACCCTAGC
	Scaffold_130_ac_5_R					GCTATGCGTAACCTGGAAGG
UFLA_170	Scaffold_170_ag_5_F	Scaffold_170	253	ag	5	CTTCTGGCCTCTATCATGCTG
	Scaffold_170_ag_5_R					CTGCTGACTCCAGCTCAATG
UFLA_147	Scaffold_147_ag_9_F	Scaffold_147	251	ag	9	ACGAAACCTGGAAAAGGTTG
	Scaffold_147_ag_9_R					AGCAACGCGTAATGTAAAGC
UFLA_121	Scaffold_121_ag_11_F	Scaffold_121	245	ag	11	CCCTGCTTCATGTCATTTCTG
	Scaffold_121_ag_11_R					TCTACCACAAGGGATTCTTGC
UFLA_121_b	scaffold_121.102938.at.9F	Scaffold_121	280	at	9	AAACTTTGCAACCTTGTCCA
	scaffold_121.102938.at.9R					TCATAACTCGATTTTGAATCCCTA
gw97_2	gw1.97.119.1_2L	Scaffold_97	300	n/a	n/a	GCGAATAATTGGAGAACCC
	gw1.97.119.1_2R					CCAATCTCGTCATCAACCTT
UFLA_122	scaffold_122.698025.ct.12F	Scaffold_122	240	ct	12	CACCATGCCAAGCATCATAG
	scaffold_122.698025.ct.12R					TCCATCATTTGTGTGTGTGC
UFLA_250	scaffold_250.163555.gat.5F	Scaffold_250	210	gat	5	CTCATGGTATTGGTGAGGGAAT
	scaffold_250.163555.gat.5R					CAGAGGTAGGGTCGGATTCA
UFLA_269	scaffold_269.103504.tc.5F	Scaffold_269	245	tc	5	TAATTCCACGGAATGGATGG
	scaffold_269.103504.tc.5R					TGAATTCTCTCGCTTAGCTTTG
UFLA_395	scaffold_395.21632.ta.5F	Scaffold_395	237	ta	5	CGCGACTCGAATCATGAAAT
	scaffold_395.21632.ta.5R					CATATTCACCTGCATGAAAGC
UFLA_728	scaffold_728.5575.ct.12F	Scaffold_728	256	ct	12	ACCGAAATGTGGGCTATGAA
	scaffold_728.5575.ct.12R					TTGTCAGGCTCAATCTATGATG
UFLA_13508	scaffold_13508.142.ctt.8F	Scaffold_13508	246	ctt	8	ACAGGTTGTGGGAGGCTGAT
	scaffold_13508.142.ctt.8R					TGCTTTCCACTTTTGATCCAG

[a.] Predicted SSR product size derived from the 'Nisqually-1' genome sequence are provided.

Table 2-2. Summary of F-tests for fixed effects in the mixed ANOVA conducted on parent tree microarray data.

	Genotype	Tissue	Tissue*genotype	Probe	Genotype*probe
Significant[a]	7,909	34,326	18,470	51,821	3,355
Non-Significant	47,884	21,467	37,323	3,972	52,438

[a.] Significance judged at FDR < .025.

Table 2-3. Primer sequences for amplification of sequence-verified SFP alleles. Gene models are as reported in v.1.1 of the *P. trichocarpa* 'Nisqually-1' genome sequence.

P. trichocarpa gene model	Primer forward sequence	Primer reverse sequence
estExt_Genewise1_v1.C_LG_III2262	GGTTGGTTCGGTATTGCTGT	GTATCGCACAACAGGCATTG
estExt_Genewise1_v1.C_LG_XVII1215	GTTCGGGTTATGGGAGGAAT	AGTGCCATGAATCCCATTTT
estExt_Genewise1_v1.C_LG_XVIII1445	CTCTGGTGGAAGGCTCAAAG	GCTGCTCAACTGGAAAAATCA
grail3.0005006601	CACATGGCTGGACACAAAAC	TTGGCTGGTCACTCCTCTCT
grail3.0016013002	TTGAATCTGGTGGTGGTGAA	AAGGTGACAACGAGCAGAGAA

Table 2-4. Summary statistics for *P. trichocarpa* X *P. deltoides* clone '52-225' microarray- and SSR-based linkage map.

Linkage group	Framework SSR loci	GEM loci mapped	SFP loci mapped	v.1.1 scaffolds mapped	Total mapped loci	Map length (cM)	Average marker spacing (cM)	Average recombinations (Std. Dev.)
LG_I	19	6	25	16	68	415.0	6.10	4.09 (1.51)
LG_II	11	10	25	4	52	193.2	3.72	2.18 (1.05)
LG_III	11	3	8	6	30	150.5	5.02	1.83 (0.87)
LG_IV	7	2	10	7	29	144.5	4.98	1.72 (0.79)
LG_V	11	2	12	9	41	154.3	3.76	1.74 (0.84)
LG_VI	8	4	11	4	31	189.1	6.10	1.96 (0.83)
LG_VII	7	3	6	8	31	118.6	3.83	1.57 (0.84)
LG_VIII	14	5	13	2	34	147.7	4.34	1.66 (0.79)
LG_IX	8	4	7	10	30	111.7	3.72	1.50 (0.64)
LG_X	10	8	23	4	45	157.7	3.50	1.84 (0.93)
LG_XI	8	3	11	10	33	131.6	3.99	1.73 (1.03)
LG_XII	4	4	5	3	20	84.3	4.22	1.41 (0.58)
LG_XIII	7	4	10	3	25	115.9	4.64	1.58 (0.69)
LG_XIV	10	0	6	7	27	141.8	5.25	1.65 (0.81)
LG_XV	6	1	3	6	20	117.7	5.89	1.43 (0.60)
LG_XVI	9	3	7	0	19	95.7	5.04	1.38 (0.59)
LG_XVII	5	1	5	10	28	132.1	4.72	1.69 (0.81)
LG_XVIII	4	1	9	2	16	76.1	4.76	1.62 (0.93)
LG_XIX	8	3	11	5	29	121.0	4.17	1.58 (0.86)
Unlinked Scaffolds	n/a	50	117	n/a	n/a	n/a	n/a	n/a
Genome Total	167	117	324	116	608	2798.5	4.62	1.80 (0.59)

Table 2-5. Summary of WGS scaffold sequences placed based on SFP and GEM markers, and resultant estimated coverage.

Linkage Group	Assembled size (Kbp)[a]	Estimated coverage (%)	Scaffold sequence added (Kbp)	Estimated coverage revised (%)
LG_I	35,500	80	5146.0	91.6
LG_II	24,500	91	98.5	91.4
LG_III	19,100	79	1526.0	85.3
LG_IV	16,600	95	1387.0	102.9
LG_V	18,000	78	2834.0	90.3
LG_VI	18,500	92	4295.0	113.4
LG_VII	12,800	85	582.4	88.5
LG_VIII	16,100	73	8.4	73.0
LG_IX	12,500	85	136.2	86.0
LG_X	21,100	100	137.8	100.6
LG_XI	15,100	82	1386.0	89.5
LG_XII	14,100	102	703.4	107.1
LG_XIII	13,100	107	2908.6	130.7
LG_XIV	14,700	85	3162.8	103.3
LG_XV	10,600	79	1792.2	92.3
LG_XVI	13,700	81	0.0	81.0
LG_XVII	6,000	56	5601.0	108.3
LG_XVIII	13,500	77	983.4	82.6
LG_XIX	12,000	65	2424.4	78.1
Mean	n/a	83.8	1848.1	94.5

[a.] Original assembled size and estimated coverage as reported [4]. Revised estimated coverage based on these previously reported statistics. Revised estimated coverages exceeding 100% may be due to erroneous estimation of linkage group size due to the assumption of uniform genetic:physical distance ratio, or may result from the map-based linear reassembly of highly divergent haplotypes that should, in fact, be collinear and distinct.

Table 2-6. Verification of scaffold map location for nine sequence scaffolds using SSR markers and the framework SSR map.

JGI v.1.1 sequence scaffold	Mapped SFP/GEM genes	SFP/GEM location	Anchored SSR flanking scaffold in GEM/SFP map	Verification SSR ID	Verification SSR location in framework map	Anchored SSR flanking scaffold in framework map
Scaffold_29	eugene3.00290072	LG_I, 85.6cM	G833, G2837	UFLA_29	LG_I, 119.7cM	G833, G3784
	estExt_fgenesh4_pg.C_290162	LG_I, 86.3cM				
Scaffold_130	gw1.130.59.1	LG_IV, 90.3cM	G1809, O545	UFLA_130	LG_IV, 109.2cM	G1809, O545
	eugene3.01300051	LG_IV, 96.8cM				
Scaffold_166	eugene3.01660055	LG_IV, 0.0cM	O349	UFLA_166	LG_IV, 0.0cM	O349
Scaffold_181	eugene3.01810009	LG_VII, 65.0cM	G354, P2794	UFLA_181	LG_VII, 52.4cM	G354, P2794
Scaffold_118	fgenesh4_pg.C_scaffold_118000002	LG_III, 79.7cM	G1629, P2611	UFLA_118	LG_III, 97.4cM	G1629, P2611
	eugene3.01180022	LG_III, 81.8cM				
Scaffold_170	eugene3.01700010	LG_XVII, 0.0cM	G125	UFLA_170	LG_XVII, 0.0cM	G125
	eugene3.01700027	LG_XVII, 0.0cM				
	fgenesh4_pg.C_scaffold_170000022	LG_XVII, 7.2cM				
	eugene3.01700033	LG_XVII, 13.7cM				
Scaffold_147	estExt_Genewise1_v1.C_1470180	LG_VI, 189.9cM	O50	UFLA_147	LG_VI_b, 94.5cM	O50
Scaffold_121	gw1.121.10.1	LG_XVIII, 76.1cM	O534, G79	UFLA_121	LG_VI_b, 7.1cM	P2221, W12
				UFLA_121_b	LG_VI_b, 7.1cM	P2221, W12
Scaffold_97	gw1.97.119.1	LG_I, 134.2cM	G3784, G937	gw97_2	LG_I, 142.5cM	G3784, G2837

Table 2-7. Verification of scaffold map location for six sequence scaffolds originally mapped based on GEM markers. Verification was conducted using sequence-identified SSR markers and the framework SSR map.

JGI v.1.1 sequence scaffold	Mapped GEM genes	GEM location	Anchored SSR flanking scaffold in GEM/SFP map	Verification SSR ID	Verification SSR location in framework map	Anchored SSR flanking scaffold in framework map
Scaffold_250	fgenesh4_pg.C_scaffold_250000006	LG_I, 397.5cM	G3688	UFLA_250	LG_I, 342.6cM	P2786b, P2385
Scaffold_395	fgenesh4_pg.C_scaffold_395000004	LG_II, 9.0cM	G734	UFLA_395	LG_II, 0.0cM	G734
Scaffold_13508	eugene3.135080001	LG_V, 152.4cM	G1255	UFLA_13508	LG_V, 166.6cM	G1255
Scaffold_269	fgenesh4_pg.C_scaffold_269000014	LG_XIV, 18.8cM	G1866, O59	UFLA_269	LG_XIV, 1.2cM	G1866, O59
Scaffold_728	estExt_Genewise1_v1.C_7280001	LG_XV, 0.0cM	G1454	UFLA_728	LG_XV, 7.4cM	G1454, G1245
Scaffold_122	gw1.122.45.1	LG_XV,26.0cM	G1454, G1245	UFLA_122	LG_XV, 28.0cM	G1454, G1245

CHAPTER 3
THE ONTOGENY OF THE GENETIC REGULATION OF GENE EXPRESSION AND TRANSCRIPTIONAL NETWORKS IN THE WOODY PERENNIAL MODEL *Populus*

This chapter will be submitted to a general biology or plant sciences journal for publication

Introduction

In humans [87], mice [83], yeast [62, 92], and several higher plants [84, 90, 91, 93, 166], large scale eQTL mapping studies demonstrated that genetic control of gene expression is highly complex, with many genes being regulated by a combination of *cis*-acting loci of generally large effect and numerous *trans*-acting elements with smaller contributions to mRNA abundance. These studies have also detected *trans*-eQTL hotspots – regions of the genome coordinately contributing to variable expression in large numbers of unlinked transcripts. The underlying genetic basis of eQTL hotspots has been a topic of current discussion [167], leading to hypotheses that hotspots may correspond to key regulators of gene expression, biochemical pathways, or developmental transitions. For example, in yeast and mouse, *trans*-acting regulators of small suites (<100) of genes have been cloned, verified [98, 106], and shown to control expression of genes in common signaling, growth, and metabolic pathways. eQTL mapping, in conjunction with traditional trait QTL analysis, has also identified polymorphisms responsible for phenotypic variation in several species [83, 112, 168]. Such results reinforce the role of transcriptional regulation in evolution [36] and suggest that pleiotropic *trans*-acting eQTL hotspots may be critical for intra- and inter-specific diversity in both gene expression and whole plant phenotypes.

Genome-wide gene expression and eQTL mapping studies have also been leveraged to reconstruct transcriptional networks contributing to biochemical and developmental pathways [78]. Network analysis can identify regulators of pathway flux and implicate previously uncharacterized members. Analysis of *a priori* defined pathways has demonstrated extensive

genetic control of underlying gene expression networks in *Arabidopsis* [169], a finding intricately explored for the glucosinolate biosynthesis [97, 101] and flowering time [93]. Transcriptional networks can also be identified *a posteriori* from eQTL data, and their biological roles defined by identifying overrepresented metabolic and regulatory functions among network members [106]. *A posteriori* approaches have proven particularly useful to demonstrate metabolic and molecular signatures associated with complex phenotypes and disease in model species [83, 110, 112], and we have previously utilized them to define a role for the lignin biosynthesis pathway in tree growth [84, 100]. Additional genomic information, including metabolomic, transcription factor binding site (TFBS), and protein-protein interaction data have been incorporated into a few transcriptional network studies [97, 114, 115], increasing the power to identify key network members [115].

Despite the insights into the role of transcriptional networks in shaping plant diversity and evolution, eQTL and transcriptional network studies have generally focused either on the genetic control of gene expression variation in the entire plant [85, 91] or in only single tissues/organs [84, 90, 170]. Thus, the ontogeny and diversity of the genetic control of transcription and gene expression networks between tissues in plants (and with few exceptions also in animals [109, 171, 172]), is largely unknown. The perennial woody plant *Populus* is an ideal model to compare the genetic architecture of gene expression between tissue types because organs are highly differentiated (for instance, woody stems vs. leaves) and the species has a rapidly expanding genomic toolbox founded upon the genome sequence of *Populus trichocarpa* [4]. In this study, we employ a pseudo-backcross pedigree of *P. trichocarpa* and *Populus deltoides* to analyze genome-wide gene expression variation among three distinct tissues (differentiating xylem, expanding leaves, and mature roots). Genetic regulation of mRNA abundance is highly

differentiated among the tissues, and variation in tissue-specific expression programs is common. We utilize tissue-specific eQTL to generate co-transcriptional networks *a posteriori* on the basis of eQTL hotspots, and predict network regulators on the basis of *cis*-acting expression regulation. We demonstrate these networks to be significantly enriched for biologically coherent groups of genes and well-characterized promoter *cis*-elements that may play key roles in tissue-specific developmental programs and phenotypic diversity between *P. trichocarpa* and *P. deltoides*.

Materials and Methods

Plant Material and Growth Conditions

A pseudo-backcross progeny (Family 52-124) of 396 individuals from a cross of *P. trichocarpa* X *P. deltoides* (genotype 52-225) and *P. deltoides* (genotype D-124) were propagated and grown as described [75]. From a common set of 192 randomly selected individuals grown under high nitrogen conditions (25mM N as NH_4NO_3; [75]), we collected 180 samples of differentiating xylem, 183 expanding leaves, and 163 whole-roots for gene expression analysis. Collected tissues were immediately flash-frozen in liquid nitrogen and stored at -80^{o}C until lyophilization and RNA extraction. We favored using of a single biological replicate of progeny to maximize the size of the population and meiotic events sampled [128]. Because the experiment reflects the analysis of a segregating population, each allele is biologically replicated in approximately half of the individuals of the population [78].

RNA Isolation and Microarray Analysis

RNA was extracted from each lyophilized sample by a standard protocol [148], converted to double-stranded cDNA, labeled with cy3, and hybridized to microarrays [173]. Hybridizations were carried out using a previously described four-plex NimbleGen (Madison, WI) microarray platform (Gene Expression Omnibus Accession# GPL7234) using probes designed to minimize

the effects of sequence polymorphism on the estimates of gene expression [173]. The microarray comprised one probe per gene for 55,793 previously described gene models derived from the annotation of the genome sequence of *P. trichocarpa* clone 'Nisqually-1' (version 1.1; [4]), and a set of non-annotated ESTs. Raw data from all hybridizations was background subtracted, $\log_2$ transformed, and quantile normalized as described [173].

eQTL Analysis

Each quantile-normalized gene expression value was analyzed as a quantitative trait using composite interval mapping analysis [66, 67] implemented in QTL Cartographer ([174]; walk speed = 2cM), based on our genetic map of Family 52-124 that is physically linked and oriented to the genome sequence of 'Nisqually-1' [75]. Significance of Log of Odds (LOD) values was estimated for each tissue using a global permutation threshold [91]. Global permutation thresholds for each tissue are reported in the footnotes of Table 3-1. eQTL were declared on the basis of a strategy wherein eQTL composed of unimodal LOD curves are located by the peak position [90]. Polymodal peaks were declared as separate eQTL if the trough between them exceeded 2 LOD.

The eQTL were classified as *cis*- or *trans*-regulated based on co-localization of the eQTL LOD peak for the gene model with the genetic map marker bin containing the gene model in the 'Nisqually-1' sequence. While the Family 52-124 map encompasses >85% of the assembled genome sequence [75], 23,116 of the 55,793 gene models/ESTs assessed by our microarray are located on unassembled genomic scaffolds (17,726) or assembled chromosome telomeres outside of the coverage of the genetic map (5,390). eQTL for these probes were designated as "ambiguous" for the purposes of declaring *cis*- vs. *trans*- regulation.

eQTL Hotspot Detection and Analysis

To identify significant eQTL hotspots, we permuted the per-bin total eQTL peak counts for each tissue across the 1840 ~2cM bins of the genetic map 1000 times, and determined the 95th percentile of these permutations. Each 2cM bin with a total absolute eQTL peak count greater than this permutation threshold was declared an eQTL hotspot. Bins with eQTL counts surpassing the permutation thresholds calculated independently for each respective tissue were compared and classified as "unique" or "shared" eQTL hotspots between the three tissues. To eliminate differential gene density as an explanatory factor for eQTL hotspots (i.e. more genes per genetic distance) we implemented a Chi-squared testing strategy [90]. When applied, 238/255 bins continued to be significantly enriched for eQTL (or 93.3% of bins identified based on permutation threshold approach).

Hotspot-Based Co-Expression Network Construction

We constructed co-expression networks conditioned on the bins declared as eQTL hotspots in the previous analysis. For each ~2cM map bin identified as an eQTL hotspot we selected all genes whose eQTL LOD values surpassed the tissue specific permutation thresholds for eQTL significance (Table 1 footnotes). From the $\log_2$ transformed, normalized expression values for these genes in the respective tissue of interest, we computed pairwise Pearson correlations of the expression values. Networks were declared when no fewer than 10 genes in a given hotspot bin demonstrated a Pearson correlation of $|r| > .80$ (uncorrected $P < 5.5 \times 10^{-3}$ for $n = 10$). Network edges were constructed and tallied between network members displaying correlations surpassing this threshold.

GO Annotation and Enrichment Testing

For each co-expression network constructed, we annotated member genes for Gene Ontology (GO) [175] categories by conducting a BLASTx search of poplar gene model

transcripts against The *Arabidopsis* Information Resource (TAIR) proteins Release v.8.0. A significant BLAST match was declared at an *E*-value threshold of $< 1x10^{-5}$, with transcripts returning $E > 1x10^{-5}$ annotated as "no hits". The GO annotation of the closest *Arabidopsis* ortholog was assigned to the respective poplar gene. We identified putative orthologs for 45,648 genes on the microarray, of which 36,688 included at least one GO designation in TAIR's database.

Overrepresentation of GO categories was tested within each network by producing 2x2 contingency tables for each GO category represented within the network, followed by a right-tailed Fishers Exact Test (calculating the probability of observing an equal or higher frequency of the category in the network, relative to the genome frequency of the GO category). Because each network was tested for a distinct number of GO enrichments, a Bonferroni correction for multiple testing was applied separately for each network tested, which was computed using the formula $P_{corr} = .05/n$ for networks comprising n distinct GO categories. To further control the frequency of Type I error, we did not consider GO categories "enriched" when only one gene in a network was assigned to that category, even if the enrichment was determined to be significant with respect to P_{corr}.

Cis-Element Detection and Enrichment Testing

To annotate the presence and absence of common plant *cis*-acting elements in the promoters of the gene models from the *P. trichocarpa* genome, we extracted the promoter sequences upstream of the start codon for the 55,793 genes represented in the microarray. Uninterrupted sequence of 1,500 bp of length could be isolated for 49,066 genes, as the position of the remaining 6,727 was less than 1,500bp from an unresolved sequence region, whole genome shotgun scaffold, or contig end. To avoid bias associated with *cis*-motifs that may be located at preferential distances from the start codon, we did not consider these 6,727 genes in

our statistical analysis. We downloaded the Plant *Cis*-Acting Regulatory Element (PLACE) sequence database [176] and determined the presence and absence of motifs within all 49,066 gene promoters by using Patmatch [177]. Among 469 *cis*-elements deposited in the PLACE database, we detected 360 in at least one gene promoter region included in our analysis.

For each of the 360 motifs, we tested each co-expression network for enrichment of genes bearing the motif in question using a right-tailed Fishers Exact test. Multiple testing was corrected using a Bonferroni threshold of $P < .05/360 = 1.389 \times 10^{-4}$ to judge significance of resulting enrichments.

Results

eQTL Detection and Genome Distribution

To analyze the role of interspecific variation on the genetic architecture of gene expression in *Populus*, mRNA levels in 180 xylem, 183 leaf, and 163 root tissue samples isolated from 192 progeny of Family 52-124 [75] were assayed using customized microarrays [173] comprised of one 60-mer probe that represents each of the 45,555 gene models derived from the genome sequence of *P. trichocarpa* 'Nisqually-1' [4] and 10,238 additional EST sequences. QTL analysis of normalized signal intensities identified 36,071 significant eQTL in xylem, 13,403 eQTL in leaf, and 9,137 eQTL in roots representing 30,313, 12,392, and 8,534 genes/ESTs, respectively (Table 3-1). eQTL were classified as *cis*- or *trans*-acting, contingent upon the overlap of the eQTL peak with the marker interval to which the gene model represented by that eQTL was located in the genome. We could not determine *cis*- or *trans*- regulation for genes located in scaffolds that have not been mapped to chromosomes in the physical assembly, and therefore classified them as "ambiguously" regulated (Table 3-1). Within each of the three tissues, c*is*-acting eQTL were detected at a relatively constant rate (~8-10% of genes) independent of the linkage group or tissue, when normalized to account for the varying number of genes per

chromosome (Figure 3-1). In contrast, *trans*-acting eQTL frequency varied widely between different linkage groups and tissues. For instance, in xylem the number of eQTL ranged from a low of 76 *trans*-eQTL on LG XII to a high of 20,935 on LG IX, suggesting that some chromosomes contain major loci implicated in regulation of mRNA level that segregate in this poplar hybrid population. When eQTL corresponding to ambiguously positioned genes were included, linkage group dependent differences in eQTL frequency increased even more so (Figure 3-2).

Genetic Regulation of Gene Expression Is Largely Tissue Specific

Given the large degree of cellular differentiation among vegetative tissues in poplar, yet the seemingly small degree of differences in expressed genes separating them [4, 35], varying mechanisms of transcriptional regulation are likely to be an important component of tissue-specific morphological and developmental differences [79]. To address this ascertation, we determined the degree of overlap between genes with eQTL in multiple tissues and the location of their eQTL peaks (Figure 3-3). In general, sharing of eQTL across tissues was infrequent – despite the fact that 4,631 unambiguoulsly placed gene models produced eQTL in both leaf and xylem (Figure 3-3A) only 1,389 of these were regulated by the same genomic interval in the two tissues (Figure 3-3B). The degree of eQTL sharing between xylem and roots was higher, as expected by the fact that both share similar types of tissue (e.g. secondary xylem). Among 2,105 unambiguously localized genes producing eQTL in both xylem and roots, 1,317 were regulated by the same genomic region. *Cis*-acting eQTL (Figure 3-3C) were shared between tissues at a substantially higher rate than *trans*-acting eQTL (Figure 3-3D) in all pairwise comparisions, with the exception of xylem and root where sharing was approximately equivalent for both eQTL types. This outcome is consistent with the hypothesis that local, *cis*-acting variants have a more significant effect on gene expression control throughout development than second-order, *trans*-

acting effects [91]. Contrastingly, our data suggest that *trans*-acting regulatory mechanisms might more commonly govern tissue specific gene expression and subsequently, developmental differentiation.

