Discovery of Cell Wall and Cuticle Gene Candidates for Biofuel Feedstock Improvement in *Populus trichocarpa* (Black Cottonwood) Using Systems Biology Approaches

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Anna Furches
August 2025

Copyright © 2025 by Anna Furches
All rights reserved.

For Leo and Orion,
my constellations

Acknowledgements

I would like to thank Leo and Ry for coming into existence during this process (and greatly complicating it in all the best ways), and Steve Furches for support. I would like to thank the members of the Jacobson Lab for offering endless help and rewarding collaborations. I would like to thank my committee members for advice and feedback: Steve DiFazio, Avat Shekoofa, Elena Shpak, and Meg Staton. I would like to thank Dan Jacobson for giving me the chance to enter the world of computational and systems biology, the opportunity to participate in many exciting collaborations, and for understanding of how difficult it is to balance the roles of parent and scientist.

Abstract

Climate change threatens agriculture and bioenergy production, necessitating the development of resilient, sustainable feedstocks that can thrive on marginal lands with limited resources. Understanding the genetic architecture underlying complex traits such as biomass accumulation and stress tolerance is essential for crop improvement. *Populus trichocarpa* (black cottonwood) is an important biofuel feedstock and model woody plant with extensive genomic resources and natural populations spanning diverse environmental gradients. This dissertation uses systems and network biology methods to characterize genomic variation in *P. trichocarpa* and link genetic loci with potentially impacted traits, furthering our understanding of the contribution of natural variation to environmental adaptation and identifying targets for feedstock improvement.

Chapter 2 leverages a genome-wide association study (GWAS) to identify single nucleotide polymorphisms (SNPs) associated with cell wall metabolite concentrations. Then, network methods are used to identify novel candidate genes involved in regulation of cell wall biosynthesis – a key trait for plant growth, stress response, and biofuel conversion efficiency. High-priority candidates have since been experimentally validated, demonstrating the power of network-based approaches for gene discovery.

In Chapter 3, I use two detection algorithms to call 794,619 and 877,216 large SVs in 1,484 wild *P. trichocarpa* accessions. Population structure analyses revealed distinct subgroups reflecting post-glacial demographic history, with northern populations showing reduced diversity consistent with rapid expansion from refugia. This SV catalog provides a resource for further studies on the impacts of large-scale genomic rearrangements.

Chapter 4 investigates associations between SVs and climate variables such as water deficit, soil moisture, and temperature. Using relaxed GWAS thresholds combined with network filtering strategies, this approach identified 13,124 and 11,297 SV-climate associations, respectively, with different subtypes capturing distinct evolutionary signals.

Network methods GRIN and MENTOR retained 5,471 biologically relevant genes, three of which (*PtKCS*) were identified as high-priority candidates. Located within a complex SV region, *PtKCS* genes encode enzymes essential for cuticular wax biosynthesis and contribute to population-level variation in drought and pathogen response.

Collectively, this work demonstrates how integrating population genomics, association mapping, and network biology can identify genetic candidates relevant for environmental adaptation in forest trees and improvement of feedstock sustainability.

Table Of Contents

1 Chapter 1: Introduction 1
1.1 Structural Variation in Plant Genomes 3
1.1.1 SV Biology and Formation 3
1.1.2 SV Detection Methods 4
1.1.3 Technical Challenges 6
1.2 *Populus trichocarpa* as a Model System 9
1.2.1 Natural Range and Diversity 9
1.2.2 Population Structure Methods 11
1.2.3 Adaptive Genomics 14
1.3 Climate and Environmental Associations 17
1.3.1 Pacific Northwest Climate 17
1.3.2 GWAS in Forest Trees 18
1.3.3 Climate-Genome Relationships 21
1.4 Network Biology Methods 23
1.4.1 Multi-omics Integration 23
1.4.2 GRIN and MENTOR Algorithms 24
1.4.3 Functional Networks in Plants 27
1.5 Structural Variation and Bioenergy Traits 28
1.5.1 SV Impacts on Bioenergy-Relevant Traits 28
1.5.2 Genomic Approaches for Bioenergy Crop Improvement 31
1.6 Synthesis and Research Gaps 34
1.7 Conclusion 36
References 38
2 Chapter 2: Finding New Cell Wall Regulatory Gene Candidates in *Populus trichocarpa* Using Multiple Lines of Evidence 55
2.1 Abstract 57
2.2 Introduction 58
2.3 Methods 62
2.3.1 Phenotypes 62

2.3.2 Genotypes 62
2.3.3 Genome Wide Association Layer 63
2.3.4 Rare Variant GWAS Layer 63
2.3.5 Co-expression Layer 63
2.3.6 Co-methylation Layer 64
2.3.7 Custom Correlation Coefficient Layer 64
2.3.8 Lines of Evidence Scoring and Network Analysis 65
2.3.9 Gene Annotation, Functional Enrichment, and Expression Analysis 66
2.3.10 Network Validation 67
2.4 Results and Discussion 69
2.4.1 Evaluation of Expression and Methylation Data 69
2.4.2 Construction of LOE Network 69
2.4.3 Tier 1: Highest Priority Candidates for Cell Wall Regulation 75
2.4.4 GROWTH-REGULATING FACTOR 9: Putative Master Regulator 77
2.4.5 Future Directions 86
2.4.6 Recent Work 88
2.5 Funding 88
2.6 Acknowledgements 89
2.7 Data Availability Statement 90
2.8 Supplemental Materials 90
2.8.1 Supplemental Text 90
2.8.2 Supplemental Tables 91
2.8.3 Supplemental Figures 93
References 105
3 Chapter 3: Structural Variation and Population Structure in *Populus trichocarpa* 119
3.1 Abstract 120
3.2 Introduction 120
3.2.1 Populus trichocarpa as a model system 121
3.2.2 Structural variation impacts phenotypic architecture and environmental adaptation 122

3.2.3 Technical challenges in short read SV detection 125
3.2.4 Group methods identify SV and population structure in P. trichocarpa 125
3.3 Methods 127
3.3.1 Genomic Data 127
3.3.2 Population subgroupings 128
3.3.3 SV identification, Genotyping, and Quality Filtering 129
3.3.4 Descriptive Statistics 130
3.3.5 Population Analyses 130
3.4 Results 131
3.4.1 Population subgroupings 131
3.4.2 SV Identification, Genotyping, and Quality Filtering 133
3.4.3 Population Structure 133
3.5 Discussion 141
3.5.1 Future Directions 143
3.6 Supplemental Materials 145
3.6.1 Supplemental Tables 145
3.6.2 Supplemental Text 145
References 148
4 Chapter 4: Identifying Climate-Associated SV Candidates in *Populus trichocarpa* using GWAS and Network Biology Methods 157
4.1 Abstract 158
4.2 Introduction 159
4.3 Methods 162
4.3.1 Genomic Data 162
4.3.2 Climate Data 162
4.3.3 SV-Climatype GWAS Analyses 162
4.3.4 Network Biology Analyses 164
4.3.5 KCS expression profiles 167
4.4 Results 168
4.4.1 Genomic and Climate Data Sets 168

4.4.2 SV-Climatype GWAS Analyses 168
4.4.3 GRIN and MENTOR Analyses 175
4.4.4 PtKCS gene region 179
4.5 Discussion 181
4.5.1 Impact of SV calling methods 181
4.5.2 Using SV candidates to improve biofuel feedstocks 185
4.5.3 SV candidates mapped to PtKCS gene region 186
4.5.4 Future Directions 188
4.5.5 Recent Work 188
4.6 Supplemental Materials 189
4.6.1 Supplemental Tables 189
4.6.2 Supplemental Files 189
4.6.3 Supplemental Figures 191
References 192
5 Chapter 5: Conclusion 198
Vita 203

List Of Tables

Table 2.1: Tier 1 high LOE genes. See Supplemental Table 2.9 for additional score and annotation information. 76

Table 3.1: Accession subgroup clustering levels and summary statistics. “Number” refers to the total number of clusters at that level. Summary statistics (mean, median, minimum, maximum, standard deviation) were performed on the number of samples within clusters for each level. Each level contains the full set of 1484 accessions. 132

Table 3.2: Total number of SV calls by type for each calling algorithm. DEL: deletions; BND: breakends, which are an alternative notation for translocations, both intrachromosomal and interchromosomal; DUP: duplications; INS: insertions; INV: inversions. 137

Table 4.1: Climate variables used in GWAS analyses. 163

Table 4.2: Twenty most extreme SV-climatype effect sizes. The “_m” suffix indicates averaged month (e.g., “_m1” indicates January). 177

List Of Figures

Figure 2.1: Overview of the method for identifying new candidate genes involved in cell wall regulation through data layering and calculation of LOE scores. 61

Figure 2.2: Histograms of network intersect scores calculated by intersecting the observed and randomized LOE networks with GO-term functional networks. (A) Intersection with the global GO-term functional network resulted in a score of 0.4953 for the observed LOE network; intersect scores for randomized networks were ≤0.3701. (B) Intersection with the cell wall-specific GO-term functional network resulted in a score of 0.4806 for the observed LOE network; intersect scores for randomized networks were ≤0.3470. 72

Figure 2.3: Tier 1 high LOE genes with regulatory annotations in the context of the LOE network arranged to highlight each gene's connection to lignin/cell wall biosynthesis pathways. Orange and green circles represent cell wall anchor genes and high LOE genes, respectively. Numbers within high LOE genes (green circles) indicate an entry within Table 2.1. Green circles that do not contain numbers represent a subset of the high LOE genes that were filtered from the final results set due to having prior evidence of cell-wall-related functions in the literature. The size of circles corresponds to their LOE breadth score. Gene symbols are Arabidopsis Best-hit matches. 73

Figure 2.4: Genome-wide neighborhood of *PtGRF9* paralogs in the global input layer networks. Gene symbols are Arabidopsis Best-hit matches. See Supplemental Table 2.10 for detailed node information; for functional enrichment details, see Supplemental Tables 2.11 (negatively co-expressed genes) and 2.12 (positively co-expressed genes). 79

Figure 2.5: Literature evidence linking *AtGRF7* genes to *AtKNOX* genes and the cell wall, overlaid on LOE edges from the *PtGRF9* subnetwork. See Supplemental Table 2.13 for *P. trichocarpa* orthologs. 83

Figure 2.6: Two-hop network created by merging a 1-hop eQTN network around the *PtGRF9* paralogs and 1-hop metabolite-GWAS networks around anchor

metabolites. See Supplemental Table 2.14 for node information. 87
Figure 3.1: Examples of the types of large genomic SV and strategies for detection using short-read paired-end data. Detection of (A) DEL and DUP using variable read depth; (B) DEL, INS, INV, TRA, and DUP (tandem) using discordant read pairs; (C) DEL using split reads; and (D) INS using de novo assembly. Note: Figure reproduced from Escaramís et al. (2015) with permission from the publisher. 123
Figure 3.2: Average pairwise kinship within sc clusters. A) bar graph depicting total number of accessions within each sc cluster; B) violin plot depicting average pairwise kinship within sc clusters. 134
Figure 3.3: Clustered heatmap with attached dendrogram of pairwise kinship among accessions. The colorized column on the left side of the heatmap ("cluster") indicates sc cluster membership for each accession. 135
Figure 3.4: Geographic distribution of *P. trichocarpa* accessions included in Chapters 3 and 4 colorized by membership in SNP pairwise kinship-based sc clusters. 136
Figure 3.5: SV counts by type and SV caller. 138
Figure 3.6: SV size distribution by type and caller. 139
Figure 3.7: Population genetic analyses of SV call sets. Visualizations in the left column (a, c, e) are Delly data and in the right column (b, d, f) are Lumpy data. Panels a, b depict PCA results; panels c, d depict UMAP and DBSCAN results; panels e, f depict the geographic distribution of accessions colorized by membership to UMAP and DBSCAN determined subgroups. 140
Figure 4.1: Total number of significant SV-climatype GWAS associations by GWAS method and SV caller. 169
Figure 4.2: Selected example of Manhattan and QQ plots for MLM GWAS analysis of the "soil" climate variable. Manhattan and QQ plots are standard diagnostic tools to assess potential bias and parameter settings in GWAS analyses. Here, I selected one of many representative examples to demonstrate that SV-climatype analyses produced numerous highly significant results (data points

above horizontal line in left most plot) and exhibit some evidence of *p* value inflation despite the incorporation of control for population structure and cryptic relatedness in the GWAS model (blue data points rising above the diagonal red line in the right most plot). 170
Figure 4.3: Overlap of significant associations among GWAS methods by SV caller. . 172
Figure 4.4: Number of significant GWAS associations per climate variable and SV type for each variant caller. 173
Figure 4.5: Distribution of significant GWAS associations across averaged months per variant caller (left column: Delly, right column: Lumpy), climate variable, and SV type. Note that the following climatypes were removed during preprocessing: swe months 3 – 10, def months 1, 11 – 12. 174
Figure 4.6: Heatmaps of median standardized effect sizes estimated during GWAS analysis by variant caller (left column: Delly, right column: Lumpy), climate variable, and SV type. Note that the following climatypes were removed during preprocessing: swe months 3 – 10, def months 1, 11 – 12. 176
Figure 4.7: Selected clusters from the MENTOR cladogram that was produced from the union of Delly and Lumpy significant mapped-to-genes SV-climatype GWAS associations. Summaries of MENTOR-IA functional profiles are provided to the left of each cluster. The genes indicated with arrows were identified as high priority candidates based on having the highest numbers of significant climate associations and from both SV call sets, and were used to select cladogram clusters for follow up analyses. The SV types that each gene was mapped from are listed below the bold gene ID. 178
Figure 4.8: *Populus trichocarpa* reference genome Nisqually-1 gene atlas expression heatmap for *PtKCS* genes. 180
Figure 4.9: Clustered heatmap and dendrogram of *PtKCS* gene expression in mature leaf tissue collected from the *P. trichocarpa* GWAS population. Rows represent accessions. Variation in gene expression levels are visualized in columns labeled with gene identifiers as a gradient from black (lowest or absent expression) to white (highest expression). Metabolite concentrations for the

alkenes heptadecosene and pentacosene are visualized in the left-most columns as a gray-scale gradient from white (low or absent concentration) to black (highest concentration). The morphological classification of leaf phenotypes based on overlapping accessions with Gonzales-Vigil et al. (2017) is visualized in the blue column, with alkene “plus” accessions represented by a green bar, alkene “minus” by a yellow bar, and accessions with unknown morphology that were not present in the 2017 study represented by blue bars. 182

List of Attachments

Attachment 2.1: Supplemental tables for Chapter 2, "Furches_Chapter2_Supplemental_Tables.xlsx".

Attachment 3.1: Supplemental table for Chapter 3, "Furches_Chapter3_SuppTab3.1.xlsx".

Attachment 4.1: Supplemental tables for Chapter 4, "Furches_Chapter4_SuppTables.xlsx".

Attachment 4.2 (Supplemental File 4.1): Delly significant SV-climatype GWAS association SV-to-genes mappings, "Furches_Chapter4_SuppFile4.1_snp2genes_DELLY.txt".

Attachment 4.3 (Supplemental File 4.2): Lumpy significant SV-climatype GWAS association SV-to-genes mappings, "Furches_Chapter4_SuppFile4.1_snp2genes_LUMPY.txt".

Attachment 4.4 (Supplemental File 4.3): Significant SV-climatype GWAS association SV mapped-to-genes unique gene set retained by GRIN for input to MENTOR, "Furches_Chapter4_SuppFile4.3_GRINretained.txt".

Attachment 4.5 (Supplemental File 4.4): MENTOR dendrogram of 5,471 GRIN-retained genes containing 270 functionally related clades, "Furches_Chapter4_SuppFile4.4_dendrogram_heatmap.pdf".

Attachment 4.6 (Supplemental File 4.5): MENTOR-IA synthesized results for cladogram cluster 4, "Furches_Chapter4_SuppFile4.5.pdf".

Attachment 4.7 (Supplemental File 4.6): MENTOR-IA synthesized results for cladogram cluster 115, "Furches_Chapter4_SuppFile4.6.pdf".

Attachment 4.8 (Supplemental File 4.7): MENTOR-IA synthesized results for cladogram cluster 130, "Furches_Chapter4_SuppFile4.7.pdf".

1 Chapter 1: Introduction

Populus trichocarpa (black cottonwood) has emerged as a premier model system for advancing our understanding of woody plant genomics and developing sustainable bioenergy feedstocks. As the first tree genome to be sequenced, *P. trichocarpa* provides an unparalleled foundation for investigating the genetic mechanisms underlying adaptive traits and biomass quality characteristics essential for biofuel production (Tuskan et al., 2006). The species' native range across the Pacific Northwest of North America, spanning diverse climate gradients from northern California to Alaska, offers a natural laboratory for studying climate adaptation and the genetic basis of stress tolerance—traits critical for developing robust bioenergy crops capable of thriving on marginal agricultural lands. The U.S. Department of Energy's investment in *P. trichocarpa* research through initiatives such as the Center for Bioenergy Innovation (CBI) reflects the strategic importance of this species for achieving national bioenergy goals. Unlike food crops, bioenergy feedstocks must be cultivated on lands unsuitable for agriculture to avoid competition with food production systems. The economic viability of lignocellulosic biofuels depends critically on both biomass yield and feedstock composition, with lignin content and structure serving as key determinants of bioconversion efficiency (Happs et al., 2021).

Recent advances in genomic technologies have revolutionized our ability to dissect the genetic architecture of complex traits in forest trees. Genome-wide association studies (GWAS) using natural populations of *P. trichocarpa* have identified numerous loci associated with adaptive traits, biomass characteristics, and cell wall chemistry (Bryant et al., 2023; Evans et al., 2014). However, these studies have primarily focused on single nucleotide polymorphisms (SNPs), overlooking the substantial contribution of structural variants (SVs) to phenotypic variation. SVs, including deletions, insertions, inversions, translocations, and copy number variations, can have profound effects on gene expression and function through various mechanisms. This dissertation addresses critical knowledge gaps in our understanding of structural variation and its role in climate adaptation and bioenergy trait variation in *P. trichocarpa*. By developing and applying cutting-edge computational approaches to identify, genotype, and functionally annotate structural

variants in natural populations, this research aims to advance both fundamental understanding of adaptive genomics in forest trees and applied goals of bioenergy feedstock improvement.

1.1 Structural Variation in Plant Genomes

1.1.1 SV Biology and Formation

SVs represent a major source of genetic diversity that significantly impacts genome function and evolution. Defined as genomic alterations involving DNA segments of 50 base pairs or larger, structural variants encompass deletions (DEL), insertions (INS), duplications (DUP), inversions (INV), translocations (TRA or BND), and complex rearrangements (BND) (Escaramís et al., 2015). These variants can have profound effects on the phenotype through multiple mechanisms, including gene disruption, altered gene expression through changes in regulatory landscapes, and dosage effects resulting from copy number variation. In plant genomes, SVs contribute substantially to intraspecific diversity and play crucial roles in adaptation to environmental stress and agricultural trait variation (McCouch et al., 2016; Saxena et al., 2014). The formation of SVs occurs through several distinct molecular mechanisms. Non-allelic homologous recombination (NAHR) between repetitive sequences such as transposable elements (TEs) can generate deletions, duplications, and complex rearrangements, particularly in regions with high repeat content (Carvalho & Lupski, 2016; Girirajan & Eichler, 2010). Non-homologous end joining (NHEJ) repair of double-strand breaks frequently results in small deletions and insertions at breakpoints, often accompanied by microhomology signatures (Lieber, 2010). Replication errors, including fork stalling and template switching (FoSTeS) and microhomology-mediated break-induced replication (MMBIR), can produce complex SVs with multiple breakpoints (J. A. Lee et al., 2007; Yang et al., 2013).

Population-level studies in diverse plant species have shown that SVs are more prevalent in pericentromeric and heterochromatic regions, likely due to the abundance of repetitive sequences that promote recombination errors (McCouch et al., 2016; Springer et al., 2009). The size distribution of SVs follows a characteristic pattern, with small variants

(50-1000 bp) being most abundant and large variants (>10 kb) being relatively rare but potentially having greater phenotypic impact (Yang et al., 2019). In forest tree populations, which typically maintain high levels of genetic diversity due to large effective population sizes and extensive gene flow, SVs can reach appreciable frequencies even when mildly deleterious (Alberto et al., 2013; Neale & Kremer, 2011). The functional impact of SVs depends critically on their genomic context and the specific genes or regulatory elements they affect. Gene-disrupting variants typically have strong phenotypic effects, while many SVs exert their effects through more subtle mechanisms, such as altering three-dimensional chromatin structure or changing the copy number of regulatory elements (Spielmann et al., 2018). The study of SV in plant genomes has been revolutionized by advances in sequencing technologies, though the repetitive nature of many plant genomes continues to pose challenges for accurate SV detection and genotyping (Ho et al., 2020; Sedlazeck et al., 2018).

1.1.2 SV Detection Methods

Accurately detecting and genotyping SVs from short-read sequencing data represents one of the most challenging problems in computational genomics (Cameron et al., 2019). Paired-end mapping approaches in which algorithms searched for discordant read pairs were amongst the first methods used to detect SV and form the foundation of many detection algorithms (K. Chen et al., 2009). However, many algorithmic approaches have since been developed that use multiple types of evidence from paired-end sequencing data. For example, LUMPY uses a probabilistic framework that integrates evidence from discordant read pairs and split reads to identify breakpoints with high sensitivity across diverse SV types (Layer et al., 2014). The algorithm's strength lies in its ability to handle complex genomic regions and maintain sensitivity even at low coverage, making it particularly suitable for population-level studies. DELLY also integrates a third signal, integrating paired-end, split-read, and soft-clip evidence through a two-stage approach that first identifies potential SV sites using discordant read pairs, then refines breakpoint predictions and filters false positives using split-read and soft-clip alignments (Rausch et al., 2012). This integrated approach enables single-nucleotide resolution of breakpoints

and improves specificity compared to methods relying solely on paired-end evidence. Recent benchmarking studies have demonstrated that DELLY performs particularly well for deletions and tandem duplications (Cameron et al., 2019). Local assembly approaches typically provide the highest precision and sensitivity by reconstructing the sequence structure around putative SV sites. GRIDSS represents a sophisticated implementation of this strategy, using a positional de Bruijn graph to assemble reads that provide evidence for SV (Cameron et al., 2019). While computationally intensive, assembly-based methods can potentially resolve complex rearrangements and novel insertions that are difficult to detect using mapping-based approaches alone.

Benchmarking studies highlight substantial differences in performance among SV detection algorithms, with accuracy varying significantly by variant type, size, and genomic context (Kosugi et al., 2019). LUMPY and DELLY typically perform well for deletions and duplications, while newer tools like Manta show improved performance for insertions and complex rearrangements. These findings have led to increased interest in ensemble approaches that combine predictions from multiple algorithms. However, an increasing number of studies have shown that ensemble methods are subject to significant bias, due in part to the SV call merging methods that are typically used. Population-based SV calling strategies have emerged as an important alternative (Zheng et al., 2024). Rather than analyzing samples independently, joint calling approaches can improve accuracy by leveraging information shared across individuals. This is particularly relevant for forest tree populations, where high levels of diversity and complex demographic histories can complicate SV detection. Subpopulation-based calling strategies, where samples are grouped by genetic similarity before SV analysis, represent a promising approach for managing population structure effects while maintaining computational tractability. Furthermore, population methods reduce the bias associated with merging single-sample based SV calls.

Quality control and filtering remain critical steps in SV analysis pipelines. Although best practices still employ manual filtering, tools like Duphold provide specialized filtering

based on read depth patterns around SV breakpoints, which can remove many false positive calls while retaining true variants (Pedersen & Quinlan, 2019). Manual curation using visualization tools such as Samplot has become standard practice for validating high-impact variants, though this approach is not scalable to population-level studies (Belyeu et al., 2018, 2021). The development of machine learning approaches for automated SV filtering represents an active area of research, with promising results for reducing false positive rates while maintaining sensitivity.

1.1.3 Technical Challenges

The accurate detection and genotyping of SVs from short-read sequencing data faces numerous technical challenges that significantly impact the reliability and completeness of SV catalogs. These challenges are particularly acute in plant genomes, which often contain high levels of repetitive sequences, complex polyploid structures, and extensive gene family duplications. Short-read sequencing biases represent a fundamental constraint on SV detection capabilities. The limited read length of current Illumina platforms creates inherent blind spots for variants larger than the insert size distribution, complex rearrangements spanning multiple breakpoints, and novel insertions containing repetitive sequences (Mahmoud et al., 2019; Sedlazeck et al., 2018). PCR amplification during library construction can also introduce systematic biases, particularly affecting GC-rich and repetitive regions (Y. C. Chen et al., 2013). Repetitive regions pose extraordinary challenges for SV detection algorithms. Transposable elements, tandem repeats, and segmental duplications create ambiguous mapping environments in which reads may align equally well to multiple genomic locations (Treangen & Salzberg, 2011). This mapping ambiguity leads to both false positive and false negative SV calls, as algorithms may incorrectly interpret repetitive sequences as SVs or fail to detect true variants embedded within repetitive contexts. The problem is exacerbated in plant genomes, where repetitive content can exceed 80% in species like maize and wheat, creating vast regions that are effectively invisible to short-read SV detection methods.

In addition, the presence of unaccounted for population structure introduces systematic

biases that can confound SV detection and interpretation. Differences in genome structure between divergent populations can lead to mapping artifacts when reads from one population are aligned to a reference genome derived from another population (Garrison et al., 2018). These reference bias effects are particularly problematic in forest tree species, which often exhibit substantial population differentiation and may harbor population-specific insertions, deletions, or rearrangements not represented in the reference genome. The result is systematic under-calling of SVs in divergent populations and potential false positive calls due to alignment artifacts. The use of reference-free pangenome based methods will ameliorate these issues, but at the moment primarily rely on long-read sequencing data.

Quality control and filtering strategies face fundamental trade-offs between sensitivity and specificity. Conservative filtering approaches that require high-confidence evidence from multiple sources may eliminate many true positive variants, particularly rare or population-specific SVs that lack strong supporting evidence. Conversely, lenient filtering strategies that maximize sensitivity often retain substantial numbers of false positive calls that can overwhelm downstream analyses (Cameron et al., 2019; Mahmoud et al., 2019). The optimal filtering strategy depends critically on the specific research objectives, population characteristics, tolerance for different types of errors, and is often also affected by computational limits since more stringently filtered, smaller data sets can be analyzed more quickly and with fewer resources. Computational scalability becomes increasingly challenging as dataset sizes grow to population-scale proportions. Many SV detection algorithms exhibit super-linear scaling with sample size, making analysis of large populations computationally prohibitive. Memory requirements for storing and processing intermediate results can become limiting factors, particularly for algorithms that maintain global data structures or perform all-versus-all comparisons. The development of efficient parallelization strategies and memory-conscious algorithms represents an ongoing area of research, though fundamental algorithmic limitations may require novel approaches for truly large-scale studies.

Validation and benchmarking of SV detection methods face inherent limitations due to the lack of comprehensive truth sets for complex genomes. Unlike single nucleotide variants, where high confidence call sets have been developed through extensive validation efforts, structural variant truth sets remain limited in scope and may not adequately represent the full spectrum of variation present in natural populations (De Coster et al., 2019; Zook et al., 2020). Simulation-based evaluations, while valuable for algorithm development, may not capture the full complexity of real genomic contexts, including the interaction effects between multiple variants and the impact of diverse demographic histories (English et al., 2015).

As discussed above, the integration of diverse short-read evidence types presents both opportunities and challenges for SV detection. While combining paired-end, split-read, and read-depth signals can improve accuracy, the appropriate weighting and integration of these evidence types remains an active area of research. Different evidence types may be more or less reliable depending on the specific SV type, genomic context, and sequencing characteristics, requiring sophisticated models that can adapt to local conditions while maintaining global consistency. Computational limitations again become an important consideration because with increasing analysis complexity greater resources and computational time are required. This becomes even more of an issue in polyploid species. Polyploidy introduces additional complications for SV detection in many plant species. Standard diploid models may be inappropriate for polyploid genomes, where copy number changes may reflect complex allelic relationships rather than simple deletions or duplications. The presence of homologous chromosomes in allopolyploids creates additional mapping ambiguities and complicates the interpretation of read depth signals (Gabur et al., 2019). Specialized methods for polyploid SV detection remain underdeveloped, limiting the applicability of current approaches to many economically important crop species.

1.2 *Populus trichocarpa* as a Model System

1.2.1 Natural Range and Diversity

1.2.1.1 Geographic Distribution and Population Structure

Populus trichocarpa has an extensive natural range along the Pacific coast of North America, spanning from Alaska to Baja California Norte (latitude 31°N to 62°N) (Piot et al., 2020). This broad latitudinal gradient encompasses diverse ecological zones from temperate rainforests to semi-arid regions, creating a natural laboratory for studying population genomics and local adaptation. The species primarily occupies riparian corridors and bottomlands along major rivers and streams, though it extends into upland sites in some regions.

Population genetic structure within this vast range has been extensively characterized using both traditional markers and genome-wide approaches. Landscape genomic analyses using 498 accessions genotyped with a 34K *Populus* SNP array revealed that three processes contribute to observed patterns: (1) introgression from the sister species *P. balsamifera*, (2) isolation by distance (IBD), and (3) natural selection (Geraldes et al., 2014). Introgression was detected only at the margins of the distribution. While IBD was significant across the sampled area, no evidence of restricted gene flow was detected in a core of drainages from southern British Columbia (Geraldes et al., 2014). This pattern suggests that gene flow is generally effective throughout most of the range, with demographic and selective forces operating at different spatial scales.

1.2.1.2 Genetic Diversity Patterns Across the Range

Large-scale population genomic studies using 544 whole-genome sequences have revealed extensive genetic diversity, with 17 million variants identified across the species range (Evans et al., 2014). Analysis of 461 genotypes from 136 provenance localities using 29,354 SNPs demonstrated high phenotypic variation and moderate to high broad-sense heritability for many traits, with heritability ranging from 0.3 to 0.9 in phenology, 0.3 to 0.8 in biomass, and 0.1 to 0.8 in ecophysiology traits (Mckown et al., 2014). The distribution of genetic diversity is not uniform across the range, with northern

populations generally showing reduced diversity compared to southern populations, consistent with post-glacial colonization patterns. Principal component analysis revealed that the first component correlated strongly with photoperiod and showed enrichment for genes relating to circadian rhythm. This finding highlights the fundamental role of photoperiodic adaptation in structuring genetic variation across the latitudinal gradient. The genetic architecture underlying this variation appears to involve many loci of small effect, consistent with polygenic adaptation to environmental gradients.

1.2.1.3 Historical Demography and Refugia

The current distribution and genetic structure of *P. trichocarpa* populations reflect complex demographic processes during the Pleistocene glacial cycles. Coalescent analyses of multilocus DNA sequence data from *P. trichocarpa* and its sister species *P. balsamifera* suggest that their divergence occurred during the Pleistocene, providing direct evidence for recent tree speciation associated with climatic events (Levsen et al., 2012). Paleodistribution modeling combined with genetic data indicates that populations persisted in multiple refugia during glacial maxima, likely in areas that remained ice-free along the Pacific coast and in interior valleys. The rapid post-glacial expansion from these refugia has left distinctive signatures in the genome. Northern populations show evidence of founder effects and demographic bottlenecks consistent with rapid range expansion, while southern populations maintain higher genetic diversity reflecting their proximity to potential refugial areas (Piot et al., 2020). These demographic processes have important implications for contemporary population structure and the distribution of adaptive variation across the range.

1.2.1.4 Clinal Variation and Local Adaptation

Strong clinal variation in adaptive traits provides compelling evidence for local adaptation across the *P. trichocarpa* range. Extensive phenotyping across altitudinal and latitudinal gradients has revealed clinal variation in traits related to seasonal growth, cold hardiness, and regeneration ability, with timing of bud phenology serving as a key adaptive response to seasonal variation in temperature (M. Zhang et al., 2019). When

climate variables were used as predictors for 12 candidate adaptive traits in multivariate regression tree analysis, evapotranspiration explained the most variation, with subsequent splits related to mean temperature of the warmest month (Oubida et al., 2015). Robust relationships between traits, population structure and geoclimate variables reflect patterns suggesting that range-wide geographical and environmental gradients have shaped both genotypic and phenotypic variability, with most traits correlating strongly with latitude, maximum daylength and temperature of tree origin (McKown et al., 2014). Provenances from high latitudes typically require longer critical day length to induce bud formation and thus show earlier growth cessation and timing of bud set (Zhang et al., 2019), demonstrating the tight coupling between geographic origin and adaptive phenotypes. Population differentiation for many traits exceeded the 99th percentile of the genome-wide distribution of F_{ST}, providing strong evidence for local adaptation (Oubida et al., 2015). The genetic architecture underlying these adaptive clines involves both parallel evolution at shared loci and unique evolutionary responses at different spatial scales. Up to 51% of outlier loci overlapped between latitudinal and altitudinal transects, with genomic clustering of adaptive loci being more pronounced for altitudinal than latitudinal transects (Holliday et al., 2016), suggesting that physical proximity of genes in coadapted complexes may buffer against gene flow of maladapted alleles.

1.2.2 Population Structure Methods

1.2.2.1 Kinship Estimation and Relatedness

Accurate estimation of kinship and relatedness is fundamental for *P. trichocarpa* population genomics. Modern approaches employ marker-inferred relatedness estimation to obtain narrow-sense estimates of population differentiation in wild populations, replacing traditional pedigree-based methods that are impractical for forest trees (Porth et al., 2015). These SNP marker-based kinship matrices have proven essential for controlling population effects in genome-wide association studies and enabling quantitative genetic parameter estimation in natural populations. The implementation of kinship estimation in *P. trichocarpa* studies typically involves constructing family relatedness using kinship (K) models within mixed-model frameworks such as the

Unified Mixed Model (McKown et al., 2014). This approach allows researchers to separate the effects of individual loci from confounding population effects, which is particularly important given the extensive population structure observed across the species' range. The kinship matrices derived from thousands of SNPs provide more accurate estimates of realized relatedness than genealogical expectations, accounting for the stochastic nature of inheritance and enabling more powerful association analyses.

1.2.2.2 PCA and UMAP for Dimensionality Reduction

Principal Component Analysis (PCA) remains the most widely used method for characterizing population structure, providing linear dimensionality reduction that identifies directions of maximal genetic variance (McKown et al., 2014). Spatial PCA (sPCA) incorporates geographical information to investigate local and global genetic structures, particularly valuable for detecting isolation-by-distance patterns (Porth et al., 2015). Spatial PCA (sPCA) represents an important extension for forest tree populations, incorporating geographical information to investigate the presence of local and global genetic structures (Porth et al., 2015). This method has been particularly valuable for detecting isolation-by-distance patterns in *P. trichocarpa*, where large positive eigenvalues indicate significant spatial genetic structure. The sPCA approach helps disentangle demographic processes from selective forces by explicitly modeling the geographic component of genetic variation.

While PCA remains standard, newer non-linear methods such as UMAP offer potential advantages for detecting subtle population structure that linear methods might miss (Diaz-Papkovich et al., 2019). UMAP excels at preserving local neighborhood structures while maintaining global relationships, making it particularly useful for detecting subtle population structure that linear methods might miss (Diaz-Papkovich et al., 2019). The method's ability to create clear clusters while preserving hierarchical relationships could prove valuable for characterizing complex population structure in forest trees.

1.2.2.3 Subpopulation Identification Strategies

Multiple complementary approaches have been developed for identifying discrete subpopulations within continuous distributions. Model-based clustering using programs like STRUCTURE and ADMIXTURE provides probabilistic assignment of individuals to ancestral populations while estimating admixture proportions (Alexander et al., 2009). These Bayesian approaches have revealed important insights into population structure, including evidence for limited introgression from sister species *P. balsamifera* at range margins and the identification of climate-based population clusters that reflect adaptive divergence (Geraldes et al., 2014).

The challenge of defining biologically meaningful subpopulations in continuously distributed forest trees has led to innovative approaches combining genetic and environmental data. Climate-based clustering has proven particularly effective in *P. trichocarpa*, where populations are grouped based on temperature and precipitation variables of their origin sites rather than solely on genetic similarity (Porth et al., 2015). This approach recognizes that adaptive differentiation may follow environmental gradients more closely than geographic distance, leading to more biologically relevant population definitions for downstream analyses.

1.2.2.4 Linkage Disequilibrium Patterns in Forest Trees

Linkage disequilibrium (LD) patterns in *P. trichocarpa* differ markedly from those in agricultural crops and model organisms, reflecting the unique demographic and reproductive characteristics of forest trees. Genome-wide studies using 29,213 SNPs across 120 trees from 10 subpopulations revealed that LD decays slower than expected from previous studies, with r^2 dropping below 0.2 within 3-6 kb (Slavov et al., 2012). This extensive LD suggests that genome-wide association studies and genomic selection may be more feasible in undomesticated *Populus* populations than previously assumed. The slower LD decay in *P. trichocarpa* reflects several biological factors unique to forest trees, including recent demographic bottlenecks during post-glacial expansion, limited effective population sizes relative to census sizes, and the influence of population

structure on genome-wide LD patterns. Estimates of recent effective population size from linkage disequilibrium (Ne ≈ 4,000-6,000) were remarkably low relative to the large census sizes of *P. trichocarpa* stands, suggesting that genetic drift has played a significant role in recent evolutionary history (Slavov et al., 2012). Fine-scale recombination rates vary widely across the *P. trichocarpa* genome but are largely predictable based on DNA sequence and methylation features, providing important insights for marker selection in association studies. The extensive LD has practical implications for genome-wide association studies, as between 67,000-134,000 SNPs would be required to capture all common variants throughout the genome at $r^2 = 0.2$, influencing study design and power calculations (Porth et al., 2015).

1.2.3 Adaptive Genomics

1.2.3.1 Signatures of Selection in P. trichocarpa

Large-scale population genomic studies have revealed extensive signatures of selection across the *P. trichocarpa* genome, providing compelling evidence for widespread local adaptation. Genome-wide selection scans using 544 whole-genome sequences identified 397 genomic regions showing evidence of recent positive and/or divergent selection, with these regions showing enrichment for associations with adaptive traits that displayed patterns consistent with natural selection (Evans et al., 2014). This comprehensive analysis revealed that climate-driven selection has acted on genes related to drought tolerance, photoperiod response, and stress resistance, reflecting the diverse environmental challenges across the species' extensive latitudinal range. Selection signatures detected through multiple outlier detection methods, including F_{ST}-based approaches and spatial analysis, have identified 2,855 SNPs showing signals of diversifying selection (Porth et al., 2015). Of these, 118 SNPs within 81 genes were associated with adaptive traits based on significant QST values, indicating their direct involvement in climate adaptation. Many of these SNPs appear to be pleiotropic, affecting functionally uncorrelated adaptive traits such as autumn phenology, height, and disease resistance, suggesting complex regulatory networks underlying adaptive responses.