Identification of Tissue-Specific eQTL Hotspots

Genomic regions regulating accumulation or turnover of large numbers of transcripts, or eQTL hotspots, were identified based on permutation thresholds derived for each tissue. Map bins containing≥ 39 eQTL peaks in xylem,≥ 21 eQTL peaks in leaf, and≥ 16 peaks eQTL in root were identified as significant hotspots. Using these critieria, we detected 67 unique bins corresponding to statistically significant eQTL hotspots in xylem, 97 in leaf, and 88 in root (Table 3-2). While a large number of unique bins were enriched for eQTL relative to chance, many of these bins were adjacent to one another (Figure 3-4) and thus likely correspond to a single hotspot – a result of the limited mapping resolution of QTL in small populations [128]. The eQTL hotspots we detected resulted primarily from the hyper-accumulation of *trans*-acting eQTL, and were generally not enriched for *cis*-acting eQTL (Table 3-3). Furthermore, to test that hotspots did not correspond to regions of high gene density we used a strategy proposed to normalize for number of genes per map bin [90], and found that >90% of the original bins corresponding to eQTL hotspots remained significantly enriched for eQTL (data not shown).

Construction of Tissue-Specific, Hotspot-Based Co-Expression Networks

Some eQTL hotspots have been shown to correspond to co-transcribed gene sets that are enriched for common functional groups or known biochemical and regulatory pathways [98, 106]. Therefore, these hotspots can serve as a foundation to build transcriptional networks that are connected by common genetic regulation. To create networks, the mRNA abundance data for genes in each eQTL hotspot were used to generate co-expression matrices based on correlations among gene expression measurements across the population. Among the 97 leaf eQTL hotspots

detected, we constructed 51 gene co-expression networks within 38 of the original genetic map bins (Table 3-4). The leaf co-expression networks encompassed 1678 distinct genes and ranged in size from 11 to 945 genes (median = 36 genes). While 38 bins were found to contain expression networks, many of these bins neighbored one another in the genetic map, and resulting networks were highly redundant [128]. Nonetheless, at least nine independent leaf co-expression networks were detected within seven bona fide unique loci (Figure 3-4, Table 3-5). Similar results were obtained for xylem and root tissues (Table 3-4).

Gene Ontology-Based Annotation of Tissue-Specific Gene Co-Expression Networks

eQTL hotspots in the segregating population revealed a number of co-expression networks, but their biological relevance remained largely unknown. To elucidate how global gene expression variation might translate into biologically meaningful differences in tissues and genotypes, we tested each co-expression network for enrichment of GO categories represented within each network. Poplar genes were categorized based on the GO designation of the most similar *Arabidopsis* homolog (BLASTx *E*-value cutoff < $1x10^{-5}$ for significant hits to TAIR release v8.0) resulting in the GO annotation of 36,688 of the 55,793 gene models represented in the microarray. For the 51 co-expression networks identified in leaf tissue, we identified 42 with at least one significant GO category enrichment (median $P_{enrichment} = 7.35x10^{-7}$). Analogous results were obtained for xylem (75/98 networks, $P_{enrichment(median)} = 1.24x10^{-5}$) and root (63/75 networks, $P_{enrichment(median)} = 9.58x10^{-6}$). Among the most significant enrichments, the "blue" hotspot locus on linkage group (LG) VI (Figure 3-4, Bin 734) revealed a network with an overrepresentation of genes associated with chloroplast biogenesis and function ($P_{enrichment} = 1.37x10^{-40}$). Of the 49 genes present within the co-expressed network, 35 (68.6%) were GO annotated as being localized to the chloroplast (Figure 3-5, Table 3-6), a >5.5x enrichment over the number of chloroplast-localized genes expected by chance in a network of this size

(4504/36,688 ≈ 12.27% or 6.25 chloroplast localized genes). Chloroplast stroma ($P_{enrichment}$ = $3.16x10^{-22}$), thylakoid lumen ($P_{enrichment}$ = $1.97x10^{-6}$), and chloroplast envelope ($P_{enrichment}$ = $5.68x10^{-10}$) cellular components were also significantly enriched in this network (Table S4), reinforcing the notion that this LG VI locus plays an important role in regulating genes related to chloroplast biogenesis and/or function. Using this strategy on all 180 networks with at least one GO categorical enrichment, we identified 183 distinct GO category enrichments, representing 1212 combinations of significant tissue-specific networks and enriched GO categories.

eQTL-Based Prediction of Putative Network Regulators

Developing transcriptional networks represents an initial step towards understanding the relationships between genes in a biological system. However, if one wishes to modify regulatory relationships influencing phenotypes, key network regulators must be identified. Co-expression networks developed on the basis of eQTL hotspots facilitate regulator identification, because differential transcript accumulation is predicted to occur due to a genetic variant underlying the eQTL hotspot position. Therefore, *cis*-regulated genes belonging to a network defined by an eQTL hotspot represent strong *a priori* candidate regulators. While this strategy is incomplete in that it will not identify network regulators differentially controlled outside the realm of transcript abundance, it has previously offered direct evidence of putative network regulators for downstream investigation [115, 171]. We identified putative network regulators for 43 of the 62 leaf co-expression networks, 38 of the 75 root networks, and 50 of the 98 xylem networks. Frequently, more than one putative regulator was identified for each network. For the LG VI leaf co-expression network previously shown to be associated with chloroplast function, we identified six network members with gene models and eQTL in *cis*- to the eQTL hotspot. Among these six genes, five code for chloroplast-localized protein products, three of which are known structural components of the chloroplast. These three genes represent the best candidates for

regulators of this co-transcriptional network (Figure 3-5, Table 3-6). Of particular interest is *FtsZ2*, a gene with a well described role in chloroplast structure, biogenesis, and division [178, 179].

In some cases, however, no candidate regulator was associated with GO categories enriched within the network in question. It is possible that a variant not regulated at the level of transcription (i.e., structural, post-transcriptional, or DNA sequence-level) is responsible for the eQTL hotspot and co-expression network. Our method of analysis would not identify the regulator in this case. Alternatively, candidate regulators might not be included in a GO category due to lack of BLAST similarity with the *Arabidopsis* gene set used for GO annotation of the poplar genes. A number of the co-expression networks include candidate regulatory genes with no hits, indicating this possibility may be quite common. Finally, incomplete GO annotation of genes may mean that candidate regulators identified by our method, while correct, have yet to be assigned a GO annotation with a network-associated process because of lack of evidence.

Enrichment of Transcription Factor Binding Sites in Co-Expression Networks

In addition to the information obtained from GO annotation enrichment and putative network regulator analysis detailed above, the functional roles of transcriptional networks may be inferred from conserved TFBS in co-regulated genes [115]. We utilized the publicly-available Plant *Cis*-acting Regulatory Element (PLACE; [176]) database to define presence or absence of 360 *cis*-element motifs in the 1500bp upstream of the start codon in 49,066 gene models represented on the microarray. Subsequently, each co-expression network was tested for significant enrichment of genes bearing each of the PLACE motifs (Fisher's exact test *P*-value < 1.389×10^{-4}, after Bonferroni correction for 360 tests per network). Enrichment was detected for 27 motifs in 35 of the 62 leaf gene co-expression networks, 32 motifs in 21 of the 75 root networks, and 36 motifs in 29 of the 98 xylem networks. Networks were significantly enriched

for as few as one to as many as 26 *cis*-elements (median of three motifs enriched per network). These conserved motifs encompass highly variable functions, including light-induced transcriptional modulation, calcium-responsive expression, sucrose-responsive elements, and hormone-responsive elements. Additionally, elements known to bind specific families or classes of transcription factor (Myb-domain, bZIP, AP2) were enriched among several networks. A total of 419 combinations of 85 tissue-specific networks and 72 enriched motifs were detected amongst the dataset.

To determine whether additional biological insight could be gained based on motif enrichment, we considered a network detected in xylem on LG IX (map bin 1064), comprised of 110 genes. Among the 108 network member genes for which upstream sequence could be characterized, we detected significant enrichment for 18 motifs, including motifs common to cytokinin-enhanced binding sites ($P = 4.84 \times 10^{-5}$, 84/108 genes), a binding site for *ATHB-2* (an Arabidopsis homeobox ZIP transcription factor; $P = 1.18 \times 10^{-4}$, 35/108 genes), and two W-box motifs previously associated with *ERF3* ($P = 5.73 \times 10^{-5}$, 105/108 genes) and *SUSIBA2* ($P = 1.30 \times 10^{-5}$, 101/108 genes) transcription factors. *SUSIBA2* is a WRKY family transcription factor originally described in barley [180] that is known to regulate amylases and other genes in the starch catabolism pathway whose transcription are responsive to sugars. Among the 101 genes carrying the *SUSIBA2* W-box motif, several are associated with sugar metabolism, including two amylase homologs, a hexose transporter, and a pyruvate kinase homolog. Further supporting the hypothesis that this network functions in sugar metabolism, the *Arabidopsis* homologs of 25 of the genes exhibiting the *SUSIBA2* binding motif were previously found to be differentially expressed in response to exogenous sucrose ([181] Table 3-7). While sucrose-associated GO categories were not significantly overrepresented in this network, chloroplast associated

functions ($P = 5.30x10^{-12}$) and amylase activities ($P = 5.38x10^{-5}$) were enriched. Therefore, although the biological interpretation of this and other networks remains complex, additional levels of data beyond the information on gene expression can add to the analysis of co-regulated eQTL hotspot networks by identifying common links in otherwise seemingly disconnected data. These links will be vital for understanding network functions in the poplar genome, which is still undergoing refined functional annotation.

Similarly, in the network affecting co-expression of the chloroplast structural genes (LG VI), four TFBS were enriched, including a pyrimidine box implicated in sugar-induced transcriptional repression present in 47 genes ($P = 1.012x10^{-4}$) and a light-responsive transcription element represented in each gene in the network ($P = 1.178x10^{-4}$). These results are perhaps not surprising given the network's close association with chloroplast GO categories and the established relationship of chloroplast function and activity to feedback regulate chloroplast-associated gene expression through the metabolic products of photosynthesis [182].

Transcriptional Networks Shared Between Tissues Are Regulated By Distinct Loci

The ontogeny and conservation of transcriptional networks among different plant tissues is largely unknown. Interestingly, by contrasting the localization of eQTL hotspots among vegetative tissues and we observed that hotspot overlap was minimal – only 9 hotspot bins were conserved between leaf and xylem, 11 between xylem and root, and 9 between root and leaf (Table 3-3). Only two hotspot bins were shared between all three tissues, suggesting that large transcriptional networks are generally tissue specific. Furthermore, very few genes exhibited eQTLs controlled by the same hotspot in different tissues (Table 3-3). Therefore, eQTL hotspots could correspond to important genetic regulators governing major differences in structure or development between tissues in poplar.

Our analysis of eQTL and eQTL hotspot sharing between tissues suggested that individual genes and networks are commonly regulated by tissue-specific loci. However, even if networks are regulated by distinct loci in separate tissues, they could still share network members. Significant gene sharing between tissue specific networks was tested with a Chi-squared analysis, accounting for the number of probes in each network and shared between networks. Several instances of shared network membership were observed after correction for multiple testing at a Bonferroni threshold of .01/*n*, for *n* cross-tissue network-to-network comparisons where at least 5 genes were shared between networks. For instance, a large leaf network of 428 genes on LG XI (interval 1226, network 1) shared 171 gene members with networks in xylem. Among the set of shared genes, 91 were located in a LG I network of 928 genes in xylem (interval 252, network 1; $P \approx 0$). An additional 53 genes in the leaf network on LG XI were located in a distinct xylem co-expression network on LG XV (interval 1557, network 1, 200 genes; $P \approx 0$) and 25 in a LG XIV xylem network (interval 1386, network 1, 1057 genes; $P < 1.832 \times 10^{-9}$). Thus, a total of 156 of the 171 shared network genes from the leaf LG XI network were restricted to just three networks in xylem, including 13 leaf network genes appearing in both the LG XV and LG I xylem networks. In total, we found 414 significant network pairs that were statistically enriched for shared genes among leaf and xylem. Similar patterns of cross-tissue network member conservation were obtained for the comparisons of leaf with root and xylem with root. Interestingly, however, these comparisons identified only 31 networks in leaves significantly overlapping for membership with 33 networks from xylem. Furthermore, the 31 networks from leaf were restricted to four genomic regions – residing on linkage groups VI, IX, XIV, and XVI. Similarly, the 32 xylem networks were derived from only six unique genomic regions, containing networks ranging from 30 to 5787 genes. These outcomes are mirrored in the

other pairwise comparisons between tissues – our analysis identified significantly overlapping network memberships between 31 leaf, 64 root, and 43 xylem networks. These results suggest that a handful of distinct *trans*-acting factors might control expression of coordinate groups of genes depending on the tissue being considered, and also indicate that different combinations of biological subnetworks could be combined to potentially drive tissue diversification.

Discussion

Unraveling the orchestrated action of genes and modeling their interactions in a biological system is among both the most significant challenges and ultimate goals of biology. Well-described gene networks should help predict growth and development, as well as the outcome of perturbations from biotic and abiotic stresses [183]. Here, a traditional quantitative genetic analysis was used to assemble networks that describe gene expression patterns, and infer biological function, as well as mechanisms of regulation by identifying putative regulators and overrepresented *cis*- elements. Networks were identified within three developmentally diverse plant tissues by analyzing an interspecific pseudo-backcross progeny of the model hardwood forest tree genus *Populus*. Our results indicate that genetic variation of gene expression accounted for by allelic differences between *P. trichocarpa* and *P. deltoides* are complex and abundant with a large number of transcripts regulated by combinations of local, *cis*-acting variants of generally large effects, and distant *trans*-acting variants of generally less significant effects.

Interestingly, among leaf, root, and xylem the prevalence of eQTL was markedly distinct, with approximately three-fold more genes exhibiting genetic control over gene expression in xylem than either root or leaf. This is consistent with our previous observation that stem tissues have the greatest diversity of expressed genes among poplar vegetative tissues [35]. Genes that are tissue- or organ-specific may also evolve more rapidly, making them more likely to undergo

diversification in expression [36, 184]. However, such factors alone cannot fully explain the vast increase in eQTL prevalence within xylem, since previous analysis showed <3500 genes expressed exclusively in stem tissue at a common FDR threshold [35]. The difference between eQTL frequencies among tissues was almost completely accounted for by a prevalence of *trans*-acting eQTL in xylem compared to leaf and root (Table 3-1). However, this striking difference noted in xylem is primarily due to eQTL comprising a few major hotspots, particularly one on LG IX encompassing >14,000 genes, or around 25% of the coding genome. We interpret the results associated with this particular hotspot cautiously, as other major pleiotropic loci detected within this study resulted in multiple subnetworks which were enriched for numerous, distinct GO categories. However, the major xylem locus on LG IX produced a single connected network comprising nearly 5,800 genes (Table 3-4), and with relatively few enriched GO categories given its size. Nonetheless, hotspots of similar orders of magnitude (comprising up to 10-15% of the coding genome) are detected on chromosome 2 of *Arabidopsis*, in crosses between accessions Bayreuth-0 and Shahdara [91] and Landsberg *erecta* x Cape Verde Islands [114]. One of these hotspots has been attributed to the *ERECTA* locus, a known pleiotropic regulator of morphology and development that varies between different accessions [114]. Our pedigree represents an interspecific cross from a dioecious and obligate outcrossing organism, encompassing levels of sequence variation and genome complexity (i.e., chromosome structural variation, [125]) that may not be adequately modeled by intraspecific pedigrees produced in self pollinating model plants such as *Arabidopsis*. Therefore, we believe our overall results support the view that most eQTL hotspots detected in *Populus* are biologically relevant. It is likely that fundamental differences in gene expression regulatory cascades, especially those mediated through *trans*-acting factors, play key roles in development and morphological differences in *Populus*. Still, the

specific mechanisms by which these effects occur clearly requires a more detailed case-by-case investigation.

A second and related outcome of our eQTL analysis was the surprising frequency with which tissue-specificity was associated with gene expression regulation. While thousands of genes exhibited eQTL in multiple tissues, the conservation associated with the location of the genomic region leading to differential regulation was limited, ranging from ~30.3-60.3% depending on the tissues considered (Figure 3-3). Conserved regulation between tissues was strongly biased towards *cis*-acting eQTL, whereas *trans*-eQTL were significantly more likely to be tissue-specific. Interestingly, this phenomenon extended not only to the analysis of eQTL hotspots, but also to gene expression networks. We found that a small number of tissue-specific hotspots frequently regulated expression subnetworks that, in other tissues, were regulated by distinct and unlinked loci. These results build upon those generated for barley, wherein limited pleiotropy associated with tissue-specific *cis*-acting eQTL was investigated [185] and found to be relatively common. Here, we have demonstrated that the tissue-specific modulation of single genes, as well as interconnected networks and biologically coherent co-transcriptional modules, may represent a key component of differentiation between tissues and individuals in higher plants. This finding greatly increases the potential for regulatory complexity to play a key role in diversification of species and tissues. A seemingly small number of genes, each regulated in a tightly controlled, tissue-specific manner by complex assortments of multiple *cis*-acting elements and *trans*-acting regulatory factors exponentially increase the number of combinations that can act in concert to generate phenotypic diversity. The regulatory landscape becomes even more complex when we consider that each *trans*-acting factor is likely subjected to similar patterns and modes of regulation. These results and hypotheses are in sound agreement with previous

findings, which demonstrated the overall presence or absence of expression of specific genes is fundamentally similar between tissues in poplar and *Arabidopsis* [35, 186], despite the fact that molecular signatures associated with transcript abundance estimates clearly distinguish organ and tissue types [186]. Our approach in this study begins to reconcile regulatory mechanisms by which these results may be explained. As information for the *Populus* model genetic system continue to grow, additional levels of data including miRNA activity, protein-protein interaction, and epigenetic regulation will continue to improve our understanding of regulatory mechanisms driving tissue specialization in higher plants. Nonetheless, our current effort provides a foundation upon which these new levels of data can be laid.

A key goal of systems biology is the ability to generate testable predictions of system-wide behavior in response to specific perturbation(s). Generating networks of co-expressed genes represents an initial step toward this goal in *Populus*, as co-expressed gene sets are expected to be responsive to perturbations among highly connected network members and putative network regulators. Our co-assembly of gene expression QTL into hotspots has revealed no less than 50 tissue-specific co-expression networks and associated candidate regulators. Previous studies in the simpler yeast genetic system showed that co-expression networks developed from gene expression and eQTL data are more predictive than those produced based on expression data alone [115]. For example, eQTL-based networks in this yeast study more accurately predicted associations between genes regulated by common transcription factors, and showed increased ability to predict expression signatures associated with genetic and pharmacological perturbation. Accordingly, we anticipate that the networks produced in our study are similarly predictive of system behavior in response to specific genetic “disruptions”. A number of testable hypotheses can be generated from these data, including the response of specific co-expression networks to

perturbation of their putative regulator(s), the role of previously described transcription factors in coordinating networks exhibiting enrichment for specific *cis*-element motifs, and the contribution of co-expressed networks to variation in phenotypic traits associated with network enriched GO categories. We have explored only a very limited sample of these hypotheses within the context of this article, by considering the role of statistically enriched *cis*-elements in a xylem network that may be regulated by sugar signaling, and identifying a network heavily enriched for genes that are likely to be associated with variation in chloroplast function, morphology, and/or biogenesis. While the experiments to directly test and verify individual hypotheses are beyond the scope of the current effort, the results enhance our understanding of the interconnectedness of the genome, transcriptome, and regulatory elements in the poplar genetic system. Furthermore, experiments to validate the role of specific candidate regulators and networks on molecular and morphological phenotypes will be the focus of subsequent efforts from our group.

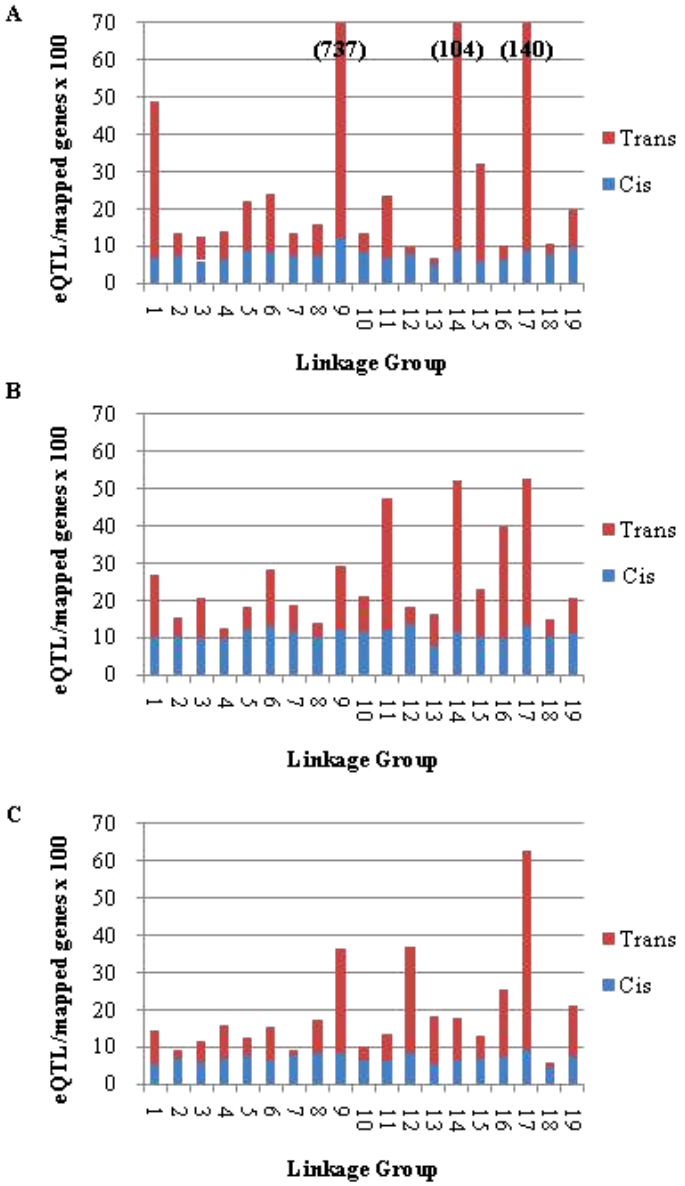


Figure 3-1. Global distribution of eQTL across linkage groups for A) xylem, B) leaf, and C) root tissues expressed as the fraction of mapped gene models. In cases where values on the eQTL/mapped genes axis exceed 70, values are provided in brackets. See Table 1-1, footnotes [b] and [c] for description of *cis*- and *trans*-eQTL categorization procedure.

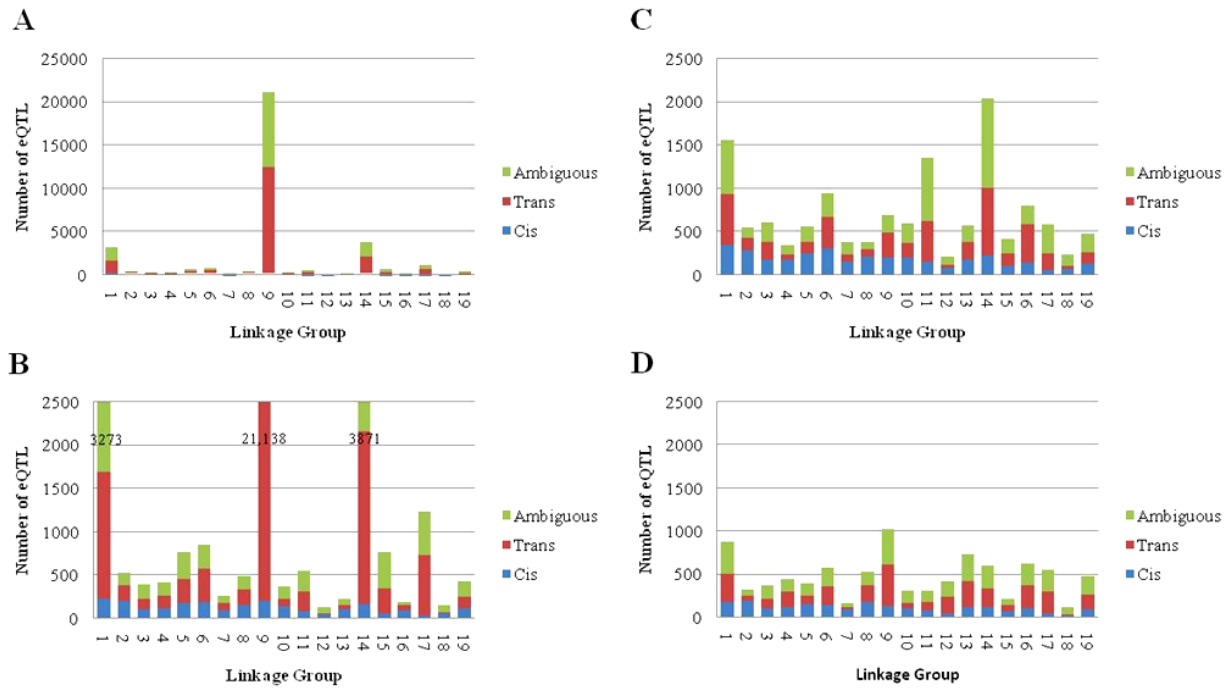


Figure 3-2. Global distribution of all eQTL (ambiguous and unambiguous) across linkage groups for A) xylem, B) xylem rescaled for comparison with C) leaf, and D) root tissues. See Table 3-1 footnotes for a description of *cis*-, *trans*-, and ambiguous eQTL categorization procedure.