The evolutionary dynamics of duplicated genes provide additional insights into adaptive evolution in *P. trichocarpa*. Selection scans have revealed that duplicated gene pairs show distinct patterns of selection, with some copies evolving under relaxed constraint while others maintain strong purifying selection (Evans et al., 2014). This observation suggests that gene duplication events have contributed to adaptive potential by providing genetic material for neofunctionalization and subfunctionalization processes that enhance environmental responsiveness.

1.2.3.2 Local Adaptation to Climate Gradients

Populus trichocarpa populations exhibit strong signatures of local adaptation to climate gradients. Quantitative genetic analyses revealed that 53% of distinct field traits show adaptive divergence, indicating widespread local adaptation to climate variables (Porth et al., 2015). This adaptive differentiation is particularly pronounced for traits related to phenology, growth, and stress tolerance, which show strong correlations with temperature and precipitation gradients across the species range. Genome-wide association studies using multi-trait approaches have identified key polymorphisms controlling morphological and physiological traits that contribute to climate adaptation (Chhetri et al., 2019). Analysis of 882 trees with over 6.78 million SNPs revealed that broad-sense heritability ranged from 0.30 to 0.56 for morphological traits and 0.08 to 0.36 for physiological traits, with 24 gene models detected through combined single-trait and multi-trait association methods. These findings demonstrate the polygenic nature of climate adaptation while identifying specific loci of major effect.

Evidence for parallel adaptation along both latitudinal and altitudinal gradients provides insights into the evolutionary constraints and opportunities shaping adaptive responses. Comparative analyses across environmental gradients have revealed both shared and unique adaptive loci, with up to 51% of outlier loci overlapping between latitudinal and altitudinal transects (Zhang et al., 2019). This pattern suggests that while some adaptive solutions are constrained by available genetic variation and developmental pathways, multiple evolutionary routes to climate adaptation exist within the species.

1.2.3.3 Genomic Offset and Vulnerability Predictions

Genomic offset approaches represent a powerful framework for predicting population vulnerability to future climate change by quantifying the genetic changes required for populations to remain adapted to shifting environmental conditions. These methods integrate current genotype-environment relationships with climate projections to estimate the magnitude of genetic change needed for populations to track optimal climate conditions (Fitzpatrick & Keller, 2015). For long-lived forest trees like *P. trichocarpa* that face rapid climate change within individual lifespans, such predictions are essential for conservation planning and assisted migration strategies.

The risk of non-adaptedness (RONA) approach provides single-locus estimates of vulnerability by modeling expected allele frequency changes required under future environmental conditions through linear regression of putatively adaptive variants (Rellstab et al., 2015). When applied to *P. trichocarpa* populations, RONA analyses reveal spatial patterns of vulnerability that align with projected climate velocity and severity of environmental change. Populations in rapidly warming regions with limited dispersal opportunities show higher RONA values, indicating greater risk of maladaptation to future conditions.

Gradient forest (GF) methods offer complementary machine-learning approaches that capture non-linear genotype-environment relationships and provide multilocus estimates of genomic offset (Feng & Du, 2022). These approaches have proven particularly valuable for forest trees, where complex trait architectures and genotype-by-environment interactions create non-linear responses to environmental variation. GF analyses in related *Populus* species have successfully predicted population responses to climate change and identified geographic regions requiring conservation priority.

Recent advances integrate genomic offset predictions with ecological niche modeling to provide comprehensive vulnerability assessments that consider both genetic and habitat suitability constraints (Capblancq & Forester, 2021). These integrated approaches

recognize that climate change impacts involve both direct effects on population genetic composition and indirect effects through habitat modification and species interactions. For *P. trichocarpa*, analyses reveal that southeastern populations may face particularly high vulnerability due to the combination of large projected genomic offsets and declining habitat suitability.

1.3 Climate and Environmental Associations

1.3.1 Pacific Northwest Climate

The Pacific Northwest region exhibits exceptional climate diversity across the native range of *P. trichocarpa*, creating a natural laboratory for investigating climate-genome relationships. The region's complex topography interacts with marine and continental air masses to produce strong west-east moisture and temperature gradients that have shaped tree population differentiation (Rupp et al., 2013). These steep environmental gradients provide the statistical power necessary for detecting climate-associated genomic variants, as populations experience markedly different selective pressures across relatively short geographic distances.

Recent climatological analyses reveal significant temporal trends in Pacific Northwest climate that are directly relevant to forest genetic studies. Annual mean temperature has increased by approximately 0.6°–0.8°C from 1901 to 2012, with corresponding changes in growing-season potential evapotranspiration and freeze-free season length (Rupp et al. 2013). These warming trends show substantial seasonal heterogeneity, with variable patterns across shorter time scales that create complex selection mosaics for tree populations.

High-resolution climate data essential for genetic association studies are primarily available through PRISM and ClimateNA. PRISM produces gridded climate data at 800-meter resolution by incorporating complex terrain effects, temperature inversions, rain shadows, and coastal proximity (Daly et al., 2008). ClimateNA provides a complementary approach by downscaling PRISM and WorldClim data to scale-free point

locations, generating over 200 climate variables including biologically relevant derived variables such as growing degree days and moisture deficit indices (T. Wang et al., 2016).

The exceptional depth of Pacific Northwest climate records, spanning 60+ years at many locations, provides sufficient temporal coverage for detecting meaningful climate-genome associations while accounting for inter-annual climate variability. Long-term records from weather station networks, combined with gridded interpolations, enable researchers to calculate climatological normals and temporal trends that capture the environmental variation experienced by tree populations over multiple generations (Rupp et al. 2013). Monthly climate variables have emerged as particularly important predictors, as they capture seasonal variation in temperature and precipitation that directly influences tree phenology, growth allocation, and stress responses. Studies across multiple forest species demonstrate that monthly temperature and precipitation variables often show stronger associations with adaptive traits than annual summaries, reflecting the importance of seasonal climate cues for tree development (Guerra et al., 2019; McKown et al., 2018). For *P. trichocarpa* populations distributed across large latitudinal and elevational gradients, monthly climate variables effectively capture the variation in growing season length, winter chill accumulation, and seasonal drought stress that drive local adaptation.

1.3.2 GWAS in Forest Trees

Genome-wide association studies in forest trees present unique methodological challenges. The long generation times, extensive gene flow, and large effective population sizes create a genomic architecture fundamentally different from domesticated species, requiring specialized analytical approaches (Howe et al., 2022). Unlike annual crops that have undergone centuries of directional selection creating major-effect QTLs, forest tree populations closely resemble wild populations where quantitative trait variation follows a highly polygenic, micro-mutational architecture with numerous small-effect loci (Du et al., 2018). This evolutionary backdrop necessitates statistical methods

optimized for detecting subtle associations across thousands of loci rather than seeking major-effect variants.

Population structure correction represents the most critical methodological consideration in forest tree GWAS, as the extensive gene flow and weak population differentiation typical of these species can generate both false positives and false negatives if not properly addressed (Müller et al., 2019). Linear mixed models (LMMs) have emerged as the standard approach for simultaneously controlling population structure and kinship effects while maintaining statistical power (Müller et al., 2019) but their application to forest trees requires careful consideration of the correlation structure between individuals (Slavov et al., 2025). The challenge is particularly acute because most phenotypic variation (60-81%) often occurs among populations and is strongly associated with climate. Advanced methods such as compressed MLM and SUPER models address the tendency of standard approaches to be overly conservative (Q. Wang et al., 2014; Z. Zhang et al., 2010).

Environmental association studies represent a powerful complementary approach. Genotype-environment association (GEA) methods such as latent factor mixed models (LFMM) and redundancy analysis (RDA) enable direct testing of associations between genomic variants and environmental variables while accounting for population structure (Capblancq & Forester, 2021; Frichot et al., 2013). These approaches have proven particularly valuable for forest trees because they can detect signatures of local adaptation even when adaptive phenotypes are difficult to measure or when environmental gradients are closely correlated with population structure (Pluess et al., 2016; Rellstab et al., 2015). The ability to incorporate multiple environmental variables simultaneously through multivariate models provides enhanced power to detect polygenic adaptation across complex environmental gradients (Sang et al., 2022).

Multiple testing and statistical power represent ongoing challenges in forest tree GWAS due to the combination of large numbers of markers, complex population structure, and

typically modest sample sizes. The standard Bonferroni correction is often overly conservative when applied to millions of correlated markers, leading to substantial loss of statistical power (Han et al., 2009). More sophisticated approaches such as the false discovery rate (FDR) and permutation-based methods provide better balance between Type I and Type II errors (Joo et al., 2016; Storey & Tibshirani, 2003).

The high genetic diversity, rapid linkage disequilibrium decay, and minimal population structure of forest trees can provide some advantages for GWAS despite the methodological challenges (Du et al., 2018). Low levels of linkage disequilibrium mean that detected associations are likely to be physically close to causal variants, potentially even representing the causal variants themselves rather than linked markers (Ingvarsson et al., 2008). This characteristically positions forest tree GWAS as particularly powerful for gene discovery and functional validation compared to species with extensive linkage disequilibrium. However, the rapid LD decay also means that much larger marker densities are required to achieve genome-wide coverage, and traditional SNP-based approaches may miss important structural variants that contribute to adaptive variation.

Successful GWAS applications in *Populus* and other forest trees have demonstrated the potential for identifying loci underlying ecologically and economically important traits, while also highlighting the importance of appropriate study design and statistical methods. Notable successes include the identification of PtFT2 as a major locus controlling bud set timing in European aspen, explaining 65% of genetic variation in this critical adaptive trait (J. Wang et al., 2018). In *P. trichocarpa*, GWAS studies have successfully identified loci associated with wood chemistry traits, growth characteristics, and stress responses, including the identification of PtHCT2 controlling chlorogenic acid levels through combined GWAS and expression QTL analyses (J. Zhang et al., 2018). Multi-trait and multilocus GWAS approaches have proven particularly effective, with studies in Norway spruce demonstrating that multilocus models can identify more associations than single-locus approaches for wood property traits (Z. Q. Chen et al., 2022).

1.3.3 Climate-Genome Relationships

The relationship between climate and genome variation in forest trees is mediated through complex genotype-by-environment interactions that shape patterns of local adaptation across environmental gradients. These interactions represent fundamental mechanisms by which tree populations maintain fitness across heterogeneous landscapes, with genetic variants showing differential performance under varying climatic conditions (Y. Li et al., 2017). Understanding these climate-genome relationships is essential for predicting population responses to climate change and identifying structural variants that contribute to adaptive capacity. The magnitude and direction of genotype-by-environment interactions can vary substantially among traits, populations, and environmental contexts, creating complex patterns of adaptive variation that require sophisticated analytical approaches to dissect.

Genotype-by-environment interactions in forest trees frequently manifest as changes in genetic correlations across environmental gradients, indicating that different sets of genes or alleles contribute to fitness under different climatic conditions (Poupon et al., 2021). Studies have documented genetic correlations ranging from strongly negative to highly positive depending on environmental contrasts. These patterns suggest that populations harbor extensive conditional genetic variation that enables adaptive responses to environmental heterogeneity.

Climate variables driving selection operate through direct physiological constraints and indirect effects on competitive interactions and resource availability. Temperature-related variables frequently show the strongest associations, particularly minimum temperature of the coldest month influencing growth cessation timing and cold hardiness (Aitken et al., 2008). Precipitation seasonality and drought stress indices represent major selective pressures, particularly under climate change (T. Li et al., 2024). The interaction between temperature and moisture availability creates complex selective landscapes where optimal genotypes must balance competing demands for growth, survival, and reproduction.

Seasonal adaptation represents one of the most clearly documented examples of climate-driven genomic differentiation in forest trees, reflecting the critical importance of synchronizing growth and dormancy cycles with local temperature regimes (Aitken et al. 2008). Timing of bud break and growth cessation represents a fundamental life-history trade-off between maximizing growth and avoiding frost damage. Recent studies have identified major-effect loci such as PtFT2 that explain substantial proportions of phenological variation. Growth-defense trade-offs intensify under climate change as trees must allocate limited resources between growth opportunities and stress resistance (Carnicer et al., 2011). Climate change is intensifying these trade-offs by simultaneously increasing growth opportunities through longer growing seasons and elevated CO_2, while also increasing stress from drought, heat, and pest outbreaks (Carnicer et al. 2011). Genomic studies of growth-defense trade-offs in *Populus* species have revealed complex networks of genes involved in carbon allocation, secondary metabolite production, and stress signaling that show differential expression and allelic variation across environmental gradients.

Temperature response mechanisms in forest trees involve both short-term acclimation processes and long-term evolutionary adaptations that operate through distinct but interconnected genomic pathways. Cold acclimation pathways regulate the biochemical and physiological changes that prepare trees for winter survival, involving the upregulation of protective compounds, modification of membrane composition, and adjustment of metabolic processes (Satyakam et al., 2022). Warming temperatures affect these processes by altering the timing and intensity of cold acclimation signals, potentially disrupting the synchronization between seasonal cues and physiological preparation. Heat tolerance mechanisms involve different sets of genes controlling protein stability, membrane fluidity, and osmotic adjustment, with genetic variation in these pathways contributing to differential performance under elevated temperatures (Saxe et al., 2001).

Precipitation response mechanisms operate through complex physiological and developmental processes that influence water uptake, storage, and utilization efficiency across varying moisture conditions. Trees have evolved sophisticated hydraulic systems that must balance the competing demands of maximizing carbon gain through transpiration while avoiding catastrophic xylem cavitation during drought periods (Li et al. 2024). Genetic variation in hydraulic architecture, including xylem vessel dimensions, pit membrane properties, and hydraulic safety margins, contributes to differential drought tolerance among populations and species. Water use efficiency represents another key adaptive trait, with genetic variation in photosynthetic biochemistry and stomatal regulation enabling populations to optimize carbon gain per unit of water lost under different precipitation regimes.

The molecular mechanisms underlying climate responses in forest trees increasingly appear to involve complex regulatory networks rather than simple Mendelian traits, with extensive pleiotropy and epistatic interactions among climate-responsive genes (Aitken et al. 2008). Transcriptomic studies have revealed that climate stress responses involve the coordinated regulation of hundreds to thousands of genes across multiple biological pathways, suggesting that adaptive variation may involve regulatory changes that alter entire gene networks rather than coding sequence variation in individual genes. Epigenetic mechanisms also appear to play important roles in climate responses, providing mechanisms for environmental memory and transgenerational plasticity that may enhance adaptive capacity. As genomic tools continue to advance, the integration of SV detection with climate association studies promises to reveal additional layers of adaptive variation that have been missed by SNP-based approaches.

1.4 Network Biology Methods

1.4.1 Multi-omics Integration

The interpretation of GWAS results remains one of the most significant challenges in modern genomics. While GWAS has successfully identified thousands of genetic loci, functional interpretation typically requires integration of multiple biological information

layers (Lessard et al., 2024). The challenge is particularly acute as most GWAS signals fall within non-coding regions, necessitating sophisticated analytical frameworks. Network-based approaches have emerged as powerful tools for functional annotation, leveraging the principle that genes associated with similar phenotypes tend to interact within common biological networks (Hasin et al., 2017). Recent developments have demonstrated improved accuracy over traditional expression-based approaches, particularly when incorporating multiple types of molecular interactions (Agamah et al., 2022). By incorporating protein-protein interaction networks, gene co-expression patterns, and regulatory relationships, network-based annotation can identify functional modules and pathways that may be disrupted by genetic variation. This approach is particularly valuable for complex traits where multiple genetic variants contribute small individual effects but collectively impact shared biological processes.

Multiomics integration strategies have been categorized into early, intermediate, and late fusion approaches (Ballard et al., 2024). Early integration concatenates datasets into a single feature matrix, intermediate integration uses canonical correlation analysis or joint matrix factorization, and late integration analyzes each layer separately before combining results (Kang et al., 2022). Finally, pathway enrichment analysis serves as a critical bridge between statistical associations and biological interpretation. Over-representation analysis tests whether genes occur more frequently than expected by chance (Huang et al., 2009), while Gene Set Enrichment Analysis considers entire ranked gene lists (Subramanian et al., 2005). Pathway topology-based methods incorporate gene interactions for mechanistically informed interpretations (Khatri et al., 2012).

1.4.2 GRIN and MENTOR Algorithms

The challenge of distinguishing true positive associations from false positives in genome-wide association studies represents a fundamental limitation in translating statistical genetics findings into biological insights. While relaxing significance thresholds can increase statistical power and capture additional biologically relevant associations, this approach inevitably introduces noise that can obscure genuine biological signals. GRIN

(Gene set Refinement through Interacting Networks) addresses this challenge by leveraging biological network topology to systematically identify and remove false positive genes from GWAS-derived gene sets, allowing researchers to confidently relax significance thresholds while maintaining analytical precision (Sullivan, Lane, et al., 2024). The algorithm is based on the principle that true positive genes underlying complex traits are more likely to be functionally interconnected through biological networks compared to randomly associated false positive genes, which lack systematic relationships to the trait of interest.

The GRIN methodology operates through a two-stage process that combines network-based gene scoring with statistical refinement to optimize gene set composition. In the first stage, all genes within the experimental gene set are ranked based on their connectivity within a comprehensive multiplex biological network, with connectivity scores calculated using random walk with restart algorithms that capture both direct and indirect functional relationships. These connectivity rankings are compared against empirical null distributions generated from randomly sampled gene sets of equivalent size, providing a baseline for assessing the biological coherence of the experimental gene set. The second stage employs a sliding window approach combined with Mann-Whitney U tests to identify the optimal cutoff point where the experimental gene set's network connectivity becomes statistically indistinguishable from random expectation, indicating the transition from true positive to false positive genes (Sullivan, Lane, et al., 2024). The power of GRIN lies in its utilization of multiplex networks that integrate diverse types of biological evidence, ranging from protein-protein interactions and gene co-expression relationships to regulatory networks and metabolic pathways. These multiplex networks capture millions of empirically derived relationships that reflect the cumulative knowledge of the biological research community, providing a comprehensive foundation for assessing functional relatedness among genes. The incorporation of tissue-specific networks, including novel predictive networks generated through explainable AI methodologies, further enhances GRIN's ability to identify biologically relevant connections that may not be captured in traditional literature-derived networks.

MENTOR (Multiplex Embedding of Networks for Team-Based Omics Research) represents a complementary but distinct approach to network-based functional interpretation, focusing on mechanistic understanding rather than gene set refinement (Sullivan, Miller, et al., 2024). While GRIN emphasizes the identification and removal of false positive associations, MENTOR aims to reveal the organizational structure and functional relationships within validated gene sets through sophisticated network embedding techniques. The algorithm employs random walk with restart exploration across multiplex networks to generate embedding vectors for each gene, with these vectors capturing the network neighborhood and functional context of individual genes within the broader biological system (Sullivan, Miller, et al., 2024). The MENTOR workflow begins with the construction of gene-specific embedding vectors through systematic exploration of multiplex network topology, where random walks starting from each gene generate rank-ordered lists of all other genes based on their network proximity and functional similarity. These exploration profiles are truncated at statistically determined elbow points to focus on the most functionally relevant connections, and pairwise correlations between the rank order gene profiles are calculated using Spearman correlation coefficients. The resulting correlation matrix is transformed into a distance matrix and subjected to hierarchical clustering to generate dendrograms that reveal functional relationships and modular organization within the gene set.

Network-based gene set refinement and false positive reduction represent critical advances in addressing the precision-recall tradeoff that has long limited GWAS interpretation. Traditional approaches often force researchers to choose between high precision with limited discovery potential or increased sensitivity with elevated false positive rates. The GRIN algorithm aims to change this paradigm by providing a principled method for maintaining precision while expanding the scope of discovery through relaxed significance thresholds. Validation studies across multiple complex traits and independent cohorts have demonstrated that GRIN-refined gene sets exhibit improved replication rates and biological coherence compared to traditional approaches, confirming the method's ability to enhance both statistical and biological validity of

GWAS findings.

Multiplex network approaches for biological discovery represent a fundamental shift from single-network analyses toward more comprehensive and realistic representations of biological systems. Traditional network analyses typically focus on single interaction types, such as protein-protein interactions or gene co-expression, potentially missing critical functional relationships that span multiple biological scales and interaction modalities. Multiplex networks address this limitation by simultaneously incorporating multiple types of biological relationships within a unified analytical framework, allowing algorithms like GRIN and MENTOR to leverage the complementary information provided by different evidence types (Valous et al., 2024). The integration of GRIN and MENTOR within comprehensive analytical workflows represents a paradigm shift toward evidence-based functional interpretation of genomic variation. By combining statistical refinement with mechanistic interpretation, these algorithms provide a systematic approach for moving from genetic association to biological understanding, addressing fundamental challenges that have limited the translation of GWAS findings into actionable biological knowledge.

1.4.3 Functional Networks in Plants

Gene regulatory networks in forest trees encompass intricate transcriptional control governing long-lived organisms adapted to diverse environments. Recent advances have begun illuminating regulatory architecture underlying key traits, revealing sophisticated networks controlling wood formation, stress responses, and reproductive development (Feng et al., 2024). Research has identified key transcriptional regulators coordinating entire biosynthetic pathways involved in lignin and cellulose production. Co-expression networks and module detection have emerged as powerful approaches for understanding functional genome organization. Weighted Gene Co-expression Network Analysis (WGCNA) has become a common approach, leveraging the principle that genes with similar expression patterns likely share functional relationships (Langfelder & Horvath, 2008). This approach identifies modules corresponding to biological pathways, enabling

gene function prediction through guilt-by-association. In forest trees, co-expression networks have been instrumental in identifying genes involved in wood formation, stress responses, and seasonal adaptation. Advanced implementations like multiWGCNA enable condition-specific network construction, revealing how co-expression relationships change across biological contexts (Van den Broeck et al., 2020).

Protein-protein interaction networks provide the physical foundation for signal transduction and regulatory control. Plant networks encompass diverse molecular interactions, from stable structural complexes to transient regulatory interactions mediating hormone signaling (Struk et al., 2019). Comprehensive interactome mapping has revealed network properties reflecting modular organization of plant biology.

Network validation requires multiple criteria including statistical significance, reproducibility, and biological plausibility (Agamah et al., 2022). Effective interpretation strategies combine computational analysis with domain expertise, leveraging gene ontology annotations and pathway databases. As networks become increasingly comprehensive, they enable new approaches to crop improvement and forest management.

1.5 Structural Variation and Bioenergy Traits

1.5.1 SV Impacts on Bioenergy-Relevant Traits

SVs represent a significant but understudied source of genetic variation affecting bioenergy crop performance. SVs can have large phenotypic effects due to their potential to disrupt gene function, alter gene dosage, or modify regulatory landscapes across multiple genes simultaneously. In the context of bioenergy applications, SVs may be particularly important given the complex, polygenic nature of key feedstock traits such as biomass yield, stress tolerance, and cell wall composition.

1.5.1.1 SV Effects on Lignin Composition and Cell Wall Chemistry

Cell wall chemistry represents one of the most economically important trait categories, as lignin content and composition directly impact bioconversion efficiency (Happs et al., 2021). Natural genetic variation studies in *P. trichocarpa* have demonstrated substantial heritability for cell wall traits, with some variants showing large effect sizes comparable to transgenic approaches (Bhagia et al., 2016). GWAS has identified novel candidate genes for lignin structure suggesting SVs may contribute to missing heritability (Bryant et al., 2023). High-resolution mapping revealed that allelic variants often cluster in genomic regions containing secondary cell wall biosynthesis gene families (Muchero et al., 2015).

1.5.1.2 Copy Number Variation in Stress Tolerance Pathways

CNV represents a particularly important SV class for stress tolerance. Gene duplication events provide raw material for functional divergence, allowing plants to maintain essential functions while evolving specialized stress responses. Studies of abiotic stress tolerance have revealed that key regulatory gene families such as DREB transcription factors, LEA proteins, and stress-responsive kinases are frequently organized in tandem arrays that facilitate copy number variation (Shinozaki & Yamaguchi-Shinozaki, 2007). In cotton, genome-wide analyses have demonstrated that copy number variations can be particularly useful for identifying phenotypic variations in complex abiotic stress tolerance traits that are not captured by single nucleotide polymorphisms (Mahmood et al., 2019). Natural genetic variation studies across crop species have shown that stress tolerance genes often exhibit higher copy number variability among diverse lines compared to housekeeping genes, suggesting ongoing evolutionary pressure for dosage optimization in stress response pathways (Xu et al., 2014).

1.5.1.3 SVs Affecting Water Use Efficiency and Drought Tolerance

Water use efficiency and drought tolerance are critical traits for bioenergy crops intended for marginal land cultivation, as these environments often experience water limitation and unpredictable precipitation patterns. SVs may affect these traits through impacts on root

architecture genes, stomatal development pathways, or osmotic adjustment mechanisms. Recent modeling approaches have identified overexpression targets for improving drought tolerance in *P. trichocarpa*, suggesting that copy number increases in specific gene families could enhance water stress responses (Barbosa et al., 2024).

The complex genetic architecture of drought tolerance traits makes them particularly susceptible to SV effects, as coordinated expression changes across multiple pathways may be required for effective stress responses. Recent transcriptomic studies in maize have shown that early drought-responsive genes display higher copy number variation among diverse lines and exhibit stronger associations with drought tolerance than late-responsive genes (He et al., 2020). In wheat, comprehensive reviews of genetic engineering approaches for drought tolerance have highlighted the importance of regulatory SVs that can coordinate expression changes across entire stress response networks rather than affecting individual genes (Trono & Pecchioni, 2022). Natural genetic variation studies have demonstrated that SVs in transcription factor genes, particularly those in the DREB, NAC, and MYB families, can have large effects on water use efficiency through their regulation of stomatal development and osmotic adjustment pathways (Bailey-Serres et al., 2019).

1.5.1.4 SV Contribution to Growth and Biomass Production Traits

Biomass yield represents the primary economic driver for bioenergy crop adoption, making growth-related SVs of particular interest for crop improvement programs. Growth traits are typically controlled by large numbers of small-effect loci, but SVs may contribute disproportionately to trait variation due to their potential for large phenotypic effects. Economic analyses have demonstrated that both yield and composition variation significantly impact the viability of bioenergy production systems, highlighting the importance of identifying genetic variants that affect these traits (Happs et al., 2021).

Recent integrated QTL and association mapping studies in *Populus* have revealed that SVs, including copy number variations and large indels, often co-locate with major

biomass QTL hotspots and can explain substantial proportions of phenotypic variation in growth traits (Du et al., 2016). Analysis of natural and induced variation in hybrid *Populus* has demonstrated that dosage-related SVs can have distinct effects on quantitative traits compared to allelic sequence variation, with most growth-related QTLs being associated with dosage-related groups rather than single nucleotide changes (Piot et al., 2024). Genome-wide studies have shown that QTLs detected for biomass-related traits in *Populus* explain an average of 26% of observed phenotypic variation, suggesting that SVs may contribute to the "missing heritability" problem commonly observed in quantitative trait mapping (Rae et al., 2009).

1.5.1.5 Regulatory Variation Affecting Biofuel-Relevant Metabolic Pathways

Regulatory SVs may have particularly broad impacts due to their potential to affect multiple genes within metabolic networks. Studies reveal hierarchical genetic networks where SVs in master regulatory genes can control large transcriptome portions (Dash et al., 2015). Global analyses show regulatory SVs affecting transcription factors can reprogram entire metabolic pathways (Luo et al., 2015). Systems biology approaches have identified regulatory modules where CNV in key transcription factors can modulate the expression of hundreds of downstream target genes, providing a mechanism by which SVs can have broad impacts on bioenergy-relevant traits.

1.5.2 Genomic Approaches for Bioenergy Crop Improvement

The identification and characterization of SVs affecting bioenergy traits provides a foundation for developing targeted crop improvement strategies. Unlike traditional breeding approaches that rely on phenotypic selection across multiple generations, genomic approaches enable precise manipulation of the genetic variants underlying key traits. This precision is particularly valuable for bioenergy applications, where the complex, multigenic nature of feedstock quality and stress tolerance traits makes them difficult to improve through conventional methods alone.

1.5.2.1 Marginal Land Adaptation and Stress Tolerance Breeding

The cultivation of bioenergy crops on marginal agricultural lands represents a critical strategy for avoiding competition with food production while expanding renewable energy feedstock availability. In addition, recent advances have redefined beyond biophysical limitations to include economic and environmental factors (Khanna et al., 2021), Understanding adaptation to marginal land requires discovering the genetic mechanisms that underly multiple stress tolerances. Traditional breeding approaches for stress tolerance have been limited by the complex genetic architecture of stress response traits and the difficulty of accurately phenotyping stress tolerance under field conditions. Genomic approaches offer significant advantages by enabling the identification of specific genes and regulatory networks that control stress responses, allowing for more targeted and efficient improvement strategies (Quinn et al., 2015). Comprehensive reviews have identified multiple bioenergy crops capable of thriving on marginal lands, including poplar species that can tolerate drought, flooding, salinity, and temperature extremes (Quinn et al., 2015).

1.5.2.2 Genetic Engineering Targets for Improved Feedstock Quality

Identifying SVs affecting cell wall composition and biomass yield provides specific targets for genetic engineering approaches to improve feedstock quality. Unlike broad-spectrum approaches that modify entire pathways, targeting specific SVs allows for precise manipulation of individual traits while potentially minimizing unintended consequences on plant development or stress tolerance. Recent advances in genome editing technologies, particularly CRISPR-Cas systems, have made it possible to precisely modify SVs in crop species. CRISPR-mediated editing of lignin biosynthesis genes in poplar has demonstrated significant reductions in lignin content (up to 29%) while maintaining normal growth, leading to improved saccharification efficiency for biofuel production (Jang et al., 2021). Similar approaches in other crops have shown that CRISPR knockout of lignin-related genes like COMT can improve lignocellulosic biofuel production by reducing lignin content and enhancing biomass digestibility (J. H. Lee et al., 2021). Recent reviews of CRISPR applications highlight the technology's potential

for creating multiplex modifications that simultaneously target multiple genes in lignin biosynthetic pathways (Sulis et al., 2023). Advanced genome editing approaches, including prime editing and base editing, continue to expand the precision and scope of genetic modifications possible in crop improvement programs (F. Chen et al., 2024).

1.5.2.3 Drought Tolerance Mechanisms and Water Stress Responses

Water limitation represents one of the most significant constraints for bioenergy crop production on marginal lands. Understanding the genetic mechanisms that underly drought tolerance provides molecular targets that can be manipulated to enhance water stress tolerance in bioenergy crops.

Multi-tissue modeling approaches have identified overexpression targets for improving *P. trichocarpa* drought tolerance (Barbosa et al., 2024), and studies of drought responses have revealed genotype-specific transcriptome and metabolome changes (Hamanishi et al., 2015). Studies in *P. nigra* have shown that drought tolerance involves complex interactions between stomatal regulation, water use efficiency, and genomic plasticity (Viger et al., 2016). Reviews of drought stress responses in poplar highlight the complex molecular networks involved in water deficit responses, including embolism formation and repair, long-term drought effects, and phenotypic plasticity mechanisms (Rosso et al., 2023).

1.5.2.4 Growth-Defense Tradeoffs in Bioenergy Applications

One of the fundamental challenges in developing bioenergy crops is balancing growth and biomass production with stress tolerance and defense responses. Understanding the genetic basis of growth-defense tradeoffs provides opportunities for developing bioenergy crops that can maintain high productivity while retaining adequate stress tolerance.

Recent reviews highlight how hormone crosstalk involving jasmonic acid, salicylic acid, and gibberellin signaling mediates growth-defense tradeoffs and provides a foundation

for breeding strategies that optimize this balance (Huot et al., 2014). Multiomic approaches have revealed coordinated resource allocation patterns that are enabled by phytohormone crosstalk and transcriptional signal cascades, providing mechanistic insights into how plants balance competing demands for growth and defense (Monson et al., 2022; Züst & Agrawal, 2017).

1.5.2.5 Translating Genomic Insights to Crop Improvement Strategies

One of the goals of SV research in bioenergy crops is to translate genomic discoveries into practical crop improvement strategies that can be implemented by breeding programs and biotechnology companies. This translation requires integrating multiple types of genetic and genomic information, including SV discoveries, gene expression data, and phenotypic trait associations. Successful translation of genomic insights requires close collaboration between basic researchers, crop breeders, and biotechnology developers to ensure that research priorities align with practical improvement goals. The Center for Bioenergy Innovation (CBI) represents one example of this collaborative approach, bringing together researchers across multiple institutions to develop climate-resilient bioenergy crops for sustainable fuel production. Advanced genomics-assisted breeding approaches, including genomic selection, marker-assisted selection, and haplotype-based breeding are revolutionizing crop improvement by enabling more precise and efficient selection strategies (Crain et al., 2018; Varshney et al., 2021). The development of comprehensive genomic resources provides new opportunities for identifying climate-adaptive alleles and translating this knowledge into practical breeding applications (Campbell et al., 2025).

1.6 Synthesis and Research Gaps

The intersection of SV, population genomics, and climate adaptation represents a rapidly evolving frontier in forest tree genomics. While significant progress has been made in understanding how SNPs contribute to local adaptation (Evans et al., 2014), the role of SVs in adaptive processes remains largely unexplored. This knowledge gap is particularly pronounced in *P. trichocarpa*, where population structure and demographic history create

complex patterns extending beyond SNP-based analyses (Geraldes et al., 2014). Current evidence suggests SVs may play disproportionately important roles in adaptive evolution due to their potential for large phenotypic effects, tendency to occur in functionally important genomic regions, and the ability of some SV types to avoid fitness valleys. However, integration of SV detection with population-level sampling and climate association analyses has been limited by technical challenges and conceptual frameworks traditionally focused on point mutations.

Several critical technical barriers impede comprehensive SV analysis in forest trees. The repetitive nature of tree genomes creates substantial challenges for accurate detection, particularly with short-read sequencing (Sedlazeck et al., 2018). Population structure effects on SV calling have not been systematically addressed, despite evidence that demographic history can significantly influence detection accuracy (Garrison et al., 2018). Computational challenges persist in functional interpretation, as SVs require specialized approaches accounting for complex effects on gene structure and regulation. Forest tree genomics has lagged in adopting comprehensive SV analysis approaches. Most population genomic studies focus exclusively on SNP variation, leaving gaps in understanding genome-wide structural diversity (Neale & Kremer, 2011). The relationship between SV and quantitative traits relevant to climate adaptation and bioenergy applications remains largely unknown (Bryant et al., 2023).

Long-read sequencing technologies represent a promising avenue for advancing SV research. These approaches enable comprehensive detection of complex rearrangements and provide necessary resolution for repetitive regions. Pangenome approaches designed for tree species will be important for capturing population-level structural diversity (Golicz et al., 2020). Integration of multiomics data with SV information will provide new insights into functional consequences. Network-based approaches will be essential for translating discoveries into actionable knowledge (Sullivan, Miller, et al., 2024).

1.7 Conclusion

The literature reveals a significant opportunity at the intersection of SV biology, population genomics, and climate adaptation research in forest trees. While substantial progress has been made understanding population structure and climate associations in *P. trichocarpa* using SNP-based approaches, the role of SVs remains largely unexplored. Current SV detection methods have not been systematically applied to natural tree populations with necessary rigor and scale. Technical challenges associated with SV calling in complex, repetitive genomes have created knowledge gaps extending beyond methodology to fundamental questions about evolutionary dynamics and functional importance of SV in long-lived perennial species. This gap is particularly pronounced for bioenergy applications, where understanding the full spectrum of genomic variation is crucial for developing climate-resilient feedstock varieties.

This dissertation addresses knowledge gaps by developing subpopulation-based SV calling strategies accounting for complex demographic history. Integration of comprehensive SV datasets with climate data using network-based functional analysis represents a novel contribution bridging fundamental population genomics with applied breeding applications. There is an increasing recognition of SV as important contributors to adaptive evolution and the growing need for climate-resilient bioenergy crops. As climate change accelerates and the demand for sustainable energy sources increases, understanding the genetic basis of climate adaptation becomes ever more critical. By developing computational frameworks to effectively integrate SV data with climate associations and functional networks, this work provides methodological advances and biological insights informing future breeding strategies. Here, the characterization of SV in natural *P. trichocarpa* populations will provide insights into the genomic basis of climate adaptation and the evolutionary dynamics of complex genomic rearrangements. The methodological contributions, particularly the development of subpopulation-aware SV calling strategies and network-based functional interpretation frameworks, will be broadly applicable to other forest tree species and complex genome systems. The identification of climate-associated SVs and their functional interpretation will provide

new targets for bioenergy crop improvement and inform breeding strategies for developing climate-resilient varieties.