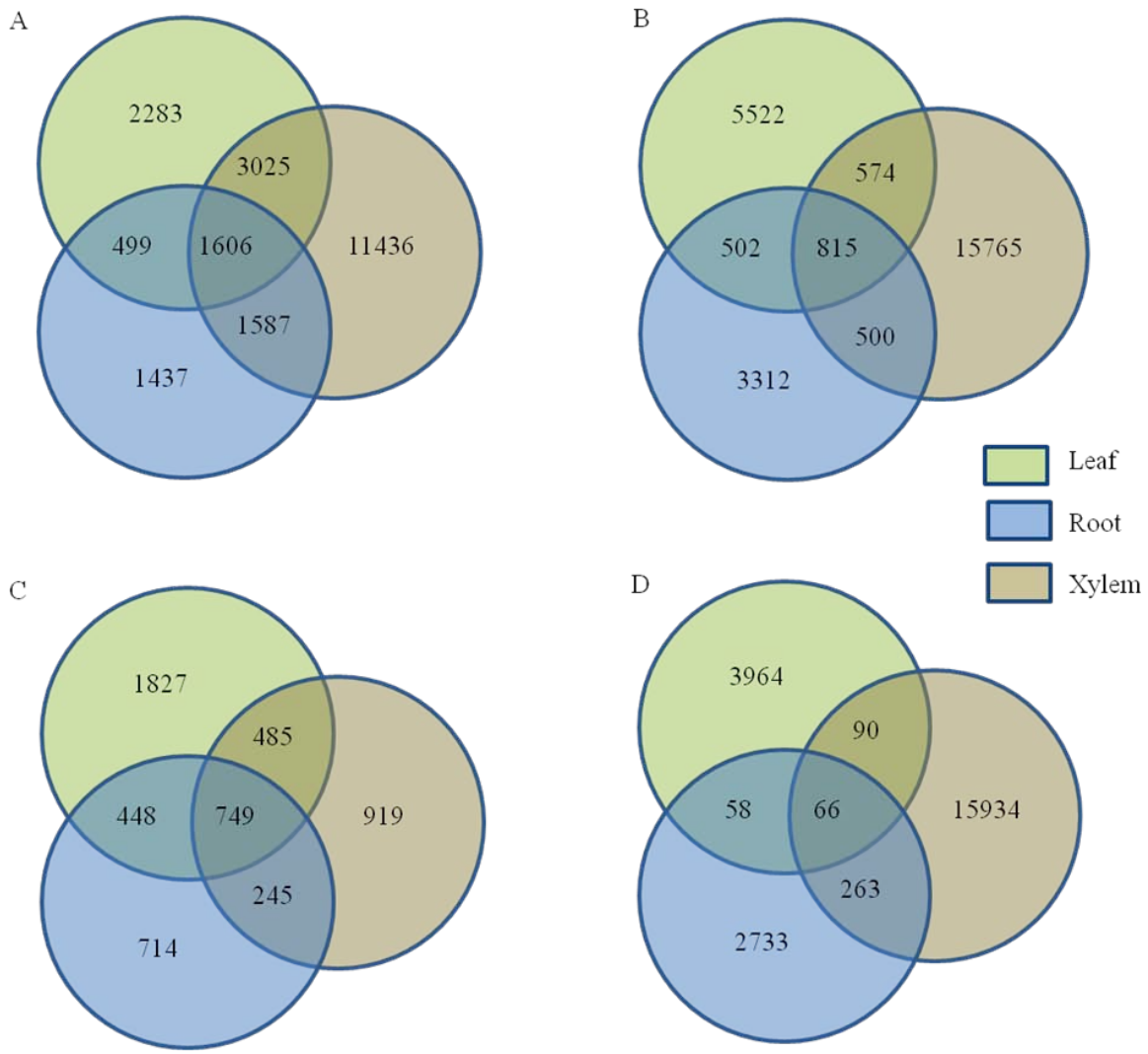


Figure 3-3. Overlap between probes and eQTL detected among the three tissues considered. All figure subparts *exclude* probes with ambiguously positioned eQTL. A) Overlap of probes (gene models) producing eQTL in each of the three tissues. B) Cross-tissue conservation of the genomic regulatory region (eQTL location) for probes in 'A'. C) Cross-tissue conservation of genomic regulatory position for *cis*-eQTL in 'B'. D) Cross-tissue conservation of genomic regulatory position for *trans*-eQTL in 'B'. Totals for sections in 'C' and 'D' do not equal that in 'B', as a small number of probes produce conserved eQTLs mapping in both *cis*- and *trans*-.

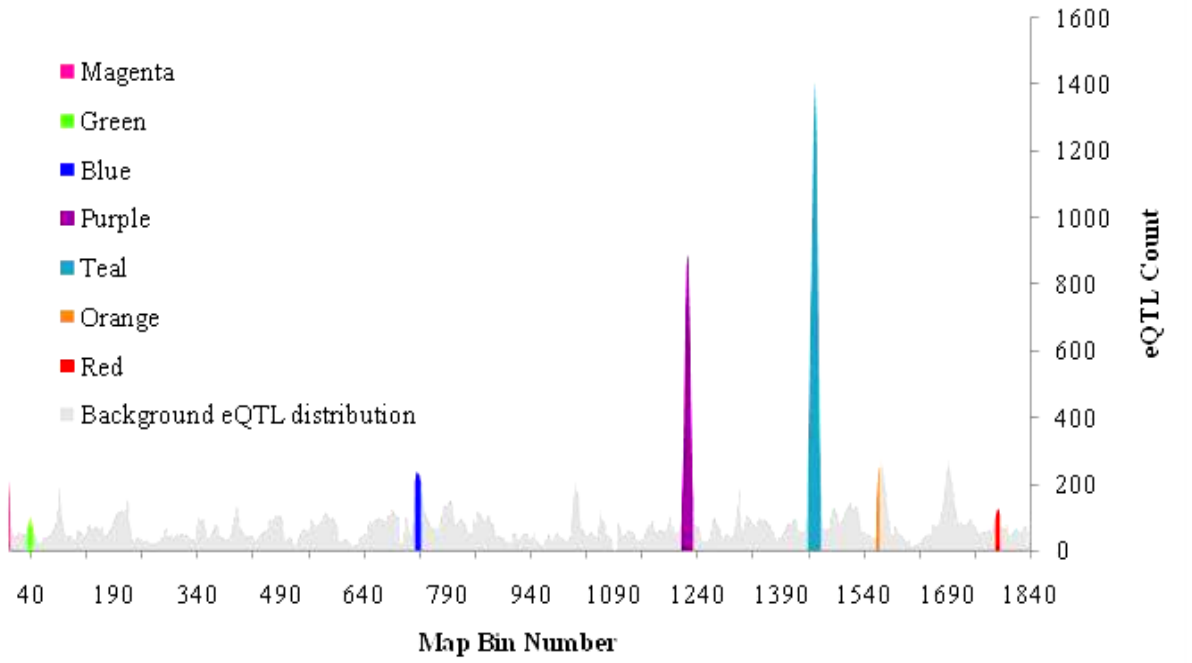


Figure 3-4. Genome-wide linkage scan of expression traits and demarcation of eQTL hotspots producing co-expressed gene networks in leaf tissue. Similar results were obtained for root and xylem tissue, detailed in Table 3-4 (linkage scans for other tissues not shown).

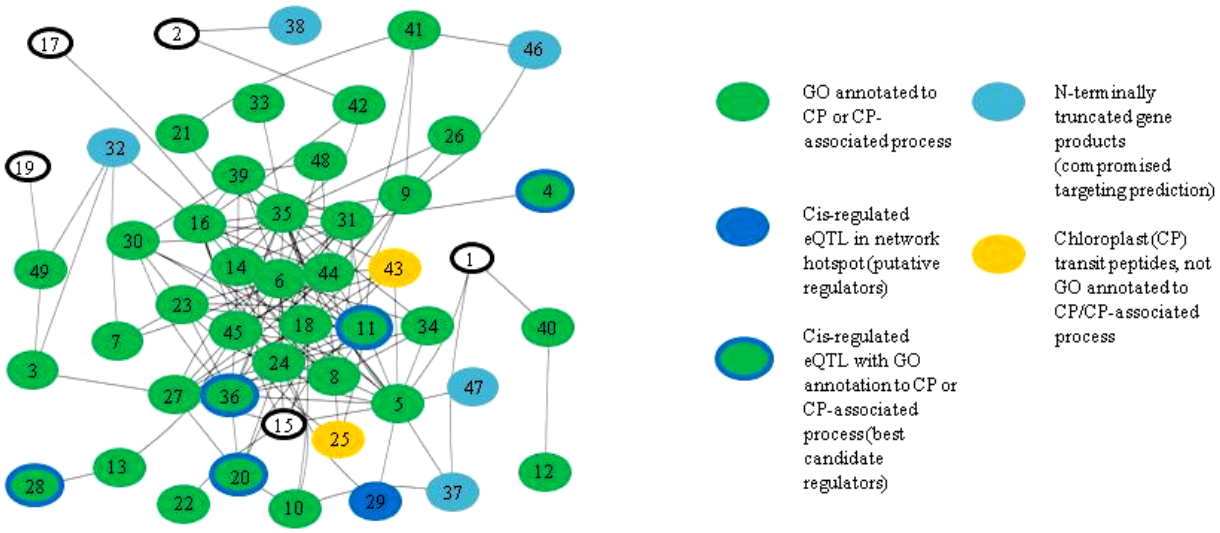


Figure 3-5. Leaf co-expression network constructed from the "blue" eQTL hotspot in Figure 3-4 and enriched for chloroplast (CP)-related Gene Ontogeny categories. Hub colors indicate annotation nature of network members, and edges indicate Pearson correlation between co-expressed genes of $|r| > .80$. Gene IDs relate to gene models presented in Table 3-6.

Table 3-1. Summary of eQTL detected for each of the three poplar tissues.

Tissue	Category	Total number[a]	Number *cis*[b]	Number *trans*[c]	Number ambiguous[d]
Xylem	Genes with eQTL	30,313	2,398	16,353	11,562
	eQTL	36,071	2,408	18,684	14,979
Leaf	Genes with eQTL	12,392	3,509	4,178	4,705
	eQTL	13,403	3,528	4,500	5,375
Root	Genes with eQTL	8,534	2,156	3,120	3,258
	eQTL	9,137	2,161	3,321	3,655

[a.] The permutation-derived LOD thresholds for eQTL significance were 2.89 for xylem, 2.92 for leaf, and 2.93 for root. [b.] A gene was declared as having a *cis*-eQTL if the eQTL peak overlapped at any point with the genetic interval containing the corresponding gene model in the genome, as defined by its flanking SSR, SFP or GEM markers. [c.] A gene was declared as having a *trans*-eQTL if the eQTL did not overlap at any point with the genetic interval containing the corresponding gene model in the genome, as defined by its flanking SSR, SFP or GEM markers. [d.] A gene was declared as having an ambiguous eQTL if the position of the corresponding gene model was not flanked on both sides by SSR, SFP, or GEM markers, or if the gene resided on an unplaced genomic sequence scaffold from the genome assembly.

Table 3-2. Significant eQTL hotspots by linkage group in each of the three tissues.

Linkage group	Total gene models (GMs)	Total GMs covered by map	Total xylem hotspot bins[a]	Total leaf hotspot bins[b]	Total root hotspot bins[c]
1	4181	3469	12	10	6
2	2940	2856	1	2	0
3	2097	1854	0	2	3
4	2040	1881	0	0	2
5	2402	2081	3	2	5
6	2526	2412	3	6	5
7	1466	1302	0	1	0
8	2169	2148	1	2	3
9	1749	1667	21	9	11
10	2586	1749	0	6	1
11	1624	1300	3	14	3
12	1478	648	0	1	6
13	1703	2323	0	5	9
14	2391	1916	6	14	10
15	1377	1076	10	4	0
16	1528	1501	0	14	9
17	1074	489	7	1	8
18	1425	736	0	1	1
19	1311	1269	0	3	6
Total	38067	32677	67	97	88

[a.] The permutation-derived hotspot threshold value for xylem was 39 eQTL. [b.] The permutation-derived hotspot threshold value for leaf was 21 eQTL. [c.] The permutation-derived hotspot threshold value for root was 16 eQTL.

Table 3-3. Summary of tissue-specific eQTL hotspot-based co-expression network construction in Family 52-124.

Linkage group	Bin number	Xylem eQTL		Leaf eQTL		Root eQTL		Probes shared
		Cis	Trans/Am	Cis	Trans/Am	Cis	Trans/Am	
Hotspots shared by xylem, leaf and root								
5	708	7	205	4	21	0	24	0
8	1014	6	35	13	9	6	22	3
Hotspots shared by xylem and leaf only								
1	1	6	77	10	204			2
2	276	4	59	9	23			4
9	1015	8	114	8	19			3
9	1020	1	52	2	20			0
9	1023	4	39	4	25			0
9	1028	2	61	4	29			0
9	1030	6	35	3	22			1
Hotspots shared by xylem and root only								
1	131	6	34			1	41	0
9	1096	22	14132			0	65	14
11	1263	9	68			2	17	2
14	1395	13	799			1	57	0
14	1396	6	140			0	19	0
14	1397	1	76			2	23	0
14	1398	0	52			0	20	0
14	1399	0	180			0	19	0
17	1664	2	83			0	33	1
Hotspots shared by leaf and root only								
5	682			10	12	16	1	0
7	838			9	16	3	15	1
11	1222			12	156	7	11	9
13	1333			9	27	3	63	1
14	1498			12	25	13	37	20
16	1628			9	21	7	14	3
19	1783			12	25	8	33	4

Table 3-4. Summary of tissue-specific eQTL hotspot-based co-expression network construction in Family 52-124.

Tissue	eQTL hotspot bins detected	Bin co-expression networks	Bins with networks	Max network size	Median network size	Total genes in networks	Minimum independent co-expression networks	Minimum independent genomic regions
Leaf	97	51	38	945	36	1678	9	7
Root	88	75	55	217	33	1188	16	11
Xylem	67	97	62	5787	99	9369	28	16

Table 3-5. Network-producing tissue-specific eQTL hotspots detected in leaf tissue and characteristics of their associated co-expression networks.

Hotspot locus color	Hotspot eQTL count	Number of networks constructed	Network size(s)	Most significantly enriched GO category	GO category type	GO enrichment nominal *P* value	Most significantly enriched *cis*-element	PLACE enrichment nominal *P* value	LG	Genomic physical interval
Magenta	275	1	31	tubulin complex	CC	1.615×10^{-5}	AGATC	1.236×10^{-4}	1	.22-1.52Mb
Green	114	1	18	n/a	n/a	n/a	n/a	n/a	1	6.87-?Mb
Blue	239	1	51	chloroplast	CC	1.370×10^{-40}	AATAAT	6.129×10^{-5}	6	2.27-3.63Mb
Purple	892	2	428	chloroplast	CC	5.703×10^{-5}	AGCGGG	1.979×10^{-7}	11	4.87-12.24Mb
			36	protein amino acid phosphorylation	MF	4.458×10^{-8}	TGCAAAG	1.923×10^{-5}	11	4.87-12.24Mb
Teal	1421	2	945	DNA recombination	MF	1.161×10^{-18}	CCACGTCATC	5.394×10^{-6}	14	4.71-7.1Mb
			27	n/a	n/a	n/a	CGCGGCAT	4.426×10^{-5}	14	4.71-7.1Mb
Orange	316	1	21	chloroplast envelope	CC	2.297×10^{-4}	AATAGAAAA	2.328×10^{-5}	16	.04-.67Mb
Red	126	1	11	manganese ion binding	MF	1.043×10^{-17}	n/a	n/a	19	0-2.31Mb

Table 3-6. Annotation of members and putative regulators of a leaf-specific coexpression network enriched for chloroplast biogenesis and functionality.

P. triochcarpa gene model	Network edges	Description/function	Chloroplast associated[a]	Cis-regulated	Candidate regulator	Gene ID number[b]
estExt_Genewise1_v1.C_LG_I7940	4	glycyl-tRNA synthetase / glycine--tRNA ligase	N	N		1
estExt_Genewise1_v1.C_LG_IV2768	2	RNA recognition motif (RRM)-containing protein	N	N		2
estExt_Genewise1_v1.C_LG_IX1928	2	cytochrome c biogenesis protein family	Y	N		3
estExt_Genewise1_v1.C_LG_VI1166	1	FTSZ2-2 (FtsZ2-2); structural molecule	Y	Y	√	4
estExt_Genewise1_v1.C_LG_X3683	15	DEAD/DEAH box helicase	Y	N		5
estExt_Genewise1_v1.C_LG_XIII0363	21	similar to hypothetical protein [Vitis vinifera]	Y	N		6
estExt_Genewise1_v1.C_LG_XV1465	3	phosphoglycerate/bisphosphoglycerate mutase family protein	Y	N		7
estExt_fgenesh4_kg.C_LG_I0088	14	SWIB complex BAF60b domain-containing protein	Y	N		8
estExt_fgenesh4_kg.C_LG_XV0016	7	GAMMA CAL2 (GAMMA CARBONIC ANHYDRASE-LIKE 2); acyltransferase/ transferase	Y	N		9
estExt_fgenesh4_pg.C_LG_I2552	3	GTP-binding protein LepA	Y	N		10
estExt_fgenesh4_pg.C_LG_VI0491	15	chloroplast thylakoid lumen protein	Y	Y	√	11
estExt_fgenesh4_pg.C_LG_VI1540	1	SQD1 (sulfoquinovosyldiacylglycerol 1); UDPsulfoquinovose synthase	Y	N		12
estExt_fgenesh4_pg.C_LG_XIX0756	2	50S ribosomal protein L21	Y	N		13
estExt_fgenesh4_pg.C_LG_XVII0327	11	porin	Y	N		14
estExt_fgenesh4_pm.C_LG_IX0206	7	mannose 6-phosphate reductase (NADPH-dependent)	N	N		15
estExt_fgenesh4_pm.C_LG_IX0706	11	SLP	Y	N		16
estExt_fgenesh4_pm.C_LG_V0614	1	similar to unnamed protein product [Vitis vinifera] (GB:CAO61051.1); contains InterPro domain Low temperature viability protein (InterPro:IPR007307)	N	N		17
eugene3.00011927	17	heavy-metal-associated domain-containing protein	Y	N		18
eugene3.00031286	1	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G27730.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO68737.1)	N	N		19

Table 3-6. Continued

P. triochcarpa gene model	Network edges	Description/function	Chloroplast associated[a]	Cis-regulated	Candidate regulator	Gene ID number [b]
eugene3.00060837	5	similar to unknown [Populus trichocarpa x Populus deltoides] (GB:ABK96465.1)	Y	Y		20
eugene3.00111342	3	similar to hypothetical protein OsI_013284 [Oryza sativa (indica cultivar-group)] (GB:EAY92051.1); similar to unknown protein [Oryza sativa (japonica cultivar-group)] (GB:AAK09232.1); contains InterPro domain Protein of unknown function DUF858	Y	N		21
eugene3.00190981	2	TIC110	Y	N		22
fgenesh4_pg.C_LG_VI000277	9	thylakoid lumenal 20 kDa protein	Y	N		23
fgenesh4_pg.C_LG_VIII001084	20	RPL15 (ribosomal protein L15)	Y	N		24
fgenesh4_pm.C_LG_VIII000641	3	CAAX amino terminal protease family protein	N	N		25
grail3.0003038902	2	APG1	Y	N		26
grail3.0013033101	8	oxygen-evolving complex-related	Y	N		27
grail3.0023008101	1	macrophage migration inhibitory factor family protein / MIF family protein	Y	Y		28
grail3.0024031501	2	ATSFGH (ARABIDOPSIS THALIANA S-FORMYLGLUTATHIONE HYDROLASE); S-formylglutathione hydrolase/ hydrolase	N	Y		29
gw1.I.1156.1	7	sugar transporter family protein	Y	N		30
gw1.I.5571.1	12	PRPL11 (PLASTID RIBOSOMAL PROTEIN L11); structural constituent of ribosome	Y	N		31
gw1.I.976.1	4	inorganic phosphate transporter	Y	N		32
gw1.IV.11.1	1	similar to hypothetical protein MtrDRAFT_AC147482g2v2 [Medicago truncatula] (GB:ABD32485.1)	Y	N		33
gw1.IV.1864.1	5	phosphoprotein phosphatase/ protein kinase	Y	N		34
gw1.IV.3048.1	22	ribosomal protein S5 family protein	Y	N		35
gw1.VI.2404.1	11	thioredoxin reductase	Y	Y	√	36

Table 3-6. Continued

P. triochcarpa gene model	Network edges	Description/function	Chloroplast associated[a]	Cis-regulated	Candidate regulator	Gene ID number [b]
gw1.VII.4002.1	2	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT4G24530.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO61608.1); similar to unknown [Populus trichocarpa] (GB:ABK95421.1); similar to hypothetical protein [Vitis vinifera] (GB:CAN78778.1); contains InterPro domain Protein of unknown function DUF246	N	N		37
gw1.X.2172.1	1	SCL30 (SC35-like splicing factor 30); RNA binding	N	N		38
gw1.X.2264.1	7	OVA1 (OVULE ABORTION 1); ATP binding / aminoacyl-tRNA ligase	Y	N		39
gw1.XIII.1619.1	2	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G28500.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO65032.1)	Y	N		40
eugene3.00440183	5	similar to hypothetical protein [Vitis vinifera] (GB:CAN83711.1); similar to unnamed protein product [Vitis vinifera] (GB:CAO22167.1); similar to unknown [Populus trichocarpa] (GB:ABK94059.1)	N	N		41
eugene3.00860035	3	CPHSC70-1 (chloroplast heat shock protein 70-1); ATP binding / unfolded protein binding	Y	N		42
eugene3.01180078	6	BAM3	Y	N		43
eugene3.01570036	13	lactoylglutathione lyase	Y	N		44
grail3.0107000401	15	chloroplast 30S ribosomal protein S20	Y	N		45
gw1.16424.1.1	2	60S ribosomal protein L18A (RPL18aC)	N	N		46
gw1.210.7.1	1	protein arginine N-methyltransferase family protein	N	N		47
gw1.129.1.1	3	similar to unknown [Populus trichocarpa x Populus deltoides] (GB:ABK96654.1); contains InterPro domain Protein of unknown function DUF1118 (InterPro:IPR009500)	Y	N		48
gw1.XVI.617.1	3	chloroplastic RNA-binding protein P67	Y	N		49

[a.] Chloroplast association judged by membership in GO category GO:009507 (chloroplast) or its associated daughter terms in the GO heirarchy. [b.] Gene ID number corresponds to network gene identifiers shown in Figure 3-5.

Table 3-7. Poplar genes from the sugar-associated co-expression network on linkage group IX in xylem that were also regulated by exogenous sucrose in a previous study [181] conducted in *Arabidopsis*.