References

Agamah, F. E., Bayjanov, J. R., Niehues, A., Njoku, K. F., Skelton, M., Mazandu, G. K., Ederveen, T. H. A., Mulder, N., Chimusa, E. R., & ’t Hoen, P. A. C. (2022). Computational approaches for network-based integrative multi-omics analysis. *Frontiers in Molecular Biosciences*, *9*, 967205. https://doi.org/10.3389/FMOLB.2022.967205

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

Alberto, F. J., Aitken, S. N., Alía, R., González-Martínez, S. C., Hänninen, H., Kremer, A., Lefèvre, F., Lenormand, T., Yeaman, S., Whetten, R., & Savolainen, O. (2013). Potential for evolutionary responses to climate change - evidence from tree populations. *Global Change Biology*, *19*(6), 1645–1661. https://doi.org/10.1111/GCB.12181

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/GR.094052.109

Bailey-Serres, J., Parker, J. E., Ainsworth, E. A., Oldroyd, G. E. D., & Schroeder, J. I. (2019). Genetic strategies for improving crop yields. *Nature*, *575*(7781), 109–118. https://doi.org/10.1038/S41586-019-1679-0

Ballard, J. L., Wang, Z., Li, W., Shen, L., & Long, Q. (2024). Deep learning-based approaches for multi-omics data integration and analysis. *BioData Mining*, *17*(1), 1–29. https://doi.org/10.1186/S13040-024-00391-Z

Barbosa, J. S. C., Schroeder, W. L., Suthers, P. F., Jawdy, S. S., Chen, J. G., Muchero, W., & Maranas, C. D. (2024). A multi-tissue genome-scale model of Populus trichocarpa elucidates overexpression targets for improving drought tolerance. *In Silico Plants*, *6*(1). https://doi.org/10.1093/INSILICOPLANTS/DIAE007

Belyeu, J. R., Chowdhury, M., Brown, J., Pedersen, B. S., Cormier, M. J., Quinlan, A. R., & Layer, R. M. (2021). Samplot: a platform for structural variant visual

validation and automated filtering. *Genome Biology*, *22*(1), 1–13. https://doi.org/10.1186/S13059-021-02380-5

Belyeu, J. R., Nicholas, T. J., Pedersen, B. S., Sasani, T. A., Havrilla, J. M., Kravitz, S. N., Conway, M. E., Lohman, B. K., Quinlan, A. R., & Layer, R. M. (2018). SV-plaudit: A cloud-based framework for manually curating thousands of structural variants. *GigaScience*, *7*(7), 1–7. https://doi.org/10.1093/GIGASCIENCE/GIY064

Bhagia, S., Muchero, W., Kumar, R., Tuskan, G. A., & Wyman, C. E. (2016). Natural genetic variability reduces recalcitrance in poplar. *Biotechnology for Biofuels*, *9*(1), 1–12. https://doi.org/10.1186/S13068-016-0521-2

Bryant, N., Zhang, J., Feng, K., Shu, M., Ployet, R., Chen, J. G., Muchero, W., Yoo, C. G., Tschaplinski, T. J., Pu, Y., & Ragauskas, A. J. (2023). Novel candidate genes for lignin structure identified through genome-wide association study of naturally varying Populus trichocarpa. *Frontiers in Plant Science*, *14*, 1153113. https://doi.org/10.3389/FPLS.2023.1153113

Cameron, D. L., Di Stefano, L., & Papenfuss, A. T. (2019). Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nature Communications*, *10*(1), 3240. https://doi.org/10.1038/s41467-019-11146-4

Campbell, Q., Bedford, J. A., Yu, Y., Halpin-McCormick, A., Castaneda-Alvarez, N., Runck, B., Neyhart, J., Ewing, P., Ortiz-Barrientos, D., Gao, L., Wang, D., Chapman, M. A., Rieseberg, L. H., & Kantar, M. B. (2025). Agricultural landscape genomics to increase crop resilience. *Plant Communications*, *6*(2). https://doi.org/10.1016/J.XPLC.2025.101260

Capblancq, T., & Forester, B. R. (2021). Redundancy analysis: A Swiss Army Knife for landscape genomics. In *Methods in Ecology and Evolution* (Vol. 12, Issue 12, pp. 2298–2309). John Wiley & Sons, Ltd. https://doi.org/10.1111/2041-210X.13722

Carnicer, J., Coll, M., Ninyerola, M., Pons, X., Sánchez, G., & Peñuelas, J. (2011). Widespread crown condition decline, food web disruption, and amplified tree mortality with increased climate change-type drought. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(4), 1474–1478.

https://doi.org/10.1073/PNAS.1010070108

Carvalho, C. M. B., & Lupski, J. R. (2016). Mechanisms underlying structural variant formation in genomic disorders. *Nature Reviews Genetics 2016 17:4*, *17*(4), 224–238. https://doi.org/10.1038/nrg.2015.25

Chen, F., Chen, L., Yan, Z., Xu, J., Feng, L., He, N., Guo, M., Zhao, J., Chen, Z., Chen, H., Yao, G., & Liu, C. (2024). Recent advances of CRISPR-based genome editing for enhancing staple crops. *Frontiers in Plant Science*, *15*, 1478398. https://doi.org/10.3389/FPLS.2024.1478398/XML

Chen, K., Wallis, J. W., McLellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., McGrath, S. D., Wendl, M. C., Zhang, Q., Locke, D. P., Shi, X., Fulton, R. S., Ley, T. J., Wilson, R. K., Ding, L., & Mardis, E. R. (2009). BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. *Nature Methods*, *6*(9), 677–681. https://doi.org/10.1038/nmeth.1363

Chen, Y. C., Liu, T., Yu, C. H., Chiang, T. Y., & Hwang, C. C. (2013). Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly. *PLOS ONE*, *8*(4), e62856. https://doi.org/10.1371/JOURNAL.PONE.0062856

Chen, Z. Q., Zan, Y., Zhou, L., Karlsson, B., Tuominen, H., García-Gil, M. R., & Wu, H. X. (2022). Genetic architecture behind developmental and seasonal control of tree growth and wood properties in Norway spruce. *Frontiers in Plant Science*, *13*, 927673. https://doi.org/10.3389/FPLS.2022.927673

Chhetri, H. B., Macaya-Sanz, D., Kainer, D., Biswal, A. K., Evans, L. M., Chen, J., Collins, C., Hunt, K., Mohanty, S. S., Rosenstiel, T., Ryno, D., Winkeler, K., Yang, X., Jacobson, D., Mohnen, D., Muchero, W., Strauss, S. H., Tschaplinski, T. J., Tuskan, G. A., & DiFazio, S. P. (2019). Multitrait genome-wide association analysis of Populus trichocarpa identifies key polymorphisms controlling morphological and physiological traits. *New Phytologist*, *223*(1), 293–309. https://doi.org/10.1111/nph.15777

Crain, J., Mondal, S., Rutkoski, J., Singh, R. P., & Poland, J. (2018). Combining High-Throughput Phenotyping and Genomic Information to Increase Prediction and Selection Accuracy in Wheat Breeding. *The Plant Genome*, *11*(1), 170043.

https://doi.org/10.3835/PLANTGENOME2017.05.0043

Daly, C., Halbleib, M., Smith, J. I., Gibson, W. P., Doggett, M. K., Taylor, G. H., Curtis, J., & Pasteris, P. P. (2008). Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *International Journal of Climatology*, *28*(15), 2031–2064. https://doi.org/10.1002/JOC.1688

Dash, M., Yordanov, Y. S., Georgieva, T., Kumari, S., Wei, H., & Busov, V. (2015). A systems biology approach identifies new regulators of poplar root development under low nitrogen. *Plant Journal*, *84*(2), 335–346. https://doi.org/10.1111/TPJ.13002;WGROUP:STRING:PUBLICATION

De Coster, W., De Rijk, P., De Roeck, A., De Pooter, T., D'Hert, S., Strazisar, M., Sleegers, K., & Van Broeckhoven, C. (2019). Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. *Genome Research*, *29*(7), 1178–1187. https://doi.org/10.1101/gr.244939.118

Diaz-Papkovich, A., Anderson-Trocmé, L., Ben-Eghan, C., & Gravel, S. (2019). UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS Genetics*, *15*(11), e1008432. https://doi.org/10.1371/journal.pgen.1008432

Du, Q., Lu, W., Quan, M., Xiao, L., Song, F., Li, P., Zhou, D., Xie, J., Wang, L., & Zhang, D. (2018). Genome-wide association studies to improve wood properties: Challenges and prospects. *Frontiers in Plant Science*, *871*, 412335. https://doi.org/10.3389/FPLS.2018.01912

English, A. C., Salerno, W. J., Hampton, O. A., Gonzaga-Jauregui, C., Ambreth, S., Ritter, D. I., Beck, C. R., Davis, C. F., Dahdouli, M., Ma, S., Carroll, A., Veeraraghavan, N., Bruestle, J., Drees, B., Hastie, A., Lam, E. T., White, S., Mishra, P., Wang, M., … Gibbs, R. A. (2015). Assessing structural variation in a personal genome-towards a human reference diploid genome. *BMC Genomics*, *16*(1), 286. https://doi.org/10.1186/s12864-015-1479-3

Escaramís, G., Docampo, E., & Rabionet, R. (2015). A decade of structural variants: Description, history and methods to detect structural variation. *Briefings in Functional Genomics*, *14*(5), 305–314. https://doi.org/10.1093/bfgp/elv014

Evans, L. M., Slavov, G. T., Rodgers-Melnick, E., Martin, J., Ranjan, P., Muchero, W., Brunner, A. M., Schackwitz, W., Gunter, L., Chen, J. G., Tuskan, G. A., & Difazio, S. P. (2014). Population genomics of Populus trichocarpa identifies signatures of selection and adaptive trait associations. *Nature Genetics*, *46*(10), 1089–1096. https://doi.org/10.1038/ng.3075

Feng, L., & Du, F. K. (2022). Landscape Genomics in Tree Conservation Under a Changing Environment. *Frontiers in Plant Science*, *13*, 822217. https://doi.org/10.3389/FPLS.2022.822217

Fitzpatrick, M. C., & Keller, S. R. (2015). Ecological genomics meets community-level modelling of biodiversity: Mapping the genomic landscape of current and future environmental adaptation. *Ecology Letters*, *18*(1), 1–16. https://doi.org/10.1111/ELE.12376;SUBPAGE:STRING:FULL

Frichot, E., Schoville, S. D., Bouchard, G., & François, O. (2013). Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models. *Molecular Biology and Evolution*, *30*(7), 1687–1699. https://doi.org/10.1093/MOLBEV/MST063

Gabur, I., Chawla, H. S., Snowdon, R. J., & Parkin, I. A. P. (2019). Connecting genome structural variation with complex traits in crop plants. *Theoretical and Applied Genetics*, *132*(3), 733–750. https://doi.org/10.1007/s00122-018-3233-0

Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., Jones, W., Garg, S., Markello, C., Lin, M. F., Paten, B., & Durbin, R. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, *36*(9), 875–881. https://doi.org/10.1038/NBT.4227;TECHMETA=45;SUBJMETA=114,208,2785,514,631,794;KWRD=COMPUTATIONAL+BIOLOGY+AND+BIOINFORMATICS,GENOME+INFORMATICS,SEQUENCING,SOFTWARE

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

https://doi.org/10.1111/evo.12497

Girirajan, S., & Eichler, E. E. (2010). Phenotypic variability and genetic susceptibility to genomic disorders. *Human Molecular Genetics*, *19*(R2), R176–R187. https://doi.org/10.1093/HMG/DDQ366

Golicz, A. A., Bayer, P. E., Bhalla, P. L., Batley, J., & Edwards, D. (2020). Pangenomics Comes of Age: From Bacteria to Plant and Animal Applications. *Trends in Genetics*, *36*(2), 132–145. https://doi.org/10.1016/J.TIG.2019.11.006

Guerra, F. P., Suren, H., Holliday, J., Richards, J. H., Fiehn, O., Famula, R., Stanton, B. J., Shuren, R., Sykes, R., Davis, M. F., & Neale, D. B. (2019). Exome resequencing and GWAS for growth, ecophysiology, and chemical and metabolomic composition of wood of Populus trichocarpa. *BMC Genomics*, *20*(1), 875. https://doi.org/10.1186/s12864-019-6160-9

Hamanishi, E. T., Barchet, G. L. H., Dauwe, R., Mansfield, S. D., & Campbell, M. M. (2015). Poplar trees reconfigure the transcriptome and metabolome in response to drought in a genotype- and time-of-day-dependent manner. *BMC Genomics*, *16*(1), 1–16. https://doi.org/10.1186/S12864-015-1535-Z

Han, B., Kang, H. M., & Eskin, E. (2009). Rapid and Accurate Multiple Testing Correction and Power Estimation for Millions of Correlated Markers. *PLOS Genetics*, *5*(4), e1000456. https://doi.org/10.1371/JOURNAL.PGEN.1000456

Happs, R. M., Bartling, A. W., Doeppke, C., Harman-Ware, A. E., Clark, R., Webb, E. G., Biddy, M. J., Chen, J. G., Tuskan, G. A., Davis, M. F., Muchero, W., & Davison, B. H. (2021). Economic impact of yield and composition variation in bioenergy crops: Populus trichocarpa. *Biofuels, Bioproducts and Biorefining*, *15*(1), 176–188. https://doi.org/10.1002/BBB.2148

Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biology 2017 18:1*, *18*(1), 1–15. https://doi.org/10.1186/S13059-017-1215-1

He, C., Du, Y., Fu, J., Zeng, E., Park, S., White, F., Zheng, J., & Liu, S. (2020). Early Drought-Responsive Genes Are Variable and Relevant to Drought Tolerance. *G3 Genes|Genomes|Genetics*, *10*(5), 1657–1670. https://doi.org/10.1534/G3.120.401199

Ho, S. S., Urban, A. E., & Mills, R. E. (2020). Structural variation in the sequencing era. In *Nature Reviews Genetics* (Vol. 21, Issue 3, pp. 171–189). Nature Publishing Group. https://doi.org/10.1038/s41576-019-0180-9

Holliday, J. A., Zhou, L., Bawa, R., Zhang, M., & Oubida, R. W. (2016). Evidence for extensive parallelism but divergent genomic architecture of adaptation along altitudinal and latitudinal gradients in Populus trichocarpa. *New Phytologist*, *209*(3), 1240–1251. https://doi.org/10.1111/nph.13643

Howe, G. T., Klápště, J., & Grattapaglia, D. (2022). Twelve Years into Genomic Selection in Forest Trees: Climbing the Slope of Enlightenment of Marker Assisted Tree Breeding. *Forests 2022, Vol. 13, Page 1554*, *13*(10), 1554. https://doi.org/10.3390/F13101554

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, *4*(1), 44–57. https://doi.org/10.1038/NPROT.2008.211

Huot, B., Yao, J., Montgomery, B. L., & He, S. Y. (2014). Growth–Defense Tradeoffs in Plants: A Balancing Act to Optimize Fitness. *Molecular Plant*, *7*(8), 1267–1287. https://doi.org/10.1093/MP/SSU049

Ingvarsson, P. K., Garcia, M. V., Luquez, V., Hall, D., & Jansson, S. (2008). Nucleotide Polymorphism and Phenotypic Associations Within and Around the phytochrome B2 Locus in European Aspen (Populus tremula, Salicaceae). *Genetics*, *178*(4), 2217–2226. https://doi.org/10.1534/GENETICS.107.082354

Jang, H. A., Bae, E. K., Kim, M. H., Park, S. J., Choi, N. Y., Pyo, S. W., Lee, C., Jeong, H. Y., Lee, H., Choi, Y. I., & Ko, J. H. (2021). Crispr-knockout of cse gene improves saccharification efficiency by reducing lignin content in hybrid poplar. *International Journal of Molecular Sciences*, *22*(18), 9750. https://doi.org/10.3390/IJMS22189750/S1

Joo, J. W. J., Hormozdiari, F., Han, B., & Eskin, E. (2016). Multiple testing correction in linear mixed models. *Genome Biology*, *17*(1), 62. https://doi.org/10.1186/S13059-016-0903-6

Kang, M., Ko, E., & Mersha, T. B. (2022). A roadmap for multi-omics data integration

using deep learning. In *Briefings in Bioinformatics* (Vol. 23, Issue 1, pp. 1–16). https://doi.org/10.1093/bib/bbab454

Khanna, M., Chen, L., Basso, B., Cai, X., Field, J. L., Guan, K., Jiang, C., Lark, T. J., Richard, T. L., Spawn-Lee, S. A., Yang, P., & Zipp, K. Y. (2021). Redefining marginal land for bioenergy crop production. *GCB Bioenergy*, *13*(10), 1590–1609. https://doi.org/10.1111/GCBB.12877

Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLOS Computational Biology*, *8*(2), e1002375. https://doi.org/10.1371/JOURNAL.PCBI.1002375

Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M., & Kamatani, Y. (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biology*, *20*(1), 8–11. https://doi.org/10.1186/s13059-019-1720-5

Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*(1), 1–13. https://doi.org/10.1186/1471-2105-9-559

Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). LUMPY: A probabilistic framework for structural variant discovery. *Genome Biology*, *15*(6), 1–19. https://doi.org/10.1186/gb-2014-15-6-r84

Lee, J. A., Carvalho, C. M. B., & Lupski, J. R. (2007). A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders. *Cell*, *131*(7), 1235–1247. https://doi.org/10.1016/j.cell.2007.11.037

Lee, J. H., Won, H. J., Hoang Nguyen Tran, P., Lee, S. mi, Kim, H. Y., & Jung, J. H. (2021). Improving lignocellulosic biofuel production by CRISPR/Cas9-mediated lignin modification in barley. *GCB Bioenergy*, *13*(4), 742–752. https://doi.org/10.1111/GCBB.12808

Lessard, S., Chao, M., Reis, K., Beauvais, M., Rajpal, D. K., Sloane, J., Palta, P., Klinger, K., de Rinaldis, E., Shameer, K., & Chatelain, C. (2024). Leveraging large-scale multi-omics evidences to identify therapeutic targets from genome-wide association studies. *BMC Genomics*, *25*(1), 1–16. https://doi.org/10.1186/S12864-

024-10971-2

Levsen, N. D., Tiffin, P., & Olson, M. S. (2012). Pleistocene Speciation in the Genus Populus (Salicaceae). *Systematic Biology*, *61*(3), 401–401. https://doi.org/10.1093/SYSBIO/SYR120

Li, T., He, B., Chen, D., Chen, H. W., Guo, L., Yuan, W., Fang, K., Shi, F., Liu, L., Zheng, H., Huang, L., Wu, X., Hao, X., Zhao, X., & Jiang, W. (2024). Increasing Sensitivity of Tree Radial Growth to Precipitation. *Geophysical Research Letters*, *51*(16), e2024GL110003. https://doi.org/10.1029/2024GL110003;WGROUP:STRING:PUBLICATION

Li, Y., Suontama, M., Burdon, R. D., & Dungey, H. S. (2017). Genotype by environment interactions in forest tree breeding: review of methodology and perspectives on research and application. *Tree Genetics and Genomes*, *13*(3), 1–18. https://doi.org/10.1007/S11295-017-1144-X

Lieber, M. R. (2010). The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annual Review of Biochemistry*, *79*(Volume 79, 2010), 181–211. https://doi.org/10.1146/ANNUREV.BIOCHEM.052308.093131/1

Luo, J., Zhou, J., Li, H., Shi, W., Polle, A., Lu, M., Sun, X., & Luo, Z. Bin. (2015). Global poplar root and leaf transcriptomes reveal links between growth and stress responses under nitrogen starvation and excess. *Tree Physiology*, *35*(12), 1283–1302. https://doi.org/10.1093/TREEPHYS/TPV091

Mahmood, T., Khalid, S., Abdullah, M., Ahmed, Z., Shah, M. K. N., Ghafoor, A., & Du, X. (2019). Insights into Drought Stress Signaling in Plants and the Molecular Genetic Basis of Cotton Drought Tolerance. *Cells 2020, Vol. 9, Page 105*, *9*(1), 105. https://doi.org/10.3390/CELLS9010105

Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., & Sedlazeck, F. J. (2019). Structural variant calling: The long and the short of it. In *Genome Biology* (Vol. 20, Issue 1, pp. 1–14). BioMed Central. https://doi.org/10.1186/s13059-019-1828-7

McCouch, S. R., Wright, M. H., Tung, C. W., Maron, L. G., McNally, K. L., Fitzgerald,

M., Singh, N., DeClerck, G., Agosto-Perez, F., Korniliev, P., Greenberg, A. J., Naredo, M. E. B., Mercado, S. M. Q., Harrington, S. E., Shi, Y., Branchini, D. A., Kuser-Falcão, P. R., Leung, H., Ebana, K., … Mezey, J. (2016). Open access resources for genome-wide association mapping in rice. *Nature Communications*, *7*(1), 1–14. https://doi.org/10.1038/NCOMMS10532

Mckown, A. D., Guy, R. D., Klápště, J., Geraldes, A., Friedmann, M., Cronk, Q. C. B., El-Kassaby, Y. A., Mansfield, S. D., & Douglas, C. J. (2014). Geographical and environmental gradients shape phenotypic trait variation and genetic structure in Populus trichocarpa. *New Phytologist*, *201*(4), 1263–1276. https://doi.org/10.1111/nph.12601

McKown, A. D., Klápště, J., Guy, R. D., El-Kassaby, Y. A., & Mansfield, S. D. (2018). Ecological genomics of variation in bud-break phenology and mechanisms of response to climate warming in Populus trichocarpa. *New Phytologist*, *220*(1), 300–316. https://doi.org/10.1111/NPH.15273

Monson, R. K., Trowbridge, A. M., Lindroth, R. L., & Lerdau, M. T. (2022). Coordinated resource allocation to plant growth–defense tradeoffs. *New Phytologist*, *233*(3), 1051–1066. https://doi.org/10.1111/NPH.17773

Muchero, W., Guo, J., DiFazio, S. P., Chen, J. G., Ranjan, P., Slavov, G. T., Gunter, L. E., Jawdy, S., Bryan, A. C., Sykes, R., Ziebell, A., Klápště, J., Porth, I., Skyba, O., Unda, F., El-Kassaby, Y. A., Douglas, C. J., Mansfield, S. D., Martin, J., … Tuskan, G. A. (2015). High-resolution genetic mapping of allelic variants associated with cell wall chemistry in Populus. *BMC Genomics*, *16*(1), 24. https://doi.org/10.1186/s12864-015-1215-z

Müller, B. S. F., de Almeida Filho, J. E., Lima, B. M., Garcia, C. C., Missiaggia, A., Aguiar, A. M., Takahashi, E., Kirst, M., Gezan, S. A., Silva-Junior, O. B., Neves, L. G., & Grattapaglia, D. (2019). Independent and Joint-GWAS for growth traits in Eucalyptus by assembling genome-wide data for 3373 individuals across four breeding populations. *New Phytologist*, *221*(2), 818–833. https://doi.org/10.1111/NPH.15449

Neale, D. B., & Kremer, A. (2011). Forest tree genomics: Growing resources and

applications. *Nature Reviews Genetics*, *12*(2), 111–122. https://doi.org/10.1038/NRG2931

Oubida, R. W., Gantulga, D., Zhang, M., Zhou, L., Bawa, R., & Holliday, J. A. (2015). Partitioning of multivariate phenotypes using regression trees reveals complex patterns of adaptation to climate across the range of black cottonwood (Populus trichocarpa). *Frontiers in Plant Science*, *6*(MAR), 103144. https://doi.org/10.3389/FPLS.2015.00181

Pedersen, B. S., & Quinlan, A. R. (2019). Duphold: scalable, depth-based annotation and curation of high-confidence structural variant calls. *GigaScience*, *8*(4), 1–5. https://doi.org/10.1093/GIGASCIENCE/GIZ040

Piot, A., Prunier, J., Isabel, N., Klápště, J., El-Kassaby, Y. A., Villarreal Aguilar, J. C., & Porth, I. (2020). Genomic Diversity Evaluation of Populus trichocarpa Germplasm for Rare Variant Genetic Association Studies. *Frontiers in Genetics*, *10*, 499249. https://doi.org/10.3389/FGENE.2019.01384

Pluess, A. R., Frank, A., Heiri, C., Lalagüe, H., Vendramin, G. G., & Oddou-Muratorio, S. (2016). Genome-environment association study suggests local adaptation to climate at the regional scale in Fagus sylvatica. *New Phytologist*, *210*(2), 589–601. https://doi.org/10.1111/NPH.13809

Porth, I., Klápště, J., McKown, A. D., La Mantia, J., Guy, R. D., Ingvarsson, P. K., Hamelin, R., Mansfield, S. D., Ehlting, J., Douglas, C. J., & El-Kassaby, Y. A. (2015). Evolutionary quantitative genomics of Populus trichocarpa. *PLoS ONE*, *10*(11), e0142864. https://doi.org/10.1371/journal.pone.0142864

Poupon, V., Chakraborty, D., Stejskal, J., Konrad, H., Schueler, S., & Lstibůrek, M. (2021). Accelerating Adaptation of Forest Trees to Climate Change Using Individual Tree Response Functions. *Frontiers in Plant Science*, *12*, 758221. https://doi.org/10.3389/FPLS.2021.758221

Quinn, L. D., Straker, K. C., Guo, J., Kim, S., Thapa, S., Kling, G., Lee, D. K., & Voigt, T. B. (2015). Stress-Tolerant Feedstocks for Sustainable Bioenergy Production on Marginal Land. *Bioenergy Research*, *8*(3), 1081–1100. https://doi.org/10.1007/S12155-014-9557-Y

Rae, A. M., Street, N. R., Robinson, K. M., Harris, N., & Taylor, G. (2009). Five QTL hotspots for yield in short rotation coppice bioenergy poplar: The Poplar Biomass Loci. *BMC Plant Biology*, *9*(1), 1–13. https://doi.org/10.1186/1471-2229-9-23

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., & Korbel, J. O. (2012). DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, *28*(18), i333–i339. https://doi.org/10.1093/bioinformatics/bts378

Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., & Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. In *Molecular Ecology* (Vol. 24, Issue 17, pp. 4348–4370). https://doi.org/10.1111/mec.13322

Rosso, L., Cantamessa, S., Bergante, S., Biselli, C., Fricano, A., Chiarabaglio, P. M., Gennaro, M., Nervo, G., Secchi, F., & Carra, A. (2023). Responses to Drought Stress in Poplar: What Do We Know and What Can We Learn? *Life 2023, Vol. 13, Page 533*, *13*(2), 533. https://doi.org/10.3390/LIFE13020533

Rupp, D. E., Abatzoglou, J. T., Hegewisch, K. C., & Mote, P. W. (2013). Evaluation of CMIP5 20th century climate simulations for the Pacific Northwest USA. *Journal of Geophysical Research Atmospheres*, *118*(19), 10,884-10,906. https://doi.org/10.1002/JGRD.50843

Sang, Y., Long, Z., Dan, X., Feng, J., Shi, T., Jia, C., Zhang, X., Lai, Q., Yang, G., Zhang, H., Xu, X., Liu, H., Jiang, Y., Ingvarsson, P. K., Liu, J., Mao, K., & Wang, J. (2022). Genomic insights into present local adaptation and future climate change vulnerability of a keystone forest tree species in East Asian. *BioRxiv*, 1–40. https://doi.org/10.1101/2022.04.04.486908

Satyakam, Zinta, G., Singh, R. K., & Kumar, R. (2022). Cold adaptation strategies in plants—An emerging role of epigenetics and antifreeze proteins to engineer cold resilient plants. *Frontiers in Genetics*, *13*, 909007. https://doi.org/10.3389/FGENE.2022.909007

Saxe, H., Cannell, M. G. R., Johnsen, Ø., Ryan, M. G., & Vourlitis, G. (2001). Tree and forest functioning in response to global warming. *New Phytologist*, *149*(3), 369–

399. https://doi.org/10.1046/J.1469-8137.2001.00057.X

Saxena, R. K., Edwards, D., & Varshney, R. K. (2014). Structural variations in plant genomes. *Briefings in Functional Genomics*, *13*(4), 296–307. https://doi.org/10.1093/BFGP/ELU016

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Von Haeseler, A., & Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, *15*(6), 461–468. https://doi.org/10.1038/s41592-018-0001-7

Shinozaki, K., & Yamaguchi-Shinozaki, K. (2007). Gene networks involved in drought stress response and tolerance. *Journal of Experimental Botany*, *58*(2), 221–227. https://doi.org/10.1093/JXB/ERL164

Slavov, G. T., Difazio, S. P., Martin, J., Schackwitz, W., Muchero, W., Rodgers-Melnick, E., Lipphardt, M. F., Pennacchio, C. P., Hellsten, U., Pennacchio, L. A., Gunter, L. E., Ranjan, P., Vining, K., Pomraning, K. R., Wilhelm, L. J., Pellegrini, M., Mockler, T. C., Freitag, M., Geraldes, A., … Tuskan, G. A. (2012). Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree Populus trichocarpa. *New Phytologist*, *196*(3), 713–725. https://doi.org/10.1111/j.1469-8137.2012.04258.x

Slavov, G. T., Macaya-Sanz, D., DiFazio, S. P., & Howe, G. T. (2025). Population structure limits the use of genomic data for predicting phenotypes and managing genetic resources in forest trees. *Proceedings of the National Academy of Sciences*, *122*(26), e2425691122. https://doi.org/10.1073/PNAS.2425691122

Spielmann, M., Lupiáñez, D. G., & Mundlos, S. (2018). Structural variation in the 3D genome. *Nature Reviews Genetics 2018 19:7*, *19*(7), 453–467. https://doi.org/10.1038/s41576-018-0007-0

Springer, N. M., Ying, K., Fu, Y., Ji, T., Yeh, C. T., Jia, Y., Wu, W., Richmond, T., Kitzman, J., Rosenbaum, H., Iniguez, A. L., Barbazuk, W. B., Jeddeloh, J. A., Nettleton, D., & Schnable, P. S. (2009). Maize Inbreds Exhibit High Levels of Copy Number Variation (CNV) and Presence/Absence Variation (PAV) in Genome Content. *PLOS Genetics*, *5*(11), e1000734.

https://doi.org/10.1371/JOURNAL.PGEN.1000734

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(16), 9440–9445. https://doi.org/10.1073/PNAS.1530509100

Struk, S., Jacobs, A., Sánchez Martín-Fontecha, E., Gevaert, K., Cubas, P., & Goormachtig, S. (2019). Exploring the protein–protein interaction landscape in plants. *Plant Cell and Environment*, *42*(2), 387–409. https://doi.org/10.1111/PCE.13433

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545–15550. https://doi.org/10.1073/PNAS.0506580102

Sulis, D. B., Jiang, X., Yang, C., Marques, B. M., Matthews, M. L., Miller, Z., Lan, K., Cofre-Vega, C., Liu, B., Sun, R., Sederoff, H., Bing, R. G., Sun, X., Williams, C. M., Jameel, H., Phillips, R., Chang, H. M., Peszlen, I., Huang, Y. Y., … Wang, J. P. (2023). Multiplex CRISPR editing of wood for sustainable fiber production. *Science*, *381*(6654), 216–221. https://doi.org/10.1126/SCIENCE.ADD4514

Sullivan, K. A., Lane, M., Cashman, M., Miller, J. I., Pavicic, M., Walker, A. M., Cliff, A., Romero, J., Qin, X., Mullins, N., Docherty, A., Coon, H., Ruderfer, D. M., Garvin, M. R., Pestian, J. P., Ashley-Koch, A. E., Beckham, J. C., McMahon, B., Oslin, D. W., … Kainer, D. (2024). Analyses of GWAS signal using GRIN identify additional genes contributing to suicidal behavior. *Communications Biology*, *7*(1), 1–16. https://doi.org/10.1038/s42003-024-06943-7

Sullivan, K. A., Miller, J. I., Townsend, A., Morgan, M., Lane, M., Pavicic, M., Shah, M., Cashman, M., & Jacobson, D. A. (2024). MENTOR: Multiplex Embedding of Networks for Team-Based Omics Research. *BioRxiv*, 2024.07.17.603821. https://doi.org/10.1101/2024.07.17.603821

Treangen, T. J., & Salzberg, S. L. (2011). Repetitive DNA and next-generation

sequencing: computational challenges and solutions. *Nature Reviews Genetics 2011 13:1*, *13*(1), 36–46. https://doi.org/10.1038/nrg3117

Trono, D., & Pecchioni, N. (2022). Candidate Genes Associated with Abiotic Stress Response in Plants as Tools to Engineer Tolerance to Drought, Salinity and Extreme Temperatures in Wheat: An Overview. *Plants 2022, Vol. 11, Page 3358*, *11*(23), 3358. https://doi.org/10.3390/PLANTS11233358

Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., others, Putnam, M., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R. R. P. R., … Rokhsar, D. (2006). The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science*, *313*(5793), 1596–1604. https://doi.org/10.1126/science.1128691

Valous, N. A., Popp, F., Zörnig, I., Jäger, D., & Charoentong, P. (2024). Graph machine learning for integrated multi-omics analysis. *British Journal of Cancer*, *131*(2), 205–211. https://doi.org/10.1038/S41416-024-02706-7

Van den Broeck, L., Gordon, M., Inzé, D., Williams, C., & Sozzani, R. (2020). Gene Regulatory Network Inference: Connecting Plant Biology and Mathematical Modeling. *Frontiers in Genetics*, *11*, 510276. https://doi.org/10.3389/FGENE.2020.00457/EPUB

Varshney, R. K., Bohra, A., Yu, J., Graner, A., Zhang, Q., & Sorrells, M. E. (2021). Designing Future Crops: Genomics-Assisted Breeding Comes of Age. *Trends in Plant Science*, *26*(6), 631–649. https://doi.org/10.1016/J.TPLANTS.2021.03.010

Viger, M., Smith, H. K., Cohen, D., Dewoody, J., Trewin, H., Steenackers, M., Bastien, C., & Taylor, G. (2016). Adaptive mechanisms and genomic plasticity for drought tolerance identified in European black poplar (Populus nigra L.). *Tree Physiology*, *36*(7), 909–928. https://doi.org/10.1093/TREEPHYS/TPW017

Wang, J., Ding, J., Tan, B., Robinson, K. M., Michelson, I. H., Johansson, A., Nystedt, B., Scofield, D. G., Nilsson, O., Jansson, S., Street, N. R., & Ingvarsson, P. K. (2018). A major locus controls local adaptation and adaptive life history variation in a perennial plant. *Genome Biology*, *19*(1), 1–17. https://doi.org/10.1186/S13059-

018-1444-Y

Wang, Q., Tian, F., Pan, Y., Buckler, E. S., & Zhang, Z. (2014). A SUPER Powerful Method for Genome Wide Association Study. *PLOS ONE*, *9*(9), e107684. https://doi.org/10.1371/JOURNAL.PONE.0107684

Wang, T., Hamann, A., Spittlehouse, D., & Carroll, C. (2016). Locally Downscaled and Spatially Customizable Climate Data for Historical and Future Periods for North America. *PLOS ONE*, *11*(6), e0156720. https://doi.org/10.1371/JOURNAL.PONE.0156720

Xu, J., Yuan, Y., Xu, Y., Zhang, G., Guo, X., Wu, F., Wang, Q., Rong, T., Pan, G., Cao, M., Tang, Q., Gao, S., Liu, Y., Wang, J., Lan, H., & Lu, Y. (2014). Identification of candidate genes for drought tolerance by whole-genome resequencing in maize. *BMC Plant Biology*, *14*(1), 1–15. https://doi.org/10.1186/1471-2229-14-83

Yang, L., Luquette, L. J., Gehlenborg, N., Xi, R., Haseley, P. S., Hsieh, C. H., Zhang, C., Ren, X., Protopopov, A., Chin, L., Kucherlapati, R., Lee, C., & Park, P. J. (2013). Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell*, *153*(4), 919–929. https://doi.org/10.1016/J.CELL.2013.04.010

Zhang, J., Yang, Y., Zheng, K., Xie, M., Feng, K., Jawdy, S. S., Gunter, L. E., Ranjan, P., Singan, V. R., Engle, N., Lindquist, E., Barry, K., Schmutz, J., Zhao, N., Tschaplinski, T. J., LeBoldus, J., Tuskan, G. A., Chen, J. G., & Muchero, W. (2018). Genome-wide association studies and expression-based quantitative trait loci analyses reveal roles of HCT2 in caffeoylquinic acid biosynthesis and its regulation by defense-responsive transcription factors in Populus. *New Phytologist*, *220*(2), 502–516. https://doi.org/10.1111/nph.15297

Zhang, M., Suren, H., Holliday, J. A., & Gaut, B. (2019). Phenotypic and Genomic Local Adaptation across Latitude and Altitude in Populus trichocarpa. *Genome Biology and Evolution*, *11*(8), 2256–2272. https://doi.org/10.1093/GBE/EVZ151

Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M., & Buckler, E. S. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, *42*(4), 355–360. https://doi.org/10.1038/NG.546

Zheng, Z., Zhu, M., Zhang, J., Liu, X., Hou, L., Liu, W., Yuan, S., Luo, C., Yao, X., Liu, J., & Yang, Y. (2024). A sequence-aware merger of genomic structural variations at population scale. *Nature Communications*, *15*(1), 1–9. https://doi.org/10.1038/s41467-024-45244-9

Zook, J. M., Hansen, N. F., Olson, N. D., Chapman, L., Mullikin, J. C., Xiao, C., Sherry, S., Koren, S., Phillippy, A. M., Boutros, P. C., Sahraeian, S. M. E., Huang, V., Rouette, A., Alexander, N., Mason, C. E., Hajirasouliha, I., Ricketts, C., Lee, J., Tearle, R., … Salit, M. (2020). A robust benchmark for detection of germline large deletions and insertions. *Nature Biotechnology*, *38*(11), 1347–1355. https://doi.org/10.1038/S41587-020-0538-8

Züst, T., & Agrawal, A. A. (2017). Trade-Offs Between Plant Growth and Defense Against Insect Herbivory: An Emerging Mechanistic Synthesis. *Annual Review of Plant Biology*, *68*(Volume 68, 2017), 513–534. https://doi.org/10.1146/ANNUREV-ARPLANT-042916-040856

2 Chapter 2: Finding New Cell Wall Regulatory Gene Candidates in *Populus trichocarpa* Using Multiple Lines of Evidence

This chapter was originally published as:

Furches, A., Kainer, D., Weighill, D., Large, A., Jones, P., Walker, A. M., Romero, J., Gazolla, J. G. F. M., Joubert, W., Shah, M., Streich, J., Ranjan, P., Schmutz, J., Sreedasyam, A., Macaya-Sanz, D., Zhao, N., Martin, M. Z., Rao, X., Dixon, R. A., DiFazio, S., Tschaplinski, T.J., Muchero, W.M., Chen, J-G., Tuskan, G.A., and Jacobson, D. (2019). Finding New Cell Wall Regulatory Genes in Populus trichocarpa Using Multiple Lines of Evidence. *Frontiers in Plant Science*, *10*. https://doi.org/10.3389/fpls.2019.01249

Author Contributions: XR and RD provided anchor gene IDs, TT provided anchor metabolite IDs; WJ developed the Parallel GPU CCC application code; DK performed the rare variant GWAS analysis; PJ performed the standard GWAS and outlier analysis and constructed the standard GWAS and eQTN networks; DW calculated methylation TPM values; DW constructed the observed co-expression, co-methylation and SNP correlation networks and calculated LOE scores for the observed LOE network; AF performed expression and methylation randomizations, JG performed random subsampling, AF calculated correlation values and performed significance tests; AF performed LOE input layer randomizations, constructed randomized LOE networks, and calculated randomized LOE scores; DK and JR created the GO-term network, and AF calculated GO intersect scores; AF performed candidate ranking, generated the clustered expression heatmap, and performed the phylogenetic analyses; DK performed the functional enrichment and TF-binding analyses; AF created eQTN subnetworks; MS mapped gene expression atlas reads and RNAseq reads and calculated gene expression TPM values; SD and GT led the effort on constructing the GWAS population; TT led the leaf sample collection for GCMS-based metabolomic analysis, identified the peaks, and summarized the metabolomics data; PR did automated extraction of metabolite intensity from GCMS; MM collected the leaf samples and manually extracted the metabolite data; NZ performed leaf sample preparation, extracted, derivatized, and analyzed the metabolites by GCMS; JS and AS generated the gene expression atlas data; SD and DM-S generated the SNP calls; WM and J-GC provided guidance regarding RNAseq data;

AF, AW, DK and AL performed the in-depth literature searches; AL performed the domain annotation; AF, DK, DW, and AL and wrote the manuscript; AF, DK, DW, AL, AW, PJ, DJ, SD, GT, JS, WM, J-GC, and RD edited the manuscript; DJ conceived of and supervised the study, participated in the network analysis, and generated MapMan annotations.