P. trichocarpa gene model	*Arabidopsis* homolog[a]	Functional annotation
fgenesh4_pg.C_LG_IX001309	AT1G07670	calcium-transporting ATPase
estExt_fgenesh4_pg.C_LG_VIII0870	AT1G48170	similar to expressed protein [Oryza sativa (GB:ABA95965.1)]
gw1.XIII.1587.1	AT1G64260	zinc finger protein-related
gw1.158.130.1	AT1G72440	EDA25 (embryo sac development arrest 25); binding
grail3.0015003801	AT1G74560	NRP1 (NAP1-RELATED PROTEIN 1); DNA binding / chromatin binding / histone binding
gw1.IX.4200.1	AT2G29050	ATRBL1 (ARABIDOPSIS THALIANA RHOMBOID-LIKE 1)
fgenesh4_pm.C_LG_X000013	AT3G06483	PDK (pyruvate dehydrogenase kinase)
estExt_Genewise1_v1.C_LG_IX3773	AT3G47160	protein binding / zinc ion binding
estExt_fgenesh4_pm.C_LG_VIII0091	AT3G52990	pyruvate kinase
eugene3.00051316	AT3G54670	ATSMC1
estExt_Genewise1_v1.C_LG_II1288	AT3G61140	CSN1
estExt_fgenesh4_pm.C_LG_IX0496	AT3G63490	ribosomal protein L1 family protein
estExt_Genewise1_v1.C_LG_I3371	AT4G13850	GRP2
eugene3.00100701	AT4G14660	RNA polymerase Rpb7 N-terminal domain-containing protein
eugene3.00120829	AT4G24770	CP31
gw1.IX.3070.1	AT4G29430	RPS15AE (ribosomal protein S15A E); structural constituent of ribosome
gw1.VII.1484.1	AT4G33470	HDA14 (histone deacetylase 14); histone deacetylase
fgenesh4_pm.C_LG_IX000514	AT4G34890	ATXDH1 (XANTHINE DEHYDROGENASE 1); xanthine dehydrogenase
grail3.0011007002	AT5G09250	KIWI; DNA binding / transcription coactivator
estExt_fgenesh4_pm.C_LG_IX0400	AT5G09590	mtHSC70-2
estExt_fgenesh4_pm.C_LG_V0048	AT5G10360	EMB3010 (EMBRYO DEFECTIVE 3010); structural constituent of ribosome
eugene3.00090483	AT5G11880	diaminopimelate decarboxylase
estExt_fgenesh4_pg.C_LG_XI0654	AT5G39850	40S ribosomal protein S9 (RPS9C)
grail3.0060007502	AT5G46160	ribosomal protein L14 family protein / huellenlos paralog (HLP)
estExt_fgenesh4_pg.C_LG_IX1183	AT5G59030	COPT1 (COPPER TRANSPORTER 1); copper ion transmembrane transporter

[a.] *Arabidopsis* homolog from genome annotation by The *Arabidopsis* Information Resource, v.8.0

CHAPTER 4
UTILIZING GENETICAL GENOMICS TO IDENTIFY AN ADP-RIBOSYLATION FACTOR, *PtARF1*, AS A CANDIDATE GENE FOR LEAF SHAPE VARIATION IN *Populus*

Portions of this chapter will be submitted to a plant biology-related journal for publication

Introduction

Throughout the plant kingdom, leaf morphology demonstrates remarkable phenotypic diversity, making it a favorable system in which to study the evolution of variations in form. A number of genes and networks have been described that affect initial leaf development and pattern formation (reviewed thoroughly by Byrne [187]) in both simple and complex leaves [188, 189]. Similarly, mutagenesis screens have identified genes that appear to function in the regulation of leaf blade shape, specifically width [131, 134, 138], length [131-133], and their ratios. An emerging paradigm from these discoveries is that two-dimensional leaf shape can be regulated both by differential cell elongation (polarity) or differential cell proliferation favoring one dimension versus the other [129]. Despite this molecular genetic framework of leaf initiation, development, and shape established through mutagenesis of model plants, comparably little is known about whether the same molecular mechanisms underlie evolutionary differences in leaf morphological variation. For instance, to our knowledge no study has directly addressed whether alternative alleles at these genes underlie variation in leaf shape within or between different plant species (although Street *et al.* [123] take steps toward answering this question in the *Populus* genetic system). In light of this shortcoming and the extensive diversity in leaf form, additional studies exploiting naturally occurring variation are needed to illuminate the role of previously discovered genes and networks in evolutionary variation for leaf traits.

Quantitative trait locus (QTL) based approaches are one method that can be readily applied to identify genetic loci responsible for existing, natural diversity in leaf traits. QTL have been identified for leaf morphological traits in several species including tomato [190], poplar [70], oak

[191], maize [192], and *Arabidopsis* [193] showing that several aspects of leaf morphological variation are subject to multigenic control. Thus, it is apparent from these studies that loci thus far discovered on the basis of mutagenesis might only explain a portion the naturally occurring variation in leaf shape. However, molecular characterization of these regulatory loci has been generally prohibitive because of the historical challenges associated with moving from QTL to gene [194]. Recently, novel experimental approaches developed from advances in genome sequence production and genomic technology have eased QTL cloning [78, 89, 195], making the QTL-based approach a powerful tool to elucidate molecular mechanisms underlying naturally occurring phenotypic variation [83, 100]. The genus *Populus* is a particularly favorable system in which to apply these approaches, given its extensive genetic and phenotypic variation, the availability of several well established interspecific pedigrees, and rapidly growing genomic toolbox founded on the genome sequence of *P. trichocarpa* [196]. The genus is comprised of five evolutionary sections, and leaf morphology is widely regarded as diagnostic of evolutionary relationships at the sectional level [8]. In addition, several studies have shown that leaf morphological characters are predictive of long term clonal performance and growth [119, 121, 197]. Therefore, detailed study of intersectional poplar hybrids may provide an approach to identify loci associated variation in leaf morphology, and indirectly, clonal productivity.

Here, we analyzed an intersectional pseudo-backcross pedigree of narrow-leaf *P. trichocarpa* and wide-leaf *P. deltoides* for variation in leaf lamina shape. From a major QTL for leaf lamina width and a genome-wide analysis of gene expression in expanding leaf tissue, we utilized expression QTL (eQTL) analysis to identify an ADP-ribosylation factor (ARF) GTPase, which we designate *ARF1*, as a candidate gene regulating leaf morphology in this pedigree by a genetical genomics approach. We subsequently develop evidence to demonstrate the role of

differential expression of *ARF1* leaf width variation between *P. trichocarpa* and *P. deltoides*. ARF proteins have been shown to control the localization of auxin efflux carriers [198], which function to establish auxin gradients and apical-basal cell polarity in developing plant organs, suggesting that evolution of differential cellular polarity and auxin flux may play a significant role in leaf morphological variation observed in subgenera of genus *Populus*.

Materials and Methods

Plant Material and Phenotyping

A previously described [75, 173] pseudo-backcross pedigree Family 52-124 of *P. trichocarpa* X *P. deltoides* (clone 52-225) and *P. deltoides* (clone D-124) was utilized for phenotyping and subsequent identification of QTL and eQTL. Plants were grown as described [75, 173]. Leaf morphological QTL were identified from traits measured digitally using Image Pro Plus software (Media Cybernetics, Inc., Bethesda, MD) from an image scan of the leaf closest to one-half the live crown height in three biological replicates of 396 individuals. Traits measured included leaf lamina length, width, and their ratio. Leaf length was measured along the midvein from the junction of the lamina and petiole to the distal tip of the leaf. Blade width was measured at the widest point of the lamina. An analysis of variance [75] was applied to the phenotypic measurements and least-square means (LSM) were estimated for each of the individuals in the population and were utilized for QTL analysis. Clonal repeatability was calculated as described [75], except excluding the effects of row and column position in the experimental design.

Genotyping and Genetic Mapping of Progeny

A previously described microsatellite (SSR) and microarray-based genetic map [75] was utilized for whole-genome QTL mapping of the leaf width trait. Additional SSR markers within the primary QTL interval were identified from the 'Nisqually-1' genome sequence (Table 4-1,

[4]) using MsatFinder v.2.0 software. Primers were designed for these loci within the MsatFinder interface, and loci were amplified and genotyped in 96 recombinant progeny (as judged by flanking framework markers), using 1% agarose gels (w/v) as described [173].

Microarray Analysis

RNA was isolated [148] from one leaf immediately apical and basal to the phenotyped leaf in one biological replicate of 183 individuals. RNA was converted to double-stranded cDNA, labeled, and hybridized to a customized NimbleGen microarray platform. The RNA manipulations and microarray design have been previously described [173]. Resulting signal data was quantile normalized and $\log_2$ transformed [173] prior to being used as input for expression QTL analysis.

QTL and eQTL Analysis

Leaf width trait QTL were initially identified using the LSM estimates for phenotypic measurements (described above) in QTL Cartographer v.4.0 [199], using composite interval mapping [66, 67] with a standard threshold of the 95^{th} percentile of 1000 permutations. Subsequently, gene expression QTL were identified, measured for significance, and classified as *cis*- or *trans*-acting as described in Chapter 3.

To determine the relationship between gene expression and phenotype, phenotypic values from the same biological replicate measured for gene expression were correlated with the normalized gene expression values using the multivariate Pearson correlation function of JMP 7.0 (SAS Institute, Cary, NC) and significance measured by the associated *t*-statistic corrected for multiple testing as described in the Results. Finally, trait QTL position was confirmed for the single biological replicate on which gene expression was measured, using the composite interval mapping and permutation approach described above.

ARF1 cDNA Cloning and Sequencing

Full-length *ARF1* cDNA sequence was isolated from Nisqually-1, clone 52-225 and clone D-124 leaf cDNAs by PCR amplification. Briefly, RNA was extracted from intact, fully unfurled leaves of each genotype by a standard laboratory protocol [148]. RNA was DNAse treated following manufacturer's protocol (Invitrogen USA, Carlsbad, CA) and purified in an RNeasy purification column (Qiagen USA, Valencia, CA). First-strand cDNA was produced from 2µg purified RNA using: 500ng oligo-dT (Promega USA, Madison, WI), 100ng random hexamer (Promega), 1µL dNTP mix (10mM), 4µL first-strand synthesis buffer, 1µL M-MLV reverse transcriptase (Promega), and 1µL RNAsin (Promega) in a total volume of 20µL. cDNA was synthesized for 2hr at 37C followed by 15min at 70C to terminate synthesis. PCR was conducted using 5µL cDNA template, 5µL each of ARF1 forward and reverse primer (10mM, Table 4-2), 5µL dNTP mix (10mM), 5µL Advantage2 polymerase buffer (10x, Clontech Laboratories Inc., Mountain View, CA), 1µL Advantage2 polymerase mix (Clontech), .65µL DMSO, and 13.5µL betaine (6.5M) in a total volume of 50µL. The thermocyle utilized comprised initial denaturing at 93C for 5min, followed by denaturing at 93°C for 30sec, touchdown annealing for 30sec from 50°C-44°C (one cycle each), and extension at 68°C for 2min 30sec. Subsequently, 20 cycles of denaturing, annealing at 44°C and extension as above were utilized. A final extension of 68°C for 30min completed the thermocycle program. A secondary PCR was completed as above, replacing the cDNA template with an equivalent volume of primary PCR diluted in a 1:1000 ratio. The gene-specific band resulting from the secondary PCR was isolated in a 1.2% w/v agarose gel and purified using the ZymoClean gel purification system (Zymo Research, Orange, CA) as directed by manufacturer's protocol. Purified PCR product was directionally cloned into the pENTR/d-TOPO vector (Invitrogen) and transformed into TOP-10 competent cells (Invitrogen) per manufacturer's protocols. Selected colonies were grown overnight in liquid

culture and prepared for sequencing from the M13 forward and reverse promoters using the Qiaprep Miniprep Spin Kit (Qiagen). Resulting sequences were reassembled and analyzed using Sequencher (Gene Codes Corporation, Ann Arbor, MI) software.

ARF1 Promoter Isolation and Sequencing

To generate promoter sequence information, Genome Walker (Clontech) libraries were constructed against the DNA of the parent trees of Family 52-124 according to manufacturer's protocol, using *Eco*RV, *Pvu*II, *Sma*I, and *Sca*I restriction endonucleases. Sequences were amplified according to manufacturer's protocol with the primary reaction consisting of 7 cycles of denaturing at 94°C for 25sec and annealing/extension at 70°C for 3min, 32 cycles of denaturing at 94°C for 25sec and annealing/extension at 65°C for 3min, and a final extension at 65°C for 7min. The secondary reaction consisted of 5 cycles of denaturing at 94°C for 25sec and annealing/extension at 72°C for 4min, 20 cycles of denaturing at 94°C for 25sec and annealing/extension at 67°C for 4min, with a final extension at 67°C for 4min. Primers utilized in the genome walker steps can be found in Table 4-3. Products were visualized in 1% agarose gels stained with ethidium bromide, and resulting bands were purified using Zymoclean Gel DNA Recovery Kit (Zymo Research) following the manufacturer's protocol. Gel-purified PCR products were cloned into the pGEM-T vector (Promega) and transformed into *E. coli* DH5α competent cells (Invitrogen), purified, and sequenced from the SP6/T7 promoters as described above.

ARF1 Mutagenesis

The full-length ARF1 cDNA clone was subjected to site-directed mutagenesis using the Stratagene QuikChange site-directed mutagenesis kit, in order to introduce amino acid substitutions known to inhibit the exchange of GTP and/or GDP by ARF-interacting proteins [200, 201]. 50ng pENTR/d-TOPO entry vector containing full-length *ARF1* was PCR-amplified

with mutagenic primers (Table 4-2) as directed by the manufacturer's protocol, using 15 cycles of primer extension at 68°C for 3min 30sec. Resulting mutagenized sequences were treated with *Dpn*I restriction endonuclease to degrade methylated template plasmid, and reactions were subsequently transformed into *E. coli* XL-1 blue cells as directed. Putatively mutagenized clones were verified by sequencing from the vector's M13 forward and reverse promoters and were subsequently recombined and replicated in the pZKY-1 overexpression binary vector using Invitrogen LR clonase enzyme mix and Invitrogen TOP-10 competent cells, as directed by manufacturer's protocols.

Nisqually Leaf Disc Expression Experiment

To determine whether *ARF1* expression varied across the leaf or plant within a uniform genetic background, we analyzed the quantitative expression of *ARF1* in Nisqually-1 within specific regions of several leaves at various stages of expansion. Five clones of Nisqually-1 were grown as described [75]. Whole, intact leaves were harvested and flash frozen in liquid nitrogen. After lyophilization, leaf discs (~25mg dry weight) were extracted from eight predetermined positions (pooled across the leaf blade midvein) in the first, third, sixth, and ninth unfurled leaves from three of the five clones. RNA was extracted from each disc using a cetyltrimethylammonium bromide/chloroform extraction [148], followed by purification of the aqueous phase in a Qiagen RNeasy column (Qiagen USA) as directed by manufacturer's protocol.

Purified RNA (500ng) was converted to first strand cDNA using Promega M-MLV reverse transcriptase, random hexamers, and oligo-dT as described above. Each resulting sample was analyzed for expression of *ARF1* by real-time PCR (primers in Table 4-2) using .5uL of first-strand cDNA template, 10x Brilliant SYBR Green qPCR Master Mix (Stratagene), and Mx3000P thermocycler (Stratagene) as directed by manufacturer's protocol, in a total volume of 10uL.

Similarly, expression was determined for the control genes *PtACT2*, *PtUBQ*, and *Pt18S* for each sample. Resultant data were analyzed by an efficiency-corrected relative quantification method as described [202] normalized to the geometric mean of the three internal control genes [203].

Verification of Allele-Specific Expression Effects in Segregating Population Subset

To verify the effect of allele combinations at the *ARF1* locus in *ARF1* expression values, we grew a subset of 60 progeny (30 each heterozygous and homozygous for *P. deltoides ARF1* allele, as judged by flanking SSR genotypes) from the segregating population under conditions identical to those of the original study [75]. We harvested the fourth and fifth fully unfurled leaves for *ARF1* expression analysis. RNA was extracted by the standard protocol [148], converted to cDNA using M-MLV, random primers, and oligo-dT, and subjected to qPCR analysis as described for the leaf disc experiment above.

Results

Identification of a Major QTL for Leaf Blade Width

We analyzed an interspecific pseudo-backcross pedigree (Family 52-124) for variation in leaf lamina shape (measured by lamina length, width, and their ratio) segregating between the narrow-leaf donor parent species *P. trichocarpa* and the broad-leaf recurrent parent *P. deltoides* (Figure 4-1) by composite interval QTL analysis and a previously established genetic map [75]. One to five significant QTL were identified for the various traits (Figure 4-2, Table 4-4) including a major locus on linkage group (LG) X that regulated both lamina width (phenotypic variation explained [PVE] = 6.0%) and length:width ratio (PVE = 14.2%; Figure 4-2). The major QTL interval, constrained by sequence-linked SSR markers PMGC_2855 and GCPM_2122, encompassed 625 genes and spanned ~3.5Mb of uninterrupted sequence in the genome sequence of *P. trichocarpa* [4]. To reduce the number of potential candidate genes in the trait QTL (tQTL) interval, mapping resolution was increased by genotyping recombinant individuals (n=96) for

seven additional SSR markers identified from the genome sequence (Table 4-1). Additional mapping decreased the QTL interval to ~3.0Mb and 450 candidate genes (Figure 4-3). Due to the high proportion of phenotypic variance explained and quality of the genome assembly in this tQTL region, we elected to pursue the major locus on LG X for further characterization and QTL cloning.

Gene Expression Analysis of Leaf Tissue Identifies *PtARF1* as a Candidate Gene for Lamina Shape

As additional recombinant progeny could not be identified to increase map resolution across the target interval, we elected to move forward with an integrative genomics approach to further reduce the pool of candidate genes. Because of the prominent role of transcriptional diversity in evolutionary distinction of species [36], we hypothesized that interspecific differences in expression of a key regulator of leaf expansion or cell division may explain the leaf lamina variation observed in the progeny of Family 52-124. To evaluate this hypothesis, we measured genome-wide gene expression in expanding leaf tissue from one biological replicate of 183 segregants, and mapped the resulting transcript abundance as eQTL. Our genome-wide analysis identified 13,403 statistically significant eQTL representing 12,392 unique gene models (described fully in Chapter 3).

Since our hypothesis was that lamina shape variation was a product of differential gene regulation arising from interspecific polymorphism(s) within the tQTL interval, we narrowed our focus to only the genes with eQTL surpassing the significance threshold in this region. Our analysis identified 161 eQTL, which we classified as *cis*- (n=116) or *trans*- (n=45) acting, contingent on the physical position of the gene model in the genome assembly. We classified all eQTL arising from gene models on unassembled genomic scaffolds (n=19) as *trans*-acting, since the *Populus* genome sequence assembly in the tQTL region was predicted to be contiguous [4].

We expected that if differential transcript accumulation of key regulatory factors accounted for phenotypic variation in leaf shape, abundance of such transcripts should exhibit a significant statistical correlation with leaf lamina shape. Furthermore, differential regulation of these transcripts should be accounted for by the same region regulating variation in the phenotypes – i.e., they should demonstrate eQTL controlled by the tQTL interval. Thus, we utilized a standard multivariate correlation to determine the relationship between the leaf lamina phenotypes for each transcript with an eQTL in the interval of interest. Among the 161 transcripts with eQTL co-localized to the trait interval, we identified only two with statistically significant (Bonferroni corrected $\alpha = .05$, $P < 3.106 \times 10^{-4}$) correlations to lamina width (Table 4-5). No genes were correlated with lamina length:width ratio at this level of statistical significance. Both significantly correlated genes were regulated in *cis*- by the trait QTL region – one gene encoded the poplar homolog of the *Arabidopsis* salt-inducable protein AtCP1 [204], while the other encoded an ADP-ribosylation factor (ARF) GTPase which we designated *PtARF1*.

We also analyzed the relationship of each gene expressed above array background in leaf tissue to both lamina shape phenotypes, by correlating each phenotypic trait to the gene expression measurements and ranking these correlations by their significance. Among 21,810 genes detected above background on the microarrays, 257 were significantly correlated with leaf width (FDR < .05, $P < .00061$). No genes correlated with lamina length:width ratio at an analogous level of significance. Among the 257 genes identified, only three underlie the trait QTL interval regulating leaf width (Table 4-6). Yet, among these three genes there was a clear discrepancy – just one gene, *ARF1*, was strongly related to to the leaf width phenotype while being regulated in *cis*- by the lamina phenotypic QTL interval. Conversely, the other two genes were not differentially regulated by the QTL interval despite their correlation with the phenotype

(Table 4-6) – in fact neither gene had a significant eQTL anywhere in the genome. Furthermore, it is notable that *AtCP1* ortholog regulated in *cis*- by the QTL interval and significantly associated with the leaf width trait by the previous analysis was not significantly associated when considered on a transcriptome-wide scale.

It might be expected that among a large gene family such as the ARF-type GTPases, compensatory expression by another *ARF* gene family member could occur in the *P. deltoides* background to modulate the effect of decreased *ARF1* expression. To determine whether additional members of the *ARF* gene family might respond in their expression to the varying alleles at *ARF1*, we analyzed the correlation of expression of all predicted *ARF*-type and *ARF*-like genes (n = 25) to *ARF1*. We also calculated the level of expression for each *ARF*, relative to the expression detected for *ARF1*, to determine whether a true compensatory effect could be expected. The expression of *ARF1* is positively correlated to most other *ARF*s, rather than negatively correlated as would be expected for a compensatory expression effect (Table 4-7). Furthermore, the two negatively correlated genes were detected in low relative abundance compared to *ARF1* and many of the other ARF family genes (Table 4-7) – in fact, both were below background for the microarrays as judged by the 97.5^{th} percentile of the negative control probes [173]. Finally, no significant *cis*-eQTL was detected for *ARF1* in either root or xylem tissues that were also assayed for transcript abundance (data not shown), suggesting *ARF1* may be expressed or modulated specifically in leaf tissue. Taken together, the entire body of statistical evidence, coupled with the robust tQTL/eQTL overlap (Figure 4-3) and functional annotations, led us to pursue *ARF1* for further characterization as a candidate gene for quantitative variation in leaf shape within our segregating population.

Isolation of Interspecific *ARF1* Coding and Promoter Polymorphisms

To identify polymorphisms in *ARF1* among the Family 52-124 parent trees, we utilized PCR to isolate the cDNA coding sequence corresponding to the *ARF1* gene. Sequence analysis revealed minimal polymorphism between haplotypes, including only three synonymous nucleotide substitutions and no insertions or deletions that would result in translational frame shifts (not shown). The predicted function of ARF1, in the core vesicle trafficking machinery, indicates that the sequence is likely to be strongly conserved [205], so these results were not surprising.

We speculated that differential expression of *ARF1* would most likely be explained by interspecific polymorphisms in the 5' regulatory region adjacent to the gene. To explore this possibility, we utilized 4 steps of Genome Walker PCR to sequentially isolate sequence 5' to the start codon of *ARF1* from each parent tree. Haplotypes arising from each parent were discriminated on the basis of first-exon synonymous substitutions initially detected in the full-length cDNA clones above. Extensive sequence heterogeneity between the parent trees made it challenging to readily amplify and isolate clones derived from the target sequence – as such, only 500bp could be readily isolated from all three *P. deltoides* haplotypes of the parent trees, whereas >1.2Kb was isolated from the *P. trichocarpa* haplotype of the hybrid tree. Characterization of polymorphism between the haplotypes revealed extensive interspecific differences between the haplotypes in this limited region, including at least seven small insertions/deletions and 18 SNP (Figure 4-4). Interestingly, however, the sequence from haplotypes within each species was highly conserved: no in/dels were identified in either species, and only two SNP were identified for *P. trichocarpa* (relative to the Nisqually-1 genome sequence) within the 500bp region of comparison. Sequence corresponding to the *P. deltoides* haplotypes was fully conserved relative to one another.

Localized Expression of *ARF1* in Expanding Leaves

To determine the expression pattern of *ARF1* throughout varying stages of leaf expansion, we conducted an experiment to sample localized regions of expanding leaves from *P. trichocarpa* genotype Nisqually-1. Five cuttings of Nisqually-1 were grown for 12 weeks under conditions mimicking those of the original experiment [75]. From four different expanding and intact leaves of three clones, we extracted small leaf disc samples (~50mg) from eight distinct regions of the leaf blade (Figure 4-5A) to measure *ARF1* expression by quantitative real-time PCR (qRT-PCR).

ARF1 expression, when normalized to the geometric mean of three internal control genes, was surprisingly uniform across all sampled positions (Figure 4-5B) within a leaf. Similarly, across the four leaves sampled, expression was relatively consistent within sample positions. Slight variation was noted within some sampled positions, corresponding to generally increased expression of *ARF1* near the midvein with increasing maturity (Figure 4-5B, positions 3, 4, 6, 8). However, error variance associated with the measurements and the small sample size of the experiment minimizes the number of differences that could be deemed significant. This preliminary analysis suggests that *ARF1* expression is likely to be relatively homogenous throughout the leaf lamina and that cellular position within the lamina, or leaf expansion stage, does little to alter the expression of *ARF1* in the *P. trichocarpa* genetic background.

Verification of Allele-Specific Expression Effects for *ARF1*

The presence of a *cis*-eQTL for *ARF1* in the segregating population dictates that expression variability for the gene exists in the population and associates with the *ARF1* locus. Lower expression of *ARF1* is associated with a *P. deltoides*-like leaf shape. Given the minimal variation noted in expression across the leaf blade in our previous experiment, we speculated that the low microarray signal obtained for *ARF1* in progeny homozygous for the *P. deltoides* alleles

might be accounted for by an overall lack of *ARF1* expression in the *P. deltoides* background. To test this hypothesis, we propagated 60 progeny genotypes of Family 52-124, including 30 genotypes each that are either homozygous or heterozygous for *P. deltoides ARF1* alleles (as judged by flanking microsatellite locus genotypes) and measured *ARF1* expression in expanding leaves by qRT-PCR.

The results of this experiment indicated that the expected allelic effect on *ARF1* expression due to substitution of a *P. trichocarpa* allele for a *P. deltoides* allele could be recapitulated in the segregating subpopulation using qRT-PCR technology (Figure 4-6). However, significant differences were not observed between allelic classes in the context of this experiment ($P < .253$, measured by Student's *t*-test with unequal variance) and expression was readily detected from both genetic backgrounds. Furthermore, while variation in *ARF1* expression at the original LG X eQTL interval was observed, this variation was not sufficient to surpass a standard significance threshold of LOD > 2.5 ($LOD_{observed} \approx 1.80$, data not shown). The outcome of this experiment was likely affected by the small subpopulation size, coupled with the measurement of only one biological replicate of progeny. Nonetheless, the observation that allelic effects can be readily captured on the basis of qRT-PCR is an important outcome for future testing of the effects of *ARF1* expression in independent verification populations. Furthermore, an overall lack of *ARF1* transcript within *P. deltoides* is not the explanatory factor in leaf phenotypic variation.

Discussion

A major goal of plant biology is to identify genes implicated in adaptation and evolution, as they may provide the genetic tools to develop more productive genotypes that are capable of withstanding varying sources of biotic and abiotic stress. Many of these adaptive traits are complex in nature and thus are suitable to dissect using a combination of traditional quantitative genetic approaches such as QTL analysis and gene expression phenotyping in segregating

populations [78, 89, 195]. Here, we considered leaf lamina shape among an interspecific hybrid family of *Populus* in a genetical genomics context [78]. From a series of QTL for leaf lamina shape characters, we identified a major QTL implicated in lamina morphology – pleiotropically regulating both lamina width and lamina length:width ratio. Utilizing whole-genome microarrays for expression analysis of leaf tissue, we identified a group of genes regulated in *cis*- by the phenotypic QTL interval, and/or whose expression was statistically correlated with both leaf width and length:width ratio. Considering the statistical relationships, genetic regulation of expression, and functional annotations, we selected an ADP-ribosylation factor GTPase – *ARF1* – as the primary candidate gene governing lamina shape characters in the interspecific hybrid pedigree.

Both our microarray and real-time PCR analyses clearly indicate that *ARF1* is differentially regulated in *cis*- by the phenotypic QTL region. It is also clear that among all genes regulated by this region, *ARF1* has the strongest statistical relationship with leaf lamina width. Furthermore, considering all genes in the genome, the correlation of *ARF1* is among the strongest to the phenotype regardless of whether any genetic regulation of the transcript is considered. From the sequencing data produced for *ARF1* from *P. trichocarpa* and *P. deltoides*, it is unlikely that transcript stability due to mRNA sequence characteristics profoundly affects the abundance of *ARF1 in vivo*. Rather, our promoter sequencing data (albeit limited) suggests that *cis*-regulatory polymorphisms in alleles of *ARF1* leads to differential transcript accumulation observed in our experiments. While further promoter cloning and reporter gene testing are clearly necessary to validate this hypothesis, it is interesting that haplotype structures of the promoter region cloned thusfar were highly conserved within each species. This observation might indicate that the regulation of *ARF1*, while different between species, is important for plant or leaf development

in *Populus*. However, the sample size of *ARF1* promoter haplotypes is insufficient to draw meaningful conclusions about specific polymorphisms that might direct *ARF1* expression. Sequencing *ARF1* promoter regions from a larger compendium of *P. trichocarpa* and *P. deltoides* individuals will shed additional light on the regulatory mechanism and *cis*-elements important for *ARF1* regulation. Previous evidence [206, 207] suggests that vital *cis*-elements can be located at long distances from the gene in question – thus the availability of resequenced poplar genomes from a collection of unrelated individuals could provide a key piece of data in the search for specific sequencees regulating this and other genes of interest.

Homogeneous expression of *ARF1* throughout the leaf lamina and at different stages of leaf expansion in *P. trichocarpa* indicates that *ARF1* is likely not regulated in a regionally specific manner during leaf development or expansion. It is possible that expression is regionally controlled only in *P. deltoides* in response to presence or absence of a specific *cis*-element. Similarly, it is possible that expression could be restricted to a specific tissue layer or type within the leaf lamina in one or both species. The sampling method we utilized to measure regional expression of *ARF1* would not detect tissue-specific differences, because sections were taken encompassing all layers of the leaf lamina (mesophyll, epidermis, and vasculature) in proportions in similar to which they would occur in an intact leaf. Finally, a more significant and functionally important *ARF1* expression difference might be noted among newly forming leaves within the apical meristem. Additional studies employing more sophisticated sampling methods, (i.e., laser capture microdissection) would be required to reconcile these possible scenarios of *ARF1* regulation.

Evidence recently generated in the *Arabidopsis* system suggests the underpinnings of a hypothetical mechanism by which differential expression of *ARF1* in the two *Populus* genetic

backgrounds could lead to variation in leaf lamina shape. The action of *AtARF1* (the ortholog of *PtARF1*) has been shown to modulate the kinetics of endocytosis, PINFORMED2 (PIN2) localization, and cell polarity in *Arabidopsis* roots [198]. Several PIN auxin efflux carrier proteins have also been shown to be post-translationally localized to apical and basal membranes of *Arabidopsis* epidermal cells in an endocytic process mediated in part by two Rab5 GTPase homologs [208]. The connection between cell polarity, auxin, and directional growth has been intricately studied in several model plant systems [209, 210]. Auxin has already been implicated in initial leaf formation [211, 212], lamina margin elaboration [188, 189], leaf vasculature patterning [213]. Thus, it is likely that leaf expansion is also directly affected by auxin flux. In *Populus*, our evidence supports a model whereby differential expression of *PtARF1* in *P. trichocarpa* and *P. deltoides* changes the dynamics of endocytosis-mediated PIN localization in leaf cells. Increased abundance of *ARF1* transcript, and hence ARF1 protein, in *P. trichocarpa* speeds the process of PIN polarization through the endocytic pathway during development and expansion. More rapid acquisition and maintenance of PIN polarization decreases lateral auxin flux and, subsequently, lamina expansion in the leaf width direction (Figure 4-7A). Conversely, in *P. deltoides*, lower abundance of *ARF1* transcript leads to a more limiting supply of ARF1 protein, which decreases the relative rate of endocytosis and, hence, PIN polarization. Slower PIN polarization throughout development increases lateral auxin flux and expansion of the lamina in the leaf width direction (Figure 4-7B). An important experiment to verify this model will be to determine the rate of endocytosis-mediated PIN localization in each species. A template for this experiment, providing a quantitative "cell polarity index" of fluorescently tagged PIN protein, has recently been published [208] and will provide a powerful framework in which to test the hypothesis.

Three other genes significantly associate with the lamina shape phenotypes through the two methods of statistical association used in our analysis. These genes include: (1) the poplar ortholog of *AtCP1*, a calcium-binding calmodulin-related protein inducible by salt stress [204]; (2) a chaperonin protein (3) and a gene encoding a calmodulin calcium binding protein. While a role for these genes in leaf development is not implausible, the evidence developed thusfar strongly favors *ARF1* as the causal regulator of lamina shape differences in the *Populus* segregating population. Furthermore, the known function of *ARF1* places it directly in well-described pathways affecting cell expansion and leaf development, whereas to our knowledge, no significant evidence has been developed for the other genes in leaf shape or development.

Collectively, our results provide another piece of compelling evidence for the role of transcriptional diversity in shaping variations in plant form. Similarly, we provide yet another way in which the key plant hormone auxin could shape diversity in plant morphology. We have demonstrated that natural variation for auxin response could play a key role in plant diversity we observe in nature, and our analysis suggests that natural variation in hormone response pathways clearly warrants additional investigation as we strive to learn more about the evolution of morphological differences among plant taxa.

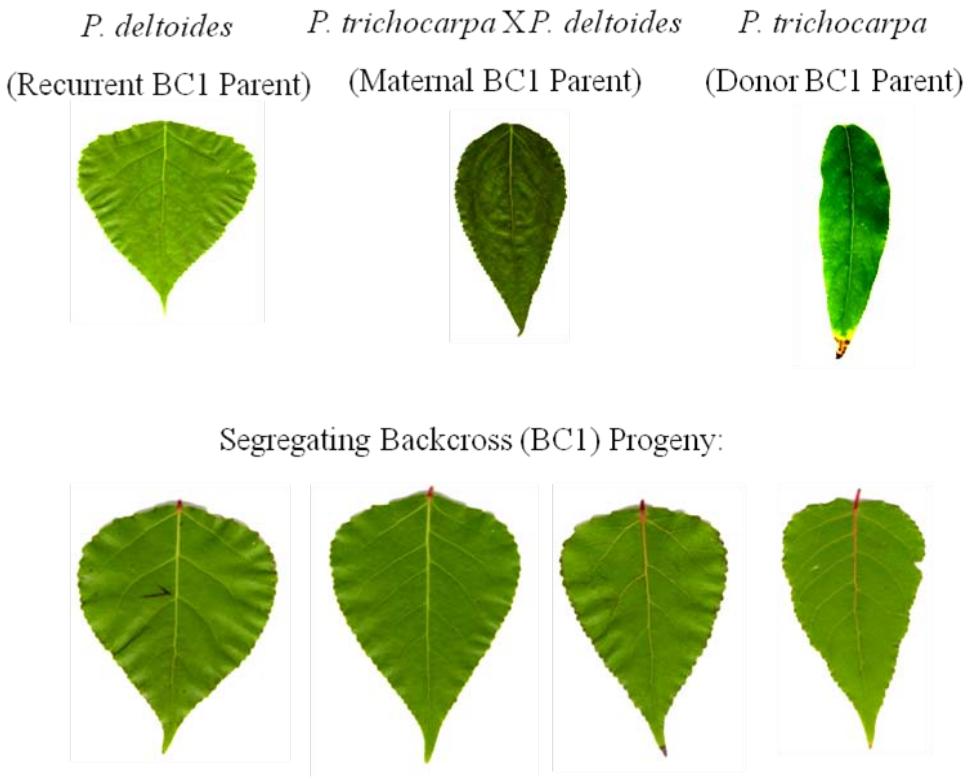


Figure 4-1. Leaf lamina shape variation among parents and progeny of the *P. trichocarpa* X *P. deltoides* pseudobackcross pedigree Family 52-124. The donor parent, *P. trichocarpa*, has a lamina length/width ratio of ~3.0, while the recurrent parent, *P. deltoides*, has a lamina ratio of ~1.0. The trait exhibits additive variation, as the hybrid parent has a ratio of ~1.5. Segregating BC1 progeny span the spectrum of phenotypic varation from the hybrid to the recurrent parent with limited transgressive segregation observed.

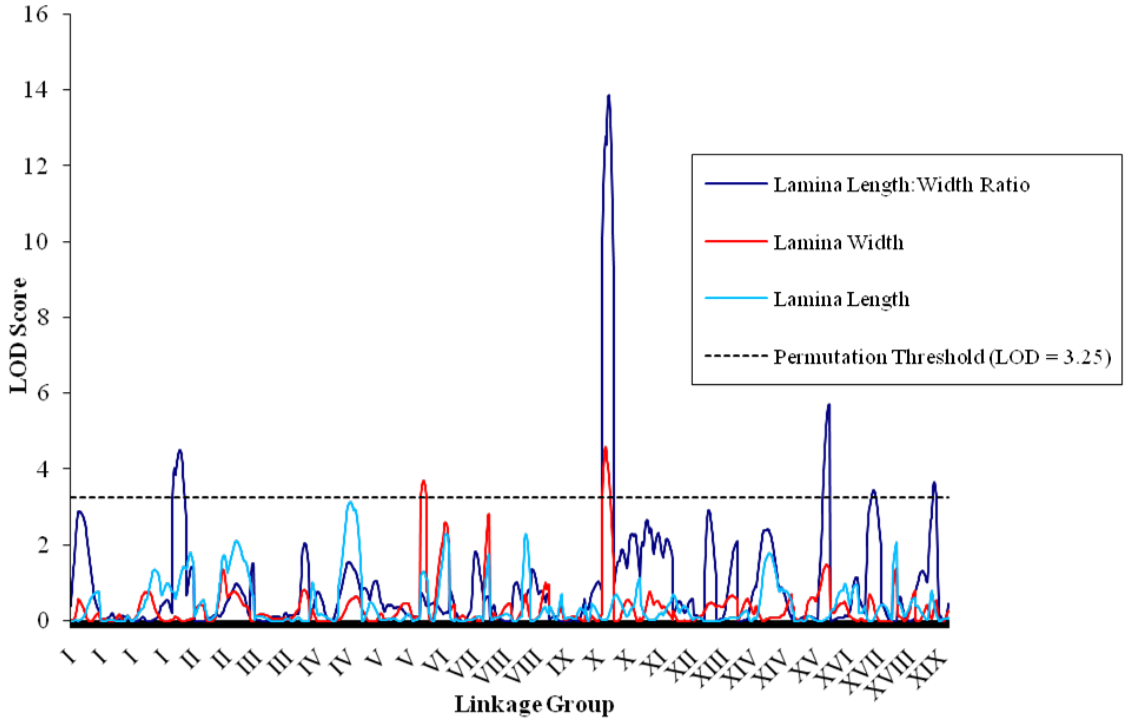


Figure 4-2. Genome-wide composite interval mapping scan for leaf lamina shape characters in Family 52-124. Least-square means of triplicate measurements from 396 progeny were utilized as the genotypic value for mapping against an SSR- and SFP-based single tree genetic map produced for the hybrid parent, genotype 52-225.

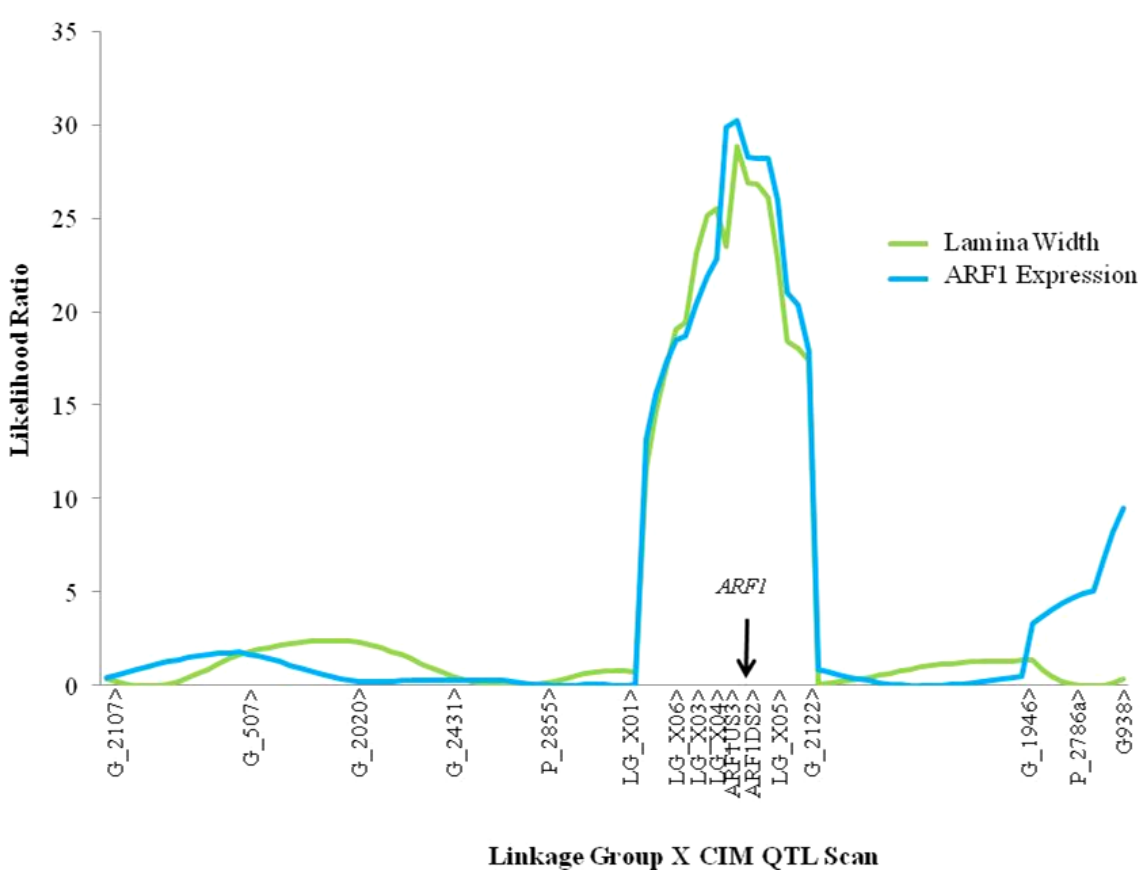


Figure 4-3. Fine-scale mapping of the major lamina width QTL on LG X in the segregating pedigree. Framework SSR loci were genotyped in 396 progeny ("G_" and "P_" loci) while additional SSR underlying the QTL were genotyped in 96 recombinant progeny. Recombinants were identified by maternally inherited marker genotypes at locus P_2855 and G_2122. The approximate genomic location of the *ARF1* gene is denoted by the downward facing arrow.

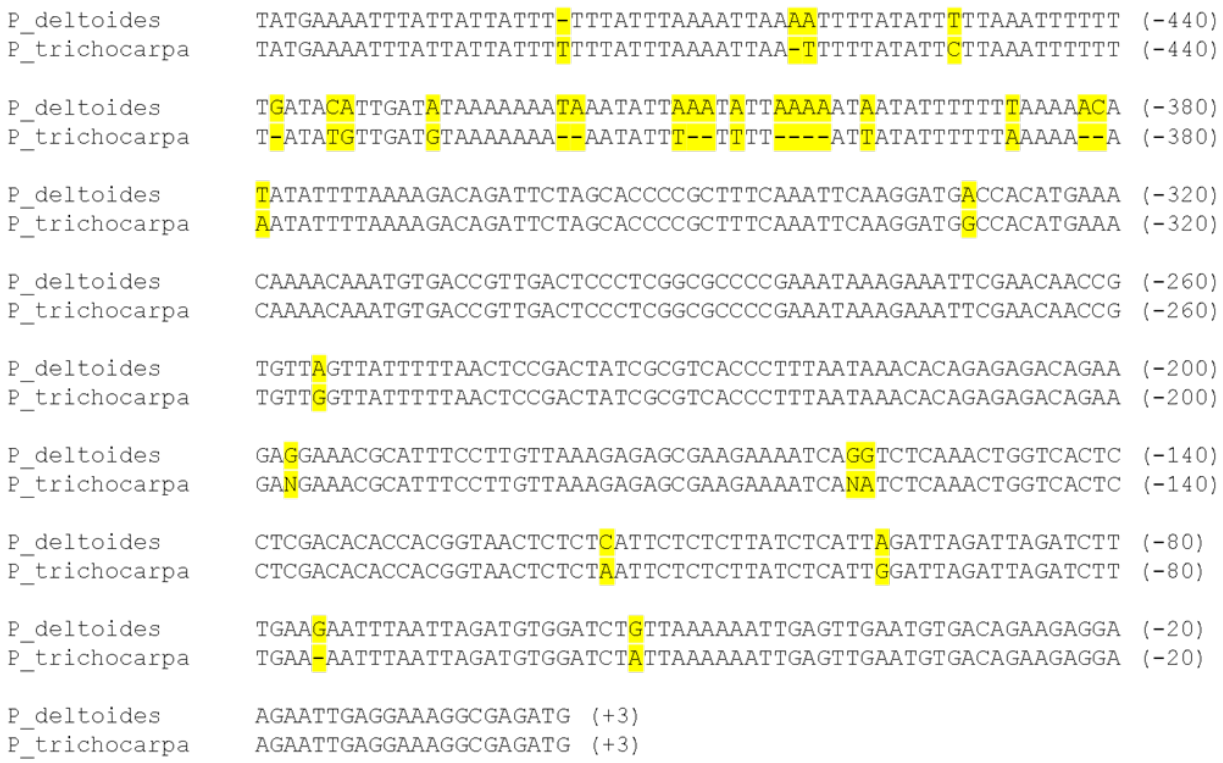


Figure 4-4. Sequence alignment for *ARF1* upstream regions isolated from *P. trichocarpa* and *P. deltoides*. Sequence was isolated by sequential steps of Genome Walker PCR as described in Materials and Methods. SNPs detected within the *P. trichocarpa* haplotype, relative to comparison to Nisqually-1, are depicted by "N". Polymorphisms between haplotypes are highlighted in yellow.

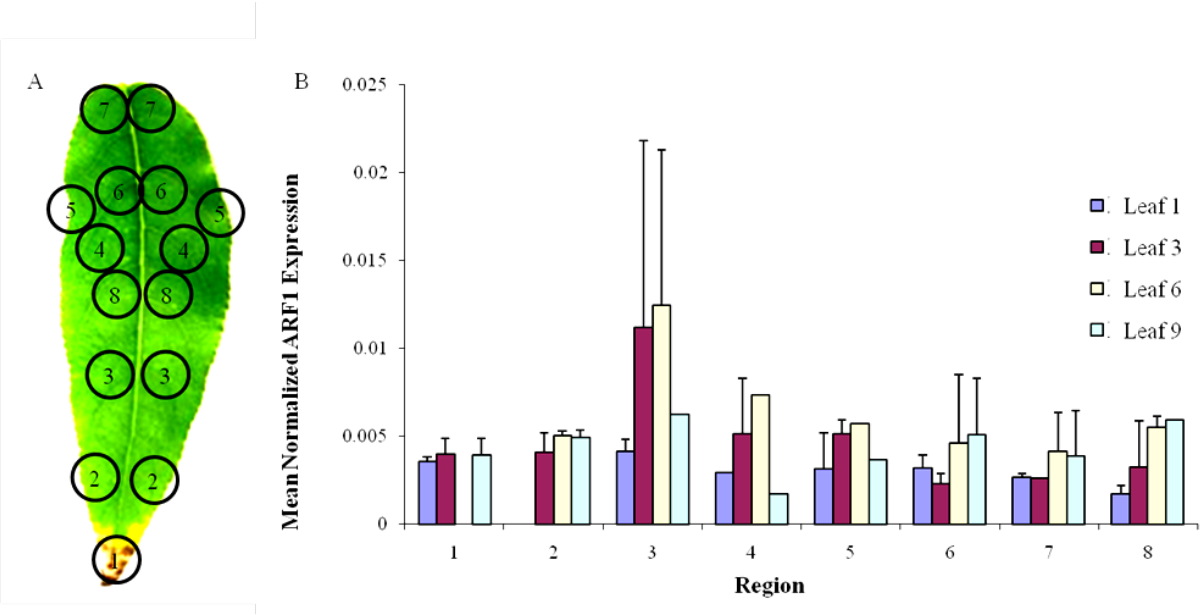


Figure 4-5. Quantitative PCR analysis of regional *ARF1* expression in differently staged expanding leaves of *P. trichocarpa*. (A) 25mg lyophilized leaf discs were harvested from the leaf regions depicted. Samples of the same number were pooled across the midvein for subsequent RNA extraction and analysis. The first, third, sixth, and ninth fully unfurled leaves were sampled in this manner for comparison of the effect of leaf maturity. (B) Results of qRT-PCR analysis. Bars represent mean and standard error of three measurements of *PtARF1* expression normalized to the geometric mean of three internal control genes. Lack of an error bar indicates that sample was not triplicated due to experimental failure at one or more stages.

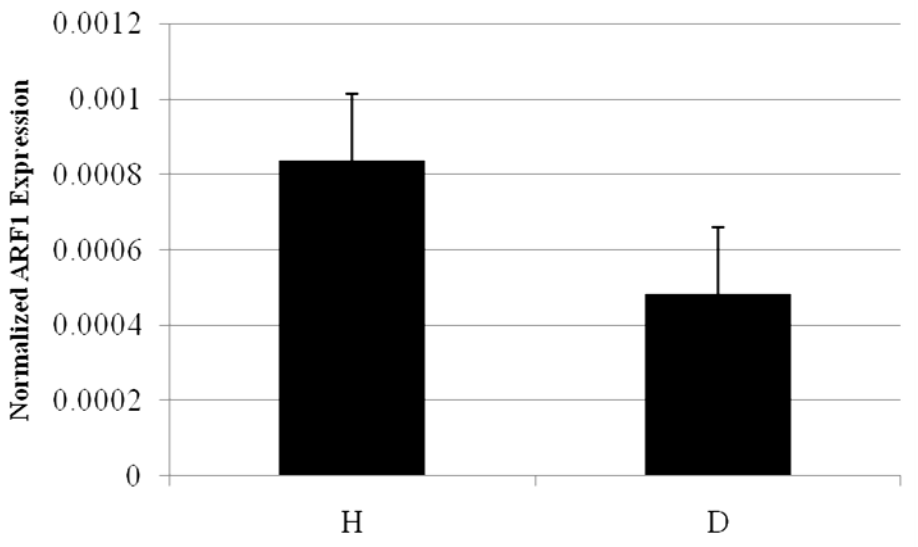


Figure 4-6. Quantitative RT-PCR verification of genetic effects on *ARF1* expression in the segregating population. Sixty progeny genotypes were grown in a randomized design, including thirty genotypes each homozygous (D) or heterozygous (H) for *P. deltoides* alleles at SSRs flanking the *ARF1* locus. Mean and standard error for each genotypic class (Student's *t*-test $P < .253$) are shown.