2.1 Abstract

Understanding the regulatory network controlling cell wall biosynthesis is of great interest in *Populus trichocarpa,* both because of its status as a model woody perennial and its importance for lignocellulosic products. We searched for genes with putatively unknown roles in regulating cell wall biosynthesis using an extended network-based Lines of Evidence (LOE) pipeline to combine multiple omics data sets in *P. trichocarpa*, including gene coexpression, gene comethylation, population level pairwise SNP correlations, and two distinct SNP-metabolite GWAS layers. By incorporating validation, ranking, and filtering approaches we produced a list of nine high priority gene candidates for involvement in the regulation of cell wall biosynthesis. We subsequently performed a detailed investigation of candidate gene GROWTH-REGULATING FACTOR 9 (*PtGRF9*). To investigate the role of *PtGRF9* in regulating cell wall biosynthesis, we assessed the genome-wide connections of *PtGRF9* and a paralog across data layers with functional enrichment analyses, predictive transcription factor binding site analysis, and an independent comparison to eQTN data. Our findings indicate that PtGRF9 likely affects the cell wall by directly repressing genes involved in cell wall biosynthesis, such as *PtCCoAOMT* and *PtMYB.41*, and indirectly by regulating homeobox genes. Furthermore, evidence suggests that *PtGRF9* paralogs may act as transcriptional co-regulators that direct the global energy usage of the plant. Using our extended pipeline, we show multiple lines of evidence implicating the involvement of these genes in cell wall regulatory functions and demonstrate the value of this method for prioritizing candidate genes for experimental validation.

2.2 Introduction

The biosynthesis and regulation of the plant cell wall has been the subject of a large body of research due to the industrial importance of lignocellulosic biomass, as well as the role of the cell wall in the function of other plant biological systems such as stress response, inter-cellular transport, and disease resistance. For industrially cultivated genera such as *Populus*, the primary cell wall constituents (*i.e.,* cellulose, lignin, hemicellulose) provide feedstock for downstream products including biofuel, lumber, paper, and advanced lignin products (Sannigrahi et al., 2010; Porth et al., 2013). There is therefore broad interest in understanding the mechanisms that regulate the biosynthesis and modification of the cell wall, both from a yield and composition perspective.

A great variety of biopolymers are synthesized and incorporated into the primary and secondary cell wall, often in response to biotic and abiotic stress, nutrient availability, and developmental and temporal switches, all of which govern the macro-scale form of the plant. A highly complex network of genetic regulation has evolved to control the rate of biosynthesis of cell wall polymers, their intrinsic monomer composition, their transport to and subsequent deposition in the cell wall, and the expansion of the wall under changing intra-cellular conditions. In the model plant *Arabidopsis thaliana*, Bischoff et al. (2010) estimated that over 1,000 genes encode proteins related to the cell wall, while Cai et al. (2014) predicted a number closer to 3,000 based on clustering of gene co-expression. Furthermore, Taylor-Teeples et al. (2015) tested a library of 1,664 transcription factors in *A. thaliana* for interaction with the promoter regions of cell wall biosynthesis genes and found 413 such interactions in root vascular tissue alone. Studies such as these highlight the immense complexity involved in cell wall regulation, much of which is yet to be elucidated.

Due to poplar's status as a model woody plant and its importance for lignocellulosic products, many studies have investigated the regulatory network of the cell wall and its components in *Populus* species or in multiple genera in combination with *Populus* (Ohtani et al., 2011; Puzey et al., 2012; Lu et al., 2013; Porth et al., 2013; Yu et al., 2013;

Ko et al., 2014; Wang et al., 2014; Zhong and Ye, 2014; Muchero et al., 2015; Lin et al., 2017; Shi et al., 2017; Xie et al., 2018a). Many of these studies have focused either on characterizing *Populus* homologs of genes that have been shown to have an effect on the cell wall chemistry or plant growth traits in mutant Arabidopsis lines, or perhaps were shown to be differentially expressed in comparisons of low and high growth genotypes. However, exploring the regulatory network controlling the cell wall in order to find new functional mechanisms is a challenging task due to the number of genes involved, extensive functional redundancy, and the multitude of transcriptional feedback loops. Such complex genetic architecture has contributed to the view that many quantitative traits are actually "omnigenic" (Boyle et al., 2017), such that virtually any expressed gene has a non-zero effect on the core biosynthetic genes at one or more transcriptional, post-transcriptional, post-translational, signaling or protein-protein interaction levels. Fisher (1919) predicted that rather than a few core genes in biosynthetic pathways, the major portion of heritability is explained by a large number of loci across the entire genome that contribute small portions of the trait heritability. Under this omnigenic model, network-theory-based methods provide an elegant approach for mining omics datasets for regulatory relationships. Any biological entity (SNP, gene, protein, metabolite, etc.) can be modeled as a node and any relationship between those entities (association, co-expression, correlation, binding) can be modeled as an edge.

The network approach has been used in several studies of cell wall regulation to date, often focusing on finding clusters of genes that co-express with each other in certain tissues, thus finding putative functional units or networks. For example, Cai et al. (2014) performed co-expression network clustering in *Populus* and found major sub-clusters enriched for primary cell wall or secondary cell wall genes. Taylor-Teeples et al. (2015) produced networks based on *A. thaliana* transcription factors and their target binding sites, providing an expanded view of the multi-tiered regulatory system with respect to secondary cell wall (SCW) biosynthesis and xylem development. Yang et al. (2011) used 121 *A. thaliana* cell wall genes obtained from text mining followed by co-expression neighbor analysis to identify 694 *A. thaliana* genes and their 817 *Populus* orthologs as

candidate genes for involvement in cell wall functions. Alejandro et al. (2012) identified the ABCG29 genes as transporting monolignol to the cell wall in *A. thaliana* by first analyzing co-expression networks followed by expression and functional analyses. These methods often produce a extensive list of candidate genes but with little more to support their involvement in cell wall regulation than the clustering or enrichment evidence.

Multiomic approaches have also been performed, which include more data types to identify candidate genes. Porth et al. (2013a) used a network-based multiomic approach to find relationships between SNP, gene expression, and wood phenotype data from *P. trichocarpa*. They constructed six phenotypic-centric networks to identify genes that most influenced the expression of their related phenotype. From this study, they were able to identify candidate genes potentially related to cell wall biogenesis. Mizrachi et al. (2017) used a network-based approach to integrate known gene interactions and eQTN data in the form of a connectivity matrix with gene expression data through matrix multiplication in order to identify genes involved in lignin-related traits.

The use of multiple layers of omics data in the identification of candidate genes related to a particular phenotype provides an increased level of confidence and context surrounding the new candidate genes. In this study, we use an extended lines of evidence (LOE) pipeline for jointly mining multiple data layers to produce a curated short list of new candidate genes putatively involved in the regulation of cell-wall-related functions (**Figure 2.1**). We use an extensive set of "anchor" genes with documented roles in cell wall biosynthetic and regulatory processes and anchor metabolomic phenotypes measured in a GWAS population of *P. trichocarpa*.

Multiomic data layers (coexpression, comethylation, pairwise SNP correlation, and two SNP-metabolite GWAS data sets) are probed to find all genes in the genome with network connectivity to the anchor set. A score is calculated for each gene with regards to the amount of evidence that the gene is involved in cell wall regulation and other cell-wall-related processes. The resulting merged LOE network of candidate genes is then

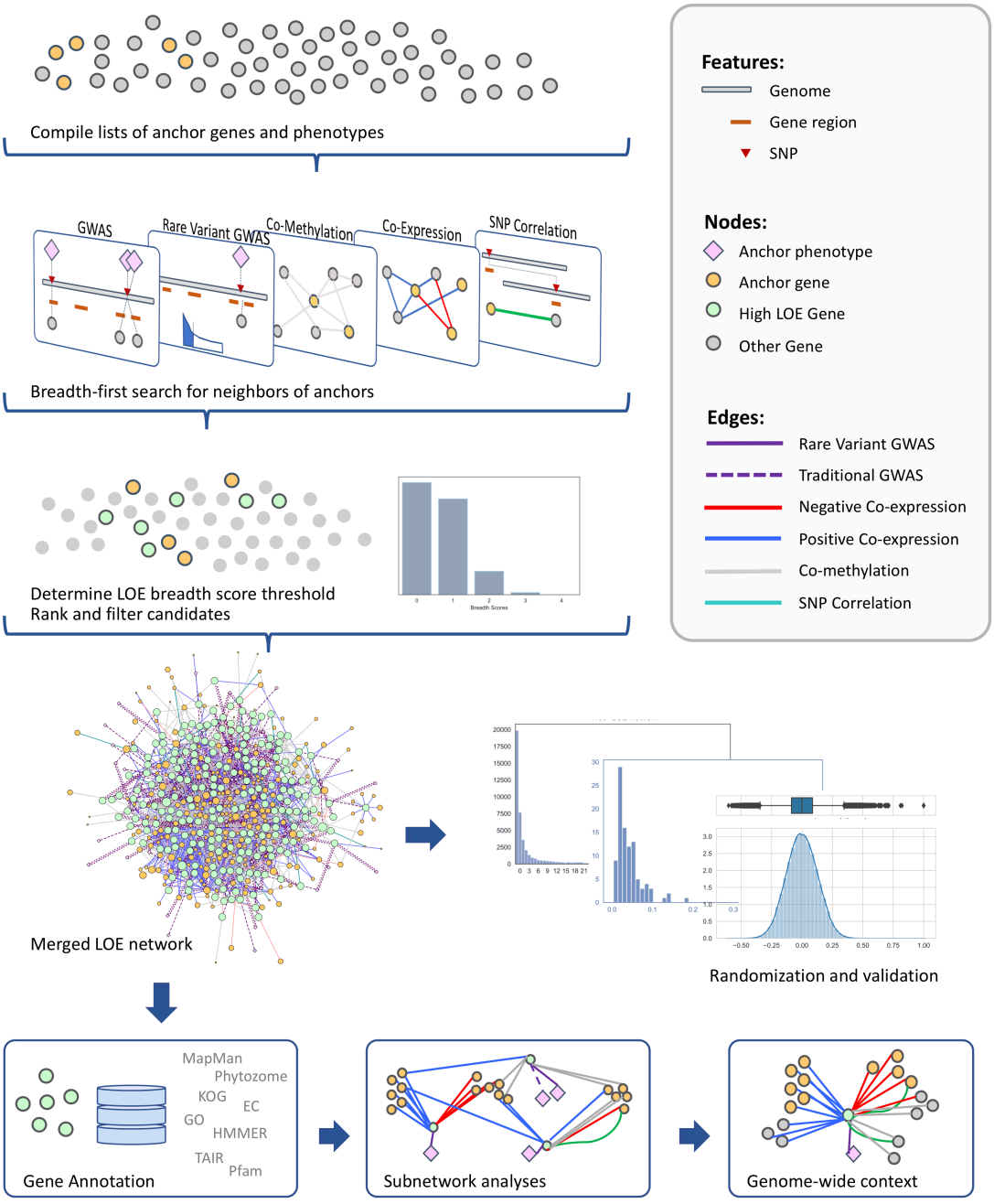


Figure 2.1: Overview of the method for identifying new candidate genes involved in cell wall regulation through data layering and calculation of LOE scores.

subjected to validation, ranking, and filtering methods, as well as post-hoc analyses. The result is a set of 330 high-ranking candidate genes, which we then filter to a subset of regulatory genes not previously discussed in the context of the cell wall biosynthesis.

2.3 Methods

This study makes use of various data accumulated for *P. trichocarpa* that have been used in previous investigations, including SNP data from a GWAS population, foliar metabolites measured in this GWAS population, and DNA methylation data across 10 different *P. trichocarpa* tissues (Vining et al., 2012), as well as the *P. trichocarpa* DOE Joint Genome Institute Plant Gene Atlas (Sreedasyam et al, unpublished data; available from phytozome.jgi.doe.gov). Each data set was considered as a separate layer for this study, and integrated though the use of LOE scores. Below, the various layers are described as well as the network analysis methods used to merge layers and identify genes with high connectivity to cell wall systems.

2.3.1 Phenotypes

We made use of metabolite data previously obtained from leaf tissue and analyzed using GC-MS. Details can be found in (Li et al., 2012b; Tschaplinski et al., 2012; Weighill et al., 2018). To prevent spurious associations, we examined each phenotype for the presence of outliers using Median Absolute Deviation (MAD). If a sample's phenotype was more than six MADs from the population median it was removed from the GWAS for that phenotype.

2.3.2 Genotypes

SNP-based variant data (see DOI 10.13139/OLCF/1411410) were obtained from https://bioenergycenter.org/besc/gwas/ and SNPs were filtered to the top 90% tranche (PASS SNPs) and call rate ≥ 0.5 using Plink (Purcell et al., 2007b) and VCFtools (Danecek et al., 2011a).

2.3.3 Genome Wide Association Layer

GWAS was performed using a linear mixed model (LMM), implemented in EMMAX (Kang et al., 2010b) and leveraging ADIOS v1.13 (Lofstead et al., 2008) for scaling, to estimate the additive effect of each SNP while accounting for population structure and cryptic relatedness between samples. The tested SNPs excluded those with minor allele frequency (MAF) <0.01, and those with a population call rate above 0.75. In addition, we used linkage disequilibrium (LD) pruning on the main set of SNPs to produce a set of independent SNPs for estimating the genomic relationship matrix, used in the LMM. The resulting p-values were corrected for multi-hypotheses bias by applying the Benjamini-Hochberg approach (Benjamini and Hochberg, 1995) with a false-discovery rate (FDR) cutoff of 0.1.

2.3.4 Rare Variant GWAS Layer

While the GWAS Linear Mixed Model (LMM) tested common and less common SNPs (MAF ≥0.01) individually for significance, rarer SNPs were tested regionally in a joint fashion. Rare SNPs (MAF <0.01) located within a given gene, or in the gene's 2-kb upstream and downstream flanking regions, were grouped as a region defined by that gene. RVtest (Zhan et al., 2016) was then used to apply the Sequence Kernel Association Test (SKAT) to each of the 41,335 regions defined from *P. trichocarpa* v3.0 annotations. SKAT tests each SNP in the region individually with an LMM and then forms a combined region score where each component SNP is weighted according to its MAF. Weights were drawn from a beta distribution with default shape parameters (1, 25), which produced a single P-value for the significance of association of each region, which were corrected for multiple testing with an FDR of 0.1.

2.3.5 Co-expression Layer

A *P. trichocarpa* gene co-expression network was constructed as described in Weighill et al. (2018). RNA-seq reads from the *P. trichocarpa* DOE Joint Genome Institute Plant Gene Atlas (Sreedasyam et al., unpublished data; available from phytozome.jgi.doe.gov; see **Supplemental Table 2.1** for sample information) were trimmed using Skewer (Jiang

et al., 2014), aligned to the version 3.0 *P. trichocarpa* reference (Tuskan et al., 2006) using Star (Dobin et al., 2013), and TPM (transcripts per million) values calculated for each gene and each sample. Star mapping was performed using the "--quantMode GeneCounts" option, which directs the program to count the number of reads per gene while performing the mapping. A read is counted if it overlaps one and only one gene. Both ends of the paired-end read are checked for overlaps. The counts coincide with those produced by htseq-count with default parameters. We then calculated the Spearman correlation coefficient between the expression profiles of all pairs of genes using the mcxarray package (van Dongen, 2008) available from https://micans.org/mcl/index.html. An absolute threshold of 0.85 was applied in order to keep only those gene-gene pairs with strong co-expression.

2.3.6 Co-methylation Layer

A *P. trichocarpa* gene co-methylation network was constructed as described in Weighill et al. (2018). MEDIP-Seq reads from the study by Vining et al. (2012) mapped to the *P. trichocarpa* V3 genome assembly, were obtained from Phytozome (Goodstein et al., 2011). The number of reads that mapped to each gene for each sample was determined using htseq-count (Anders et al., 2015). These counts were then converted to TPM values for each gene and each sample. Spearman correlation coefficients between the co-methylation profiles of all pairs of genes were then calculated in a manner similar to the co-expression layer, followed by an absolute threshold of 0.95.

2.3.7 Custom Correlation Coefficient Layer

After filtering the SNP set to remove those with MAF <0.01, the custom correlation coefficient (CCC) (Climer et al., 2014) between all pairs of remaining SNPs were calculated using a Parallel GPU implementation of the CCC (Joubert et al., 2019). In order to minimize correlation among SNPs due to linkage disequilibrium, only correlations from SNP pairs greater than 10 kb apart and with a CCC ≥0.7 were retained. SNPs were then mapped to the genes in which they were located, resulting in gene-gene correlations. Significantly correlated SNPs represent co-segregating and interacting

cellular components.

2.3.8 Lines of Evidence Scoring and Network Analysis

The LOE method calculates a score for every gene in the genome by quantifying the connectivity of a given gene to anchor genes/phenotypes from the system of interest. Each data layer described above provides one possible line of evidence. For example, if Gene A co-expresses with one or more cell wall anchor genes, then this is counted as one line of evidence for Gene A's involvement in the cell wall. A list of 295 anchor genes was compiled from the literature (Hao and Mohnen, 2014; Zhong and Ye, 2014; Nakano et al., 2015; Liu et al., 2017; Rao and Dixon, 2018) (**Supplemental Table 2.2**). Metabolites that affect cell wall development and composition, such as sugar substrates, lignin precursors, and lignin competitors, were also selected for use as cell wall anchor phenotypes (**Supplemental Table 2.3**).

To calculate LOE scores for each gene in the *P. trichocarpa* genome, each data layer was represented as a network. Each layer consisted of a list of source entities (cell wall anchor genes and phenotypes, or "anchor nodes"), target entities (potential candidate genes, or "target nodes"), and interactions between them (correlations/associations, or "edges"). From each layer's network, a breadth-first search was used to extract the neighbors of anchor nodes, resulting in a "one-hop" ("1-hop") network for each layer. LOE scores were calculated as per Weighill et al. (2018). Briefly, the LOE breadth score for a gene is the count of the different layers in which that gene has connections to anchor genes/phenotypes. An LOE depth score – the count of all connections to anchor genes/phenotypes across all data layers – was also calculated for each gene. After scoring, the 1-hop networks from all layers were thresholded based on the distribution of LOE breadth scores, then merged to form the LOE network containing cell wall anchor genes and phenotypes and all genes connected to them via one or more layers ("high LOE genes"). All genes in the merged LOE network were ranked based upon breadth and depth scores and genes with previously documented cell-wall-related roles were removed. Networks were visualized and manipulated with Cytoscape 3.6.1 (Shannon et

al., 2003).

2.3.9 Gene Annotation, Functional Enrichment, and Expression Analysis

Functional annotations for *P. trichocarpa* genes were obtained from JGI Phytozome 12 (Goodstein et al., 2011) and MapMan using the Mercator tool (Lohse et al., 2014). A number of high LOE genes were not annotated in MapMan or Phytozome. To better understand the potential functions of those genes, protein sequences were extracted from the *P. trichocarpa* v3.1 primary transcript sequence (Tuskan et al., 2006) available from Phytozome and analyzed using HMMER v3.1b2 (Eddy, 1998) to annotate both Pfam v31.0 (Punta et al., 2011) and TIGRfam v15.0 (Haft et al., 2001) domains. Domains were thresholded using an independent E-value of 0.001. GO-term enrichment was performed on selected sets of genes using the BinGO Cytoscape app (Maere et al., 2005) using the Hypergeometric Test as well as Benjamini & Hochberg False Discovery Rate Correction at a significance level of 0.1.

A clustered heatmap of gene expression data was created using the Python (v3.6.2) package seaborn (v0.8.0; https://seaborn.pydata.org/index.html). Prior to analysis, six samples were removed from the data set that were outliers relative to their tissue type and treatment subgroups. Gene expression was normalized across tissues and genes were clustered using a Euclidean distance metric and Ward clustering method.

To assess orthology for a subset of genes during post-hoc analyses in Section 4.4.1, amino acid sequences containing characteristic PFAM domains (http://pfam.xfam.org/) were obtained from UniProt (www.uniprot.org; KNOXI: PF03790 per Mukherjee et al., 2009; POX/BELL: PF07526 per (Bellaoui et al., 2001) and reciprocal BLASTp searches were performed against *P. trichocarpa* and *A. thaliana* genomes using NCBI's BLAST (https://blast.ncbi.nlm.nih.gov) with default settings.

2.3.10 Network Validation

2.3.10.1 Randomizations of Expression and Methylation Data

We assessed whether our coexpression and comethylation networks contain greater biological signal than random networks by performing analyses on multiple randomized expression and methylation datasets. First we generated 100 randomized gene expression data sets by shuffling TPM values within genes across tissues, thereby preserving the observed range of expression for each gene but destroying the associations with tissue samples. We then generated a Spearman coexpression matrix for each random dataset and randomly subsampled 100,000 correlation values from each, resulting in a total pool of 10,000,000 random coexpression samples. We then collected 10,000,000 random subsamples from our observed coexpression data set and compared the distributions of our observed values to those of the shuffled data sets using the Wilcoxon rank-sum test using the Python package SciPy stats module (docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ranksums.html). We also performed this method with the comethylation data layer.

2.3.10.2 Functional Validation of LOE Network

To assess whether our observed LOE network captured a greater amount of biological function than random networks, we intersected the observed network as well as 100 randomized LOE networks with a GO-term functional network. We first constructed a functional network from GO Biological Process terms whereby genes that share GO terms are connected and are more likely to share biological function than unconnected genes. GO annotations for *P. trichocarpa* genes were obtained from PlantRegMap (Jin et al., 2017) and we removed any term present in over 1000 genes to avoid generating an overly dense network from highly generic functions. Furthermore, we weighted edges with a score inversely proportional to the number of genes with that GO term, such that between genes due to rarer GO terms were considered more functionally valuable than edges due to broader GO terms. If two genes shared multiple GO terms then we retained only the higher scoring edge. We then generated 100 randomized networks for each input data layer by holding anchor nodes and edges constant and replacing their 1-hop

neighbors with gene labels randomly drawn from the genome, thereby ensuring that the size and structure of the randomized networks were comparable to the LOE input networks. For each set of random networks (consisting of one randomized network of each type: comethylation, coexpression, SNP correlation, traditional metabolite-GWAS, and rare variant metabolite-GWAS), LOE scoring and thresholding was performed. Each merged LOE network was then intersected with a GO-term functional network and an intersect score was recorded. The intersect score is calculated by summing the values of the GO-term network edges that are also present in the LOE scored network. We then compared the intersect score of our observed LOE network to the distribution of randomized network intersect scores.

2.3.10.3 Expression Quantitative Trait Networks

We utilized eQTN data as an independent line of evidence for investigating the putative regulatory roles of the *PtGRF9* paralogs. RNAseq sequencing data from (Zhang et al., 2018b) were obtained from the NCBI SRA database (SRA numbers: SRP097016 - SRP097036; www.ncbi.nlm.nih.gov/sra). Reads were aligned to the *Populus trichocarpa* v.3.0 reference (Tuskan et al., 2006), using STAR (Dobin et al., 2013). Transcript per million (TPM) counts were then obtained for each genotype, resulting in a genotype-transcript matrix. For each gene transcript we determined outlier values, masking TPM values that exceeded a median absolute deviation from the non-zero median threshold of 5.0. Transcripts that had a non-outlier observed TPM value in more than 20% of the population were retained for further analysis. These expression profiles were then used as phenotypes in a Genome Wide Association Study (GWAS), using EMMAX (Kang et al., 2010a). Single nucleotide polymorphisms (SNPs) data, for the same population of *P. trichocarpa* genotypes, was obtained from (DOI 10.13139/OLCF/1411410). The SNPs were processed using VCFTOOLS (Danecek et al., 2011b) and PLINK (Purcell et al., 2007a), selecting for the 90% tranche and a minor allele frequency of 0.01. A hierarchical approach (Peterson et al., 2016) was used to correct for multiple hypotheses bias associated with the number of phenotypes. The procedure involved two rounds of false discovery rate (FDR) corrections, the initial using the Benjamini-Hochberg (Benjamini

and Hochberg, 1995) procedure (q1 < 0.1), followed by the Gavrilov-Benjamini-Sarkar stepdown approach (Gavrilov et al., 2009) (q2 < 5.1e-4). SNP to phenotype association that passed the respective thresholds were determined to be statistically significant. 1-hop eQTN networks were then created around the *PtGRF9* paralogs.

2.4 Results and Discussion

2.4.1 Evaluation of Expression and Methylation Data

The Wilcoxon rank-sum test was used to determine whether the distribution of correlation values differed between our observed data set and values from randomized datasets (**Supplemental Figure 2.1**). For both the expression and methylation data sets, the observed distributions were significantly different to random ($p < 0.01$ for both data types). Our coexpression data layer was thresholded to exclude correlation values below 0.85, resulting in 16,122 values (0.19 %) being retained. In our shuffled data set, only 45 values (or 5.25e-04 %) were above the 0.85 threshold. Our comethylation data layer was thresholded to exclude correlation values below 0.95, resulting in 87,458 values (0.88%) being retained. In our shuffled data set, only 1,090 values (0.01 %) were above the 0.95 threshold.

2.4.2 Construction of LOE Network

The LOE method was used to identify new candidate genes involved in regulating the cell wall in *P. trichocarpa* by jointly probing five different omics data layers. LOE depth scores were calculated for each gene, indicating the number of lines of evidence within each layer connecting that gene to an input set of cell wall anchor genes and metabolites. An LOE breadth score was also calculated for each gene, indicating the number of types of lines of evidence that connected the gene to input cell-wall-related targets. A merged LOE network was created after determining an appropriate LOE breadth score threshold and taking the union of all thresholded input networks. Threshold criteria dictated that candidate genes have a significant association with one or more metabolites in either the traditional or rare variant data layers as well as a total breadth score of three. We required a minimum of one GWAS association for retention in the merged network because

metabolite-GWAS associations represent a measurable cell wall phenotype. A breadth score of three was selected in order to prioritize a small set of genes having strong evidence for involvement in cell-wall-related processes, and the distribution of breadth scores exhibits an inflection point at three (**Supplemental Figure 2.2A**). These criteria identified a list of 315 "high LOE genes" as potential candidates for involvement in cell-wall-related functions. Seven high LOE genes had a breadth score of four and 308 had a breadth score of three (**Supplemental Figure 2.2B**). Overall, high LOE genes were from a variety of functional categories (**Supplemental Figure 2.3**) and 80 of these genes were annotated with potential regulatory functions (**Supplemental Table 2.4**).

2.4.2.1 Candidate Gene Ranking

To prioritize candidates, we created three ranked tiers to which high LOE genes were assigned (Tier 1 is the highest priority, Tier 3 is the lowest priority). Genes were ranked by 1) breadth score and 2) total depth score minus co-methylation depth score. While our co-expression data vectors contain 64 data points per gene (64 tissues and experimental conditions), our co-methylation data vectors contain only ten data points per gene (ten tissues and experimental conditions), resulting in an increased probability for spurious correlations in the co-methylation data layer. While the distribution of comethylation correlation values (**Supplemental Figure 2.1B**) was significantly different than random, the shape of the distribution suggests a conservative approach is warranted. In order to avoid upwardly biasing gene rankings, co-methylation data was included in the first stage of the ranking process (overall rank by breadth score) but excluded from the second stage of the ranking process (ranking within breadth score bins by depth score). Genes with an LOE Breadth score of four were included in Tier 1 by default (seven genes). In addition, genes with an LOE Breadth score of three and total depth minus comethylation depth scores of five or greater were included in Tier 1, resulting assignment of 45 genes. Thirty-two genes were assigned to Tier 2 based on a total depth minus comethylation depth score of four. The remaining 238 high LOE genes had total depth minus comethylation depth scores of three or less and were assigned to Tier 3.

2.4.2.2 Functional Validation of LOE Network

Intersection of the observed thresholded LOE network with the global GO-term functional network resulted in an intersect score of 0.4953, whereas intersect scores for the 100 randomized LOE networks (also thresholded) ranged from 0 to 0.3701 (**Figure 2.2A**). Intersection of the observed LOE network with the cell wall-specific GO-term network resulted in a score of 0.4806; intersect scores for the 100 randomized networks ranged from 0 to 0.3470 (**Figure 2.2B**). These results imply that our observed LOE network captures a greater amount of biological signal than the randomized LOE networks.

2.4.2.3 Literature Evidence

Recovering genes for which cell-wall-related functions have been previously reported is an important internal validation for the LOE method. We performed an extensive literature review to find evidence of previously validated genes in our results set. Forty-four genes were recovered with previous validation regarding cell-wall-related functions in *P. trichocarpa*, Arabidopsis, or other plant species and for which there is evidence of orthology in *P. trichocarpa* (**Supplemental Table 2.5**). Fifteen of these high LOE genes were also in our anchor gene list. Genes with prior evidence of cell-wall-related functions were removed from our merged LOE network in order to present researchers with "new" candidate genes: fourteen from Tier 1, four from Tier 2, and eleven from Tier 3; seventeen of these genes are represented in **Figure 2.3**. However, the literature review process was not as thorough for Tiers 2 or 3, thus it is possible that some of the remaining genes in these tiers have prior connections to cell wall processes. The full ranked and filtered high LOE gene list can be found in **Supplemental Table 2.6**. For the remainder of the manuscript, we focus on Tier 1 genes.

A notable example of a high LOE gene with prior evidence of a cell wall regulatory role is IQ-domain 10 calcium-signaling gene *PtIQD10* (Potri.011G096500). *PtIQD10* has a breadth score of three and a depth score of 48, including rare variant metabolite-GWAS associations with syringin, coniferin, and xylulose, and significant coexpression and

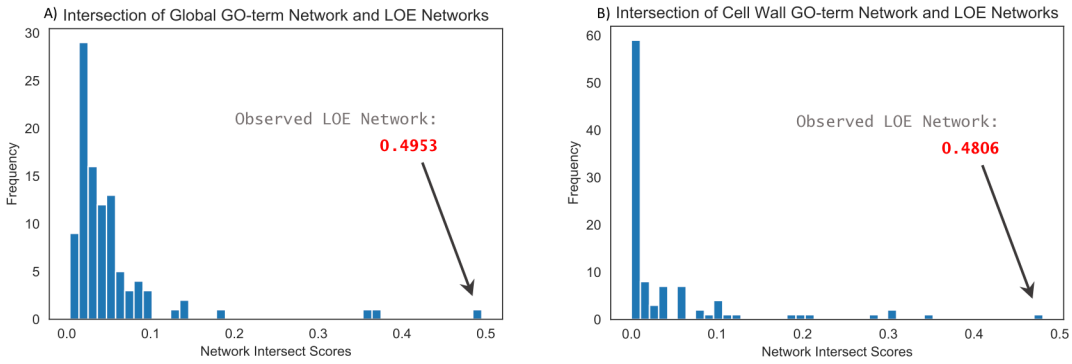


Figure 2.2: Histograms of network intersect scores calculated by intersecting the observed and randomized LOE networks with GO-term functional networks. (A) Intersection with the global GO-term functional network resulted in a score of 0.4953 for the observed LOE network; intersect scores for randomized networks were ≤0.3701. (B) Intersection with the cell wall-specific GO-term functional network resulted in a score of 0.4806 for the observed LOE network; intersect scores for randomized networks were ≤0.3470.

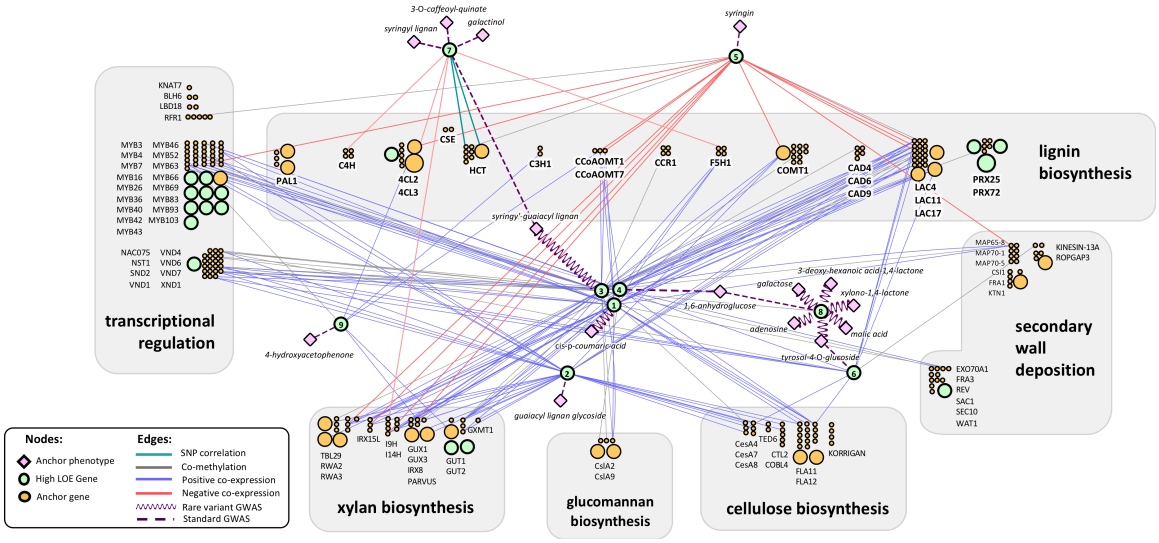


Figure 2.3: Tier 1 high LOE genes with regulatory annotations in the context of the LOE network arranged to highlight each gene's connection to lignin/cell wall biosynthesis pathways. Orange and green circles represent cell wall anchor genes and high LOE genes, respectively. Numbers within high LOE genes (green circles) indicate an entry within Table 2.1. Green circles that do not contain numbers represent a subset of the high LOE genes that were filtered from the final results set due to having prior evidence of cell-wall-related functions in the literature. The size of circles corresponds to their LOE breadth score. Gene symbols are Arabidopsis Best-hit matches.

comethylation with 41 cell wall anchor genes (**Supplemental Table 2.5**, **Supplemental Figure 2.4**, and **Supplemental Table 2.7**). The Arabidopsis ortholog *AtIQD10* (AT3G15050; orthology with *PtIQD10* and *P. deltoides PdIQD10* supported by phylogenetic analysis in Badmi et al., 2018) is differentially expressed in Arabidopsis lines overexpressing the transcription factor SECONDARY WALL-ASSOCIATED NAC DOMAIN PROTEIN 2 (*AtSND2*) (Hussey et al., 2011). Hussey et al. (2011) hypothesize *AtIQD10* activates AtSND1 NAC, followed by activation of SND2, MYBs, and cell wall polymerization functions. Consistent with this model, orthologs of these genes are present in the *PtIQD10* one-hop neighborhood (**Supplemental Figure 2.4**). Additional evidence has recently been observed in *P. trichocarpa* congeners. An ortholog of *PtIQD10* in the *P. alba* x *P. glandulosa* hybrid "84k" is differentially expressed during the transition between primary and secondary growth phases in stems (Li et al., 2017). In addition, *P. deltoides* ortholog *PdIQD10* has higher expression levels in tension-stressed xylem tissues and secondary walled cells, and RNAi repression of *PdIQD10* results in altered phenotypes such as increased cellulose, wall glucose content, plant height, stem count, and stem density (Macaya-Sanz et al., 2017; Badmi et al., 2018). *PdIQD10* is coexpressed with secondary cell wall related genes such as *SUSY*, *CESAs*, and *KOR* (Badmi et al., 2018), orthologs of which are present in our PtIQD10 subnetwork (Potri.018G103900 cellulose synthase / *PdCesA7-B* / *AtCESA7* and Potri.004G059600 *PtCESA.2* / *PdCESA8-B* / *AtCESA8*; see **Supplemental Table 2.7** for the *PtIQD10* one-hop subnetwork node information for **Supplemental Figure 2.4**).

In another example of a high LOE gene with prior evidence of a cell-wall-related role, Porth et al. (Porth et al., 2013) found that a SNP in an exostosin family protein gene (Potri.019G044600) involved in xylogalacturonan biosynthesis was correlated with xylose (hemicellulose) content. In yet another example, (Pomiès et al., 2017) found a berberine bridge enzyme gene (Potri.011G161500) with orthology to *AtEDA28/MEE23* (AT2G34790, shown to play a role in lignin monolignol metabolism) was highly up-regulated 72 hours after mechanical perturbation of stems as plants modified cell wall properties in response. Another example with growing evidence of cell-wall-related

regulatory functions is MADS-box transcription factor *PtAGL12* (Du et al., 2009, 2011; Weighill et al., 2018; see **Supplemental Text 2.8.1**, **Supplemental Figure 2.5**, **Supplemental Table 2.8**, and **Supplemental Figure 2.6** for additional evidence regarding the putative role of *PtAGL12* in regulating cell wall biosynthesis).

2.4.3 Tier 1: Highest Priority Candidates for Cell Wall Regulation

Tier 1 genes have the strongest evidence of involvement in cell wall related processes (**Table 2.1**). Of these, nine genes had regulatory annotations (via MapMan, www.arabidopsis.org, or PFAM; see **Supplemental Table 2.4** for categories considered regulatory). While the remaining 21 genes did not have regulatory annotations, our results suggest they play a role in cell wall biosynthesis.

Among Tier 1 regulatory genes, there were a total of eighteen metabolite-GWAS associations, eight of which were rare variant hits (**Figure 2.3**). Potri.013G093800 (Arabidopsis homolog AT1G71350, a eukaryotic translation initiation factor SUI1 family protein) has the highest number of rare variant metabolite-GWAS associations (six) of any high LOE gene as well as the highest number of total combined GWAS edges (seven). Most Tier 1 regulatory genes share edges with cell wall anchor genes from multiple process categories (in **Figure 2.3**, gray boxes indicating functional groupings of cell wall anchor genes). On average, Tier 1 genes were connected by multiple edges to four different functional groups, suggesting that Tier 1 genes influence multiple aspects of cell wall biosynthesis. Furthermore, eight Tier 1 regulatory genes shared edges with anchor cell wall transcriptional regulation genes (**Figure 2.3**).