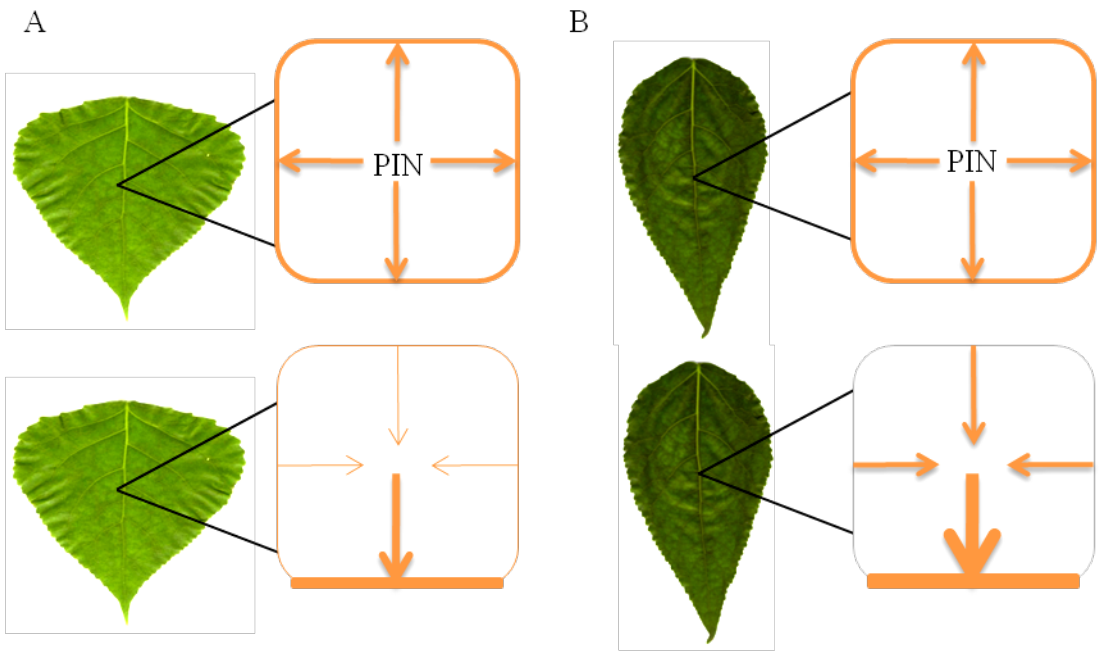


Figure 4-7. A hypothesized molecular mechanism for how differential *ARF1* expression affects leaf lamina phenotypic variation in *P. trichocarpa* and *P. deltoides*. PIN protein flux and localization in the leaf cell is depicted in orange thoughout the figure. Figure modified from the model of PIN localization presented by Dhonukshe et al. [208]. (A) In *P. deltoides*, newly synthesized PIN, secreted to the plasma membrane in a non-polar pattern, is endocytically localized to apical/basal membranes slowly due to decreased ARF1 abundance. This results in increased auxin flux from the lateral cell membranes and increased expansion/growth in the lamina width direction. (B) In *P. trichocarpa*, newly synthesized PIN is also secreted in a non-polar manner. Higher ARF1 abundance increases the rate of endocytic apical/basal PIN localization, minimizing lateral auxin flux and polarizing cell growth and expansion preferentially in the lamina length direction.

Table 4-1. Primers utilized to amplify microsatellite loci across the major linkage group X QTL for leaf lamina characters in Family 52-124.

Locus	Primer forward sequence	Primer reverse sequence	Motif	Repeats	Linkage group X physical location (bp)
PMGC_2855	GGTATCTTGTTATCCACTGCC	TTTTCCTCGTTAATTAGAGTCG	GA	n/a	11,122,293
LG_X01	GCCACCAATCCAGCAAGTAA	AATGGAGGTGTGGCAGTAGC	TA	9	11,872,486
LG_X06	GAGAAGCAGCAATGCAGGAT	ATGCAAACTGGTCCGGATAC	CTT	6	12,645,373
LG_X03	ACCTGGTCCATTTGTTGAGC	TGCAGGCAATCTCAAACTCA	GA	5	13,378,326
LG_X04	GGGATGGCAAAATACGTTCA	CATCATCGTACAACCTCACCTT	TA	5	14,126,639
ARF1US3	GGCAAGGACAGCGCATGATG	TTTCACCGAACTTCCACACTTT	n/a	n/a	14,365,406
ARF1DS2	GTGATAGCAGAGAGCCGAAA	GTAAGTATGAGAGGAATGAGGGG	n/a	n/a	14,369,387
LG_X05	TGGTAGATTGCGAGCTGAGA	AGGGCTTTCCGGCTATTAAA	CT	5	15,159,499
GCPM_2122	TCAGCAACTATCACCATGAA	GGAATGTGCAGCATATACAA	GT	12	15,661,771

Table 4-2. Primer sequences utilized for cloning and qRT-PCR experiments.

Gene	Primer forward sequence	Primer reverse sequence
ARF1 Full-length (Primary PCR)	CTTGTTAAAGAGAGCGAAGA	AATGCCAATAGTCTACACACTG
ARF1 Full-length (Secondary PCR)	CACCATGGGGCTGTCATTCAC	AATGCCAATAGTCTACACACTG
ARF1 T31N Mutant	GTCTTGATGCTGCTGGTAAAGACACCATTCTTTACAAGC	ACTTGAGCTTGTAAAGAATGGTGTCTTTACCAGCAGC
ARF1 Q71L Mutant	GGGATGTTGGCGGTCTAGACAAGATTCGTCCTTTGTGG	CGAATCTTGTCTAGACCGCCAACATCCCAAACAGTGAAGC
ARF1 qPCR	GCCTTCCAAGTGCTACGAGAG	GACAGCCCCATCTCGCCT
PtACT2 qPCR	CCCATTGAGCACGGTATTGT	TACGACCACTGGCATACAGG
PtUBQ qPCR	GTTGATTTTTGCTGGGAAGC	GATCTTGGCCTTCACGTTGT
Pt18S qPCR	AAACGGCTACCACATCCAAG	CCTCCAATGGATCCTCGTTA

Table 4-3. Primer sequences utilized for Genome Walker promoter PCRs.

Genome walker step	Primary reaction reverse primer	Secondary reaction reverse primer
A	CAATCGGCCAAGTAATTTCGTGAATGAC	GTAATGTCTCCACAAAGGACGAATCTTG
B	GCCTCACCAACTCGATCTCGGTCATTG	ATTCAACATCCTGTGCAGCTCGTCTCTG
C	CGACAATCCAATCGACCACACAAGACT	GAAAGCAAATCCATTCACGTCAACCAT
D	TCTGTCTCTCTGTGTTTATTAAAGGGTGAC	CCGAGGGAGTCAACGGTCACATTTGTT

Table 4-4. Summary of QTL detected for leaf lamina shape characters in *P. trichocarpa* X *P. deltoides* X *P. deltoides* Family 52-124.

Trait	Clonal repeatability (Std error)	QTL number	Linkage group	Flanking marker 1	Flanking marker 2	Origin of positive allele	LOD peak	Phenotypic variance explained
Lamina Length	.2186 (.0244)	1	IV	G1809	G3847	*P. deltoides*	3.14	6.31%
Lamina Width	.2407 (.0249)	1	VI	P2221	W12	*P. trichocarpa*	3.69	5.12%
		2	X	P2855	G2122	*P. deltoides*	4.59	5.99%
Lamina Length:Width Ratio	.3490 (.0276)	1	I	G124	G2903	*P. deltoides*	4.51	6.16%
		2	X	P2855	G2122	*P. trichocarpa*	12.77	14.20%
		3	XV	G1245	G1424	*P. deltoides*	5.73	5.12%
		4	XVII	G3580	G641	*P. trichocarpa*	3.45	5.00%
		5	XIX	O597	G2319	*P. deltoides*	3.69	4.00%

Table 4-5. Genes with significant phenotypic correlations to leaf lamina and eQTL regulated by the major lamina shape QTL locus. Significance was judged by a Bonferroni correction for 161 tests of expression-phenotypic correlation.

P. trichocarpa gene model	Cis/trans eQTL	Correlation lamina width	Correlation significance lamina width (*P*-value)	Ath ortholog	Ath ortholog annotation
estExt_Genewise1_v1.C_LG_X0744	cis	-0.3669	0.0001	AT1G10630	ATARFA1F; GTP binding / phospholipase activator/ protein binding
estExt_fgenesh4_pg.C_LG_X1324	cis	-0.2864	0.0032	AT5G49480	ATCP1 (CA2+-BINDING PROTEIN 1); calcium ion binding

Table 4-6. Gene models expressed above microarray background and encoded within the lamina shape QTL interval with statistically significant correlations to lamina width. Significance was judged at a modified FDR threshold $P < 6.1E-04$.

P. trichocarpa gene model	Correlation lamina width	Correlation *P*-value	Correlation rank lamina width	Cis/trans	Ath ortholog	Ath ortholog annotation
estExt_Genewise1_v1.C_LG_X0744	-0.3668554	7.4854E-05	49	cis	AT1G10630	ATARFA1F; GTP binding / phospholipase activator/ protein binding
fgenesh4_pm.C_LG_X000536	0.35563947	1.2789E-04	69	n/a	AT1G26230	chaperonin
grail3.0006040101	-0.3487108	1.7631E-04	90	n/a	AT3G25600	calmodulin

Table 4-7. Expression characteristics of ARF-type gene family members relative to *ARF1*.

P. trichocarpa gene model	Expression correlation to *ARF1*	Correlation significance (*P*-value)	Relative abundance[a] (median)	Expressed above microarray background?
estExt_Genewise1_v1.C_LG_I4117	0.39521569	1.764E-05	0.63332863	Y
estExt_Genewise1_v1.C_LG_V4094	0.33574725	3.153E-04	0.71948814	Y
estExt_Genewise1_v1.C_LG_XIX1392	0.28014634	2.901E-03	0.35397767	N
estExt_Genewise1_v1.C_LG_XVIII2916	0.2098178	2.709E-02	0.47177235	N
estExt_fgenesh4_kg.C_LG_XII0031	0.43947159	1.396E-06	0.61256805	Y
estExt_fgenesh4_kg.C_LG_XV0033	0.43443381	1.898E-06	0.35373527	N
estExt_fgenesh4_pg.C_LG_IX0748	0.37086319	6.152E-05	0.74631113	Y
estExt_fgenesh4_pg.C_LG_XVIII0541	0.26424056	5.072E-03	0.56174961	N
estExt_fgenesh4_pm.C_LG_VI0838	0.27363038	3.662E-03	0.44748339	N
estExt_fgenesh4_pm.C_LG_VII0412	0.431432	2.274E-06	1.3114264	Y
estExt_fgenesh4_pm.C_LG_VIII0406	0.30101811	1.326E-03	0.44116017	N
eugene3.00130048	0.51562433	6.988E-09	9.64183283	Y
eugene3.00131058	0.40448725	1.068E-05	0.9141763	Y
grail3.0016004802	0.2856963	2.369E-03	7.16681348	Y
grail3.0050012801	0.45146446	6.585E-07	4.38820514	Y
grail3.0061009602	0.45939209	3.944E-07	3.35093299	Y
gw1.I.3910.1	-0.1974934	3.774E-02	0.2810372	N
gw1.II.442.1	0.42146597	4.092E-06	0.99980107	Y
gw1.IV.3637.1	0.32899233	4.225E-04	0.95721298	Y
gw1.IX.2325.1	0.04655942	6.275E-01	0.28198022	N
estExt_Genewise1_v1.C_1310224	0.34693432	1.912E-04	0.29649079	N
estExt_fgenesh4_pm.C_1710005	0.02297545	8.108E-01	0.41922534	N
eugene3.12450001	-0.4747608	1.406E-07	0.33204579	N
grail3.0131005102	-0.1257081	1.886E-01	0.30736057	N
gw1.3608.1.1	0.48736645	5.806E-08	0.3344929	N

a. Relative abundance is the proportion of signal detected relative to *ARF1*.

CHAPTER 5
CONCLUSIONS

The overall goal of this project was to determine the role of genetic variation in shaping gene expression diversity in genus *Populus*, and to extend these findings to the level of morphological phenotypes to understand how transcriptional diversity affects evolutionarily conserved variation for leaf shape between the Tacamahaca and Aigeiros sections of the genus.

An important component of understanding the role of genetic variation on gene expression and phenotypes is the ability to accurately and densely map the genome of the species in question. In Chapter 2 of this study, I reported the first utilization of sequence-anchored, microarray-based markers for high density genetic map construction in an outcrossing species. The genetic map produced by this effort was important for the subsequent analysis of trait and expression QTL, while also resulting in the placement of nearly 50% of the whole-genome shotgun scaffold sequence that previously had no known location in the genome of *P. trichocarpa* [4]. This outcome paints a favorable picture for further utilization of high density genetic mapping to aid in draft assemblies of whole-genome shotgun sequences, even from highly heterozygous species. Similarly, a preliminary microarray-based genetic map for *Eucalyptus*, comprising nearly 10,000 loci, has been produced (L.G. Neves and M. Kirst, personal communication) and is likely to play an important role in the genome assembly currently in progress by the Department of Energy [214]. New approaches, including microarray-based [215] and in-solution sequence capture platforms [216] coupled to high-throughput sequencing [217, 218], will continue to refine genotyping in forest tree species and help to provide even more saturated SNP-based genetic maps for QTL and association-based experiments.

From genetic map data and whole-genome expression "phenotypes", a preliminary goal is to determine the overall ontogeny of the genetic regulation over gene expression. Early studies in model species (both plants and other eukaryotes) characterized *cis*- and *trans*-acting regulation on a genome wide scale and determined heritability and quantitative genetic effects in gene expression phenotypes [62, 84, 91]. Subsequently, it was found that eQTL studies may be useful to identify gene coexpression networks on account of key network regulators that segregate in these populations [183]. In Chapter 3 of this study, I utilized the high density genetic map together with whole-genome eQTL data to characterize the mechanism of tissue-specific genetic regulation of mRNA abundance and coexpression networks. In addition to generating a wealth of information regarding tissue-specific regulation of expression among individual loci and networked genes, two important and overarching conclusions could be reached. First, it is apparent that major loci controlling the abundance or turnover of large suites of transcripts are common in the *Populus* genetic system. Selective forces shaping the evolution of the modern poplar genome [4] may have driven these loci to accumulate on specific chromosomes (Figures 3-1 & 3-2). Secondly, major loci often control transcript levels in a tissue-specific manner, and common networks of genes that are regulated by a given locus in one tissue frequently remain co-regulated in other tissues, albeit by different major loci.

In addition to the ability to describe and define the ontogeny of genome-wide expression, eQTL studies, when allied to quantitative phenotypic data gathered from the same population, can facilitate the *a posteriori* identification of candidate genes regulating complex phenotypes. Co-localization of regulation of expression and phenotypic traits, together with a significant statistical relationships between phenotypic and gene expression measurements, can define a list of candidate genes for further investigation [78]. Frequently this list is orders of magnitude

smaller than a list that considers all genes in a tQTL interval to be equally likely as the phenotypic regulator [83, 89]. Chapter 4 describes preliminary evidence implicating an ADP-rybosylation factor GTPase, *ARF1*, as a candidate gene underlying a QTL for leaf width in the pseudobackcross pedigree. Among all genes whose regulation is controlled by the leaf width QTL interval, *ARF1* displays the most significant statistical relationship with the trait. Furthermore, allele-specific effects of gene expression are prominently displayed in qRT-PCR experiments, and the analysis identified a number of candidate *cis*-regulatory polymorphisms that may work independently or together to affect differential transcript abundance in *P. trichocarpa* and *P. deltoides*. Previous data have implicated the Arabidopsis *ARF1* ortholog in the endocytic pathway, where it influences the distribution of PIN auxin efflux carriers on the cell membrane surface. The predisposed role of *ARF1* in this pathway supports a generalized model wherein higher expression of *ARF1* from the *P. trichocarpa* background leads to stronger basal localization of auxin efflux carriers, driving cell expansion in the longitudinal plane of the leaf lamina. In *P. deltoides*, weaker *ARF1* expression leads to decreased basal localization of the auxin efflux machinery may result in increased lateral expansion of the leaf lamina.

Clearly, a number of testable hypotheses remain unaddressed in the scope of the current work. Of immediate importance is developing a more thorough understanding of the role that co-expression networks play in phenotypic variation in *Populus*. While it is apparent from the work outlined in Chapter 4 that individual genes can be implicated in vital phenotypic differences, past precedent indicates that differential regulation of biochemical pathways or transcriptional and signaling cascades is similarly, if not more important for complex phenotypic variation [97, 219-221]. An immediate step towards attainment of this goal in Family 52-124 is the alliance of previously collected phenotypic data [75] with the co-expression networks identified in Chapter

3. A preliminary connection between phenotypes and these networks can be established simply due to overlap of genomic intervals regulating phenotypic QTL and network eQTL. These comparisons indicate several cases of measured phenotypes exhibiting common genetic regulation with established *a posteriori* coexpression networks – particularly for cell wall constituent and biomass phenotypes with xylem networks (data not shown). Transgenic modulation of candidate network regulators could verify the role of these networks in the phenotypes in question. Of course, it is likely that more complex coexpression networks subject to multiple points of genetic regulation will influence some, if not most, of the measured traits. These would not have been clearly identified by the analysis described in Chapter 3, since it focused on known regions of pleiotropic transcriptional regulation. One case of a more complex regulatory network has already been identified in Family 52-124, wherein a regulatory locus encoding a putative S-adenosyl methionine synthase (SAMS) is co-expressed with several members of the lignin biosynthesis pathway. The eQTL for the putative SAMS co-localizes with a pleiotropic locus regulating wood chemistry and biomass phenotypes and SAMS expression correlates strongly with each phenotype (E. Novaes, M. Kirst; personal communication). Thus, while the initial results described in Chapter 3 give a favorable starting point to understand the role of transcriptional networks in phenotypic diversity, additional analysis may frequently be required to understand the specific transcriptional mechanisms underlying variation in individual phenotypic QTLs.

It will also be important to continue building an understanding of the role of genome evolution in shaping transcriptional diversity. It is clear from the results presented in Chapter 3 that the distribution of pleiotropic expression regulators is non-random and that specific chromosomes harbor these regulators more frequently than others. Equally obvious are the facts

that sequence polymorphisms in the poplar genome are also distributed non-randomly across chromosomes (Table 1-2), and certain chromosomes have been subjected to complex patterns of evolution in order to reach their modern structure [4]. For instance, chromosome XVII was shown to harbor one of the highest normalized frequencies of eQTL for leaf, root, and xylem tissues (Figure 3-2). In the poplar genome assembly, chromosome XVII is unique in that less than 60% of its predicted sequence has been contiguously assembled, suggesting it harbors an unusually high rate of polymorphism ([4], Table 2-5). In agreement with this suggestion, the chromosome possesses one of the highest frequencies of SNP per base pair in the Nisqually-1 genome (Table 1-2). Furthermore, chromosome XVII has a high frequency of scaffold sequence mapping to it (Table 2-5) and is one chromosome predicted to possess a more complex evolutionary history [4]. Scaffold sequence, while generally less polymorphic for SNP substitutions, is highly variable for insertion/deletion polymorphism both within and outside gene sequence relative to the genome as a whole (Table 1-2). Thus, chromosome XVII and other similar chromosomes could provide a unique "laboratory" in which to study the role of chromosome structure and sequence features in shaping interspecific diversity of gene expression. Particularly powerful studies to address these questions could be developed based on the expression data contained within this study after the genome sequences of several *P. trichocarpa* and *P. deltoides* clones become available. These sequences are currently being generated and are expected to reach the public realm within the next 2-3 years (G. Tuskan, personal communication).

A shortcoming of the work presented in Chapter 4 is the lack of conclusive functional verification of the role of *ARF1* in PIN localization and leaf shape determination. Overexpression and RNAi constructs for *ARF1* have been produced and plant transformation is currently

underway. While the evidence presented herein indicates that *ARF1* should be specialized relative to other ARFs for its function in leaf development in *Populus,* it is possible that radical perturbation of *ARF1* expression may result in compensatory expression of another *ARF* gene in leaf tissue. Accordingly, mutant constructs that repress the function of ARF1-interacting proteins may play a key role in the verification of function, as described recently in *Arabidopsis* [198]. These constructs have been generated by a standard site-directed mutagenesis protocol and are also currently being used in plant transformation. Analysis of each of these transgenic constructs should increase our understanding of whether *ARF1* affects leaf shape in *P. trichocarpa* and *P. deltoides*, and the mechanism by which the effect occurs.

A final and overarching insight that can be reached from this study relates to the role of tissue-specific regulation in shaping phenotypic diversity. There is little doubt as to the importance of transcriptional diversity, both in the realm of evolution between species and in the evolution of variations in form within an organism (i.e., tissues and organs). A commonly considered mechanism by which transcriptional diversity can effect its action is through expression of tissue-, organ-, or species-specific genes. However, we [35] and others [186] have noted that quite frequently, the contingent of genes expressed in a given species, organ, or tissue is not highly discriminatory of the phenotypic diversity between these levels. An alternate mechanism to drive diversification is the specialized regulation of genes or groups of genes common to all tissues, organs, or species. Indeed, observations supporting this mechanism have been made in higher plant species [185, 186, 222] and recently also in mammalian cell lines [207, 223]. The results outlined in Chapter 3, when viewed in light of these observations, tend to support the hypothesis that diversification of organs and tissues may be driven extensively by differential points of regulatory input in common suites of genes. The analysis we describe in

Chapter 3 found this pattern to be more prevalent for *trans*-acting regulatory loci than for *cis*-acting effects (mirrored by a recent outcome in mice [223]), but the resolution associated with eQTL mapping in our study makes it difficult to determine if the same *cis*-acting factors are at work across all tissues. High resolution association-based eQTL mapping in human cell lines indicated that even among closely related cell types, less than 30% of *cis*-acting regulatory polymorphisms could be accounted for by a common SNP variant [207]. It will be interesting to determine whether a similar pattern of transcriptional diversity is common for both *cis*- and *trans*-acting regulatory loci in poplar by employing association genetic populations that are currently being developed and propagated (G. Tuskan, M. Kirst; personal communication).

The importance of these hypotheses and outcomes must also be considered with regards to the role of *ARF1* in regulation of leaf shape. While the initial attempts to localize *ARF1* expression to specific "hotspots" of activity in expanding leaves of *P. trichocarpa* proved generally ineffective (Figure 4-5), it must be noted that leaf samples obtained in this manner still consisted of no less than four distinct cell/tissue types (spongy and palisade mesophyll, epidermis, vasculature) and may have been no more enriched for differential *ARF1* expression than a full leaf sample. The aspects of *ARF1* regulation that result in differential leaf shapes may be restricted to only a subset of these cell types and/or may function homogenously across the plane of the leaf lamina. Therefore, more specialized techniques such as laser capture microdissection or single-cell micropipette transcript isolation [224] may be required to determine the specific mechanism of *ARF1* regulation and action. Similarly, phenotypes (or apparent lack thereof) in experiments undertaken with transgenic plants expressing mutant or RNAi constructs from constitutive promoters should be interpreted with appropriate

consideration to the tissue- or cell-specific context in which *ARF1* and its interacting factors may normally be functioning *in vivo*.

LIST OF REFERENCES

1. FAO: **Global forest resources assessment 2000 - Main report**. In: *FAO Forestry paper 140*. 2000.

2. FAO: **Activities Related to Poplar and Willow Cultivation and Utilization, 2004 through 2007**. In: *Poplars, Willows, and People's Wellbeing: 23rd Session of the International Poplar Commission*. Rome: Food and Agriculture Organization of the United Nations; 2008.

3. Dickmann DI, Isebrands JG, Eckenwalder JE, Richardson JE: **Poplar Culture in North America**. Ottawa: National Research Council of Canada; 2001.

4. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A *et al*: **The genome of black cottonwood, Populus trichocarpa (Torr. & Gray)**. *Science* 2006, **313**(5793):1596-1604.

5. Taylor G: **Populus: arabidopsis for forestry. Do we need a model tree?** *Ann Bot (Lond)* 2002, **90**(6):681-689.

6. Tuskan GA, DiFazio SP, Teichmann T: **Poplar genomics is getting popular: the impact of the poplar genome project on tree research**. *Plant Biol (Stuttg)* 2004, **6**(1):2-4.

7. Wullschleger SD, Jansson S, Taylor G: **Genomics and forest biology: Populus emerges as the perennial favorite**. *Plant Cell* 2002, **14**(11):2651-2655.

8. Stettler R, Bradshaw HD, Heilman PE, Hinckley TM: **Biology of Populus and its implications for management and conservation**. Ottawa, Ontario, Canada: NRC Research Press; 1996.

9. Rottmann WH, Meilan R, Sheppard LA, Brunner AM, Skinner JS, Ma C, Cheng S, Jouanin L, Pilate G, Strauss SH: **Diverse effects of overexpression of LEAFY and PTLF, a poplar (Populus) homolog of LEAFY/FLORICAULA, in transgenic poplar and Arabidopsis**. *Plant J* 2000, **22**(3):235-245.