Notably, coexpression edges for Tier 1 regulatory genes were either strictly negative for a given gene, or strictly positive, perhaps hinting at the regulatory mechanism of each gene. Two Tier 1 regulatory genes (Potri.015G006200: GROWTH-REGULATING FACTOR 9 / *PtGRF9* and Potri.018G105600: NUCLEOID-ASSOCIATED PROTEIN YBAB) were negatively coexpressed with cell wall genes and six were positively co-expressed with cell wall genes. The negatively coexpressed genes (Potri.015G006200,

Table 2.1: Tier 1 high LOE genes. See Supplemental Table 2.9 for additional score and annotation information.

Node #	Gene ID	Arabidopsis gene/domain symbol	Description	Arabidopsis ortholog
Regulatory genes				
1	Potri.008G112300		DNA glycosylase superfamily protein	AT1G13635.2
2	Potri.001G216000	EAR1	ENHANCER OF ABA CO-RECEPTOR 1	AT5G22090.1
3	Potri.013G060500	ATCRT1	RING/U-box superfamily protein	AT5G56340.1
4	Potri.013G156300	Shisa	Wnt and FGF inhibitory regulator	
5	Potri.015G006200	AtGRF7, GRF7	growth-regulating factor 7	AT5G53660.1
6	Potri.017G053000	AMC1, ATMC1, ATMCPB1	metacaspase 1	AT1G02170.1
7	Potri.018G105600	YbaB_DNA_bd	YbaB/EbfC DNA-binding family	AT2G24020.1
8	Potri.013G093800		eukaryotic translation initiation factor SUI1 family protein	AT1G71350.1
9	Potri.010G072700		RING/U-box superfamily protein	AT5G43420.1
Other genes				
	Potri.004G085400	ATGLN1;1, ATGSR1, GLN1;1, GSR 1	glutamine synthase clone R1	AT5G37600.1
	Potri.006G256000		Phox (PX) domain-containing protein	AT4G32160.1
	Potri.012G093800	ATNDPK2, NDPK1A, NDPK2	nucleoside diphosphate kinase 2	AT5G63310.1
	Potri.010G155600		Leucine-rich repeat transmembrane protein kinase	AT1G53440.1
	Potri.001G340400	SEO_N	SEO_N--Sieve element occlusion N-terminus	
	Potri.006G153300		N-acetylated-alpha-linked acidic dipeptidase (NAALAD)	AT5G19740.1
	Potri.008G156600	AST12, SULTR3;1	sulfate transporter 3;1	AT3G51895.1
	Potri.003G079900	AW: HRGP	hydroxyproline-rich glycoprotein family protein	AT4G16790.1
	Potri.T135500	CYCP4;1	cyclin p4;1	AT2G44740.1
	Potri.018G090300	AHA1, HA1, OST2, PMA	H(+)-ATPase 1	AT2G18960.1
	Potri.017G059300	SHM4	serine hydroxymethyltransferase 4	AT4G13930.1
	Potri.004G059900		Protein of unknown function DUF1685	AT2G42760.1
	Potri.016G115200	LHCB4.2	light harvesting complex photosystem II	AT3G08940.2
	Potri.015G063400	ATATH2, ATH2	ABC2 homolog 2	AT3G47740.1
	Potri.019G087700	ATSERK1, SERK1	somatic embryogenesis receptor-like kinase 1	AT1G71830.1
	Potri.007G027400		anti-muellerian hormone type-2 receptor	AT3G50685.1
	Potri.005G067000		Protein kinase protein with adenine nucleotide alpha hydrolases-like domain	AT1G77280.1
	Potri.001G352200	ATPUP10, PUP10	purine permease 10	AT4G18210.1
	Potri.011G142200	PSBR	photosystem II subunit R	AT1G79040.1
	Potri.006G060100	CRR6	chlororespiratory reduction 6	AT2G47910.1
	Potri.010G113700	FAB1C	FORMS APLOID AND BINUCLEATE CELLS 1C	AT1G71010.1

Potri.018G105600) did not share any neighbor nodes, however they are both connected to lignin and xylan biosynthesis genes. In contrast, positively coexpressed Tier 1 regulatory genes had a large overlap in neighbor cell wall anchor genes. The overlap was even more pronounced among Potri.008G112300, Potri.001G216000, Potri.013G060500, and Potri.013G156300 despite a complete lack of overlap among metabolite-GWAS edges or MAPMAN functional annotations (**Supplemental Table 2.9**).

We conducted an in-depth investigation into the Tier 1 regulatory gene *PtGRF9* (Potri.015G006200) to assess support for *PtGRF9* playing a regulatory role in cell wall biosynthesis.

2.4.4 GROWTH-REGULATING FACTOR 9: Putative Master Regulator

The transcription factor gene GROWTH-REGULATING FACTOR 9 (*PtGRF9* / Potri.015G006200) had a breadth score of three and depth score of seventeen, including thirteen negative coexpression edges (the highest negative coexpression depth score in our analysis). *PtGRF9* shared nine edges with lignin biosynthesis genes, four edges with xylan biosynthesis genes, two edges with transcriptional regulation genes, and one edge with a secondary cell wall deposition gene.

The *P. trichocarpa* genome annotation (v3.0; available on https://phytozome.jgi.doe.gov; (Tuskan et al., 2006) indicates the best-hit Arabidopsis match for *PtGRF9* is AT5G53660 (growth-regulating factor 7, *AtGRF7*). To assess support for orthology, we performed reciprocal BLASTp searches of amino acid sequences containing the WRC (PF08879) and QLQ (PF08880) domains from *A. thaliana* and *P. trichocarpa* (obtained from UniProt; www.uniprot.org) and a phylogenetic analysis (see **Supplemental Figure 2.7**). Our results support an orthologous relationship between *PtGRF9* and *AtGRF7*, which is consistent with the phylogenetic analysis of Cao et al. (2016). While investigating support for orthology between *PtGRF9* and *AtGRF7*, we discovered a second *AtGRF7* ortholog in the *P. trichocarpa* genome (Potri.012G022600; hereafter, Potri.015G006200 is referred to as "*PtGRF9a*" and Potri.012G022600 as "*PtGRF9b*"; **Supplemental**

Figure 2.7). *PtGRF9b* was not present in our set of high LOE genes because it has a breadth score of 2 and was not associated with any cell wall phenotypes through GWAS analyses. Because *PtGRF9b* had strong positive coexpression with *PtGRF9a* and shared edges with many cell wall genes, we included *PtGRF9b* in further analyses.

We constructed genome-wide 1-hop networks around each *PtGRF9* paralog across all data layers to assess the functional annotations of nearest neighbors (**Figure 2.4**; see **Supplemental Table 2.10** for detailed information about nodes). *PtIQD10* is present in the 1-hop network, along with many other genes with documented roles in cell wall processes. *PtGRF9a* and *PtGRF9b* are jointly positively co-expressed with fourteen genes (one of which is a high LOE gene related to cell wall processes) and are jointly negatively co-expressed with 27 genes (including seven cell wall anchor genes and two high LOE genes), implying an overlap in function. However, the bulk of neighbor genes are unique to each paralog, indicating divergence and perhaps specialization for specific tissues and conditions. GO-term functional enrichment analysis of the negative co-expression nodes in the 1-hop network showed significant enrichment for cell wall biological processes, including lignin biosynthesis, xylan biosynthesis and cell wall organization or biogenesis (**Supplemental Table 2.11**). In addition, the metabolite-GWAS association between *PtGRF9a* and syringin (a monolignol glucoside) indicated this SNP is associated with an allelic effect on syringin concentration (**Supplemental Figure 2.8**), further implicating *PtGRF9a* and *PtGRF9b* as repressors of secondary cell wall formation.

In Arabidopsis, *AtGRF7* is one of nine members of the *GRF* family of transcription factors (there are 20 *GRF* homologs in *P. trichocarpa*) that affect growth via multiple mechanisms (Omidbakhshfard et al., 2015). AtGRF7 has specifically been shown to modulate drought response by repressing *DREB2A* (Joshi et al., 2016) which ensures that drought response genes normally activated by DREB2A are not expressed under non-drought conditions, thus avoiding reduced growth. In addition to stress response, *GRF* genes are involved in regulating cell proliferation and differentiation in the shoot apical

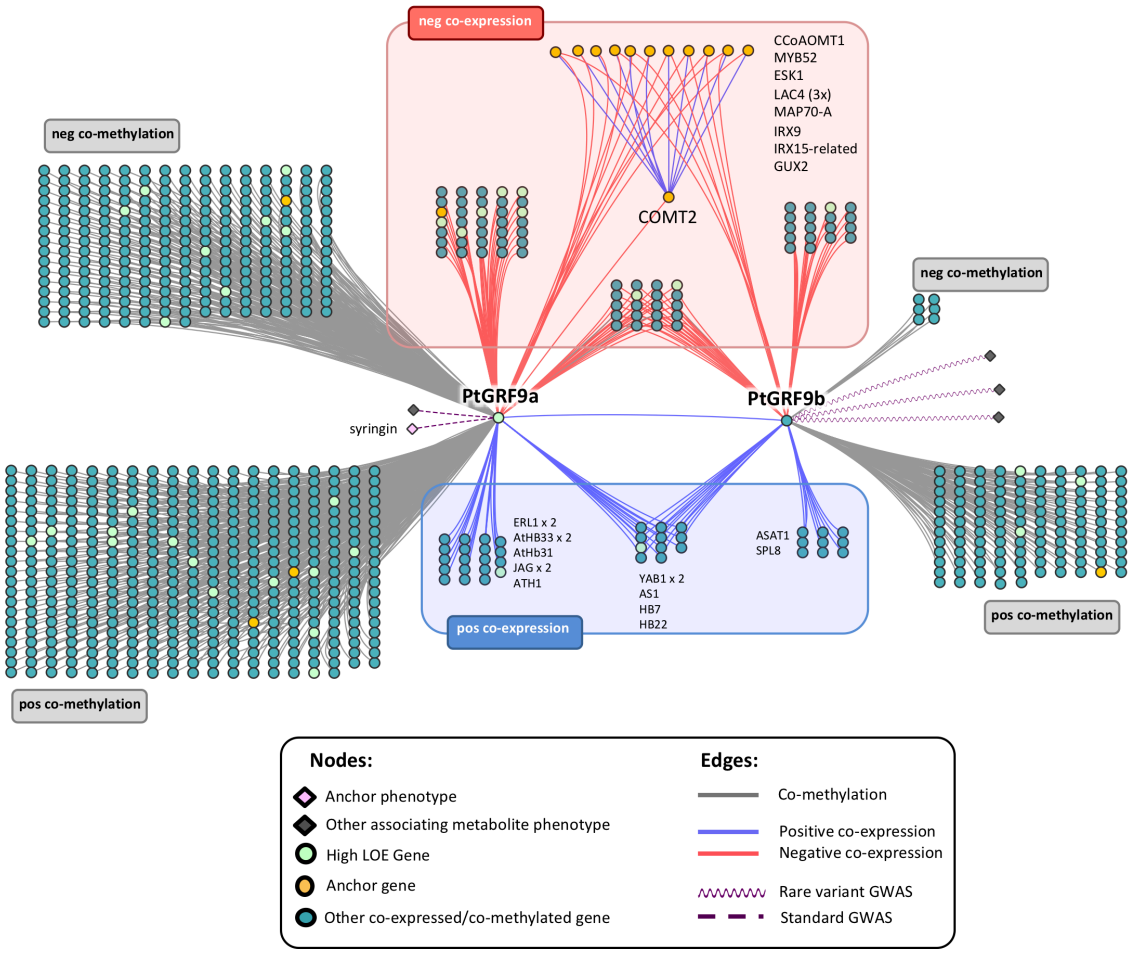


Figure 2.4: Genome-wide neighborhood of *PtGRF9* paralogs in the global input layer networks. Gene symbols are Arabidopsis Best-hit matches. See Supplemental Table 2.10 for detailed node information; for functional enrichment details, see Supplemental Tables 2.11 (negatively co-expressed genes) and 2.12 (positively co-expressed genes).

meristem (SAM). *GRF* genes therefore impact the elongation of stems, new leaf initiation, and the size and shape of leaves (Gonzalez et al., 2012). The phenotypic penetrance may occur as part of a complex formed with GRF Interacting Factor (GIF1/AN3) proteins (Hoe Kim et al., 2015), where the GRF-GIF complex serves as a transcriptional activator, recruits chromatin remodeling complexes, and regulates the meristematic state of a tissue.

GO-term enrichment analysis of the positive coexpression nodes in the *PtGRF9* 1-hop network was consistent with roles reported in the literature for *GRF* genes (**Supplemental Table 2.12**). The most significantly enriched Biological Process GO terms include specification of axis polarity, shoot system development, shoot system morphogenesis and negative regulation of cell proliferation. Numerous osmotic-stress related genes are also found in the *PtGRF9* network (e.g., *AHA1/OST2*, *ERL1*, *PIP2;2*, *TIP4;1,* and *AREB3*), reflecting the well-documented relationship between *AtGRF7* and drought response. Significant connections between the *PtGRF9* paralogs and *PtGIF1* or *PtDREB2A* are not present in our LOE network. On closer inspection of co-expression values across tissues we see that *PtGRF9a* and *PtGIF1* do coexpress strongly in bud and immature leaf, but expression diverges in mature leaf and roots which causes the strength of coexpression to fall just below our 0.85 threshold (**Supplemental Figure 2.9**). The case is less clear for *PtDREB2A* as it shows little expression in most tissues.

Evidence that the *PtGRF9* paralogs play roles in regulating growth, defense, stress response, secondary growth, and cell wall biosynthesis suggest that *PtGRF9a* and *PtGRF9b* could be transcriptional co-regulators as described by (Xie et al., 2018b), acting as master regulators that direct the global allocation of energy within a plant.

2.4.4.1 Evidence for Regulation of the Cell Wall by PtGRF9

To date, a role for the *GRF* family in cell wall regulation has not been reported, though it has been noted that cell proliferation and timing of differentiation must require control or delay of secondary cell wall deposition (Mele et al., 2003). Barros et al. (2015) noted that

lignin cannot be removed once deposited, thus, specific regulatory mechanisms are required to control lignin biosynthesis and deposition at specific stages during cell differentiation. The contrasting patterns of coexpression between cell wall biosynthesis and meristematic control in our *PtGRF9* 1-hop network (**Figure 2.4**) suggest that it could be involved in such a mechanism. Furthermore, the GWAS association with syringin suggests that allelic variation in *PtGRF9a* in this population may have an additive effect on the amount of sinapyl alcohol stored or released for cell wall lignification.

Knowledge regarding downstream targets of *GRF* genes is incomplete (see Omidbakhshfard et al., 2015 for a comprehensive review). AtGRF7 has been shown to repress *AtDREB2A* by binding to the motif *TGTCAGG* (Kim et al., 2012). Additionally, the central *CAG* sub-motif is enriched in the promoter of *KNOX* genes that are targeted by *GRFs* (Kuijt et al., 2014). We searched for the complete *TGTCAGG* motif in the promoter regions of Arabidopsis homologs of the genes that coexpress with *PtGRF9a* using the online *athamap.de* tool, revealing two potential AtGRF7 targets in our 1-hop network: caffeoyl coenzyme A *O*-methyltransferase 1 (AT4G34050 / *AtCCoAOMT1*) and MADS-box transcription factor *AtAGL12* (AT1G71692). Both genes are relevant to the cell wall, and *P. trichocarpa* homologs of these genes are negatively co-expressed with *PtGRF9a*. To further investigate these genes as potential PtGRF9a targets, we used Analysis of Motif Enrichment (AME) (McLeay and Bailey, 2010), but found no evidence for enrichment of the *TGTCAGG* motif in the 2-kb upstream or CDS regions of *PtCCoAOMT* (Potri.001G304800 and Potri.009G099800) or *PtAGL12* (Potri.013G102600). Manual examination revealed that the *TGTCAGG* motif appears inexactly in the upstream regions of *PtCCoAOMT1* and *PtAGL12* (*TGTTCAGG* in *CCoAOMT1* Potri.009G099800; *TGTCAGC* in *PtCCoAOMT* Potri.001G304800 and *PtAGL12*). Consistent with the findings of Franco-Zorrilla et al. (2014), who show that repressor TFs such as PtGRF9a are more likely than activator TFs to bind downstream of a target gene, we found 27 *Populus* genes significantly enriched for *TGTCAGG* in the 1-kb downstream region, including *PtMYB41* (Potri.012G039400, a homolog of *AtMYB52*), which is negatively coexpressed with *PtGRF9a*. AtMYB52 is a TF known to induce

secondary cell wall biosynthesis genes and its repression reduces secondary wall thickening in fibers (Zhong et al., 2008). Furthermore, *AtMYB52* overexpression has been linked with drought tolerance (Park et al., 2011). Given the established role of AtGRF7 in drought response, repression of *PtMYB41* is a potential avenue for PtGRF9a to regulate both lignification and drought tolerance, although further experimental evidence is required.

Analysis of our 1-hop network suggests that *PtGRF9* also affects cell wall biosynthesis by regulating a host of homeobox genes. Twenty homeobox genes were present in the *PtGRF9* network, including *PtATHB.12* (Potri.001G188800; homolog of *AtHB15* / AT1G52150), which has been shown to influence secondary wall formation and cambial production of xylem (Schrader, 2004; Cassan-Wang et al., 2013), and *PtAGL12* (Du et al., 2009, 2011; Weighill et al., 2018) (**Supplemental Figure 2.5**). There was also indirect evidence in the *PtGRF9* network suggesting PtGRF9 interacts with *PtKNOX* genes. *KNOX* genes are involved in meristem maintenance and are downregulated to facilitate lateral primordia development and the differentiation of cambium into xylem and phloem (Hertzberg et al., 2001; Schrader, 2004; Hay and Tsiantis, 2010). *GRF* genes are involved in specification of primordia cells and have been shown to repress *KNOX* genes by forming hairpins in targeted regions (Kuijt et al., 2014; Hoe Kim et al., 2015). Interactions between AtGRF7 and *KNOX* genes have yet to be investigated, but the primary motif of the target sequence by which AtGRF7 binds *AtDREB2A* was shown to be enriched in several *KNOX* genes, and experiments in rice, barley, and Arabidopsis have confirmed that multiple *GRF* genes bind these motifs in *KNOX* genes (Kim et al., 2012; Kuijt et al., 2014). The presence of several genes that exclusively or directly interact with *KNOX* genes in the 1-hop network strongly implies that PtGRF9 proteins influence the cell wall via interactions with the *PtKNOX1* genes *PtSTM* and *PtBP*, and likely other *PtKNOX* genes as well (**Supplemental Table 2.13** and **Figure 2.5**). Although *KNOX* family genes were not present in the *PtGRF9* network, this was likely due to highly tissue-specific expression patterns which our coexpression analysis methods were not designed to detect (see **Supplemental Figure 2.10**).

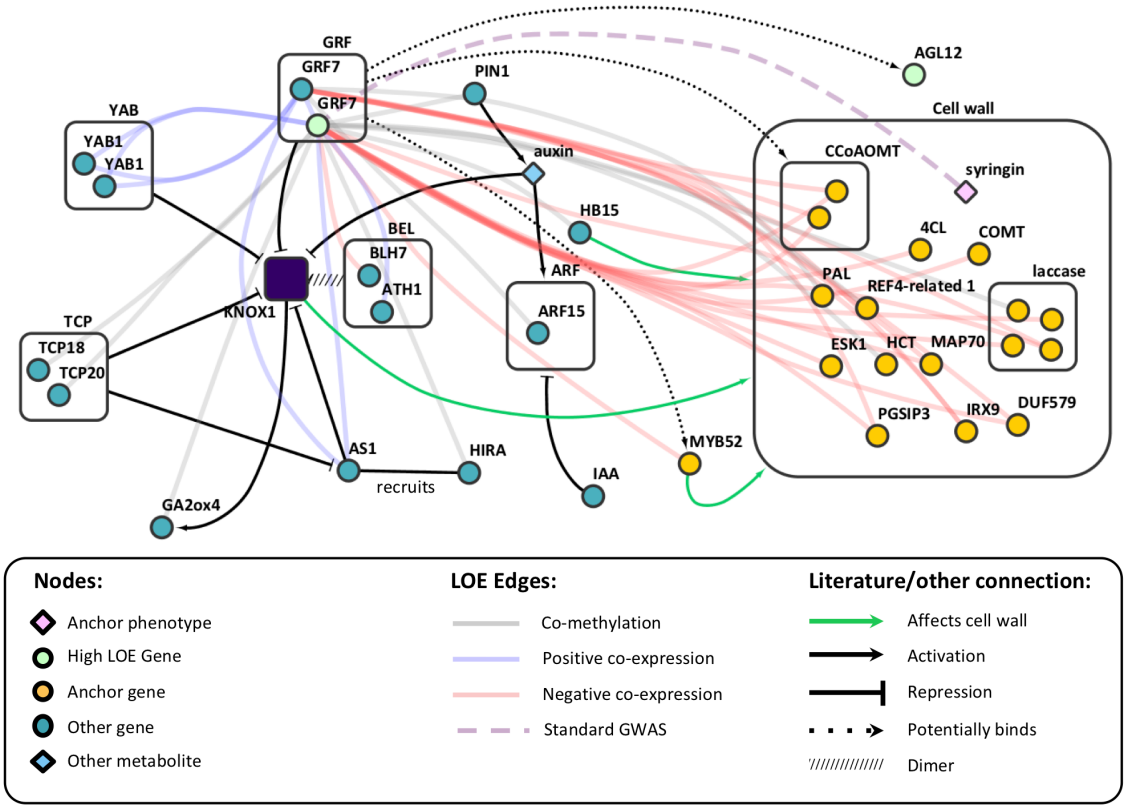


Figure 2.5: Literature evidence linking *AtGRF7* genes to *AtKNOX* genes and the cell wall, overlaid on LOE edges from the *PtGRF9* subnetwork. See Supplemental Table 2.13 for *P. trichocarpa* orthologs.

The *PtKNOX*-associated genes in the *PtGRF9* network have documented roles in cell wall and secondary growth phenotypes (**Figure 2.5**). SHOOT-MERISTEMLESS (PtSTM) downregulates gibberellic acid levels by repressing gibberellin 20-oxidase (*PtGA20ox*) biosynthesis genes and upregulating catabolism genes such as *PtGA2ox4* (positively co-methylated with *PtGRF9a*), which inhibits xylem production (Eriksson et al., 2000; Jasinski et al., 2005). Overexpression of *PtSTM*/ARBORKNOX1 (*PtSTM* / *PtARK1*) in *P. tremula* x *P. alba* has been shown to inhibit differentiation of leaf primordia, elongation of internodes, and differentiation of secondary vascular cells (Groover et al., 2006). Counterintuitively, overexpression of *PtSTM/PtARK1* in secondary meristems also results in upregulation of some lignin biosynthesis genes and increased lignin content. Long-term transcriptional repression of BREVIPEDICELLUS (*AtBP*), KNOTTED-like 2 from *A. thaliana* (*AtKNAT2)* and *AtKNAT6* outside the meristem is facilitated by chromatin remodeling carried out by the protein encoded by ASYMMETRIC LEAVES 1 (*AtAS1; PtAS1* is positively co-expressed with *PtGRF9a* and *PtGRF9b*), which dimerizes with AtAS2 and recruits the histone chaperone protein encoded by *AtHIRA* (*PtHIRA* is negatively co-methylated with *PtGRF9a*) (Guo et al., 2008; Hay and Tsiantis, 2010). AS2 is involved in controlling seasonal lignification in spruce, likely through its role in repressing *BP* (Jokipii-Lukkari et al., 2018). BP decreases lignin deposition and regulates the localization of lignification by binding the promoters of *AtCOMT1*, *AtCCoAOMT1*, laccases, and peroxidases (putative orthologs of which are all negatively co-expressed with *PtGRF9a* and *PtGRF9b*) (Mele et al., 2003). The *PtGRF9* network includes many of the cell wall biosynthesis-related genes that Mele et al. (2003) found to be differentially expressed in *bp* mutants, including five putative orthologs (*PAL1*, *OMT1*, two *CCoAOMT1* paralogs, *PME3*, and *GH9B5*; see **Supplemental Table 2.13**) and an additional 23 genes belonging to the same families as differentially expressed genes in *bp* mutants (*4CL2*, five *PME*s, *KCS19*, four peroxidases, four laccases, *ERD4*, *GAUT4*, *PUB24*, *MEE23*, *ERF1-3*, and three R2R3 MYBs: *MYB52*, *MYB93*, *MYB111*). Consistent with these observations in Arabidopsis, overexpression of *AtBP*/ARBORKNOX2 (*AtBP/AtARK2*) in *P. alba* x *P. tremula* results in downregulation of ABNORMAL FLORAL ORGANS (*PtAFO/PtYAB1)*, PIN-FORMED 1 (*PtPIN1*),

PtAGL12 (all negatively co-expressed with *PtGRF9a*) and *PtGA20ox* genes, leading to inhibition of cellular differentiation and division and decreases in biomass (Du et al., 2009). Furthermore, overexpression of *PtBP/PtARK2* results in downregulation of cell wall biosynthesis genes, decreased lignin content, reduced phloem fibers, and reduced secondary xylem in stems.

We did not find a connection between the *PtGRF9* genes and cell wall anchor genes *KNAT7* (Potri.001G112200, a *PtKNOX2* gene) and BEL1-like homeodomain 6 genes (*PtBLH6*, Potri.004G159300 and Potri.009G120800). These genes have well-documented roles in cell wall regulation (Li et al., 2012a; Cassan-Wang et al., 2013). However, the *PtGRF9* genes do not appear to be involved in their regulation, perhaps because *PtKNOX2* genes are generally more functionally diverse and broadly expressed than *PtKNOX1* genes (consistent with expression data in **Supplemental Figure 2.10**) (Furumizu et al., 2015). In addition, they are not involved in meristematic maintenance, and in some cases seem to overlap in function with genes that promote differentiation (consistent with what we recovered in our LOE analysis).

2.4.4.2 PtGRF9 *eQTN Network: An Independent Line of Evidence*

As a means of independently evaluating support for the hypothesis that *PtGRF9* paralogs regulate cell wall biosynthesis, we constructed 1-hop eQTN networks around *PtGRF9a* and *PtGRF9b* (**Figure 2.6**; detailed node information available in **Supplemental Table 2.14**). SNPs in both the *PtGRF9a* and *PtGRF9b* 1-hop networks were associated with cell wall expression phenotypes in leaf and xylem tissues, as well as expression phenotypes consistent with the previously documented roles of *AtGRF7* and other *GRF* orthologs in regulating functions such as growth, defense, and stress response. In agreement with the multiomic 1-hop network in **Figure 2.4**, the eQTN network indicated each paralog has connectivity with cell wall-related genes affecting multiple facets of cell wall biosynthesis, including transcriptional regulation, cellulose biosynthesis, lignin biosynthesis, xylan biosynthesis, and secondary cell wall deposition. Also consistent with the multiomic 1-hop network, the eQTN analysis indicated that despite a low degree of

topological overlap between the *PtGRF9a* and *PtGRF9b* neighborhoods, the paralogs still largely overlap in function.

To gain an understanding of how the *PtGRF9* paralogs potentially affect cell wall metabolites, the 1-hop eQTN network was merged with 1-hop anchor metabolite networks generated from traditional and rare variant metabolite-GWAS data layers. Beyond the direct GWAS association of *PtGRF9a* with syringin, fourteen additional anchor metabolites are present in the 2-hop eQTN to metabolite-GWAS network (**Figure 2.6**), six of which are indirectly associated with both paralogs through various intermediate genes. There appears to be a pattern of segregation regarding metabolite associations between tissue types and *PtGRF9* paralogs, perhaps indicating that these genes are diverging to fulfill different tissue-specific regulatory roles.

2.4.5 Future Directions

Our extended network analysis pipeline has provided a short list of putative cell wall regulatory genes to the scientific community for experimental validation. We performed an in-depth investigation of the *PtGRF9* paralogs, which are particularly promising candidates for regulation of cell wall biosynthesis and secondary growth. Furthermore, we show the *PtGRF9* paralogs are potential transcriptional co-regulators that coordinate the flow of energy among growth, defense, stress response, and lignification, in a manner consistent with the hypothesis of (Xie et al., 2018b). The ability to manipulate transcriptional co-regulators such as these via genetic engineering and breeding programs would provide a powerful tool for shaping bioenergy crops.

Incorporating a rare variant metabolite-GWAS data layer in the LOE analysis has proven to be a valuable asset in the identification of new candidate genes. Incorporating a genome-wide eQTN (SNP-to-expression-phenotype GWAS) data layer in future analyses would provide greater clarity regarding the mechanisms through which these genes regulate cell-wall-related functions. Furthermore, DNA affinity purification sequencing (DAP-seq) could provide further support for hypothesized transcription factor binding

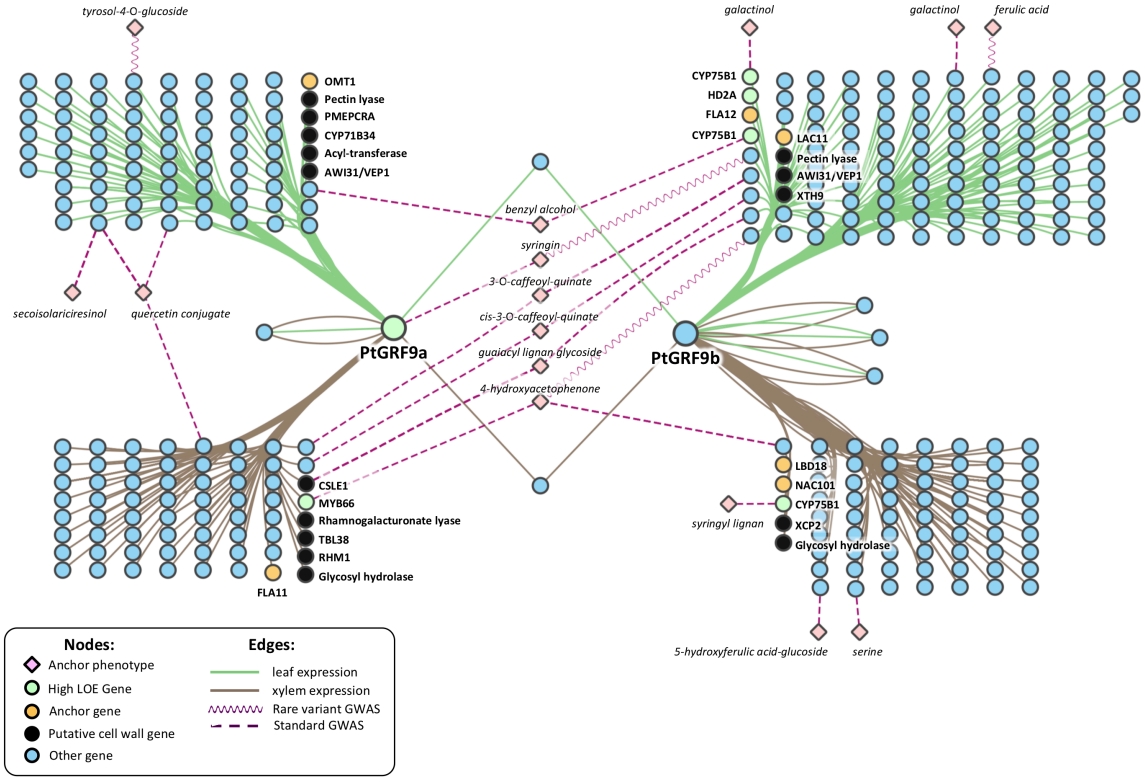


Figure 2.6: Two-hop network created by merging a 1-hop eQTN network around the *PtGRF9* paralogs and 1-hop metabolite-GWAS networks around anchor metabolites. See Supplemental Table 2.14 for node information.

sites, and thus help elucidate relevant transcription factor regulatory networks. Tissue-specific expression analysis across a GWAS population would allow for increased "tissue level resolution" of the regulatory networks. The extended network analysis pipeline will be a valuable tool to integrate these new layers with the previous networks to produce a holistic model of cell wall regulation.

2.4.6 Recent Work

Following the publication of this manuscript, we filed (2019) and were awarded a patent covering the putative activity of top ranked gene candidates in this study, particularly focused on *PtGRF9* (Tschaplinski et al, 2023, US Patent 11,725,215). I also co-led two follow-up projects to validate candidates identified in this project. The first project focused on confirming function and characterizing substrate specificities of UDP-glycosyltransferase (UGT) candidates in collaboration with Stephanie Galanie (Merck), Timothy Tschaplinksi (ORNL), and Patricia Saint-Vincent (ORNL). This culminated in a publication (Saint-Vincent and Furches et al., 2023) and a provisional patent (2022) covering glycosylation activity by these candidates. The second project is focused on confirming function and characterizing activity of candidates that putatively encode carbohydrate active enzymes (CAZymes), proteins of unknown function (PUFs), and xylan and glucomannan biosynthesis in collaboration with Breeanna Urbanowicz (Complex Carbohydrate Research Center, UGA). In August 2022, we were awarded a grant from the DOE Joint Genome Institute through the Center for Bioenergy Innovation (CBI) to fund experimental validation research.

2.5 Funding

Funding provided by The Center for Bioenergy Innovation (CBI). U.S. Department of Energy Bioenergy Research Centers supported by the Office of Biological and Environmental Research in the DOE Office of Science.

An award of computer time was provided by the INCITE program. This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of

Science User Facility supported under Contract DE-AC05-00OR22725. This research also used resources of the Compute and Data Environment for Science (CADES) at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE- AC05-00OR22725.

Support for the Poplar GWAS dataset was provided by The BioEnergy Science Center (BESC) and The Center for Bioenergy Innovation (CBI). U.S. Department of Energy Bioenergy Research Centers supported by the Office of Biological and Environmental Research in the DOE Office of Science. The Poplar GWAS Project used resources of the Oak Ridge Leadership Computing Facility and the Compute and Data Environment for Science at Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

The JGI Plant Gene Atlas project conducted by the U.S. Department of Energy Joint Genome Institute was supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

AL and AW acknowledge that this project was supported in part by appointments to the Higher Education Research Experiences (HERE) Program at Oak Ridge National Laboratory, administered by the Oak Ridge Institute for Science and Education (ORISE). ORISE is managed by Oak Ridge Associated Universities (ORAU) for the U.S. Department of Energy (DOE).

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

2.6 Acknowledgements

The authors would like to acknowledge Nancy Engle, David Weston, Ryan Aug, KC Cushman, Lee Gunter, and Sara Jawdy for metabolomics sample collection. We thank the

Department of Energy Joint Genome Institute (JGI) and collaborators for prepublication access to the Plant Gene Atlas Data.

2.7 Data Availability Statement

Populus trichocarpa genome sequence, annotation, and Gene Atlas expression data sets are available on Phytozome (http://phytozome.jgi.doe.gov). *P. trichocarpa* variant data (DOI 10.13139/OLCF/1411410) is available from https://doi.ccs.ornl.gov/ui/doi/55. Scripts used to calculate LOE scores, create GO-term networks, and calculate weighted intersect scores are available on GitHub: https://github.com/afurches/Cell_Wall_LOE.

2.8 Supplemental Materials

2.8.1 Supplemental Text

2.8.1.1 High LOE gene AGAMOUS-LIKE 12.

There is evidence that high LOE gene AGAMOUS-LIKE 12 (*PtAGL12*, Potri.013G102600) functions in a regulatory role in cell wall biosynthesis regulatory role is MADS box transcription factor (**Supplemental Table 2.4**). *PtAGL12* is significantly associated with cis-3-*O*-caffeoyl-quinate via traditional metabolite-GWAS and with fructose via rare variant metabolite-GWAS (**Supplemental Figure 2.5**) and is significantly coexpressed and comethylated with 20 cell wall anchor genes (**Supplemental Table 2.6).** AGAMOUS-LIKE genes form a large and diverse family, with many playing roles in floral organ identity, flowering time, and transition between reproductive and vegetative growth phases (Lee et al., 2000; Ng and Yanofsky, 2001; Masiero et al., 2011; Fernandez et al., 2014). There is also growing evidence that AGAMOUS-LIKE genes affect lignin content. A patent exists for altering the expression of *AtAGL8* to control lignin content in Arabidopsis (Yanofsky et al., 2004), and *AtAGL15* has been shown to regulate *PRX17*, which in turn affects lignin content (Cosio et al., 2017). In tomato, RNAi-silencing of *TAGL1* results in increased expression of lignin biosynthesis genes and lignin content (Giménez et al., 2010). Phylogenetic analysis indicates *PtAGL12* and Arabidopsis *AtAGL12* / AT1G71692 are orthologous (**Supplemental Figure 2.6**); birch *BpMADS12,* Arabidopsis *AtAGL12*, and *PtAGL12* are

SOC1/TM3 class MADS-box genes, which belong to the MIKC^{C}-type of the type II MADS-box genes (Leseberg et al., 2006; Li et al., 2016). Li et al. (2016) found that lines overexpressing the *BpMADS12* gene had significantly higher lignin content than wild-type lines, suggesting *BpMADS12* might play a role in regulating the expression of lignin biosynthesis genes, possibly via brassinosteroid signaling.

2.8.2 Supplemental Tables

See Attachment "Furches_Chapter2_Supplemental_Tables.xlsx" for the following:

Supplemental Table 2.1: Additional information about Gene Atlas expression data.

Supplemental Table 2.2: Cell wall anchor genes used as input for LOE analysis.

Supplemental Table 2.3: Cell wall anchor metabolites used as input for LOE analysis.

Supplemental Table 2.4: Number of high LOE genes with regulatory-like MapMan annotations.

Supplemental Table 2.5: High LOE genes with documented cell wall-related roles, including anchor genes retrieved as high LOE genes.

Supplemental Table 2.6: LOE scores, Priority Tiers, and functional annotations for ranked and filtered high LOE genes (metabolite-GWAS score ≥1, breadth score ≥3).

Supplemental Table 2.7: Node information for **Supplemental Figure 2.4** (one-hop LOE subnetwork surrounding *P. trichocarpa PtIQD10*).

Supplemental Table 2.8: Node information for **Supplemental Figure 2.5** (one-hop LOE subnetwork around *P. trichocarpa PtAGL12*).

Supplemental Table 2.9: Additional information for genes in **Table 1** (Tier 1 of ranked high LOE genes).

Supplemental Table 2.10: Node information for **Figure 2.4** (Genome-wide neighborhood of *P. trichocarpa PtGRF9* paralogs).

Supplemental Table 2.11: BinGO output for genes negatively co-expressed with *PtGRF9* paralogs.

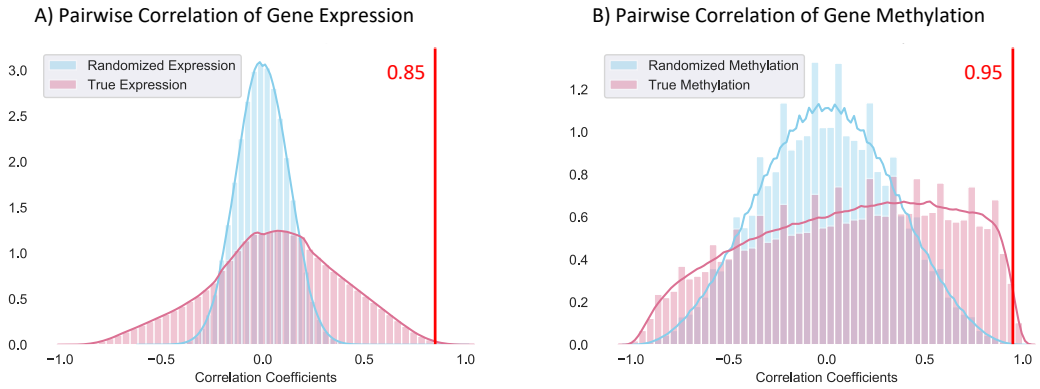
Supplemental Table 2.12: BinGO output for genes positively co-expressed with *PtGRF9* paralogs.

Supplemental Table 2.13: KNOX-associated and differentially expressed genes in bp

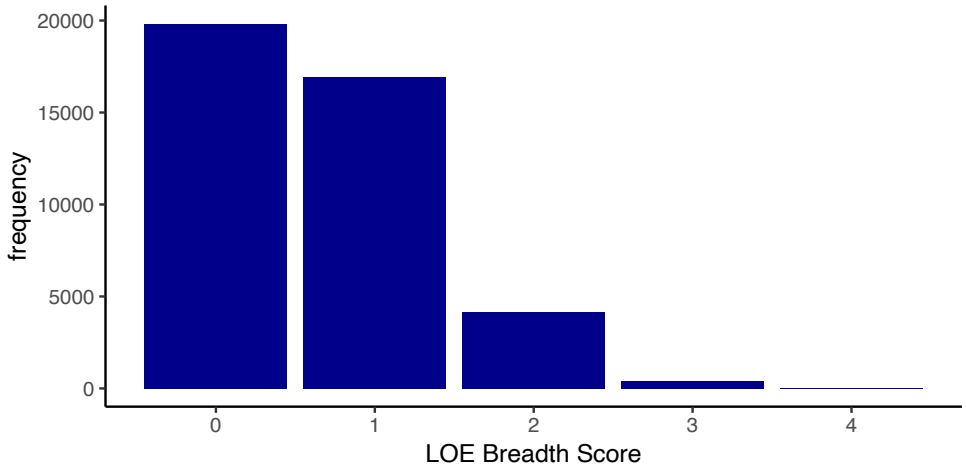
mutants (Mele et al., 2003) that are present in the *PtGRF9* genome-wide 1-hop network.

Supplemental Table 2.14: Nodes present in the eQTN one-hop network surrounding *PtGRF9* paralogs.

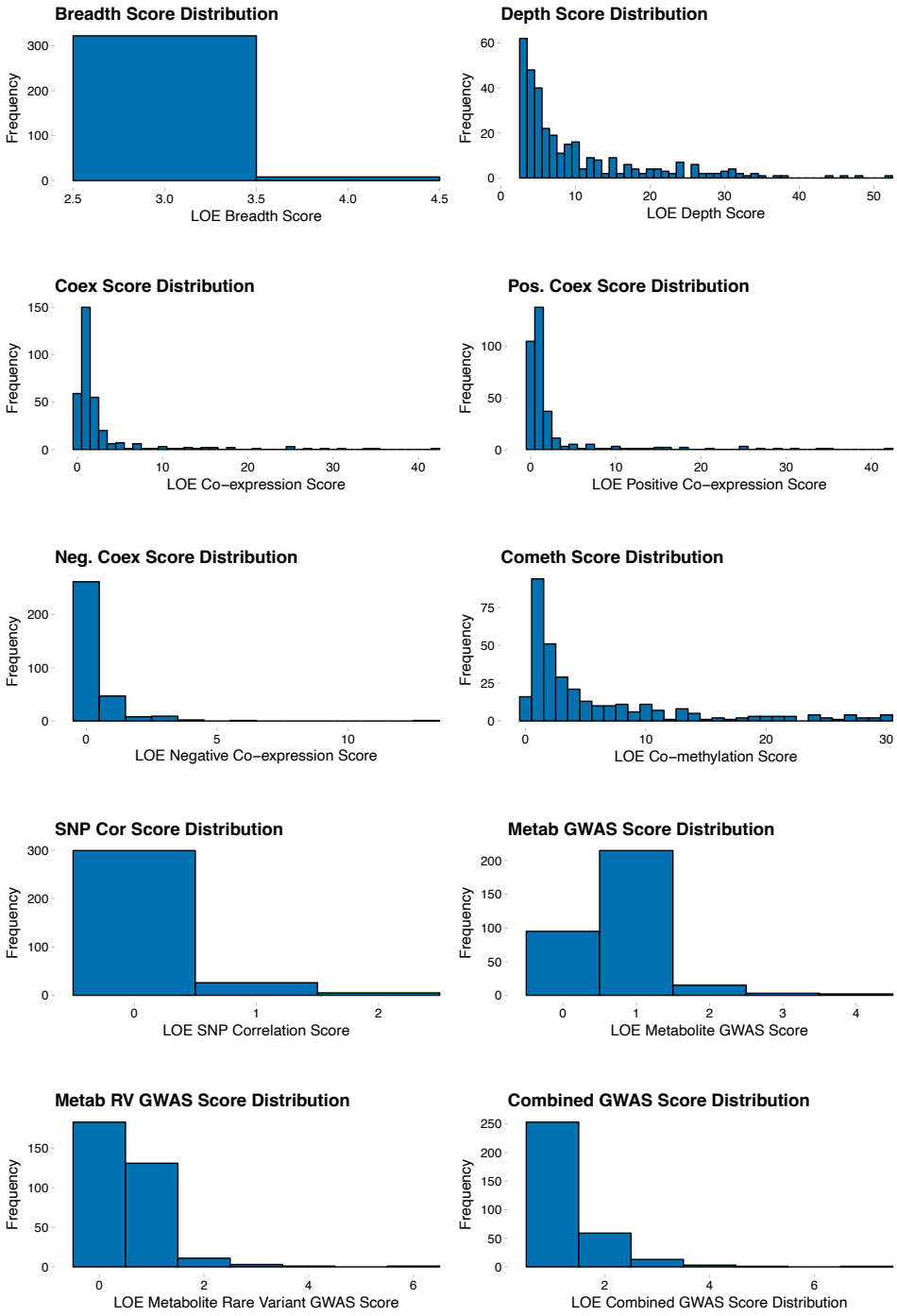
2.8.3 Supplemental Figures



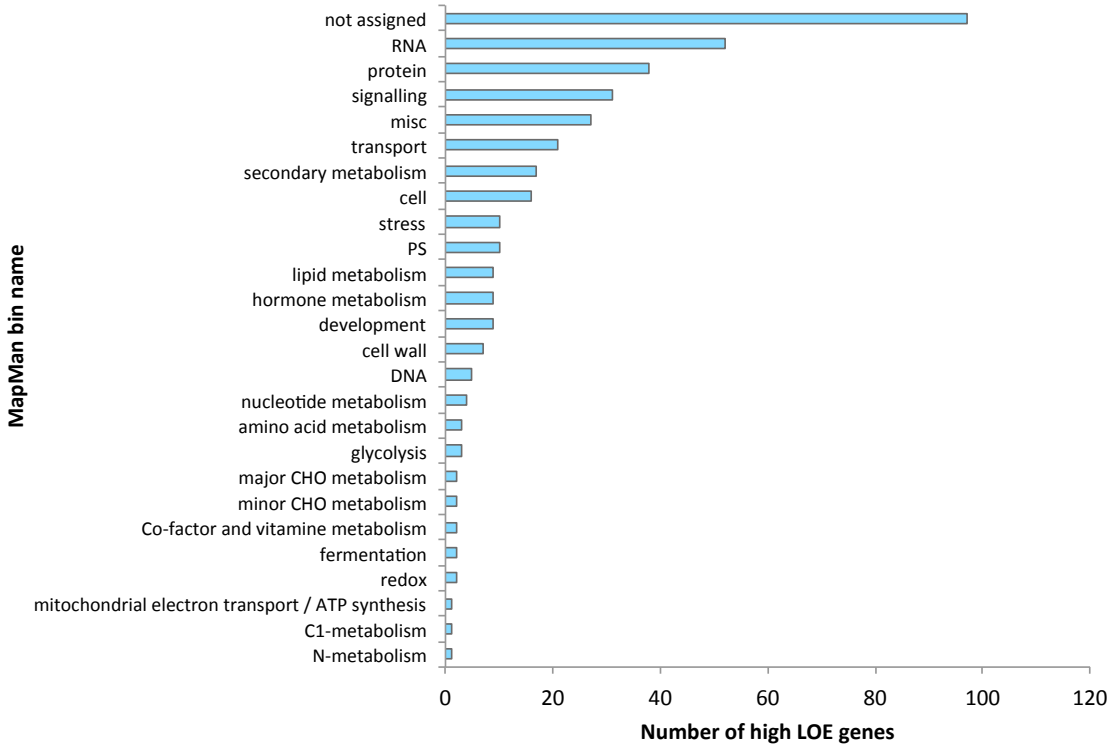
Supplemental Figure 2.1: (A) Distribution of observed gene coexpression correlation values (red) versus correlation values for randomized expression data (blue). We applied a threshold to exclude correlation values below 0.85 prior to LOE analysis (threshold indicated by vertical red line). (B) Distribution of observed gene comethylation correlation values (red) versus correlation values for randomized methylation data (blue). We applied a threshold to exclude correlation values below 0.95 prior to LOE analysis (threshold indicated by the vertical red line).



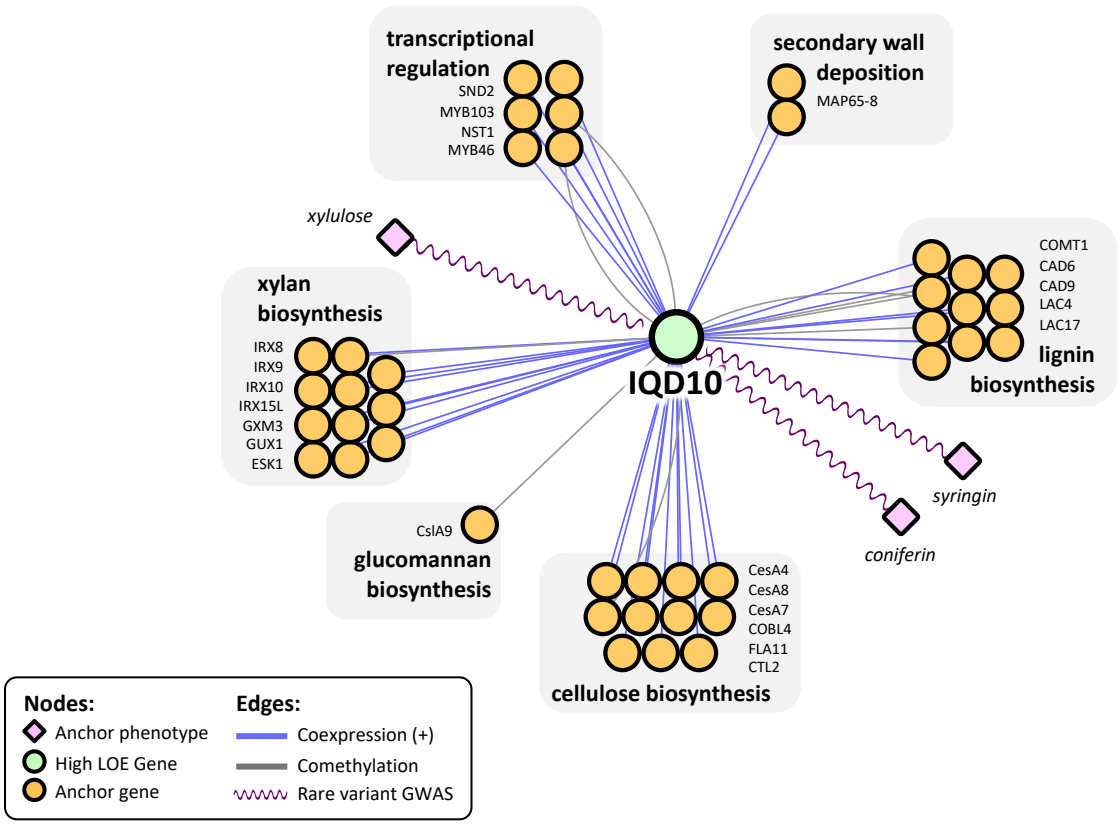
Supplemental Figure 2.2: Distributions of LOE scores. (A) Distribution of LOE breadth scores for all genes.



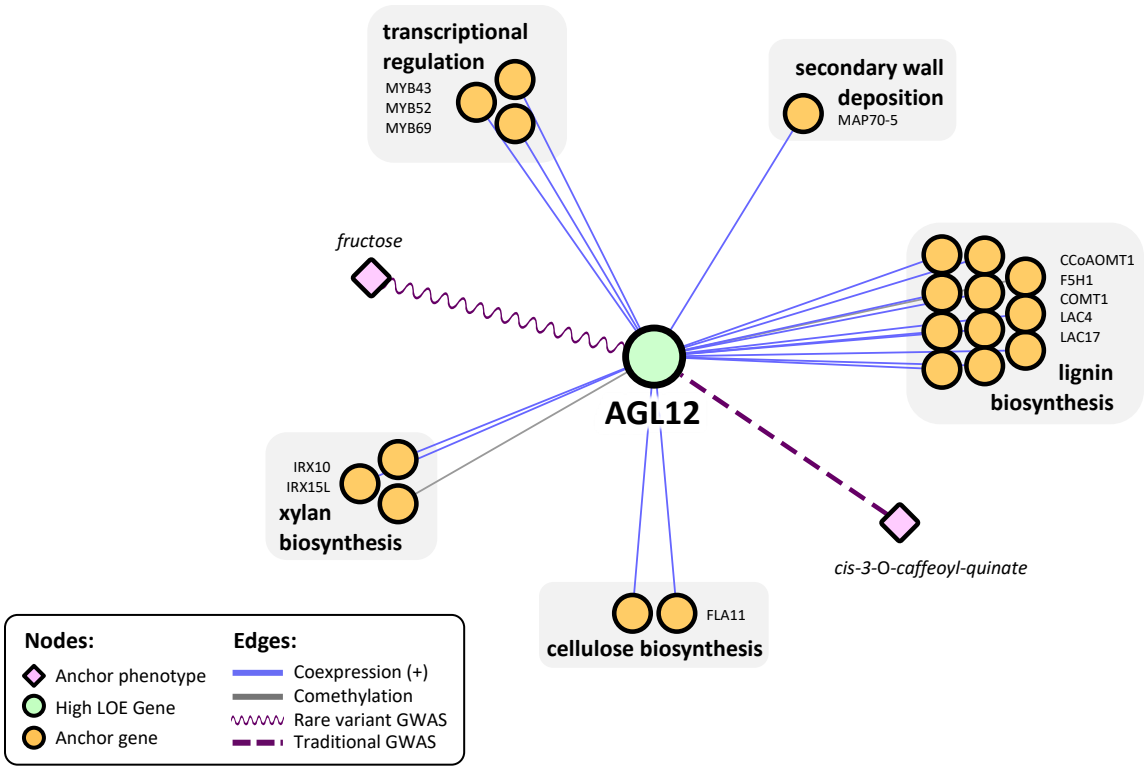
Supplemental Figure 2.2 (Continued): (B) Distributions of LOE scores for high LOE genes.



Supplemental Figure 2.3: MapMan annotations of high-LOE genes.

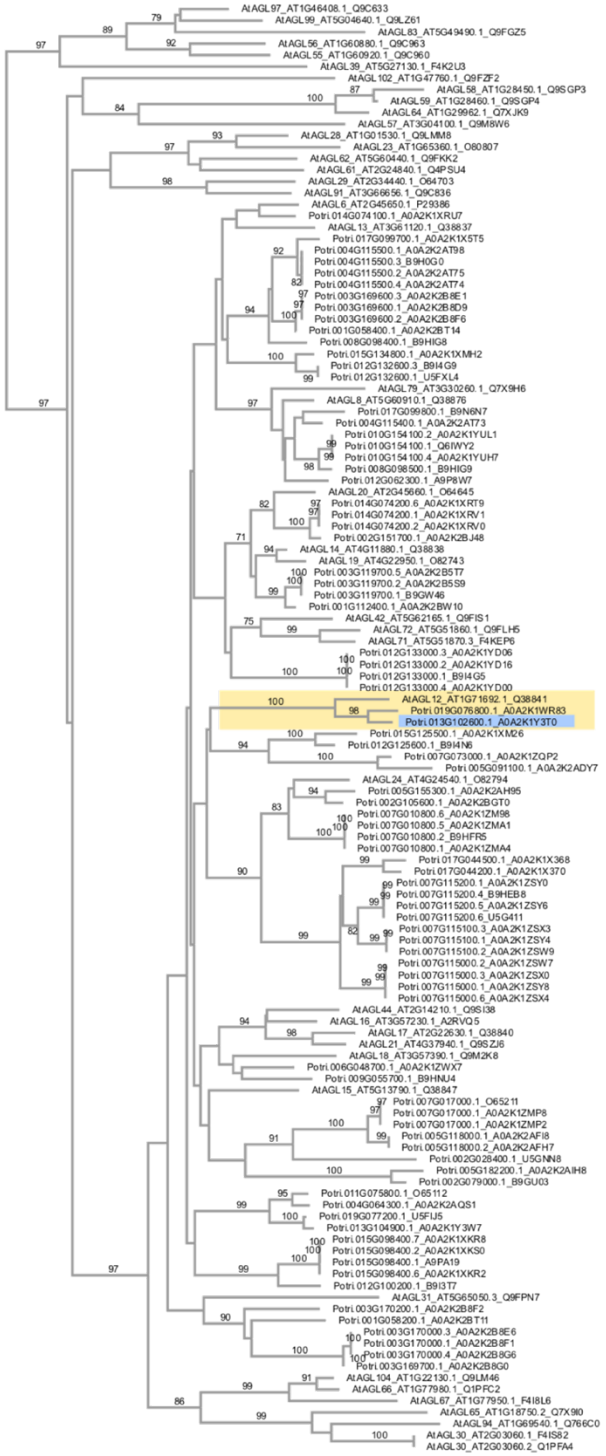


Supplemental Figure 2.4: LOE subnetwork depicting the one-hop neighbors of *PtIQD10* (Potri.011G096500). Gene symbols are Arabidopsis Best-hit matches. See Supplemental Table 2.7 for detailed node information.

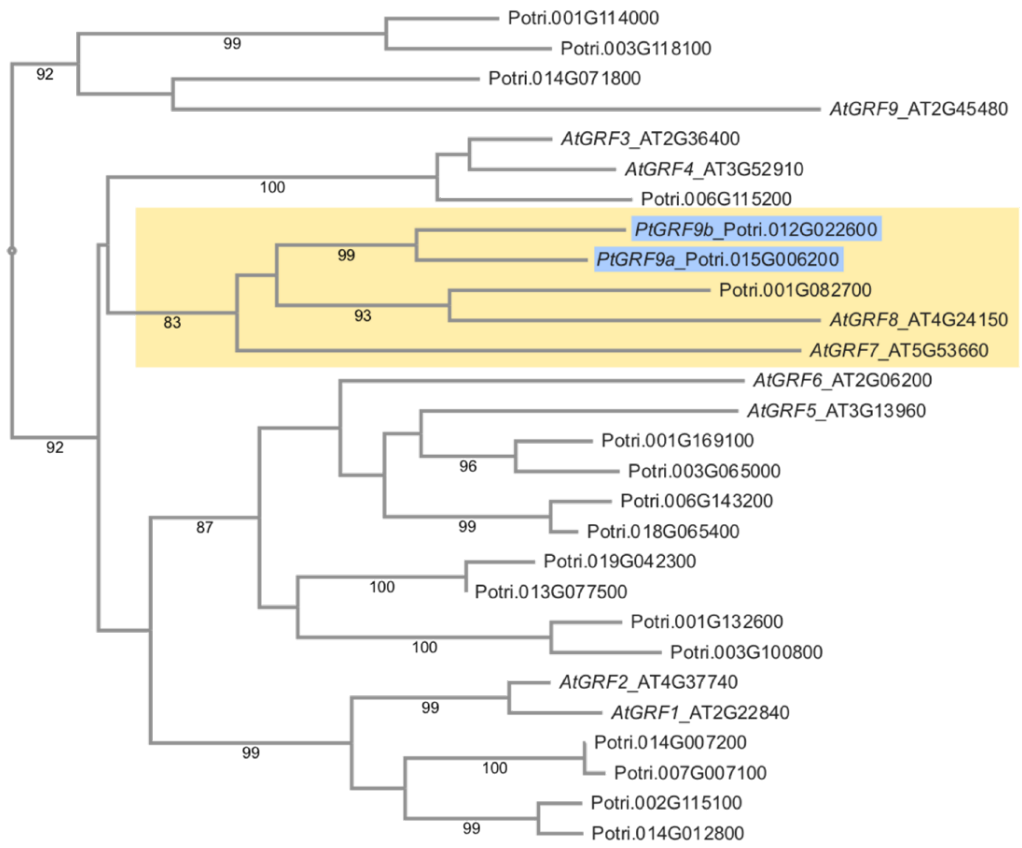


Supplemental Figure 2.5: LOE subnetwork depicting the one-hop neighbors of *P. trichocarpa PtAGL12* (Potri.013G102600). Gene symbols are Arabidopsis Best-hit matches. See Supplemental Table 2.8 for detailed node information.

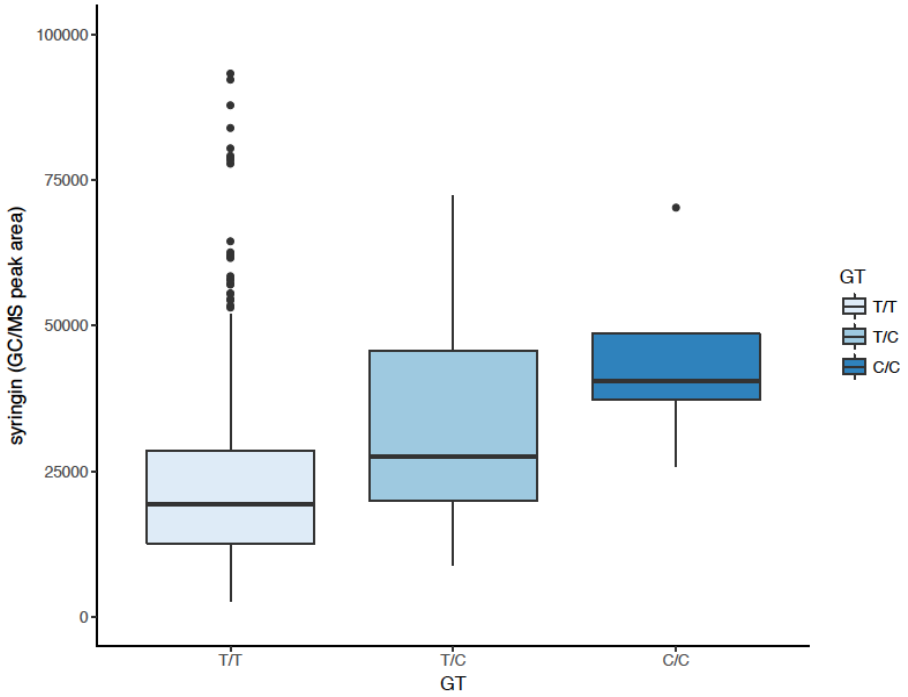
Supplemental Figure 2.6: Phylogram of *Populus trichocarpa* and *Arabidopsis thaliana* AGAMOUS-LIKE (AGL) amino acid sequences. *PtAGL12* (Potri.013G102600, indicated in blue) is a member of the clade containing *AtAGL12* (AT1G71692, in yellow). To assess support for orthology, we performed reciprocal BLASTp searches of amino acid sequences containing the SRF-type transcription factor domain (PF00319) and K-box region domain (PF01486) from *A. thaliana* and *P. trichocarpa* (obtained from UniProt; www.uniprot.org) and a phylogenetic analysis. Canonical amino acid sequences containing the SRF-type transcription factor (PF00319) domain and K-box region (PF01486) domain for Arabidopsis and P. trichocarpa were obtained from UniProt (www.uniprot.org). AGL amino acid sequences were aligned using MAFFT v7 (Kuraku et al., 2013; Katoh et al., 2017) using the E-INS-i method, BLOSUM62 scoring matrix, gap opening penalty of 1.53, offset value of 0.0, and default guide tree settings. A UPGMA tree was calculated to assess high-level relationships, in which sequences were grouped into two large clades. Sequences belonging to the same clade as *AtAGL12* were retained (which included all *P. trichocarpa* sequences) and a new alignment was created using the same parameters described above. A neighbor-joining tree was constructed from the second alignment using all "gap-free sites" (84 amino acids), a Poisson substitution model, and bootstrapping with 1000 replicates. A phylogram was created using Archaeopteryx.js (Zmasek and Eddy, 2001; Han and Zmasek, 2009). Bootstrap scores above 70 are indicated on nodes. Our results support an orthologous relationship between *PtAGL12* and *AtAGL12*, which is consistent with the phylogenetic analysis of (Leseberg et al., 2006).



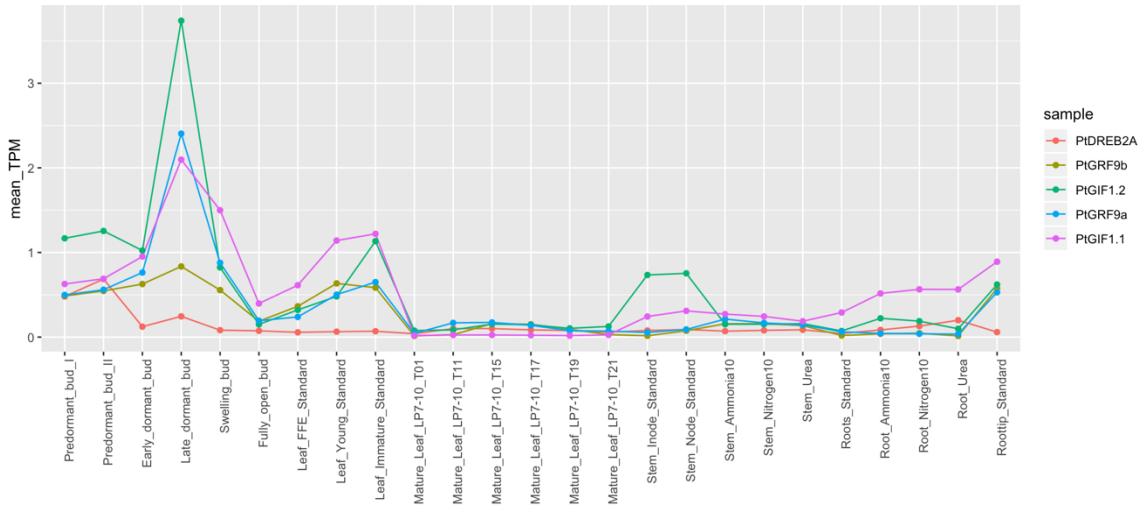
Supplemental Figure 2.6 (Continued)



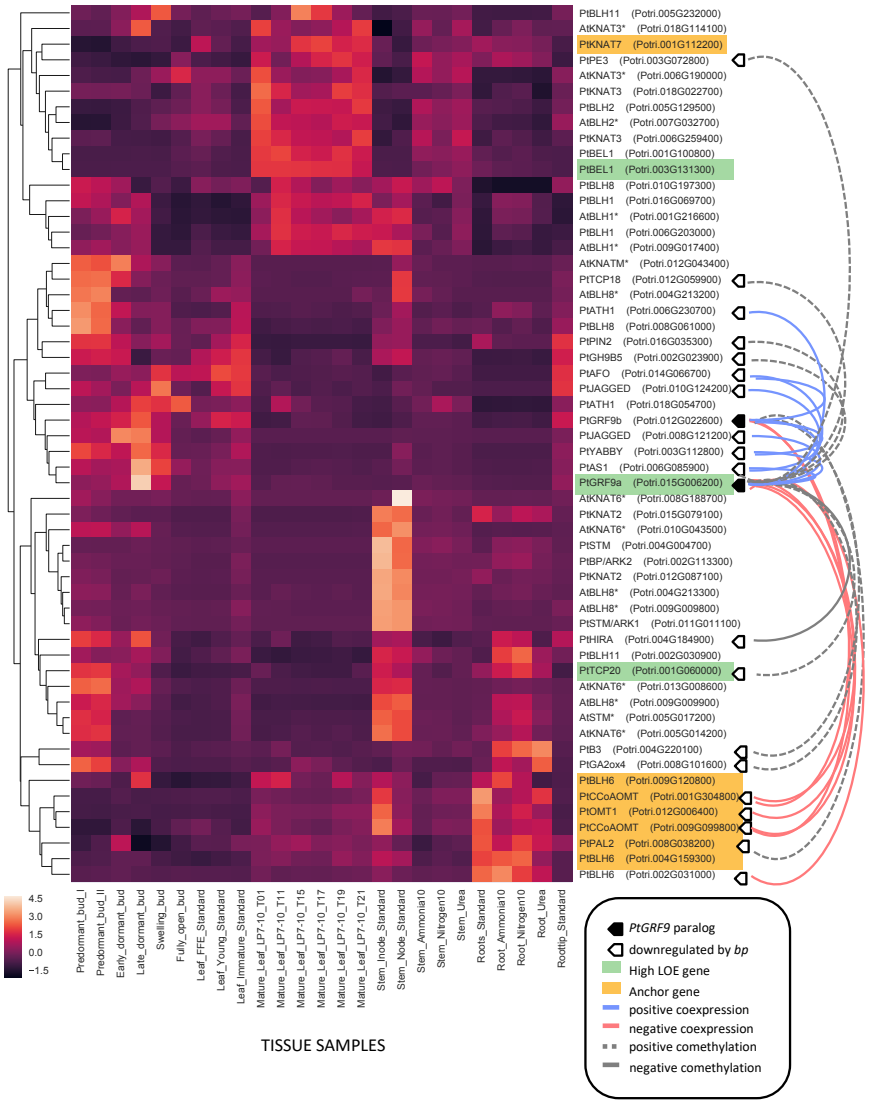
Supplemental Figure 2.7: Phylogram of *Arabidopsis thaliana* and *Populus trichocarpa* GROWTH-REGULATING FACTOR (*GRF*) amino acid sequences. *PtGRF9a* and *PtGRF9b* (indicated in blue) belong to the clade containing *AtGRF7* (in yellow). Canonical amino acid sequences containing the WRC (PF08879) and QLQ (PF08880) domains for Arabidopsis and *P. trichocarpa* were obtained from UniProt (www.uniprot.org). GRF proteins were aligned using MAFFT v7 (Kuraku et al., 2013; Katoh et al., 2017) using the E-INS-i method, BLOSUM62 scoring matrix, gap opening penalty of 1.53, offset value of 0.0, default guide tree settings, MAFFT-homologs option with 50 homologs and a threshold of E=1e-10. A neighbor-joining tree was constructed using all "gap-free sites" (92 amino acids), a Poisson substitution model, and bootstrapping with 1000 replicates. A phylogram was created using Archaeopteryx.js (Zmasek and Eddy, 2001; Han and Zmasek, 2009). Bootstrap scores above 70 are indicated on nodes.



Supplemental Figure 2.8: Allelic effect of the SNP located 3,788bp upstream of the *PtGRF9a* coding region that associates with syringin. SNP location: Chr15:400369; *PtGRF9a* location: Chr15:394635..396581 (reverse strand).



Supplemental Figure 2.9: Correlation between *P. trichocarpa PtGRF9a*, *PtGRFb*, *PtGIF1* and *PtDREB2A*. See Supplemental Table 2.1 for additional information about Gene Atlas expression data.



Supplemental Figure 2.10: Tissue-specific expression heatmap of *P. trichocarpa PtGRF9* paralogs (indicated by black pentagons), all *PtKNOX* genes, all *PtBELL* genes, and a subset of genes present in the *PtGRF9* 1-hop network: putative orthologs of *AtKNOX*-associated genes and cell wall biosynthesis genes down-regulated in *bp* mutants (Mele et al., 2003) (indicated by white pentagons; additional details are listed in Supplemental Table 2.13). Genes names with asterisks (*) do not have gene symbols or aliases in the *Populus trichocarpa* v3.0 genome annotation (Phytozome), therefore gene symbols from Arabidopsis Best-hit matches were used. See Supplemental Table 2.1 for additional information about Gene Atlas tissue samples.

References

Alejandro, S., Lee, Y., Tohge, T., Sudre, D., Osorio, S., Park, J., et al. (2012). AtABCG29 is a monolignol transporter involved in lignin biosynthesis. *Current Biology* 22, 1207–1212.

Allaby, R. (2019). Clonal crops show structural variation role in domestication. *Nat Plants* 5, 915–916. doi: 10.1038/s41477-019-0515-8.

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169.

Badmi, R., Payyavula, R. S., Bali, G., Guo, H.-B., Jawdy, S. S., Gunter, L., et al. (2018). A new calmodulin-binding protein expresses in the context of secondary cell wall biosynthesis and impacts biomass properties in Populus. *Front Plant Sci* 9, 1669. doi: 10.3389/FPLS.2018.01669.

Barros, J., Serk, H., Granlund, I., and Pesquet, E. (2015). The cell biology of lignification in higher plants. *Ann Bot* 115, 1053–1074. doi: 10.1093/aob/mcv046.

Bastiaanse, H., Zinkgraf, M., Canning, C., Tsai, H., Lieberman, M., Comai, L., et al. (2019). A comprehensive genomic scan reveals gene dosage balance impacts on quantitative traits in Populus trees. *Proc Natl Acad Sci U S A* 116, 13690–13699. doi: 10.1073/pnas.1903229116.

Bellaoui, M., Pidkowich, M. S., Samach, A., Kushalappa, K., Kohalmi, S. E., Modrusan, Z., et al. (2001). The Arabidopsis BELL1 and KNOX TALE homeodomain proteins interact through a domain conserved between plants and animals. *Plant Cell* 13, 2455–70. doi: 10.1105/tpc.010161.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57, 289–300. doi: 10.2307/2346101.

Bischoff, V., Nita, S., Neumetzler, L., Schindelasch, D., Urbain, A., Eshed, R., et al. (2010). TRICHOME BIREFRINGENCE and its homolog AT5G01360 encode plant-specific DUF231 proteins required for cellulose biosynthesis in Arabidopsis. *Plant Physiol* 153, 590–602.

Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An expanded view of complex traits:

from polygenic to omnigenic. *Cell* 169, 1177–1186.

Cai, B., Li, C.-H., and Huang, J. (2014). Systematic identification of cell-wall related genes in Populus based on analysis of functional modules in co-expression network. *PLoS One* 9, e95176.

Cameron, D. L., Baber, J., Shale, C., Valle-Inclan, J. E., Besselink, N., van Hoeck, A., et al. (2021). GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing. *Genome Biol* 22, 1–25. doi: 10.1186/s13059-021-02423-x.

Cao, Y., Han, Y., Jin, Q., Lin, Y., and Cai, Y. (2016). Comparative Genomic Analysis of the GRF Genes in Chinese Pear (Pyrus bretschneideri Rehd), Poplar (Populous), Grape (Vitis vinifera), Arabidopsis and Rice (Oryza sativa). *Front Plant Sci* 7, 1–14. doi: 10.3389/fpls.2016.01750.

Cassan-Wang, H., Goué, N., Saidi, M. N., Legay, S., Sivadon, P., Goffner, D., et al. (2013). Identification of novel transcription factors regulating secondary cell wall formation in Arabidopsis. *Front Plant Sci* 4, 1–14. doi: 10.3389/fpls.2013.00189.

Catacchio, C. R., Alagna, F., Perniola, R., Bergamini, C., Rotunno, S., Calabrese, F. M., et al. (2019). Transcriptomic and genomic structural variation analyses on grape cultivars reveal new insights into the genotype-dependent responses to water stress. *Sci Rep* 9. doi: 10.1038/s41598-019-39010-x.

Chhetri, H. B., Furches, A., Macaya-Sanz, D., Walker, A. R., Kainer, D., Jones, P., et al. (2020). Genome-Wide Association Study of Wood Anatomical and Morphological Traits in Populus trichocarpa. *Front Plant Sci* 11, 545748. doi: 10.3389/fpls.2020.545748.

Cosio, C., Ranocha, P., Francoz, E., Burlat, V., Zheng, Y., Perry, S. E., et al. (2017). The class III peroxidase PRX17 is a direct target of the MADS-box transcription factor AGAMOUS-LIKE15 (AGL15) and participates in lignified tissue formation. *New Phytologist* 213, 250–263. doi: 10.1111/nph.14127.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011a). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011b). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21.

Dolatabadian, A., Bayer, P. E., Tirnaz, S., Hurgobin, B., Edwards, D., and Batley, J. (2020). Characterization of disease resistance genes in the Brassica napus pangenome reveals significant structural variation. *Plant Biotechnol J* 18, 969–982. doi: 10.1111/pbi.13262.

Du, J., Mansfield, S. D., and Groover, A. T. (2009). The Populus homeobox gene ARBORKNOX2 regulates cell differentiation during secondary growth. *Plant Journal* 60, 1000–1014. doi: 10.1111/j.1365-313X.2009.04017.x.

Du, J., Miura, E., Robischon, M., Martinez, C., and Groover, A. (2011). The Populus Class III HD ZIP transcription factor POPCORONA affects cell differentiation during secondary growth of woody stems. *PLoS One* 6, e17458.

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics* 14, 755–763.

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., et al. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* 11, 446–450. doi: 10.1038/nrg2809.

Eriksson, M. E., Israelsson, M., Olsson, O., and Moritz, T. (2000). Increased gebberellin biosynthesis in transgenic trees promotes growth, biomass production and xylem fiber length. *Nat Biotechnol* 18, 784–788.

Fernandez, D. E., Wang, C.-T., Zheng, Y., Adamczyk, B. J., Singhal, R., Hall, P. K., et al. (2014). The MADS-Domain Factors AGAMOUS-LIKE15 and AGAMOUS-LIKE18, along with SHORT VEGETATIVE PHASE and AGAMOUS-LIKE24, Are Necessary to Block Floral Gene Expression during the Vegetative Phase. *Plant Physiol* 165, 1591–1603. doi: 10.1104/pp.114.242990.