10. Hsu CY, Liu YX, Luthe DS, Yuceer C: **Poplar FT2 shortens the juvenile phase and promotes seasonal flowering**. *Plant Cell* 2006, **18**(8):1846-1861.

11. Bohlenius H, Huang T, Charbonnel-Campaa L, Brunner AM, Jansson S, Strauss SH, Nilsson O: **CO/FT regulatory module controls timing of flowering and seasonal growth cessation in trees**. *Science* 2006, **312**(5776):1040-1043.

12. Rae AM, Street NR, Rodriguez-Acosta M: **Populus Trees**. In: *Genome Mapping and Molecular Breeding in Plants: Forest Trees*. Edited by Kole C, vol. 7. Berlin: Springer-Verlag; 2007: 1-21.

13. Fillatti JJ, Sellmer J, Mccown B, Haissig B, Comai L: **Agrobacterium mediated transformation and regeneration of Populus**. *Molecular & General Genetics* 1987, **206**:192-199.

14. **Analysis of the genome sequence of the flowering plant Arabidopsis thaliana**. *Nature* 2000, **408**(6814):796-815.

15. Goff SA, Ricke D, Lan TH, Presting G, Wang RL, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H *et al*: **A draft sequence of the rice genome (Oryza sativa L. ssp japonica)**. *Science* 2002, **296**(5565):92-100.

16. Yu J, Hu SN, Wang J, Wong GKS, Li SG, Liu B, Deng YJ, Dai L, Zhou Y, Zhang XQ *et al*: **A draft sequence of the rice genome (Oryza sativa L. ssp indica)**. *Science* 2002, **296**(5565):79-92.

17. Bhalerao R, Keskitalo J, Sterky F, Erlandsson R, Bjorkbacka H, Birve SJ, Karlsson J, Gardestrom P, Gustafsson P, Lundeberg J *et al*: **Gene expression in autumn leaves**. *Plant Physiol* 2003, **131**(2):430-442.

18. Dejardin A, Leple JC, Lesage-Descauses MC, Costa G, Pilate G: **Expressed sequence tags from poplar wood tissues--a comparative analysis from multiple libraries**. *Plant Biol (Stuttg)* 2004, **6**(1):55-64.

19. Kohler A, Delaruelle C, Martin D, Encelot N, Martin F: **The poplar root transcriptome: analysis of 7000 expressed sequence tags**. *Febs Letters* 2003, **542**(1-3):37-41.

20. Lee H, Lee JS, Noh EW, Bae EK, Choi YI, Han MS: **Generation and analysis of expressed sequence tags from poplar (Populus alba x P-tremula var. glandulosa) suspension cells**. *Plant Science* 2005, **169**(6):1118-1124.

21. Nanjo T, Sakurai T, Totoki Y, Toyoda A, Nishiguchi M, Kado T, Igasaki T, Futamura N, Seki M, Sakaki Y *et al*: **Functional annotation of 19,841 Populus nigra full-length enriched cDNA clones**. *Bmc Genomics* 2007, **8**.

22. Ralph S, Oddy C, Cooper D, Yueh H, Jancsik S, Kolosova N, Philippe RN, Aeschliman D, White R, Huber D *et al*: **Genomics of hybrid poplar (Populus trichocarpax 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* 2006, **15**(5):1275-1297.

23. Ranjan P, Kao YY, Jiang HY, Joshi CP, Harding SA, Tsai CJ: **Suppression subtractive hybridization-mediated transcriptome analysis from multiple tissues of aspen (Populus tremuloides) altered in phenylpropanoid metabolism**. *Planta* 2004, **219**(4):694-704.

24. Sterky F, Bhalerao RR, Unneberg P, Segerman B, Nilsson P, Brunner AM, Charbonnel-Campaa L, Lindvall JJ, Tandre K, Strauss SH *et al*: **A Populus EST resource for plant functional genomics**. *Proc Natl Acad Sci U S A* 2004, **101**(38):13951-13956.

25. Sterky F, Regan S, Karlsson J, Hertzberg M, Rohde A, Holmberg A, Amini B, Bhalerao R, Larsson M, Villarroel R *et al*: **Gene discovery in the wood-forming tissues of poplar: analysis of 5, 692 expressed sequence tags**. *Proc Natl Acad Sci U S A* 1998, **95**(22):13330-13335.

26. Major IT, Constabel CP: **Molecular analysis of poplar defense against herbivory: comparison of wound- and insect elicitor-induced gene expression**. *New Phytol* 2006, **172**(4):617-635.

27. Miranda M, Ralph SG, Mellway R, White R, Heath MC, Bohlmann J, Constabel CP: **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* 2007, **20**(7):816-831.

28. Rinaldi C, Kohler A, Frey P, Duchaussoy F, Ningre N, Couloux A, Wincker P, Le Thiec D, Fluch S, Martin F *et al*: **Transcript profiling of poplar leaves upon infection with compatible and incompatible strains of the foliar rust Melampsora larici-populina**. *Plant Physiol* 2007, **144**(1):347-366.

29. Zhang Y, Zhang X, Chen Y, Wang Q, Wang M, Huang M: **Function and chromosomal localization of differentially expressed genes induced by Marssonina brunnea f. sp multigermtubi in Populus deltoides**. *Journal of Genetics and Genomics* 2007, **34**(7):641-648.

30. Bogeat-Triboulot MB, Brosche M, Renaut J, Jouve L, Le Thiec D, Fayyaz P, Vinocur B, Witters E, Laukens K, Teichmann T *et al*: **Gradual soil water depletion results in reversible changes of gene expression, protein profiles, ecophysiology, and growth performance in Populus euphratica, a poplar growing in arid regions**. *Plant Physiol* 2007, **143**(2):876-892.

31. Brosche M, Vinocur B, Alatalo ER, Lamminmaki A, Teichmann T, Ottow EA, Djilianov D, Afif D, Bogeat-Triboulot MB, Altman A *et al*: **Gene expression and metabolite profiling of Populus euphratica growing in the Negev desert**. *Genome Biol* 2005, **6**(12):R101.

32. Street NR, Skogstrom O, Sjodin A, Tucker J, Rodriguez-Acosta M, Nilsson P, Jansson S, Taylor G: **The genetics and genomics of the drought response in Populus**. *Plant J* 2006, **48**(3):321-341.

33. Taylor G, Street NR, Tricker PJ, Sjodin A, Graham L, Skogstrom O, Calfapietra C, Scarascia-Mugnozza G, Jansson S: **The transcriptome of Populus in elevated CO2**. *New Phytol* 2005, **167**(1):143-154.

34. Andersson-Gunneras S, Mellerowicz EJ, Love J, Segerman B, Ohmiya Y, Coutinho PM, Nilsson P, Henrissat B, Moritz T, Sundberg B: **Biosynthesis of cellulose-enriched tension wood in Populus: global analysis of transcripts and metabolites identifies biochemical and developmental regulators in secondary wall biosynthesis**. *Plant J* 2006, **45**(2):144-165.

35. Quesada T, Li Z, Dervinis C, Li Y, Bocock PN, Tuskan GA, Casella G, Davis JM, Kirst M: **Comparative analysis of the transcriptomes of Populus trichocarpa and Arabidopsis thaliana suggests extensive evolution of gene expression regulation in angiosperms**. *New Phytol* 2008, **180**(2):408-420.

36. Levine M, Tjian R: **Transcription regulation and animal diversity**. *Nature* 2003, **424**(6945):147-151.

37. Keim P, Paige KN, Whitham TG, Lark KG: **Genetic analysis of an interspecific hybrid swarm of Populus: occurrence of unidirectional introgression**. *Genetics* 1989, **123**(3):557-565.

38. Yin T, Huang M, Wang M, Zhu LH, Zeng ZB, Wu R: **Preliminary interspecific genetic maps of the populus genome constructed from RAPD markers**. *Genome* 2001, **44**(4):602-609.

39. Yin T, Zhang X, Huang M, Wang M, Zhuge Q, Tu S, Zhu LH, Wu R: **Molecular linkage maps of the Populus genome**. *Genome* 2002, **45**(3):541-555.

40. Yin TM, DiFazio SP, Gunter LE, Riemenschneider D, Tuskan GA: **Large-scale heterospecific segregation distortion in Populus revealed by a dense genetic map**. *Theor Appl Genet* 2004, **109**(3):451-463.

41. Gupta PK, Rustgi S: **Molecular markers from the transcribed/expressed region of the genome in higher plants**. *Funct Integr Genomics* 2004, **4**(3):139-162.

42. Kelleher CT, Chiu R, Shin H, Bosdet IE, Krzywinski MI, Fjell CD, Wilkin J, Yin T, DiFazio SP, Ali J *et al*: **A physical map of the highly heterozygous Populus genome: integration with the genome sequence and genetic map and analysis of haplotype variation**. *Plant J* 2007, **50**(6):1063-1078.

43. Tuskan GA, Gunter LE, Yang ZMK, Yin TM, Sewell MM, DiFazio SP: **Characterization of microsatellites revealed by genomic sequencing of Populus trichocarpa**. *Canadian Journal of Forest Research* 2004, **34**(1):85-93.

44. Dayanandan S, Rajora OP, Bawa KS: **Isolation and characterization of microsatellites in trembling aspen (*Populus tremuloides*)**. *Theor Appl Genet* 1998, **96**:950-956.

45. Rahman MH, Dayanandan S, Rajora OP: **Microsatellite DNA markers in Populus tremuloides**. *Genome* 2000, **43**(2):293-297.

46. Smulders MJM, Van Der Schoot J, Arens P, Vosman B: **Trinucleotide repeat microsatellite markers for black poplar (*Populus nigra* L.)**. *Molec Ecol Notes* 2001, **1**(3):188-190.

47. van der Schoot J, Pospiskova M, Vosman B, Smulders MJM: **Development and characterization of microsatellite markers in black poplar (Populus nigra L.)**. *Theoretical and Applied Genetics* 2000, **101**(1-2):317-322.

48. Rahman MH, Rajora OP: **Microsatellite DNA fingerprinting, differentiation, and genetic relationships of clones, cultivars, and varieties of six poplar species from three sections of the genus Populus**. *Genome* 2002, **45**(6):1083-1094.

49. Rajora P, Rahman H: **Microsatellite DNA and RAPD fingerprinting, identification and genetic relationships of hybrid poplar (Populus x canadensis) cultivars**. *Theor Appl Genet* 2003, **106**(3):470-477.

50. Unneberg P, Stromberg M, Lundeberg J, Jansson S, Sterky F: **Analysis of 70,000 EST sequences to study divergence between two closely related Populus species**. *Tree Genetics & Genomes* 2005, **1**(3):109-115.

51. Gilchrist EJ, Haughn GW, Ying CC, Otto SP, Zhuang J, Cheung D, Hamberger B, Aboutorabi F, Kalynyak T, Johnson L *et al*: **Use of Ecotilling as an efficient SNP discovery tool to survey genetic variation in wild populations of *Populus trichocarpa***. *Molecular Ecology* 2006, **15**:1367–1378.

52. Hall D, Luquez V, Garcia VM, St Onge KR, Jansson S, Ingvarsson PK: **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* 2007, **61**(12):2849-2860.

53. Ingvarsson PK: **Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (Populus tremula L., Salicaceae)**. *Genetics* 2005, **169**(2):945-953.

54. Ingvarsson PK: **Multilocus patterns of nucleotide polymorphism and the demographic history of Populus tremula**. *Genetics* 2008, **180**(1):329-340.

55. Ingvarsson PK, Garcia MV, Hall D, Luquez V, Jansson S: **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* 2006, **172**(3):1845-1853.

56. Heuertz M, De Paoli E, Kallman T, Larsson H, Jurman I, Morgante M, Lascoux M, Gyllenstrand N: **Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce [Picea abies (L.) Karst]**. *Genetics* 2006, **174**(4):2095-2105.

57. Krutovsky KV, Neale DB: **Nucleotide diversity and linkage disequilibrium in cold-hardiness- and wood quality-related candidate genes in Douglas fir**. *Genetics* 2005, **171**(4):2029-2041.

58. Ralph SG, Chun HJ, Cooper D, Kirkpatrick R, Kolosova N, Gunter L, Tuskan GA, Douglas CJ, Holt RA, Jones SJ *et al*: **Analysis of 4,664 high-quality sequence-finished poplar full-length cDNA clones and their utility for the discovery of genes responding to insect feeding**. *BMC Genomics* 2008, **9**:57.

59. Novaes E, Drost DR, Farmerie WG, Pappas GJ, Jr., Grattapaglia D, Sederoff RR, Kirst M: **High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome**. *BMC Genomics* 2008, **9**:312.

60. Borevitz JO, Liang D, Plouffe D, Chang HS, Zhu T, Weigel D, Berry CC, Winzeler E, Chory J: **Large-scale identification of single-feature polymorphisms in complex genomes**. *Genome Research* 2003, **13**(3):513-523.

61. Winzeler EA, Richards DR, Conway AR, Goldstein AL, Kalman S, McCullough MJ, McCusker JH, Stevens DA, Wodicka L, Lockhart DJ *et al*: **Direct allelic variation scanning of the yeast genome**. *Science* 1998, **281**(5380):1194-1197.

62. Brem RB, Yvert G, Clinton R, Kruglyak L: **Genetic dissection of transcriptional regulation in budding yeast**. *Science* 2002, **296**(5568):752-755.

63. West MAL, van Leeuwen H, Kozik A, Kliebenstein DJ, Doerge RW, St Clair DA, Michelmore RW: **High-density haplotyping with microarray-based expression and single feature polymorphism markers in Arabidopsis**. *Genome Research* 2006, **16**(6):787-795.

64. Singer T, Fan Y, Chang HS, Zhu T, Hazen SP, Briggs SP: **A high-resolution map of Arabidopsis recombinant inbred lines by whole-genome exon array hybridization**. *PLoS Genet* 2006, **2**(9):e144.

65. Bradshaw HD, Villar M, Watson BD, Otto KG, Stewart S, Stettler RF: **Molecular-Genetics of Growth and Development in Populus .3. a Genetic-Linkage Map of a Hybrid Poplar Composed of Rflp, Sts, and Rapd Markers**. *Theoretical and Applied Genetics* 1994, **89**(2-3):167-178.

66. Zeng ZB: **Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci**. *Proc Natl Acad Sci U S A* 1993, **90**(23):10972-10976.

67. Zeng ZB: **Precision mapping of quantitative trait loci**. *Genetics* 1994, **136**(4):1457-1468.

68. Bradshaw HD, Stettler RF: **Molecular-Genetics of Growth and Development in Populus .4. Mapping Qtls with Large Effects on Growth, Form, and Phenology Traits in a Forest Tree**. *Genetics* 1995, **139**(2):963-973.

69. Rae AM, Ferris R, Tallis MJ, Taylor G: **Elucidating genomic regions determining enhanced leaf growth and delayed senescence in elevated CO2**. *Plant Cell Environ* 2006, **29**(9):1730-1741.

70. Wu R, Bradshaw HD, Stettler RF: **Molecular genetics of growth and development in Populus (Salicaceae) .5. Mapping quantitative trait loci affecting leaf variation**. *American Journal of Botany* 1997, **84**(2):143-153.

71. Wu R, Stettler RF: **Quantitative Genetics of Growth and Development in Populus .1. a 3-Generation Comparison of Tree Architecture During the First 2 Years of Growth**. *Theoretical and Applied Genetics* 1994, **89**(7-8):1046-1054.

72. Wu RL: **Genetic mapping of QTLs affecting tree growth and architecture in Populus: implication for ideotype breeding**. *Theoretical and Applied Genetics* 1998, **96**(3-4):447-457.

73. Grattapaglia D, Sederoff R: **Genetic linkage maps of Eucalyptus grandis and Eucalyptus urophylla using a pseudo-testcross: mapping strategy and RAPD markers**. *Genetics* 1994, **137**(4):1121-1137.

74. Myburg AA, Griffin AR, Sederoff RR, Whetten RW: **Comparative genetic linkage maps of Eucalyptus grandis, Eucalyptus globulus and their F-1 hybrid based on a double pseudo-backcross mapping approach**. *Theoretical and Applied Genetics* 2003, **107**(6):1028-1042.

75. Novaes E, Osorio L, Drost DR, Miles BL, Boaventura-Novaes CR, Benedict C, Dervinis C, Yu Q, Sykes R, Davis M *et al*: **Quantitative genetic analysis of biomass and wood chemistry of Populus under different nitrogen levels**. *New Phytol* 2009.

76. Wullschleger S, Yin TM, DiFazio SP, Tschaplinski TJ, Gunter LE, Davis MF, Tuskan GA: **Phenotypic variation in growth and biomass distribution for two advanced-generation pedigrees of hybrid poplar**. *Can J Forest Res* 2005, **35**(8):1779-1789.

77. Yin TM, DiFazio SP, Gunter LE, Jawdy SS, Boerjan W, Tuskan GA: **Genetic and physical mapping of Melampsora rust resistance genes in Populus and characterization of linkage disequilibrium and flanking genomic sequence**. *New Phytologist* 2004, **164**(1):95-105.

78. Jansen RC, Nap JP: **Genetical genomics: the added value from segregation**. *Trends Genet* 2001, **17**(7):388-391.

79. King MC, Wilson AC: **Evolution at 2 levels in humans and chimpanzees**. *Science* 1975, **188**(4184):107-116.

80. Damerval C, Maurice A, Josse JM, de Vienne D: **Quantitative trait loci underlying gene product variation: a novel perspective for analyzing regulation of genome expression**. *Genetics* 1994, **137**(1):289-301.

81. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray**. *Science* 1995, **270**(5235):467-470.

82. Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR *et al*: **Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer**. *Nature Biotechnology* 2001, **19**(4):342-347.

83. Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G *et al*: **Genetics of gene expression surveyed in maize, mouse and man**. *Nature* 2003, **422**(6929):297-302.

84. Kirst M, Basten CJ, Myburg AA, Zeng ZB, Sederoff RR: **Genetic architecture of transcript-level variation in differentiating xylem of a eucalyptus hybrid**. *Genetics* 2005, **169**(4):2295-2303.

85. Kliebenstein DJ, West MA, van Leeuwen H, Kim K, Doerge RW, Michelmore RW, St Clair DA: **Genomic survey of gene expression diversity in Arabidopsis thaliana**. *Genetics* 2006, **172**(2):1179-1189.

86. Monks SA, Leonardson A, Zhu H, Cundiff P, Pietrusiak P, Edwards S, Phillips JW, Sachs A, Schadt EE: **Genetic inheritance of gene expression in human cell lines**. *American Journal of Human Genetics* 2004, **75**(6):1094-1105.

87. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG: **Genetic analysis of genome-wide variation in human gene expression**. *Nature* 2004, **430**(7001):743-747.

88. Rockman MV, Kruglyak L: **Genetics of global gene expression**. *Nature Reviews Genetics* 2006, **7**(11):862-872.

89. Wayne ML, McIntyre LM: **Combining mapping and arraying: An approach to candidate gene identification**. *Proc Natl Acad Sci USA* 2002, **99**(23):14903-14906.

90. Potokina E, Druka A, Luo Z, Wise R, Waugh R, Kearsey M: **Gene expression quantitative trait locus analysis of 16 000 barley genes reveals a complex pattern of genome-wide transcriptional regulation**. *Plant J* 2008, **53**(1):90-101.

91. West MA, Kim K, Kliebenstein DJ, van Leeuwen H, Michelmore RW, Doerge RW, St Clair DA: **Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis**. *Genetics* 2007, **175**(3):1441-1450.

92. Brem RB, Kruglyak L: **The landscape of genetic complexity across 5,700 gene expression traits in yeast**. *Proc Natl Acad Sci U S A* 2005, **102**(5):1572-1577.

93. Keurentjes JJ, Fu J, Terpstra IR, Garcia JM, van den Ackerveken G, Snoek LB, Peeters AJ, Vreugdenhil D, Koornneef M, Jansen RC: **Regulatory network construction in**

Arabidopsis by using genome-wide gene expression quantitative trait loci. *Proc Natl Acad Sci U S A* 2007, **104**(5):1708-1713.

94. Schadt EE, Monks SA, Friend SH: **A new paradigm for drug discovery: integrating clinical, genetic, genomic and molecular phenotype data to identify drug targets**. *Biochemical Society Transactions* 2003, **31**:437-443.

95. Eaves IA, Wicker LS, Ghandour G, Lyons PA, Peterson LB, Todd JA, Glynne RJ: **Combining mouse congenic strains and microarray gene expression analyses to study a complex trait: the NOD model of type 1 diabetes**. *Genome Res* 2002, **12**(2):232-243.

96. Karp CL, Grupe A, Schadt E, Ewart SL, Keane-Moore M, Cuomo PJ, Kohl J, Wahl L, Kuperman D, Germer S *et al*: **Identification of complement factor 5 as a susceptibility locus for experimental allergic asthma**. *Nature Immunol* 2000, **1**(3):221-226.

97. Wentzell AM, Rowe HC, Hansen BG, Ticconi C, Halkier BA, Kliebenstein DJ: **Linking metabolic QTLs with network and cis-eQTLs controlling biosynthetic pathways**. *PLoS Genet* 2007, **3**(9):1687-1701.

98. Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, Mackelprang R, Kruglyak L: **Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors**. *Nat Genet* 2003, **35**(1):57-64.

99. DeRisi JL, Iyer VR, Brown PO: **Exploring the metabolic and genetic control of gene expression on a genomic scale**. *Science* 1997, **278**(5338):680-686.

100. Kirst M, Myburg AA, De Leon JP, Kirst ME, Scott J, Sederoff R: **Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cDNA microarray data in an interspecific backcross of eucalyptus**. *Plant Physiol* 2004, **135**(4):2368-2378.

101. Wentzell AM, Boeye I, Zhang Z, Kliebenstein DJ: **Genetic networks controlling structural outcome of glucosinolate activation across development**. *PLoS Genet* 2008, **4**(10):e1000234.

102. Zhang B, Horvath S: **A general framework for weighted gene co-expression network analysis**. *Statistical Applications in Genetics and Molecular Biology* 2005, **4**.

103. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I *et al*: **Transcriptional regulatory networks in Saccharomyces cerevisiae**. *Science* 2002, **298**(5594):799-804.

104. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L: **Integrated genomic and proteomic analyses of a systematically perturbed metabolic network**. *Science* 2001, **292**(5518):929-934.

105. Bing N, Hoeschele I: **Genetical genomics analysis of a yeast segregant population for transcription network inference**. *Genetics* 2005, **170**(2):533-542.

106. Wu C, Delano DL, Mitro N, Su SV, Janes J, McClurg P, Batalov S, Welch GL, Zhang J, Orth AP *et al*: **Gene set enrichment in eQTL data identifies novel annotations and pathway regulators**. *PLoS Genet* 2008, **4**(5):e1000070.

107. Zhu J, Lum PY, Lamb J, GuhaThakurta D, Edwards SW, Thieringer R, Berger JP, Wu MS, Thompson J, Sachs AB *et al*: **An integrative genomics approach to the reconstruction of gene networks in segregating populations**. *Cytogenet Genome Res* 2004, **105**(2-4):363-374.

108. Chaibub Neto E, Ferrara CT, Attie AD, Yandell BS: **Inferring causal phenotype networks from segregating populations**. *Genetics* 2008, **179**(2):1089-1100.

109. Dobrin R, Zhu J, Molony C, Argman C, Parrish ML, Carlson S, Allan MF, Pomp D, Schadt EE: **Multi-tissue coexpression networks reveal unexpected subnetworks associated with disease**. *Genome Biol* 2009, **10**(5):R55.

110. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S *et al*: **Genetics of gene expression and its effect on disease**. *Nature* 2008, **452**(7186):423-428.

111. Hood L, Heath JR, Phelps ME, Lin B: **Systems biology and new technologies enable predictive and preventative medicine**. *Science* 2004, **306**(5696):640-643.

112. Mehrabian M, Allayee H, Stockton J, Lum PY, Drake TA, Castellani LW, Suh M, Armour C, Edwards S, Lamb J *et al*: **Integrating genotypic and expression data in a segregating mouse population to identify 5-lipoxygenase as a susceptibility gene for obesity and bone traits**. *Nat Genet* 2005, **37**(11):1224-1233.

113. Shlomi T, Cabili MN, Herrgard MJ, Palsson BO, Ruppin E: **Network-based prediction of human tissue-specific metabolism**. *Nat Biotechnol* 2008, **26**(9):1003-1010.

114. Fu J, Keurentjes JJ, Bouwmeester H, America T, Verstappen FW, Ward JL, Beale MH, de Vos RC, Dijkstra M, Scheltema RA *et al*: **System-wide molecular evidence for phenotypic buffering in Arabidopsis**. *Nat Genet* 2009, **41**(2):166-167.

115. Zhu J, Zhang B, Smith EN, Drees B, Brem RB, Kruglyak L, Bumgarner RE, Schadt EE: **Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks**. *Nat Genet* 2008, **40**(7):854-861.