Fisher, R. A. (1919). XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh* 52, 399–433. doi: 10.1017/S0080456800012163.

Franco-Zorrilla, J. M., López-Vidriero, I., Carrasco, J. L., Godoy, M., Vera, P., and

Solano, R. (2014). DNA-binding specificities of plant transcription factors and their potential to define target genes. *Proceedings of the National Academy of Sciences* 111, 2367–2372. doi: 10.1073/pnas.1316278111.

Fuentes, R. R., Chebotarov, D., Duitama, J., Smith, S., de la Hoz, J. F., Mohiyuddin, M., et al. (2019). Structural variants in 3000 rice genomes. *Genome Res* 29, 870–880. doi: 10.1101/gr.241240.118.

Furches, A., Kainer, D., Weighill, D., Large, A., Jones, P., Walker, A. M., et al. (2019). Finding New Cell Wall Regulatory Genes in Populus trichocarpa Using Multiple Lines of Evidence. *Front Plant Sci* 10. doi: 10.3389/fpls.2019.01249.

Furumizu, C., Alvarez, J. P., Sakakibara, K., and Bowman, J. L. (2015). Antagonistic Roles for KNOX1 and KNOX2 Genes in Patterning the Land Plant Body Plan Following an Ancient Gene Duplication. *PLoS Genet* 11, 1–24. doi: 10.1371/journal.pgen.1004980.

Gabur, I., Chawla, H. S., Snowdon, R. J., and Parkin, I. A. P. (2019). Connecting genome structural variation with complex traits in crop plants. *Theoretical and Applied Genetics* 132, 733–750. doi: 10.1007/s00122-018-3233-0.

Gavrilov, Y., Benjamini, Y., and Sarkar, S. K. (2009). An adaptive step-down procedure with proven FDR control under independence. *The Annals of Statistics* 37, 619–629. doi: 10.1214/07-AOS586.

Geraldes, A., Difazio, S. P., Slavov, G. T., Ranjan, P., Muchero, W., Hannemann, J., et al. (2013). A 34K SNP genotyping array for Populus trichocarpa: Design, application to the study of natural populations and transferability to other Populus species. *Mol Ecol Resour* 13, 306–323. doi: 10.1111/1755-0998.12056.

Geraldes, A., Farzaneh, N., Grassa, C. J., Mckown, A. D., Guy, R. D., Mansfield, S. D., et al. (2014). Landscape genomics of populus trichocarpa: The role of hybridization, limited gene flow, and natural selection in shaping patterns of population structure. *Evolution (N Y)* 68, 3260–3280. doi: 10.1111/evo.12497.

Giménez, E., Pineda, B., Capel, J., Antón, M. T., Atarés, A., Pérez-Martín, F., et al. (2010). Functional analysis of the Arlequin mutant corroborates the essential role of the ARLEQUIN/TAGL1 gene during reproductive development of tomato. *PLoS*

One 5. doi: 10.1371/journal.pone.0014427.

Gonzalez, N., Vanhaeren, H., and Inzé, D. (2012). Leaf size control: complex coordination of cell division and expansion. *Trends Plant Sci* 17, 332–340. doi: 10.1016/j.tplants.2012.02.003.

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., et al. (2011). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* 40, D1178--D1186. doi: 10.1093/NAR/GKR944.

Groover, A. T., Mansfield, S. D., DiFazio, S. P., Dupper, G., Fontana, J. R., Millar, R., et al. (2006). The Populus homeobox gene ARBORKNOX1 reveals overlapping mechanisms regulating the shoot apical meristem and the vascular cambium. *Plant Mol Biol* 61, 917–932. doi: 10.1007/s11103-006-0059-y.

Guo, J., Cao, K., Deng, C., Li, Y., Zhu, G., Fang, W., et al. (2020). An integrated peach genome structural variation map uncovers genes associated with fruit traits. *Genome Biol* 21, 258. doi: 10.1186/s13059-020-02169-y.

Guo, M., Thomas, J., Collins, G., and Timmermans, M. C. P. (2008). Direct Repression of KNOX Loci by the ASYMMETRIC LEAVES1 Complex of Arabidopsis. *the Plant Cell Online* 20, 48–58. doi: 10.1105/tpc.107.056127.

Haft, D. H., Loftus, B. J., Richardson, D. L., Yang, F., Eisen, J. A., Paulsen, I. T., et al. (2001). TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res* 29, 41–43.

Han, M. v, and Zmasek, C. M. (2009). phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics* 10, 356. doi: 10.1186/1471-2105-10-356.

Hao, Z., and Mohnen, D. (2014). A review of xylan and lignin biosynthesis: Foundation for studying Arabidopsis irregular xylem mutants with pleiotropic phenotypes. *Crit Rev Biochem Mol Biol* 49, 212–241. doi: 10.3109/10409238.2014.889651.

Hay, A., and Tsiantis, M. (2010). KNOX genes: versatile regulators of plant development and diversity. *Development* 137, 3153–3165. doi: 10.1242/dev.030049.

Hertzberg, M., Aspeborg, H., Schrader, J., Andersson, A., Erlandsson, R., Blomqvist,

K., et al. (2001). A transcriptional roadmap to wood formation. *Proceedings of the National Academy of Sciences* 98, 14732–14737. doi: 10.1073/pnas.261293398.

Hoe Kim, J., Tsukaya, H., Kim, J. H., and Tsukaya, H. (2015). Regulation of plant growth and development by the GROWTH-REGULATING FACTOR and GRF-INTERACTING FACTOR duo. *J Exp Bot* 66, 6093–6107. doi: 10.1093/JXB/ERV349.

Hussey, S. G., Mizrachi, E., Spokevicius, A. v, Bossinger, G., Berger, D. K., and Myburg, A. A. (2011). SND2, a NAC transcription factor gene, regulates genes involved in secondary cell wall development in Arabidopsis fibres and increases fibre cell area in Eucalyptus. *BMC Plant Biol* 11, 173.

Jasinski, S., Piazza, P., Craft, J., Hay, A., Woolley, L., Rieu, I., et al. (2005). KNOX action in Arabidopsis is mediated by coordinate regulation of cytokinin and gibberellin activities. *Current Biology* 15, 1560–1565. doi: 10.1016/j.cub.2005.07.023.

Jiang, H., Lei, R., Ding, S.-W., and Zhu, S. (2014). Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics* 15, 182.

Jin, J., Tian, F., Yang, D.-C., Meng, Y.-Q., Kong, L., Luo, J., et al. (2017). PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res* 45, D1040–D1045. doi: 10.1093/nar/gkw982.

Jokipii-Lukkari, S., Delhomme, N., Schiffthaler, B., Mannapperuma, C., Prestele, J., Nilsson, O., et al. (2018). Transcriptional Roadmap to Seasonal Variation in Wood Formation of Norway Spruce. *Plant Physiol* 176, 2851–2870. doi: 10.1104/pp.17.01590.

Joshi, C. P., Alqudah, A. M., Pereira, A., Singla-Pareek, S. L., Joshi, R., Wani, S. H., et al. (2016). Transcription factors and plants response to drought stress: current understanding and future directions. *Front Plant Sci* 7, 1029. doi: 10.3389/fpls.2016.01029.

Joubert, W., Nance, J., Climer, S., Weighill, D., and Jacobson, D. (2019). Parallel accelerated Custom Correlation Coefficient calculations for genomics applications.

Parallel Comput 84, 15–23. doi: 10.1016/J.PARCO.2019.02.003.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S., Freimer, N. B., et al. (2010a). Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* 42, 348–354. doi: 10.1038/ng.548.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., et al. (2010b). Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* 42, 348–354. doi: 10.1038/ng.548.

Katoh, K., Rozewicki, J., and Yamada, K. D. (2017). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform*. doi: 10.1093/bib/bbx108.

Kim, J.-S., Mizoi, J., Kidokoro, S., Maruyama, K., Nakajima, J., Nakashima, K., et al. (2012). Arabidopsis GROWTH-REGULATING FACTOR7 Functions as a Transcriptional Repressor of Abscisic Acid- and Osmotic Stress-Responsive Genes, Including DREB2A. *Plant Cell* 24, 3393–3405. doi: 10.1105/tpc.112.100933.

Ko, J.-H., Jeon, H.-W., Kim, W.-C., Kim, J.-Y., and Han, K.-H. (2014). The MYB46/MYB83-mediated transcriptional regulatory programme is a gatekeeper of secondary wall biosynthesis. *Ann Bot* 114, 1099–1107.

Kuijt, S. J. H., Greco, R., Agalou, A., Shao, J., ’t Hoen, C. C. J., Overnas, E., et al. (2014). Interaction between the GROWTH-REGULATING FACTOR and KNOTTED1-LIKE HOMEOBOX Families of Transcription Factors. *Plant Physiol* 164, 1952–1966. doi: 10.1104/pp.113.222836.

Kuraku, S., Zmasek, C. M., Nishimura, O., and Katoh, K. (2013). aLeaves facilitates on-demand exploration of metazoan gene family trees on MAFFT sequence alignment server with enhanced interactivity. *Nucleic Acids Res* 41, W22–W28. doi: 10.1093/nar/gkt389.

Lee, H., Suh, S. S., Park, E., Cho, E., Ahn, J. H., Kim, S. G., et al. (2000). The AGAMOUS-lIKE 20 MADS domain protein integrates floral inductive pathways in Arabidopsis. *Genes Dev* 14, 2366–2376. doi: 10.1101/gad.813600.

Leseberg, C. H., Li, A., Kang, H., Duvall, M., and Mao, L. (2006). Genome-wide analysis of the MADS-box gene family in Populus trichocarpa. *Gene* 378, 84–94.

doi: 10.1016/j.gene.2006.05.022.

Li, E., Bhargava, A., Qiang, W., Friedmann, M. C., Forneris, N., Savidge, R. A., et al. (2012a). The Class II KNOX gene KNAT7 negatively regulates secondary wall formation in Arabidopsis and is functionally conserved in Populus. *New Phytologist* 194, 102–115. doi: 10.1111/j.1469-8137.2011.04016.x.

Li, H., Yang, Y., Wang, Z., Guo, X., Liu, F., Jiang, J., et al. (2016). BpMADS12 gene role in lignin biosynthesis of Betula platyphylla Suk by transcriptome analysis. *J For Res (Harbin)* 27, 1111–1120. doi: 10.1007/s11676-016-0229-y.

Li, Y., Jin, F., Chao, Q., and Wang, B.-C. (2017). Proteomics analysis reveals the molecular mechanism underlying the transition from primary to secondary growth of poplar. *J Plant Physiol* 213, 1–15.

Li, Y., Tschaplinski, T. J., Engle, N. L., Hamilton, C. Y., Rodriguez, M., Liao, J. C., et al. (2012b). Combined inactivation of the Clostridium cellulolyticum lactate and malate dehydrogenase genes substantially increases ethanol yield from cellulose and switchgrass fermentations. *Biotechnol Biofuels* 5, 2.

Lin, Y.-C. J., Chen, H., Li, Q., Li, W., Wang, J. P., Shi, R., et al. (2017). Reciprocal cross-regulation of VND and SND multigene TF families for wood formation in Populus trichocarpa. *Proceedings of the National Academy of Sciences*, 201714422.

Liu, Y., Wei, M., Hou, C., Lu, T., Liu, L., Wei, H., et al. (2017). Functional characterization of Populus PsnSHN2 in coordinated regulation of secondary wall components in tobacco. *Sci Rep* 7, 1–11. doi: 10.1038/s41598-017-00093-z.

Lofstead, J. F., Klasky, S., Schwan, K., Podhorszki, N., and Jin, C. (2008). Flexible IO and integration for scientific codes through the adaptable IO system (ADIOS). in *Proceedings of the 6th international workshop on Challenges of large applications in distributed environments - CLADE '08* (New York, New York, USA: ACM Press), 15. doi: 10.1145/1383529.1383533.

Lohse, M., Nagel, A., Herter, T., May, P., Schroda, M., Zrenner, R., et al. (2014). Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. *Plant Cell Environ.* 37, 1250–1258. doi: 10.1111/pce.12231.

Lu, S., Li, Q., Wei, H., Chang, M.-J., Tunlaya-Anukit, S., Kim, H., et al. (2013). Ptr-

miR397a is a negative regulator of laccase genes affecting lignin content in Populus trichocarpa. *Proceedings of the National Academy of Sciences* 110, 10848–10853.

Macaya-Sanz, D., Chen, J. G., Kalluri, U. C., Muchero, W., Tschaplinski, T. J., Gunter, L. E., et al. (2017). Agronomic performance of Populus deltoides trees engineered for biofuel production. *Biotechnol Biofuels* 10, 253. doi: 10.1186/s13068-017-0934-6.

Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: A Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics* 21, 3448–3449. doi: 10.1093/bioinformatics/bti551.

Masiero, S., Colombo, L., Grini, P. E., Schnittger, A., and Kater, M. M. (2011). The Emerging Importance of Type I MADS Box Transcription Factors for Plant Reproduction. *Plant Cell* 23, 865–872. doi: 10.1105/tpc.110.081737.

Mckown, A. D., Guy, R. D., Klápště, J., Geraldes, A., Friedmann, M., Cronk, Q. C. B., et al. (2014). Geographical and environmental gradients shape phenotypic trait variation and genetic structure in Populus trichocarpa. *New Phytologist* 201, 1263–1276. doi: 10.1111/nph.12601.

McLeay, R. C., and Bailey, T. L. (2010). Motif Enrichment Analysis: A unified framework and an evaluation on ChIP data. *BMC Bioinformatics* 11. doi: 10.1186/1471-2105-11-165.

Mele, G., Ori, N., Sato, Y., and Hake, S. (2003). The knotted1-like homeobox gene BREVIPEDICELLUS regulates cell differentiation by modulating metabolic pathways. *Genes Dev* 17, 2088–2093. doi: 10.1101/gad.1120003.2088.

Mizrachi, E., Verbeke, L., Christie, N., Fierro, A. C., Mansfield, S. D., Davis, M. F., et al. (2017). Network-based integration of systems genetics data reveals pathways associated with lignocellulosic biomass accumulation and processing. *Proceedings of the National Academy of Sciences* 114, 1195–1200. doi: 10.1073/pnas.1620119114.

Muchero, W., Guo, J., DiFazio, S. P., Chen, J. G., Ranjan, P., Slavov, G. T., et al. (2015). High-resolution genetic mapping of allelic variants associated with cell wall chemistry in Populus. *BMC Genomics* 16, 24. doi: 10.1186/s12864-015-1215-z.

Mukherjee, K., Brocchieri, L., Bürglin, T. R., and Burglin, T. R. (2009). A Comprehensive Classification and Evolutionary Analysis of Plant Homeobox Genes. *Mol Biol Evol* 26, 2775–2794. doi: 10.1093/molbev/msp201.

Nakano, Y., Yamaguchi, M., Endo, H., Rejab, N. A., and Ohtani, M. (2015). NAC-MYB-based transcriptional regulation of secondary cell wall biosynthesis in land plants. *Front Plant Sci* 6, 1–18. doi: 10.3389/fpls.2015.00288.

Ng, M., and Yanofsky, M. F. (2001). Function and evolution of the plant MADS-box gene family. *Nat Rev Genet* 2, 186–195. doi: 10.1038/35056041.

Ohtani, M., Nishikubo, N., Xu, B., Yamaguchi, M., Mitsuda, N., Goué, N., et al. (2011). A NAC domain protein family contributing to the regulation of wood formation in poplar. *The Plant Journal* 67, 499–512.

Omidbakhshfard, M. A., Proost, S., Fujikura, U., and Mueller-Roeber, B. (2015). Growth-Regulating Factors (GRFs): A Small Transcription Factor Family with Important Functions in Plant Biology. *Mol Plant* 8, 998–1010. doi: 10.1016/j.molp.2015.01.013.

Park, M. Y., Kang, J.-Y., and Kim, S. Y. (2011). Overexpression of AtMYB52 confers ABA hypersensitivity and drought tolerance. *Mol Cells* 31, 447–454. doi: 10.1007/s10059-011-0300-7.

Peterson, C. B., Bogomolov, M., Benjamini, Y., and Sabatti, C. (2016). Many Phenotypes Without Many False Discoveries: Error Controlling Strategies for Multitrait Association Studies. *Genet Epidemiol* 40, 45–56. doi: 10.1002/gepi.21942.

Pomiès, L., Decourteix, M., Franchel, J., Moulia, B., and Leblanc-Fournier, N. (2017). Poplar stem transcriptome is massively remodelled in response to single or repeated mechanical stimuli. *BMC Genomics* 18, 300. doi: 10.1186/s12864-017-3670-1.

Porth, I., Klápště, J., McKown, A. D., la Mantia, J., Guy, R. D., Ingvarsson, P. K., et al. (2015). Evolutionary quantitative genomics of Populus trichocarpa. *PLoS One* 10, e0142864. doi: 10.1371/journal.pone.0142864.

Porth, I., Klápště, J., Skyba, O., Friedmann, M. C., Hannemann, J., Ehlting, J., et al. (2013). Network analysis reveals the relationship among wood properties, gene

expression levels and genotypes of natural Populus trichocarpa accessions. *New Phytologist* 200, 727–742. doi: 10.1111/NPH.12419.

Prunier, J., Giguère, I., Ryan, N., Guy, R., Soolanayakanahally, R., Isabel, N., et al. (2019). Gene copy number variations involved in balsam poplar (Populus balsamifera L.) adaptive variations. *Mol Ecol* 28, 1476–1490. doi: 10.1111/mec.14836.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., et al. (2011). The Pfam protein families database. *Nucleic Acids Res* 40, D290--D301.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007a). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* 81, 559–575. doi: 10.1086/519795.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R. R., Bender, D., et al. (2007b). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* 81, 559–575. doi: 10.1086/519795.

Puzey, J. R., Karger, A., Axtell, M., and Kramer, E. M. (2012). Deep annotation of populus trichocarpa micro rnas from diverse tissue sets. *PLoS One* 7. doi: 10.1371/journal.pone.0033034.

Rao, X., and Dixon, R. A. (2018). Current Models for Transcriptional Regulation of Secondary Cell Wall Biosynthesis in Grasses. *Front Plant Sci* 9, 399. doi: 10.3389/fpls.2018.00399.

Saint-Vincent, P. M. B., Furches, A., Galanie, S., Teixeira Prates, E., Aldridge, J. L., Labbe, A., Zhao, N., Martin, M. Z., Ranjan, P., Jones, P., Kainer, D., Kalluri, U. C., Chen, J. G., Muchero, W., Jacobson, D. A., & Tschaplinski, T. J. (2023). Validation of a metabolite–GWAS network for *Populus trichocarpa* family 1 UDP-glycosyltransferases. *Frontiers in Plant Science*, *14*(July), 1–14. https://doi.org/10.3389/fpls.2023.1210146

Sang, Y., Long, Z., Dan, X., Feng, J., Shi, T., Jia, C., et al. (2022). Genomic insights into present local adaptation and future climate change vulnerability of a keystone forest

tree species in East Asian. *bioRxiv*, 1–40. doi: 10.1101/2022.04.04.486908.

Sannigrahi, P., Ragauskas, A. J., and Tuskan, G. A. (2010). Poplar as a feedstock for biofuels: A review of compositional characteristics. *Biofuels, Bioproducts and Biorefining* 4, 209–226. doi: 10.1002/bbb.206.

Schiessl, S. V., Katche, E., Ihien, E., Chawla, H. S., and Mason, A. S. (2019). The role of genomic structural variation in the genetic improvement of polyploid crops. *Crop Journal* 7, 127–140. doi: 10.1016/j.cj.2018.07.006.

Schrader, J. (2004). A High-Resolution Transcript Profile across the Wood-Forming Meristem of Poplar Identifies Potential Regulators of Cambial Stem Cell Identity. *the Plant Cell Online* 16, 2278–2292. doi: 10.1105/tpc.104.024190.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13, 2498–2504.

Shi, R., Wang, J. P., Lin, Y.-C., Li, Q., Sun, Y.-H., Chen, H., et al. (2017). Tissue and cell-type co-expression networks of transcription factors and wood component genes in Populus trichocarpa. *Planta* 245, 927–938.

Taylor-Teeples, M., Lin, L., de Lucas, M., Turco, G., Toal, T. W., Gaudinier, A., et al. (2015). An Arabidopsis gene regulatory network for secondary cell wall synthesis. *Nature* 517, 571.

Todesco, M., Owens, G. L., Bercovich, N., Légaré, J. S., Soudi, S., Burge, D. O., et al. (2020). Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature* 584, 602–607. doi: 10.1038/s41586-020-2467-6.

Tschaplinski, T. J., Standaert, R. F., Engle, N. L., Martin, M. Z., Sangha, A. K., Parks, J. M., et al. (2012). Down-regulation of the caffeic acid O-methyltransferase gene in switchgrass reveals a novel monolignol analog. *Biotechnol Biofuels* 5, 71.

Tschaplinski, T. J., Jacobson, D. A., Kainer, D., Weighill, D. A., & Furches, A. (2023). Method for Controlling Cell Wall Biosynthesis and Genetically Modified Plants (Patent No. US11725215B2).

Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., et al. (2006). The genome of black cottonwood, Populus trichocarpa (Torr. & Gray).

Science (1979) 313, 1596–1604. doi: 10.1126/science.1128691.

van Dongen, S. (2008). Graph clustering via a discrete uncoupling process. *SIAM Journal on Matrix Analysis and Applications* 30, 121–141.

Vining, K. J., Pomraning, K. R., Wilhelm, L. J., Priest, H. D., Pellegrini, M., Mockler, T. C., et al. (2012). Dynamic DNA cytosine methylation in the Populus trichocarpa genome: tissue-level variation and relationship to gene expression. *BMC Genomics* 13, 27.

Wang, S., Li, E., Porth, I., Chen, J.-G., Mansfield, S. D., and Douglas, C. J. (2014). Regulation of secondary cell wall biosynthesis by poplar R2R3 MYB transcription factor PtrMYB152 in Arabidopsis. *Sci Rep* 4, 5054.

Weighill, D. A., Jones, P., Shah, M., Ranjan, P., Muchero, W., Schmutz, J., et al. (2018). Pleiotropic and Epistatic Network-Based Discovery: Integrated Networks for Target Gene Discovery. *Front Energy Res* 6, 267997. doi: 10.3389/fenrg.2018.00030.

Xie, M., Muchero, W., Bryan, A. C., Yee, K. L., Guo, H., Zhang, J., et al. (2018a). A 5-enolpyruvylshikimate 3-phosphate synthase functions as a transcriptional repressor in Populus. *Plant Cell* 30, tpc.00168.2018. doi: 10.1105/tpc.18.00168.

Xie, M., Zhang, J., Tschaplinski, T. J., Tuskan, G. A., Chen, J.-G., and Muchero, W. (2018b). Regulation of Lignin Biosynthesis and Its Role in Growth-Defense Tradeoffs. *Front Plant Sci* 9, 1427. doi: 10.3389/fpls.2018.01427.

Yang, X., Ye, C.-Y., Bisaria, A., Tuskan, G. A., and Kalluri, U. C. (2011). Identification of candidate genes in Arabidopsis and Populus cell wall biosynthesis using text-mining, co-expression network analysis and comparative genomics. *Plant science* 181, 675–687.

Yanofsky, M. F., Liljegren, S., and Ferrandiz, C. (2004). SELECTIVE CONTROL OF LIGNIN BIOSYNTHESIS IN TRANSGENIC PLANTS.

Yu, L., Chen, H., Sun, J., and Li, L. (2013). PtrKOR1 is required for secondary cell wall cellulose biosynthesis in Populus. *Tree Physiol* 34, 1289–1300.

Zhan, X., Hu, Y., Li, B., Abecasis, G. R., and Liu, D. J. (2016). RVTESTS: An efficient and comprehensive tool for rare variant association analysis using sequence data.

Bioinformatics 32, 1423–1426. doi: 10.1093/bioinformatics/btw079.

Zhang, J., Yang, Y., Zheng, K., Xie, M., Feng, K., Jawdy, S. S., et al. (2018a). Genome-wide association studies and expression-based quantitative trait loci analyses reveal roles of HCT2 in caffeoylquinic acid biosynthesis and its regulation by defense-responsive transcription factors in Populus. *New Phytologist* 220, 502–516. doi: 10.1111/nph.15297.

Zhang, J., Yang, Y., Zheng, K., Xie, M., Feng, K., Jawdy, S. S., et al. (2018b). Genome-wide association studies and expression-based quantitative trait loci analyses reveal roles of HCT2 in caffeoylquinic acid biosynthesis and its regulation by defense-responsive transcription factors in Populus. *New Phytologist* 220, 502–516. doi: 10.1111/nph.15297.

Zhong, R., Lee, C., Zhou, J., McCarthy, R. L., and Ye, Z.-H. H. (2008). A Battery of Transcription Factors Involved in the Regulation of Secondary Cell Wall Biosynthesis in Arabidopsis. *Plant Cell* 20, 2763–2782. doi: 10.1105/tpc.108.061325.

Zhong, R., and Ye, Z.-H. (2014). Secondary cell walls: biosynthesis, patterned deposition and transcriptional regulation. *Plant Cell Physiol* 56, 195–214.

Zmasek, C. M., and Eddy, S. R. (2001). ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics* 17, 383–4. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11301314 [Accessed December 16, 2018].

Zmienko, A., Marszalek-Zenczak, M., Wojciechowski, P., Samelak-Czajka, A., Luczak, M., Kozlowski, P., et al. (2020). AthCNV: A map of DNA copy number variations in the arabidopsis genome. *Plant Cell* 32, 1797–1819. doi: 10.1105/tpc.19.00640.

3 Chapter 3: Structural Variation and Population Structure in *Populus trichocarpa*

3.1 Abstract

Structural variants (SVs) are large-scale genomic rearrangements that include insertions, deletions, duplications, inversions, and translocations. SVs represent a major source of genetic variation influencing plant adaptation and phenotypic diversity. Despite their importance, SVs remain largely uncharacterized in wild populations of *Populus trichocarpa*, a model tree species and biofuel feedstock. Here, I present the first population-scale analysis of SVs in 1,484 wild *P. trichocarpa* accessions sampled across the native range in the Pacific Northwest of North America. Using two complementary SV detection algorithms (Delly and Lumpy), I identified approximately 795,000 and 877,000 high-quality SVs, respectively. Population structure analyses using PCA, UMAP, and DBSCAN clustering revealed distinct genetic groups that reflect the species' post-glacial demographic history. The largest, most homogeneous cluster comprises a northern group that rapidly expanded following glacial retreat after the Last Glacial Maximum, while southern groups exhibited greater substructure consistent with persistence in glacial refugia or beyond the reach of the Cordilleran ice sheet. Deletions were the most abundant SV type (54% in Delly, 38% in Lumpy), followed by translocations and duplications. The two calling methods captured complementary aspects of genomic variation, with Delly detecting more fine-scale population structure and Lumpy revealing clearer large-scale geographic patterns. These datasets provide an essential foundation for investigating climate-associated adaptive variants in Chapter 4 and understanding how SV shapes environmental adaptation in a keystone forest species facing rapid climate change.

3.2 Introduction

Genetic variation underlies the capacity of organisms to respond and adapt to biotic and abiotic environmental changes. It provides the raw material for natural selection to act upon, enabling phenotypes to evolve in response to shifting conditions—a process vital for long-term survival. Analytical and experimental approaches for identifying adaptive variants offer critical insight into how organisms cope with environmental stress. As climate change accelerates, understanding these mechanisms has become increasingly

urgent to predict and support species resilience.

Identifying adaptive variants is essential to breeding and engineering more resilient agricultural crops. For example, understanding the genetic basis of stress tolerance can inform efforts to develop crops better able to withstand drought, increased temperatures, salinity, and other climate-related stresses (Evans et al., 2014; Tuskan et al., 2006). Surveying wild accessions from natural populations is key for identifying genomic variants that can improve crop performance and aid conservation efforts (Liu et al., 2020; Mérot et al., 2020). Population-scale analyses offer the additional advantage of decreasing false positive rates through frequency information and replication of variant calls across multiple individuals. However, to avoid bias and fully capture signals of local adaptation, population genetic structure must also be quantified and considered.

3.2.1 **Populus trichocarpa** *as a model system*

Populus trichocarpa serves as both an economically important biofuel feedstock and a model woody plant species (Sannigrahi et al., 2010; Tuskan et al., 2006). Quantifying standing genetic variation and identifying variants that enable certain genotypes to flourish in challenging conditions informs efforts to improve biofuel crop performance while advancing fundamental tree science.

As a model system, *P. trichocarpa* offers several key advantages for studying adaptive genetic variation. The species' extensive native range across the Pacific Northwest in North America, spanning diverse environmental gradients, provides natural variation in climate conditions ideal for genotype-by-environment studies (Evans et al., 2014). The availability of the genome wide association study (GWAS) population, which consists of 1,496 wild accessions collected across this native range, facilitates capture of substantial natural variation (Evans et al., 2014; Lagergren et al., 2023). Wild outcrossing populations such as *P. trichocarpa* exhibit rapidly decaying linkage disequilibrium, facilitating easier identification of potentially causal variants compared to agricultural or domesticated species (Slavov et al., 2012). Additionally, extensive genomic, phenotypic,

and environmental datasets are already available for this population, including common garden experiments and detailed abiotic environmental data, creating a rich resource for integrative studies.

3.2.2 Structural variation impacts phenotypic architecture and environmental adaptation

Much of the research on genetic variation in *P. trichocarpa* utilized SNP (single nucleotide polymorphism) data derived from arrays or high throughput sequencing data (Geraldes et al., 2013; Tuskan et al., 2006). SNPs are genetic variants that are one nucleotide or base pair in length, and encompass insertions and deletions (indels), transitions (a substitution from the same nucleotide class, such as purine for purine), or transversions (a substitution from differing nucleotide classes, such as purine for pyrimidine). SNP-based studies investigating growth, stress tolerance, plant-microbe interactions, and other phenomena have identified important candidate genetic variants (Chhetri et al., 2020; Furches et al., 2019; Mizrachi et al., 2017; J. Zhang et al., 2018). However, it has long been recognized that SNP- and genic-based data capture only a portion of the true genetic variation in a population (Eichler et al., 2010). An increasing number of studies have discovered differing patterns of genetic variation and new mechanisms that shape phenotypes by studying genomic structural variation (SV).

SV is typically defined as genomic variants > 50 bp (base pairs) in size and include sequence insertions (INS), deletions (DEL), duplications (DUP), copy number variation (CNV), inversions (INV), and translocations (TRA or BND) (**Figure 3.1**) (Escaramís et al, 2015; Cameron et al., 2019). Although SV has been recognized as an important source of genetic variation for over 80 years, research has rapidly accelerated in the last two decades following advances in high performance computing and high throughput sequencing technologies, particularly short-read paired-end and long-read sequencing (Dobzhansky & Sturtevant, 1938; Escaramís et al., 2015; McClintock, 1950; Wellenreuther et al., 2019). Short-read paired-end sequencing technology involves fragmenting the genome into short pieces and sequencing approximately 50 - 300 bp of

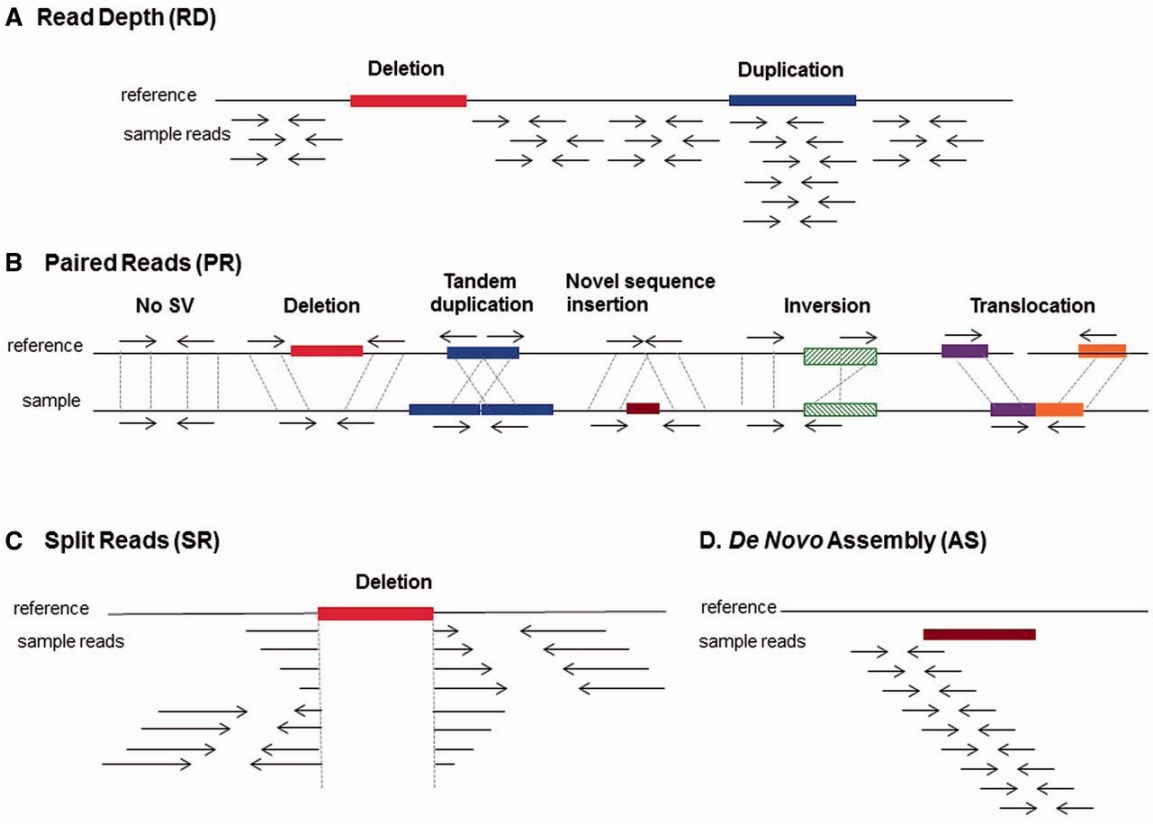


Figure 3.1: Examples of the types of large genomic SV and strategies for detection using short-read paired-end data. Detection of (A) DEL and DUP using variable read depth; (B) DEL, INS, INV, TRA, and DUP (tandem) using discordant read pairs; (C) DEL using split reads; and (D) INS using de novo assembly. Note: Figure reproduced from Escaramís et al. (2015) with permission from the publisher.

each side of the fragment. The sequenced portions are analyzed as a pair representing known sequence data for that particular genome fragment. In contrast, long-read sequencing methods analyze stretches of genomic DNA that are thousands to hundreds of thousands of base pairs in length. While long-read data provides improved resolution of genomic features, short-read methods are much cheaper and facilitate high throughput of multiple samples.

Pioneering studies in agricultural crop species have demonstrated that SV can have a significant impact on complex traits. For example, in peach fruit, an insertion was found to be associated with early fruit maturity, a deletion was associated with fruit flesh color, and a large inversion was associated with fruit shape (Guo et al., 2020). Furthermore, studies in other economically important crops such as barley, wheat, corn, and rice have shown that SV plays an important role in plant adaptation (Allaby, 2019; Fuentes et al., 2019; Gabur et al., 2019; Schiessl et al., 2019). For example, a recent study on grape cultivars found that increased tolerance of water stress was explained by higher expression levels of key metabolic, stress response, and signaling genes due to gene dosage effects associated with CNV (i.e., increased copy number of key genes led to increased gene expression) (Catacchio et al., 2019). And a study of differential disease resistance in canola lines tracked adaptive response to significant CNV and presence-absence variation (INS, DEL) in resistance genes (Dolatabadian et al., 2020).

To date, only a handful of studies have performed genome-wide characterization of SV in *Populus* species (Pinosio et al., 2016; B. Zhang et al., 2019). While most were not conducted at population scale or on wild native populations, one key study in 500 *P. nigra* x *P. deltoides* F1 hybrid offspring demonstrated how gene dosage effects traced to highly heritable INS, DEL, and CNV significantly impacted tree biomass, leaf drop, leaf color, bud burst, and canopy duration, which are traits that are of interest to biofuel engineering and breeding efforts (Bastiaanse et al., 2019). Another important insight was made in a study comparing two small populations of *P. balsamifera* from opposite ends of the species range, in which CNV exhibited differential adaptation to endemic selection

as measured by pathogen resistance, water use efficiency, and photosynthetic assimilation rates (Prunier et al., 2019). Further studies using the population- scale resources that are available for *P. trichocarpa* will be able to capture significant natural variation, facilitating new and more robust insights into the role that SV play in phenotypic architecture and natural adaptation.

3.2.3 Technical challenges in short read SV detection

While relatively inexpensive and methodologically well-established, SV detection from short-read data has inherent technical challenges and inflated type I error (Cameron et al., 2019). This is due in part to using short sequence fragments to characterize large genomic features. Read pairs are sometimes mapped to incorrect locations (particularly in repetitive genomic regions), fail to span chromosomal breakpoints, or fail to capture other critical features of variants. Although long read sequencing technologies avoid many of these issues, short-read methods will remain in use for the foreseeable future because the cost of generating long-read data sets can still be prohibitive. Furthermore, substantially more short-read datasets are currently available in public repositories for non-model organisms.

One method for mitigating short-read biases is to leverage larger sample sizes through group-based calling methods (Cameron et al., 2019; X. Wang et al., 2017). When putative SV is detected in multiple samples the false positive rate is lower than if called in a single individual because true variants will have substantially higher statistical support through replication and false positives will be filtered out in downstream quality control and data preparation steps. Group calling also provides better sensitivity through increased detection and support of rare variants, improved breakpoint accuracy, and consistent SV coordinates across the population.

3.2.4 Group methods identify SV and population structure in **P. trichocarpa**

Here, I utilized group methods implemented by two distinct callers, Delly and Lumpy, to identify SV in the *P. trichocarpa* GWAS population (Layer et al., 2014; Rausch et al.,

2012). This replication provides duplicate coverage of the major categories of large SV and ensures use of multiple types of read pair signals during identification which improves recall (Kosugi et al., 2019). Prior to calling SV it was necessary to subdivide the population into smaller groups because the full size of the population exceeded algorithmic and HPC analysis capacity. Because prior SV data for *P. trichocarpa* was lacking, I subdivided the population based on SNP-derived pairwise kinship (a practice supported by previous studies) (Göktay et al., 2021; Zhou et al., 2019). Pairwise kinship is calculated by estimating the probability that two individuals in a population share alleles due to descent from a common ancestor. If the average pairwise kinship within a group is high, those individuals are more likely to share common alleles due to shared ancestry and demographic history. Lower within group average pairwise kinship means that individuals are less likely to share common alleles. In this study, I use group-based SV calling methods to increase sensitivity and reduce the false positive rate, therefore it is logical to perform SV calling on groups with higher within group average pairwise kinship, and therefore higher frequencies of true positive SV alleles. After calling SV, the union of calls for each respective caller was used to genotype individual samples, the genotype data was merged, and the results were filtered on quality thresholds.