116. Eckenwalder JE: **Systematics and evolution of Populus**. In: *Biology of Populus and its implications for management and conservation*. Edited by Stettler RF, Bradshaw HD, Heilman PE, Hinckley TM. Ottawa, Canada: NRC Research Press; 1996: 7-32.

117. Cronk QC: **Plant eco-devo: the potential of poplar as a model organism**. *New Phytol* 2005, **166**(1):39-48.

118. Collinson ME: **The early fossil history of Salicaceae: a brief review**. *Proc Roy Soc Edinburgh* 1992, **98B**:155–167.

119. Harrington CA, Radwan MA, DeBell DS: **Leaf characteristics reflect growth rates of 2-year-old Populus trees**. *Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere* 1997, **27**(8):1321-1325.

120. Marron N, Villar M, Dreyer E, Delay D, Boudouresque E, Petit JM, Delmotte FM, Guehl JM, Brignolas F: **Diversity of leaf traits related to productivity in 31 Populus deltoides x Populus nigra clones**. *Tree Physiol* 2005, **25**(4):425-435.

121. Marron N, Ceulemans R: **Genetic variation of leaf traits related to productivity in a Populus deltoides x Populus nigra family**. *Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere* 2006, **36**(2):390-400.

122. Rae AM, Tricker PJ, Bunn SM, Taylor G: **Adaptation of tree growth to elevated CO2: quantitative trait loci for biomass in Populus**. *New Phytol* 2007, **175**(1):59-69.

123. Street NR, Sjodin A, Bylesjo M, Gustafsson P, Trygg J, Jansson S: **A cross-species transcriptomics approach to identify genes involved in leaf development**. *BMC Genomics* 2008, **9**:589.

124. Ferris R, Long L, Bunn SM, Robinson KM, Bradshaw HD, Rae AM, Taylor G: **Leaf stomatal and epidermal cell development: identification of putative quantitative trait loci in relation to elevated carbon dioxide concentration in poplar**. *Tree Physiology* 2002, **22**(9):633-640.

125. Yin T, Difazio SP, Gunter LE, Zhang X, Sewell MM, Woolbright SA, Allan GJ, Kelleher CT, Douglas CJ, Wang M *et al*: **Genome structure and emerging evidence of an incipient sex chromosome in Populus**. *Genome Res* 2008, **18**(3):422-430.

126. Morreel K, Goeminne G, Storme V, Sterck L, Ralph J, Coppieters W, Breyne P, Steenackers M, Georges M, Messens E *et al*: **Genetical metabolomics of flavonoid biosynthesis in Populus: a case study**. *Plant J* 2006, **47**(2):224-237.

127. Rae AM, Pinel MPC, Bastien C, Sabatti M, Street NR, Tucker J, Dixon C, Marron N, Dillen SY, Taylor G: **QTL for yield in bioenergy Populus: identifying GxE interactions from growth at three contrasting sites**. *Tree Genetics & Genomes* 2008, **4**(1):97-112.

128. Mackay TF: **The genetic architecture of quantitative traits**. *Annu Rev Genet* 2001, **35**:303-339.

129. Tsukaya H: **Mechanism of leaf-shape determination**. *Annu Rev Plant Biol* 2006, **57**:477-496.

130. Donnelly PM, Bonetta D, Tsukaya H, Dengler RE, Dengler NG: **Cell cycling and cell enlargement in developing leaves of Arabidopsis**. *Dev Biol* 1999, **215**(2):407-419.

131. Tsuge T, Tsukaya H, Uchimiya H: **Two independent and polarized processes of cell elongation regulate leaf blade expansion in Arabidopsis thaliana (L.) Heynh**. *Development* 1996, **122**(5):1589-1600.

132. Kim GT, Tsukaya H, Uchimiya H: **The ROTUNDIFOLIA3 gene of Arabidopsis thaliana encodes a new member of the cytochrome P-450 family that is required for the regulated polar elongation of leaf cells**. *Genes Dev* 1998, **12**(15):2381-2391.

133. Narita NN, Moore S, Horiguchi G, Kubo M, Demura T, Fukuda H, Goodrich J, Tsukaya H: **Overexpression of a novel small peptide ROTUNDIFOLIA4 decreases cell proliferation and alters leaf shape in Arabidopsis thaliana**. *Plant J* 2004, **38**(4):699-713.

134. Kim GT, Shoda K, Tsuge T, Cho KH, Uchimiya H, Yokoyama R, Nishitani K, Tsukaya H: **The ANGUSTIFOLIA gene of Arabidopsis, a plant CtBP gene, regulates leaf-cell expansion, the arrangement of cortical microtubules in leaf cells and expression of a gene involved in cell-wall formation**. *Embo J* 2002, **21**(6):1267-1279.

135. Corda D, Colanzi A, Luini A: **The multiple activities of CtBP/BARS proteins: the Golgi view**. *Trends Cell Biol* 2006, **16**(3):167-173.

136. Chinnadurai G: **CtBP, an unconventional transcriptional corepressor in development and oncogenesis**. *Mol Cell* 2002, **9**(2):213-224.

137. Stern MD, Aihara H, Cho KH, Kim GT, Horiguchi G, Roccaro GA, Guevara E, Sun HH, Negeri D, Tsukaya H *et al*: **Structurally related Arabidopsis ANGUSTIFOLIA is functionally distinct from the transcriptional corepressor CtBP**. *Dev Genes Evol* 2007, **217**(11-12):759-769.

138. Horiguchi G, Kim GT, Tsukaya H: **The transcription factor AtGRF5 and the transcription coactivator AN3 regulate cell proliferation in leaf primordia of Arabidopsis thaliana**. *Plant J* 2005, **43**(1):68-78.

139. Kumar R, Qiu J, Joshi T, Valliyodan B, Xu D, Nguyen HT: **Single feature polymorphism discovery in rice**. *PLoS ONE* 2007, **2**(3):e284.

140. Ronald J, Akey JM, Whittle J, Smith EN, Yvert G, Kruglyak L: **Simultaneous genotyping, gene-expression measurement, and detection of allele-specific expression with oligonucleotide arrays**. *Genome Res* 2005, **15**(2):284-291.

141. Coram TE, Settles ML, Wang M, Chen X: **Surveying expression level polymorphism and single-feature polymorphism in near-isogenic wheat lines differing for the Yr5 stripe rust resistance locus**. *Theor Appl Genet* 2008, **117**(3):401-411.

142. Cui X, Xu J, Asghar R, Condamine P, Svensson JT, Wanamaker S, Stein N, Roose M, Close TJ: **Detecting single-feature polymorphisms using oligonucleotide arrays and robustified projection pursuit**. *Bioinformatics* 2005, **21**(20):3852-3858.

143. Luo ZW, Potokina E, Druka A, Wise R, Waugh R, Kearsey MJ: **SFP genotyping from affymetrix arrays is robust but largely detects cis-acting expression regulators**. *Genetics* 2007, **176**(2):789-800.

144. Rostoks N, Borevitz JO, Hedley PE, Russell J, Mudie S, Morris J, Cardle L, Marshall DF, Waugh R: **Single-feature polymorphism discovery in the barley transcriptome**. *Genome Biol* 2005, **6**(6).

145. West MA, van Leeuwen H, Kozik A, Kliebenstein DJ, Doerge RW, St Clair DA, Michelmore RW: **High-density haplotyping with microarray-based expression and single feature polymorphism markers in Arabidopsis**. *Genome Res* 2006, **16**(6):787-795.

146. Kirst M, Caldo R, Casati P, Tanimoto G, Walbot V, Wise RP, Buckler ES: **Genetic diversity contribution to errors in short oligonucleotide microarray analysis**. *Plant Biotechnol J* 2006, **4**(5):489-498.

147. Cooke JEK, Brown KA, Wu R, Davis JM: **Gene expression associated with N-induced shifts in resource allocation in poplar**. *Plant Cell and Environment* 2003, **26**(5):757-770.

148. Chang S, Puryear J, Cairney J: **A simple and efficient method for isolating RNA from pine trees**. *Plant Mol Biol Rep* 1993, **11**:117-121.

149. Tuskan GA, Gunter LE, Yang ZMK, Yin TM, Sewell MM, DiFazio SP: **Characterization of microsatellites revealed by genomic sequencing of Populus trichocarpa**. *Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere* 2004, **34**(1):85-93.

150. Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newburg L: **MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations**. *Genomics* 1987, **1**(2):174-181.

151. Ma CX, Yu Q, Berg A, Drost D, Novaes E, Fu G, Yap JS, Tan A, Kirst M, Cui Y *et al*: **A statistical model for testing the pleiotropic control of phenotypic plasticity for a count trait**. *Genetics* 2008, **179**(1):627-636.

152. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.

153. Bassam BJ, Caetanoanolles G, Gresshoff PM: **Fast and Sensitive Silver Staining of DNA in Polyacrylamide Gels**. *Analytical Biochemistry* 1991, **196**(1):80-83.

154. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 2003, **19**(2):185-193.

155. Storey JD, Tibshirani R: **Statistical significance for genomewide studies**. *Proc Natl Acad Sci U S A* 2003, **100**(16):9440-9445.

156. Haldane JB, Waddington CH: **Inbreeding and Linkage**. *Genetics* 1931, **16**(4):357-374.

157. Zeileis A, F. Leish, K. Hornik, and C. Kleiber: **Strucchange: An R package for testing structural change in linear regression models**. *J Statistical Software* 2002, **7**:1-38.

158. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R *et al*: **Clustal W and Clustal X version 2.0**. *Bioinformatics* 2007, **23**(21):2947-2948.

159. Cervera MT, Storme V, Ivens B, Gusmao J, Liu BH, Hostyn V, Van Slycken J, Van Montagu M, Boerjan W: **Dense genetic linkage maps of three Populus species (Populus deltoides, P. nigra and P. trichocarpa) based on AFLP and microsatellite markers**. *Genetics* 2001, **158**(2):787-809.

160. Das S, Bhat PR, Sudhakar C, Ehlers JD, Wanamaker S, Roberts PA, Cui X, Close TJ: **Detection and validation of single feature polymorphisms in cowpea (Vigna unguiculata L. Walp) using a soybean genome array**. *BMC Genomics* 2008, **9**:107.

161. Rennie C, Noyes HA, Kemp SJ, Hulme H, Brass A, Hoyle DC: **Strong position-dependent effects of sequence mismatches on signal ratios measured using long oligonucleotide microarrays**. *BMC Genomics* 2008, **9**(1):317.

162. Ching A, Caldwell KS, Jung M, Dolan M, Smith OS, Tingey S, Morgante M, Rafalski AJ: **SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines**. *BMC Genet* 2002, **3**:19.

163. Kolkman JM, Berry ST, Leon AJ, Slabaugh MB, Tang S, Gao W, Shintani DK, Burke JM, Knapp SJ: **Single nucleotide polymorphisms and linkage disequilibrium in sunflower**. *Genetics* 2007, **177**(1):457-468.

164. Tenaillon MI, Sawkins MC, Anderson LK, Stack SM, Doebley J, Gaut BS: **Patterns of diversity and recombination along chromosome 1 of maize (Zea mays ssp. mays L.)**. *Genetics* 2002, **162**(3):1401-1413.

165. Voorrips RE: **MapChart: software for the graphical presentation of linkage maps and QTLs**. *J Hered* 2002, **93**(1):77-78.

166. Vuylsteke M, van Eeuwijk F, Van Hummelen P, Kuiper M, Zabeau M: **Genetic analysis of variation in gene expression in Arabidopsis thaliana**. *Genetics* 2005, **171**(3):1267-1275.

167. Breitling R, Li Y, Tesson BM, Fu J, Wu C, Wiltshire T, Gerrits A, Bystrykh LV, de Haan G, Su AI *et al*: **Genetical genomics: spotlight on QTL hotspots**. *PLoS Genet* 2008, **4**(10):e1000232.

168. Sonderby IE, Hansen BG, Bjarnholt N, Ticconi C, Halkier BA, Kliebenstein DJ: **A systems biology approach identifies a R2R3 MYB gene subfamily with distinct and overlapping functions in regulation of aliphatic glucosinolates**. *PLoS ONE* 2007, **2**(12):e1322.

169. Kliebenstein DJ, West MA, van Leeuwen H, Loudet O, Doerge RW, St Clair DA: **Identification of QTLs controlling gene expression networks defined a priori**. *BMC Bioinformatics* 2006, **7**:308.

170. DeCook R, Lall S, Nettleton D, Howell SH: **Genetic regulation of gene expression during shoot development in Arabidopsis**. *Genetics* 2006, **172**(2):1155-1164.

171. Chesler EJ, Lu L, Shou S, Qu Y, Gu J, Wang J, Hsu HC, Mountz JD, Baldwin NE, Langston MA *et al*: **Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function**. *Nat Genet* 2005, **37**(3):233-242.

172. Petretto E, Mangion J, Dickens NJ, Cook SA, Kumaran MK, Lu H, Fischer J, Maatz H, Kren V, Pravenec M *et al*: **Heritability and tissue specificity of expression quantitative trait loci**. *PLoS Genet* 2006, **2**(10):e172.

173. Drost DR, Novaes E, Boaventura-Novaes C, Benedict CI, Brown RS, Yin T, Tuskan GA, Kirst M: **A microarray-based genotyping and genetic mapping approach for highly heterozygous outcrossing species enables localization of a large fraction of the unassembled Populus trichocarpa genome sequence**. *Plant J* 2009.

174. Wang S. CJB, and Z.-B. Zeng **Windows QTL Cartographer 2.5**. In. Raleigh, NC: Department of Statistics, North Carolina State University; 2007: http://statgen.ncsu.edu/qtlcart/WQTLCart.htm.

175. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.

176. Higo K, Ugawa Y, Iwamoto M, Korenaga T: **Plant cis-acting regulatory DNA elements (PLACE) database: 1999**. *Nucleic Acids Res* 1999, **27**(1):297-300.

177. Yan T, Yoo D, Berardini TZ, Mueller LA, Weems DC, Weng S, Cherry JM, Rhee SY: **PatMatch: a program for finding patterns in peptide and nucleotide sequences**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W262-266.

178. McAndrew RS, Froehlich JE, Vitha S, Stokes KD, Osteryoung KW: **Colocalization of plastid division proteins in the chloroplast stromal compartment establishes a new functional relationship between FtsZ1 and FtsZ2 in higher plants**. *Plant Physiol* 2001, **127**(4):1656-1666.

179. Stokes KD, McAndrew RS, Figueroa R, Vitha S, Osteryoung KW: **Chloroplast division and morphology are differentially affected by overexpression of FtsZ1 and FtsZ2 genes in Arabidopsis**. *Plant Physiol* 2000, **124**(4):1668-1677.

180. Sun C, Palmqvist S, Olsson H, Boren M, Ahlandsberg S, Jansson C: **A novel WRKY transcription factor, SUSIBA2, participates in sugar signaling in barley by binding to the sugar-responsive elements of the iso1 promoter**. *Plant Cell* 2003, **15**(9):2076-2092.

181. Muller R, Morant M, Jarmer H, Nilsson L, Nielsen TH: **Genome-wide analysis of the Arabidopsis leaf transcriptome reveals interaction of phosphate and sugar metabolism**. *Plant Physiol* 2007, **143**(1):156-171.

182. Koch KE: **Carbohydrate-Modulated Gene Expression in Plants**. *Annu Rev Plant Physiol Plant Mol Biol* 1996, **47**:509-540.

183. Jansen RC, Tesson BM, Fu J, Yang Y, McIntyre LM: **Defining gene and QTL networks**. *Curr Opin Plant Biol* 2009.

184. Blanc G, Wolfe KH: **Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution**. *Plant Cell* 2004, **16**(7):1679-1691.

185. Potokina E, Druka A, Luo Z, Moscou M, Wise R, Waugh R, Kearsey M: **Tissue-dependent limited pleiotropy affects gene expression in barley**. *Plant J* 2008, **56**(2):287-296.

186. Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Scholkopf B, Weigel D, Lohmann JU: **A gene expression map of Arabidopsis thaliana development**. *Nat Genet* 2005, **37**(5):501-506.

187. Byrne ME: **Networks in leaf development**. *Curr Opin Plant Biol* 2005, **8**(1):59-66.

188. Barkoulas M, Hay A, Kougioumoutzi E, Tsiantis M: **A developmental framework for dissected leaf formation in the Arabidopsis relative Cardamine hirsuta**. *Nat Genet* 2008, **40**(9):1136-1141.

189. Hay A, Tsiantis M: **The genetic basis for differences in leaf form between Arabidopsis thaliana and its wild relative Cardamine hirsuta**. *Nat Genet* 2006, **38**(8):942-947.

190. Holtan HE, Hake S: **Quantitative trait locus analysis of leaf dissection in tomato using Lycopersicon pennellii segmental introgression lines**. *Genetics* 2003, **165**(3):1541-1550.

191. Gailing O: **QTL analysis of leaf morphological characters in a Quercus robur full-sib family (Q. robur x Q. robur ssp. slavonica)**. *Plant Biol (Stuttg)* 2008, **10**(5):624-634.

192. Mickelson SM, Stuber CS, Senior L, Kaeppler SM: **Quantitative trait loci controlling leaf and tassel traits in a B73 x MO17 population of maize**. *Crop Science* 2002, **42**(6):1902-1909.

193. Juenger T, Perez-Perez JM, Bernal S, Micol JL: **Quantitative trait loci mapping of floral and leaf morphology traits in Arabidopsis thaliana: evidence for modular genetic architecture**. *Evol Dev* 2005, **7**(3):259-271.

194. Kearsey M: **QTL analysis: problems and (possible) solutions**. In: *Quantitative Genetics, Genomics and Plant Breeding*. Edited by Kang MS. Wallingford, U.K.: CABI Publishing; 2002: 45–58.

195. Hansen BG, Halkier BA, Kliebenstein DJ: **Identifying the molecular basis of QTLs: eQTLs add a new dimension**. *Trends Plant Sci* 2008, **13**(2):72-77.

196. Tuskan GA, DiFazio, S.P., Teichmann, T.: **Poplar Genomics is Getting Popular: The Impact of the Poplar Genome Project on Tree Research**. *Plant Biology* 2004, **6**:2-4.

197. Marron N, Villar M, Dreyer E, Delay D, Boudouresque E, Petit JM, Delmotte FM, Guehl JM, Brignolas F: **Diversity of leaf traits related to productivity in 31 Populus deltoides x Populus nigra clones**. *Tree Physiology* 2005, **25**(4):425-435.

198. Xu J, Scheres B: **Dissection of Arabidopsis ADP-RIBOSYLATION FACTOR 1 function in epidermal cell polarity**. *Plant Cell* 2005, **17**(2):525-536.

199. Wang SC, Basten CJ, Zeng ZB: **Windows QTL Cartographer 2.5.**. Raleigh, NC.: Department of Statistics, North Carolina State University.; 2007.

200. Dascher C, Balch WE: **Dominant inhibitory mutants of ARF1 block endoplasmic reticulum to Golgi transport and trigger disassembly of the Golgi apparatus**. *J Biol Chem* 1994, **269**(2):1437-1448.

201. Pepperkok R, Whitney JA, Gomez M, Kreis TE: **COPI vesicles accumulating in the presence of a GTP restricted arf1 mutant are depleted of anterograde and retrograde cargo**. *J Cell Sci* 2000, **113 (Pt 1)**:135-144.

202. Peirson SN, Butler JN, Foster RG: **Experimental validation of novel and conventional approaches to quantitative real-time PCR data analysis**. *Nucleic Acids Res* 2003, **3**(e73).

203. Vandesompele J, De Preter K, Pattyn F, Poppe B, Van R, De Paepe A, Speleman F: **Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes**. *Genome Biol* 2002, **3**:0034.

204. Jang HJ, Pih KT, Kang SG, Lim JH, Jin JB, Piao HL, Hwang I: **Molecular cloning of a novel Ca2+-binding protein that is induced by NaCl stress**. *Plant Mol Biol* 1998, **37**(5):839-847.

205. Click ES, Stearns T, Botstein D: **Systematic structure-function analysis of the small GTPase Arf1 in yeast**. *Mol Biol Cell* 2002, **13**(5):1652-1664.

206. Clark RM, Wagler TN, Quijada P, Doebley J: **A distant upstream enhancer at the maize domestication gene tb1 has pleiotropic effects on plant and inflorescent architecture**. *Nat Genet* 2006, **38**(5):594-597.

207. Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H, Ingle C, Beazley C, Gutierrez Arcelus M, Sekowska M *et al*: **Common regulatory variation impacts gene expression in a cell type-dependent manner**. *Science* 2009, **325**(5945):1246-1250.

208. Dhonukshe P, Tanaka H, Goh T, Ebine K, Mahonen AP, Prasad K, Blilou I, Geldner N, Xu J, Uemura T *et al*: **Generation of cell polarity in plants links endocytosis, auxin distribution and cell fate decisions**. *Nature* 2008, **456**(7224):962-966.

209. Boutte Y, Ikeda Y, Grebe M: **Mechanisms of auxin-dependent cell and tissue polarity**. *Curr Opin Plant Biol* 2007, **10**(6):616-623.

210. Grebe M, Friml J, Swarup R, Ljung K, Sandberg G, Terlou M, Palme K, Bennett MJ, Scheres B: **Cell polarity signaling in Arabidopsis involves a BFA-sensitive auxin influx pathway**. *Curr Biol* 2002, **12**(4):329-334.

211. Reinhardt D, Pesce ER, Stieger P, Mandel T, Baltensperger K, Bennett M, Traas J, Friml J, Kuhlemeier C: **Regulation of phyllotaxis by polar auxin transport**. *Nature* 2003, **426**(6964):255-260.

212. Scanlon MJ: **The polar auxin transport inhibitor N-1-naphthylphthalamic acid disrupts leaf initiation, KNOX protein regulation, and formation of leaf margins in maize**. *Plant Physiol* 2003, **133**(2):597-605.

213. Scarpella E, Marcos D, Friml J, Berleth T: **Control of leaf vascular patterning by polar auxin transport**. *Genes Dev* 2006, **20**(8):1015-1027.

214. Myburg A, Grattapaglia D, Tuskan G, Schmutz J, Barry K, Bristow JD, Network TEG: **Sequencing the Eucalyptus Genome: Genomic Resources for Renewable Energy and Fiber Production; W195**. In: *International Plant and Animal Genome Conference XVI: 2008; San Diego, CA*; 2008.

215. Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME: **Microarray-based genomic selection for high-throughput resequencing**. *Nature Methods* 2007, **4**(11):907-909.

216. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C *et al*: **Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing**. *Nat Biotechnol* 2009, **27**(2):182-189.

217. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B *et al*: **Real-Time DNA Sequencing from Single Polymerase Molecules**. *Science* 2009, **323**(5910):133-138.

218. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z *et al*: **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature* 2005, **437**(7057):376-380.

219. Chesler EJ, Lu L, Shou SM, Qu YH, Gu J, Wang JT, Hsu HC, Mountz JD, Baldwin NE, Langston MA *et al*: **Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function**. *Nature Genetics* 2005, **37**(3):233-242.

220. de Haan G, Bystrykh LV, Weersing E, Sutton S, Manly K, Chesler E, Jansen RC, Vellenga E, Williams RW, Cooke MP: **Genetical genomics to identify gene networks in hematopoietic stem cells**. *Experimental Hematology* 2005, **33**(7):60-60.

221. de Haan G, Weersing E, Dontje B, Sutton S, Vellenga E, Manly KF, Williams RW, Cooke M, Bystrykh LV: **Identification of transcriptional networks affecting hematopoietic stem cell function using "genetical genomics"**. *Blood* 2003, **102**(11):343A-343A.

222. Segerman B, Jansson S, Karlsson J: **Characterization of genes with tissue-specific differential expression patterns in Populus**. *Tree Genetics & Genomes* 2007, **3**(4):351-362.

223. Gerrits A, Li Y, Tesson BM, Bystrykh LV, Weersing E, Ausema A, Dontje B, Wang X, Breitling R, Jansen RC *et al*: **Expression quantitative trait loci are highly sensitive to cellular differentiation state**. *PLoS Genet* 2009, **5**(10):e1000692.

224. Lee JY, Levesque M, Benfey PN: **High-throughput RNA isolation technologies. New tools for high-resolution gene expression profiling in plant systems**. *Plant Physiol* 2005, **138**(2):585-590.

BIOGRAPHICAL SKETCH

Derek R. Drost was born in Madison, Wisconsin, in 1983. Upon graduating from high school in 2001, he attended the University of Nebraska-Lincoln, where he obtained a Bachelor of Science degree in agricultural sciences and majored in agronomy. During that time, Derek conducted research to determine the molecular genetics of resistance to acetolactate synthase inhibitor herbicides among invasive *Sorghum bicolor* populations. In 2005, Derek joined the Plant Molecular and Cellular Biology graduate program at the University of Florida, where he researched the quantitative genetics of transcription in *Populus*, under the direction of Drs. Matias Kirst and Gary Peter. Derek completed his Ph.D. in December of 2009 and is currently employed by the Monsanto Corporation, where he contributes to a research program aimed at discovering genes influencing biotic stress tolerance in maize.