Population structure refers to the differential distribution of alleles within a population over its geographic range. Populations that do not experience any differences in demographic history, geographic barriers to reproduction, or selection pressures across their range will exhibit very low population structure, meaning the frequency of alleles will be distributed roughly evenly over the range (panmixia), and any subsampled group of accessions will bear high resemblance to any other subgroup with regards to allele frequencies (and likely morphology and other features as well). However, when variance in factors like geography (latitude, elevation, distance, water bodies), demography (historical population bottleneck or expansion), and selection pressure (differences in aridity, pathogen presence, cloud cover) occur within a population and the rate and distance of exchange of gametes across the range (migration) are insufficient to counteract local differences, population structure is observed. When populations are

structured, the presence and absence of alleles and their frequencies vary among subgroups of the global population. It is important to account for population structure in molecular analyses because it can confound results such that significant findings attributed to selection pressure, for example, could in fact be due to demographic history, or vice versus. Specifically in the case of *P. trichocarpa*, the north-south orientation of the population distribution means that latitudinally-related differences in selection pressure can be difficult to distinguish from isolation of subpopulations due to separation by geographic distance that exceeds gamete migration (IBD, isolation-by-distance).

While population structure has been observed in *P. trichocarpa* SNP, morphological, and other data sets, it has not yet been quantified using SV data (Evans et al., 2014; Geraldes et al., 2014; Prunier et al., 2019). Here, I collected summary statistics and performed population structure analyses using PCA, UMAP, and DBSCAN clustering on each call set separately. These analyses reveal the patterns of SV variation across the species' geographic range and provide essential population structure information required for the environmental association analyses in Chapter 4. The resulting high-quality SV datasets represent the foundation for investigating climate-associated variants that may contribute to understanding how *P. trichocarpa* adapts to environmental variation across its native range.

3.3 Methods

A subset of analyses for this chapter were performed in collaboration with the following colleagues: Roshan Abeyratne (dosage matrix), Hari Chhetri (PCA), Doug Hyatt (UMAP method), and Mallory Morgan (DBSCAN). All analyses were carried out on the Eos, Andes, and Frontier supercomputers at the Oak Ridge Leadership Computing Facility.

3.3.1 Genomic Data

The generation of genomic data was previously described in (Chhetri et al., 2019; Lagergren et al., 2023; Weighill et al., 2018). Briefly, short-read paired-end whole genome resequencing data were generated from DNA extracts collected from leaves of

1492 wild *P. trichocarpa* trees accessed in their native range in the Pacific Northwest of the United States (hereafter, "GWAS population") and aligned against the *P. trichocarpa* v4 reference genome "Nisqually-1" using BWA-MEM (Li, 2013). Prior to performing SV analyses, 8 samples suspected of being cryptic hybrids or otherwise having low quality data were removed, leaving a set of 1484 samples.

3.3.2 Population subgroupings

An objective of this research was to utilize multisample methods to analyze population-scale groups. However, the full GWAS population data set was too large to process in a single analysis with either of the SV callers included in this chapter. To break the population into smaller groups in a biologically meaningful way, samples were clustered into subpopulations based on pairwise kinship. Kinship was calculated using SNP data since prior SV data were not available for these samples and previous studies have shown that SNP data roughly recapitulate population structure observed in SV (Sang et al., 2022; Weissensteiner et al., 2020). Creation of the SNP data set was previously described in (Lagergren et al., 2023).

Pairwise SNP-based kinship was calculated by filtering the SNP data set to retain only the PASS filtered top tranche with 5% missing data allowed and MAF >= 0.05 using the BCFtools 'view' function (Danecek et al., 2021). The kinship matrix was calculated by first performing LD pruning on the SNP data set using the Plink 'indep-pairwise' function (Purcell et al., 2007) with window size of 50 kb, step size of 5, and r^2 threshold of 0.2. Next, the genetic relationship matrix (GRM) was calculated using the Plink 'make-rel' function.

To find the grouping that maximized average pairwise kinship within groups the optimum number, size, and membership of groups was determined using K-means and elbow methods using the Python packages pandas, scipy (scipy.spatial.distance, pdist, squareform), sklearn (sklearn.cluster, KMeans; scipy.cluster.hierarchy, linkage, dendrogram, fcluster), yellowbrick (yellowbrick.cluster, KElbowVisualizer) matplotlib,

and seaborn. Group summary statistics and visualizations and a clustered heatmap with attached dendrogram were calculated using Python pandas, matplotlib, and seaborn. Sample cluster membership was plotted geographically using Python mpl_toolkits Basemap and matplotlib.

Some analyses required that clusters were further subdivided another 1-3 times (hereafter, "big clusters," or "bc"; "small clusters," or "sc"; "tiny clusters," or "tc"; "nano clusters" or "nc"), for which the K-means and elbow methods were again used. For each analysis, the largest groupings that could be successfully analyzed were always used.

3.3.3 SV identification, Genotyping, and Quality Filtering

Two callers were used to identify SV in the GWAS population: Delly (Rausch et al., 2012) and Lumpy. Both callers identify deletions, duplications, inversions, and translocations (DEL, DUP, INV, BND), but Delly also identifies insertions (INS). Both callers utilize discordant read pairs (also referred to as "paired-ends"), split reads, and read depth signals to identify SV. For both callers, SV identification was performed across all 50 mapped chromosomes and scaffolds, including the mitochondrial and chloroplast genomes. For each call set, genotyping was performed using the native pipeline genotyper.

3.3.3.1 Delly

Following author recommended protocols (https://github.com/dellytools/delly), SV identification was performed in tc subgroups using the 'delly call' function with default parameter settings (see **Supplemental Text 3.6.2** for parameter details). A unified SV site list was created by taking the union of SV calls across subgroups using the 'delly merge' function with default parameters. The SV site list was then used to genotype samples using Delly's native genotyper ('delly call -v') with default parameter settings. Genotyped sample data was merged into a multisample variant call file using the software package BCFtools 'merge' function using default parameters. Genotyped calls were then PASS filtered. The criteria for Delly to assign a PASS value to the VCF

FILTER column is read pair support >= 3 reads for all variants except translocations, read pair support >= 5 reads for translocations, mean mapping quality >= 20, and genotyping quality >= 15.

3.3.3.2 Lumpy

Following author recommendations (https://github.com/arq5x/lumpy-sv), the Smoove wrapper was used to run Lumpy calling protocols (https://github.com/brentp/smoove; https://brentp.github.io/post/smoove/). SV identification was performed in nc subgroups using the 'smoove call' function with default parameter settings. A unified SV site list was created by taking the union of SV calls across subgroups using the 'smoove merge' function with default parameter settings. The SV site list was then used to genotype samples using the 'smoove genotype' function (which is a wrapper for SVTyper) (Chiang et al., 2015) with default parameters. Genotyped sample data was merged into a multisample variant call file using the 'smoove paste' function with default parameters. Genotyped calls were then filtered to retain only those with a QUAL score >=20, which indicates confidence in a given call.

3.3.4 Descriptive Statistics

For each call set, total SV counts per caller and type were calculated using BCFtools 'query' and custom bash scripting. Total SV counts by caller and SV type and SV size distributions by caller and type were calculated and visualized using the Python packages pandas and seaborn.

3.3.5 Population Analyses

3.3.5.1 PCA: global structure

To capture SV correlated population structure, a principal component analysis (PCA) was performed on each genotyped call set using GAPIT (J. Wang & Zhang, 2021). Prior to performing the analysis, data were filtered by removing putative clonal relatives (criteria: pairwise kinship > 0.177), SV calls located in scaffolds, sites with missing data > 15% and MAF < 0.02. A genotype dosage matrix was created by converting genotypes to

{0,1,2} format in which “0” represents homozygous reference genotype, “1” represents the heterozygous genotype, “2” represents homozygous alternate genotype, and missing genotypes are set “0”. A visualization of the two-dimensional projection of the first 2 PCs was produced by GAPIT.

3.3.5.2 UMAP and DBSCAN: regional structure

To capture finer population structure than can typically be identified using PCA alone (Diaz-Papkovich et al., 2019), Uniform Manifold Approximation and Projection (UMAP) analyses were conducted on the first 2-8 PCs for each call set with number of neighbors set to 15 and minimum distance of 0.1, using the Python packages umap-learn, pandas, matplotlib, and seaborn. The best fit number of PCs to use in downstream analyses of the UMAP results were judged by observing the splitting of clusters in the UMAP visualization into the smallest groups possible without splitting out individual samples. Once the optimal number of input PCs were determined for the UMAP analyses, the resulting coordinates were clustered using DBSCAN (Density-Based Spatial Clustering of Applications with Noise) (Ester et al., 1996) to determine the most likely number of subpopulations using an epsilon value of 2 and minimum data points >= 7, using the R package dbscan. The UMAP projections colorized by DBSCAN cluster membership and geographic distribution of DBSCAN clusters were visualized using the R packages plotly and leaflet.

3.4 Results

3.4.1 Population subgroupings

Cluster levels and summary statistics are provided in **Table 3.1** and a list of accessions included in this study and their cluster membership is provided in **Supplemental Table 3.1**. Note that only nc level membership is provided because the nested cluster ID naming pattern makes it possible to infer cluster membership at all levels. For example, accession BESC-1089 belongs to nc “cluster 1.1a.1”. From this cluster ID, we can infer bc membership (cluster 1), sc membership (cluster 1.1), and tc membership (cluster 1.1a). Average pairwise kinship within sc clusters, a clustered heatmap with attached

Table 3.1: Accession subgroup clustering levels and summary statistics. "Number" refers to the total number of clusters at that level. Summary statistics (mean, median, minimum, maximum, standard deviation) were performed on the number of samples within clusters for each level. Each level contains the full set of 1484 accessions.

Level	Number	Mean	Median	Minimum	Maximum	Std Dev.
big (bc)	4	371.0	238.5	123	884	346.98
small (sc)	14	106.0	54.5	23	423	110.77
tiny (tc)	28	53.0	51.0	23	86	19.87
nano (nc)	58	25.6	27.0	17	32	3.54

dendrogram, and geographic distribution of sc clusters are depicted in **Figures 3.2 - 3.4**. The cluster naming and colorization schemata make it possible to also infer bc level statistics, kinship, and mappings from sc figures. However, because population structure was low within sc clusters, visualizations were not created for tc or nc groupings.

3.4.2 SV Identification, Genotyping, and Quality Filtering

Delly identified 863,241 total SV (**Table 3.2**). PASS filtering reduced the Delly call set to a total of 794,619. Lumpy identified 896,882 total SV, with 877,216 remaining after QUAL filtering. Note that Lumpy does not identify insertions. "BND" is an abbreviation for "breakend." In recent VCF formatting specifications, translocations (both intrachromosomal and interchromosomal) are represented as a pair of breakend records rather than a single "TRA", "ITX", or "CTX" record. Overall, Delly and Lumpy identified a similar number of SV in the GWAS population, however the distribution of calls was different (**Figure 3.5**). The size distribution of calls was also similar, with significant overlap in call sizes per SV types between callers (**Figure 3.6**; note that BND are absent because they are represented as a single point between basepairs).

3.4.3 Population Structure

Removal from the data set of 71 accessions exceeding the threshold for classification as putative clonal relatives (pairwise kinship > 0.177) resulted in 1413 samples for input to PCA. Removal of SV calls located in scaffolds reduced to the Delly data set to 788,622 calls and the Lumpy data set to 867,643 calls. In the Delly set, filtering SV calls with > 15% missing values resulted in the removal of 54,014 sites, leaving 734,608 calls. After filtering sites with MAF < 0.02, a total of 325,372 SV call sites remained for PCA. In the Lumpy data set, filtering SV calls with > 15% missing values resulted in the removal of 16,523 sites, leaving 851,120 calls. After filtering sites with MAF < 0.02, a total of 386,170 SV call sites remained for PCA.

PCA analysis of the Delly data set indicated that PC1 and PC2 explained 3.78% and 1.6% of genetic variance, respectively (**Figure 3.7a**, **b**). PCA of the Lumpy call set

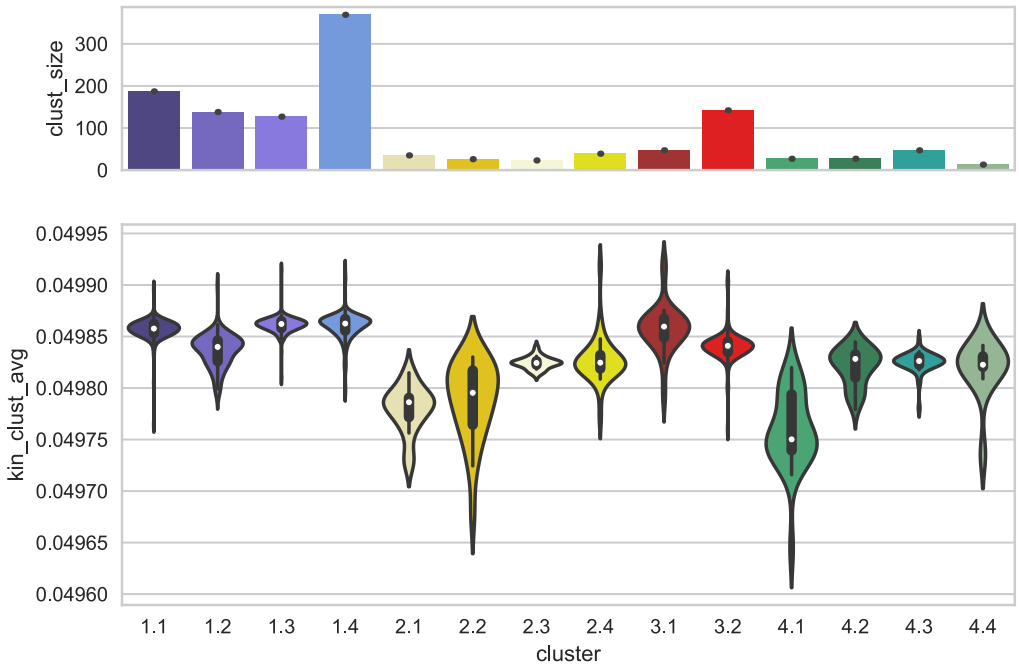


Figure 3.2: Average pairwise kinship within sc clusters. A) bar graph depicting total number of accessions within each sc cluster; B) violin plot depicting average pairwise kinship within sc clusters.

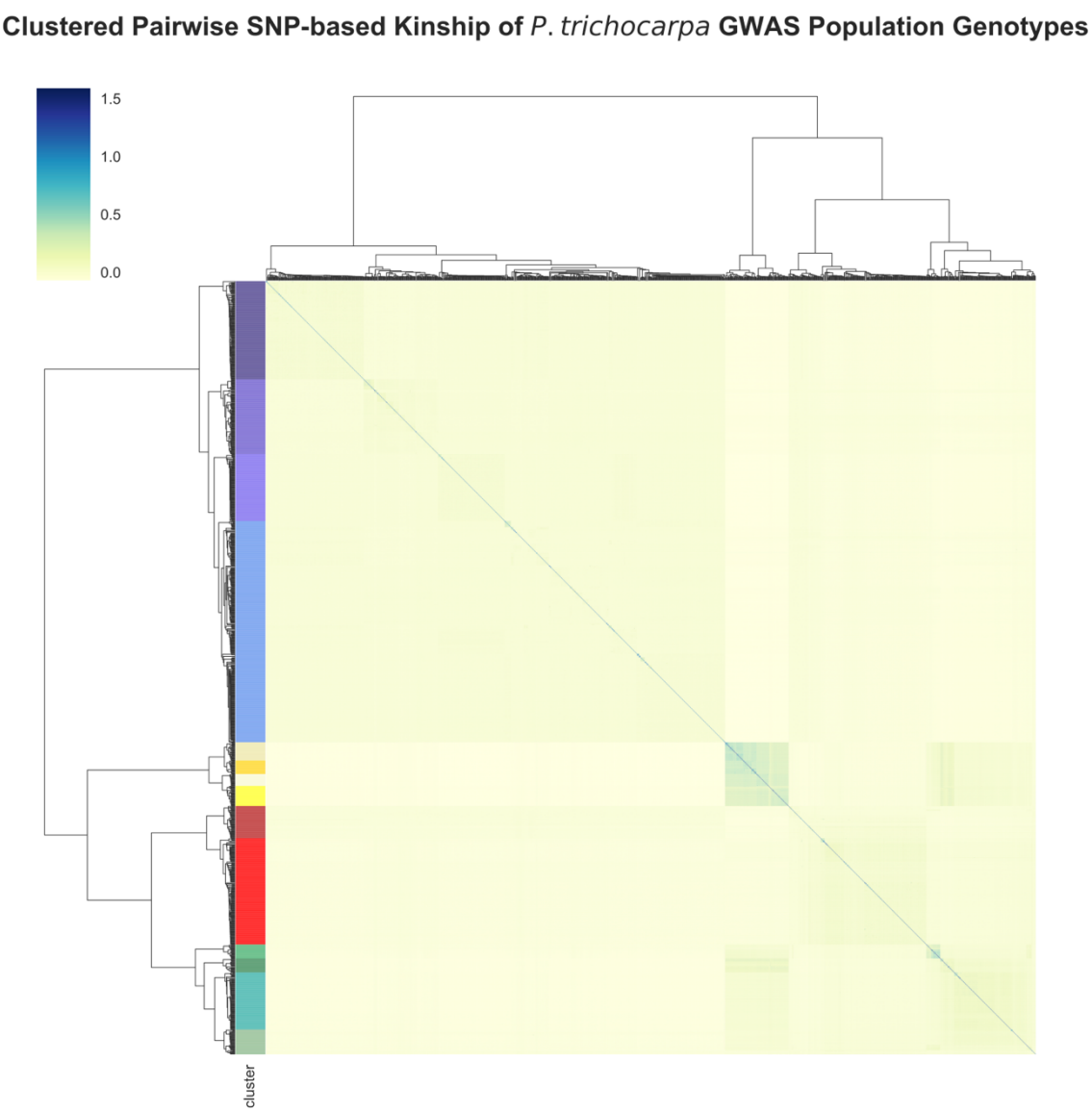


Figure 3.3: Clustered heatmap with attached dendrogram of pairwise kinship among accessions. The colorized column on the left side of the heatmap ("cluster") indicates sc cluster membership for each accession.

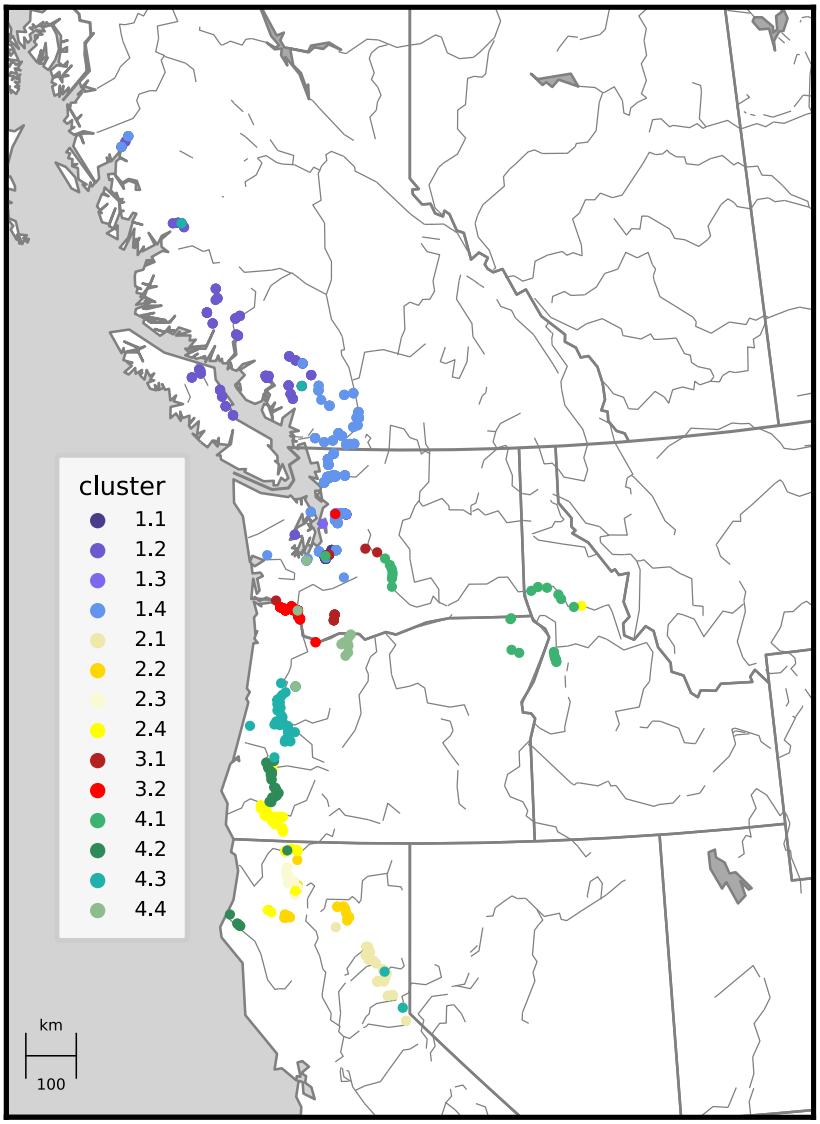


Figure 3.4: Geographic distribution of *P. trichocarpa* accessions included in Chapters 3 and 4 colorized by membership in SNP pairwise kinship-based sc clusters.

Table 3.2: Total number of SV calls by type for each calling algorithm. DEL: deletions; BND: breakends, which are an alternative notation for translocations, both intrachromosomal and interchromosomal; DUP: duplications; INS: insertions; INV: inversions.

Caller	DEL	BND	DUP	INS	INV
Delly	426,814	230,852	58,882	54,481	23,590
Lumpy	335,444	401,408	101,859	-	38,505

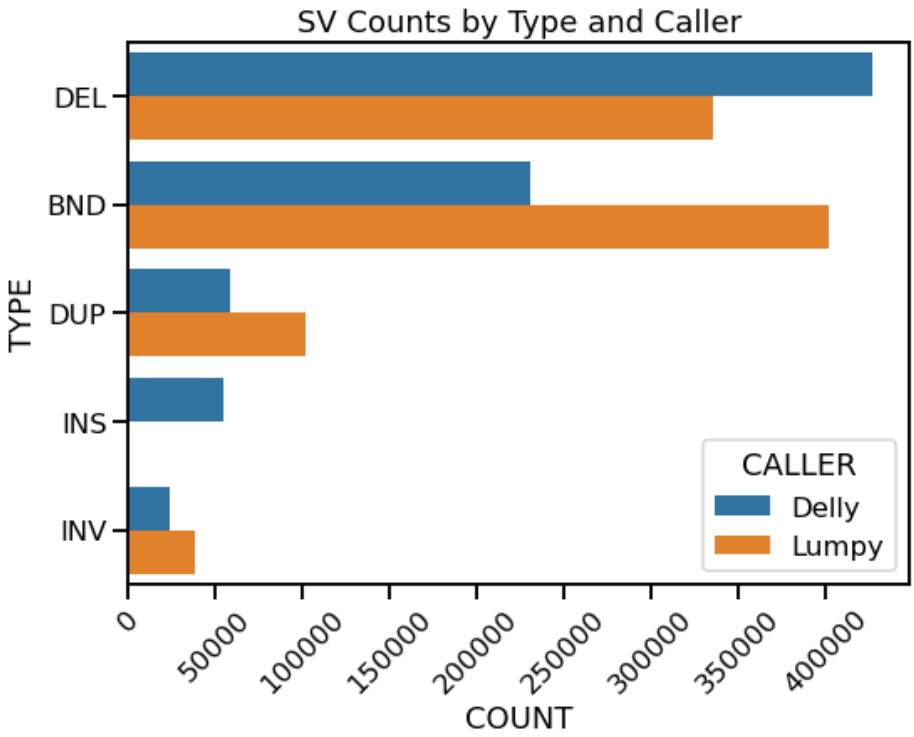


Figure 3.5: SV counts by type and SV caller.

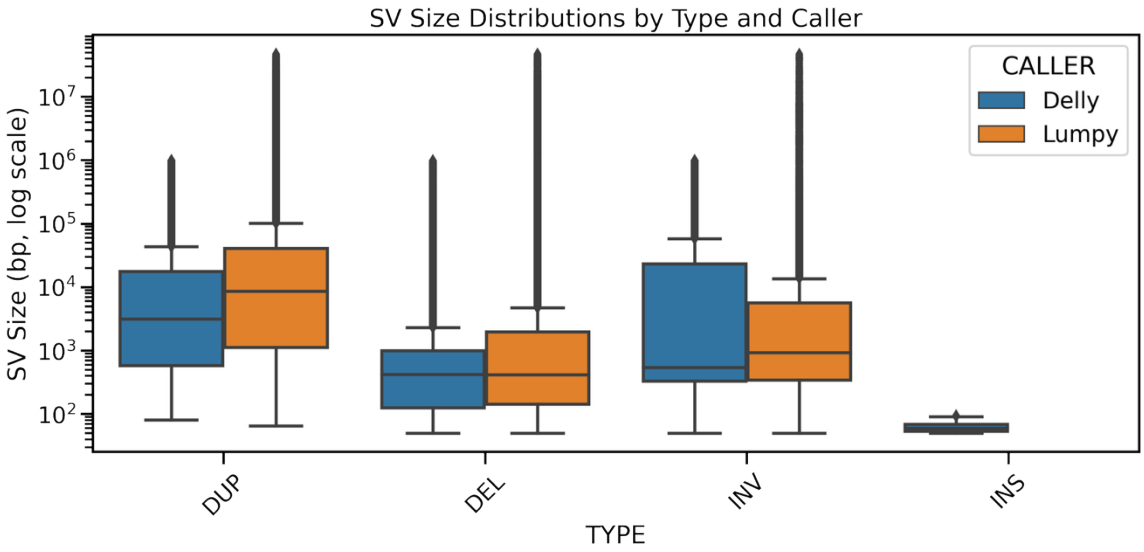


Figure 3.6: SV size distribution by type and caller.

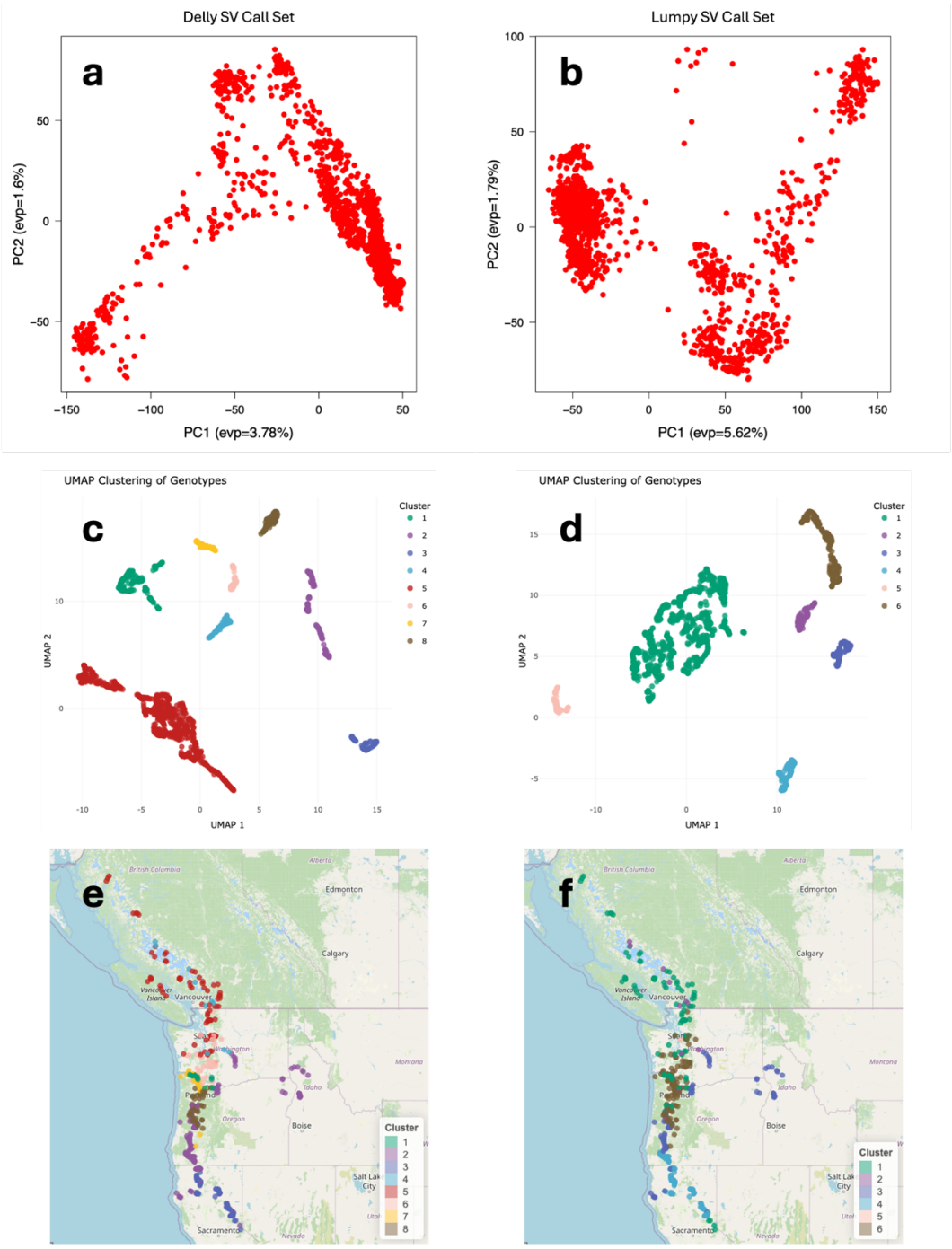


Figure 3.7: Population genetic analyses of SV call sets. Visualizations in the left column (a, c, e) are Delly data and in the right column (b, d, f) are Lumpy data. Panels a, b depict PCA results; panels c, d depict UMAP and DBSCAN results; panels e, f depict the geographic distribution of accessions colorized by membership to UMAP and DBSCAN determined subgroups.

indicated PC1 and PC2 explained 5.62% and 1.79% of genetic variance. In each of the call set projections, roughly two major groups are present, with one being more densely clustered and well-defined than the remaining points, which are somewhat less cohesive. These major groups are more distinct in the Lumpy call set and somewhat continuous in the Delly call set. These groupings approximately correspond to the patterns observed in SNP-based kinship clusters.

3.4.3.1 UMAP and DBSCAN: regional structure

The best fit number of PCs for UMAP projection was determined to be four for the Delly call set and two for the Lumpy call set (**Figure 3.7c**, **d**). These values produced well-separated groups while minimizing isolation of individual samples. One potential reason for the difference in optimum PC numbers between callers is that Delly identifies insertions, which likely contribute to additional population structure. As expected, the UMAP projections depicted much finer population structure than the PCA projections. Similar to the PCA projections, UMAP projections depicted one cluster that was much larger than the others in both Delly and Lumpy data. The DBSCAN analysis found eight subgroups in the Delly data set and six subgroups in the Lumpy data set (**Figure 3.7c** - **f**). Colorization of the geographically plotted clusters shows that in both data sets, the largest cluster is the most northern (**Figure 3.7e**, **f**). However, in the Delly data set, this large cluster (Cluster 5) ranges from British Columbia, Canada to Washington, United States. However, in the Lumpy data set, the largest group (Cluster 1) ranges down into Oregon, US. Despite differences in how sharply clusters are delineated between the two call sets, overall cluster membership is largely consistent across both the Delly and Lumpy projections. Moreover, the spatial distribution of these clusters closely mirrors the patterns observed in the SNP-based kinship map (**Figure 3.4**).

3.5 Discussion

Differences between Delly and Lumpy call sets were expected because each calling algorithm detects signals differently, which in turn is expected to lead to differences in the population structure and geographic distribution captured by each call set (Kosugi et

al., 2019). Each SV subtype is generated, maintained, or eliminated from the genome by different mechanisms, and each reflects different aspects of demographic history, migration, divergence, and local adaptation. Inversions tend to be under stronger purifying selection than other SV types because they suppress recombination between favorable combinations of alleles, consequently they are thought to reflect environmental structuring along climatic gradients rather than isolation-by-distance patterns (Wellenreuther & Bernatchez, 2018). In contrast, deletions, duplications, and translocations more closely reflect neutral demographic processes and population history, making them valuable for identifying signals associated with migration such as founder effects and population bottlenecks. While deletions are expected to provide the clearest signals of population structure and geographic separation, duplications and translocations are expected to exhibit less distinct population structure (Pokrovac & Pezer, 2022).

Here, PCA and UMAP analyses revealed that while broad geographic patterns and some fine scale patterns were consistent across the Delly and Lumpy call sets, Delly generally captured finer scale population structure while Lumpy indicated more distinct separation in large scale structure. These patterns are likely due in part to the identification of insertions and a greater number of deletions by Delly than Lumpy. Relative to the reference genome against which SV are identified, large insertions can represent novel sequence from other populations of the same species (population variation), hybridization with other species (introgression), transfer from other organisms (bacteria, viruses) or organelles (horizontal gene transfer), in addition to duplication of sequences that are present in the reference (Saxena et al., 2014; Bayer et al., 2020). Thus, depending on the ratio of inserted novel sequence versus duplicated sequence, insertions can increase population structure. Furthermore, Lumpy identified greater numbers of translocations, duplications, and inversions than Delly, which are expected to exhibit less distinct structure.

Despite these differences, both the Delly and Lumpy calls sets exhibit the same overarching pattern of population structure, which is consistent with the previously

documented demographic history and geographic patterns captured by other genetic and morphological markers in *P. trichocarpa*. The large-scale population structure observed in the PCA analyses indicates the presence of two major population subgroups in the native range of *P. trichocarpa*. The largest, most northern cluster appears more genetically homogenous and less structured, while the second group is less well-defined and contains substructure.

These subgroups are consistent with expansion out of glacial refugia following the Last Glacial Maximum (LGM) (Slavov et al., 2012). During the LGM, the Cordilleran Ice Sheet eliminated habitat in the northern range of *P. trichocarpa*, continuing as far south as Oregon (Geraldes et al., 2014). When the ice sheet retreated, the northern most subpopulations rapidly expanded along river systems to repopulate the northern territory, stretching up into British Columbia. Rapid population expansions out of refugia are characterized by genetic bottlenecks and founder effects, creating a pattern in which the expanded population contains lower genetic variation and population structure than subpopulations in the ancestral portion of the species' range.

The fine-scale structure observed in UMAP and DBSCAN results are consistent with this pattern, partitioning the second major subgroup into a series of five to seven smaller subgroups that are distributed in the southern range of *P. trichocarpa* and are more genetically distinct.

These smaller, southern subgroups are the descendants of those populations not affected by glaciation. In these groups, gene flow and population structure are affected by water sheds and river basins, which is consistent with the riparian pioneer life history of *Populus* species, as well as hybridization with *Populus* species that are not present in the northern range of *P. trichocarpa* (Bothwell et al., 2017; Evans et al., 2014; Geraldes et al., 2014; Slavov et al., 2009).

3.5.1 Future Directions

The geographic distribution of population structure observed in the SNP-based clusters

largely corresponds with observations in the SV data sets based on UMAP and DBSCAN results. Some degree of correlation was expected given that a portion of SNPs reside within SV and are transmitted within the same LD blocks, and previous studies have generally found that SNP-based population structure significantly overlaps that of SVs (S. Chen et al., 2022; Göktay et al., 2021; Zhou et al., 2019). However, the group calling methods used here could have introduced bias in the SV call sets such that they might be predisposed to align with SNP patterns. Potential bias is expected to be low since genotyping methods used the union of all SV calls and filtering (on quality, missing data, and MAF) was not performed until after genotyped calls were merged into a multisample VCF file. However, because PASS criteria and quality scores were assigned based on the original subgroup-based calling, it is possible that low frequency or low-quality calls that were filtered from call sets would have experienced greater support and potentially been retained in the call sets if they had been identified at population scale.

One approach to mitigating this bias is to use an SV caller that can identify SVs across the entire population in a single, unified analysis. Of the tools surveyed, only CNVcaller (X. Wang et al., 2017) supports this functionality, though it is limited to detecting copy number variants (CNVs) specifically. In lieu of extant software expanding maximum group capacity for calling or a new algorithm being published that can genotype at large scale, an alternative method for mitigating subgroup bias that is widely applicable to all SV types would be to use a population scale re-genotyper on unfiltered genotyped call sets such as PanPOP (Zheng et al., 2024). Furthermore, sequence-aware group-capable genotypers like BayesTyper or sequence-alignment enabled population-scale re-genotypers like PanPOP help balance the improved recall that is gained by using multiple calling methods with increased accuracy and precision. Following population scale calling or re-genotyping to eliminate potential bias, it would be valuable to investigate whether differences in genetic and geographic structure are observed.

In addition to investigating bias, further analysis to identify the relative amount of population structure and differentiation that each SV type contributes would be valuable

for future analyses to identify selection on SV. This could be accomplished by performing PCA, UMAP, DBSCAN and F_{ST} analyses on each subtype individually. It would also be valuable to explore variance partitioning among SV types relative to demographic history (i.e., expansion following the LGM and gene flow among river basins) using AMOVA (analysis of molecular variance), which accounts for evolutionary distances between variants and can hierarchically partition variance among geographic regions, populations within regions, and within populations (Excoffier et al., 1992). This approach would inform the relative role of each SV type in capturing patterns associated with local adaptation versus neutral demographic processes. For example, inversions and CNV would be expected to exhibit higher among-population variance due to their role in environmental adaptation (Wellenreuther & Bernatchez, 2018). In contrast, deletions would be expected to exhibit high overall population structure because they reflect geographic separation, and complex rearrangements (BND) would primarily reflect neutral processes such as historical migration patterns, founder effects, and population bottlenecks (Pokrovac & Pezer, 2022). This data would help inform whether differences between SV calling methods reflect genuine biological differences in how SV types track evolutionary processes (for example, differences between the total number and identity of SV identified by Delly versus Lumpy) or are primarily due to technical biases in variant detection (i.e., the approach of each caller's unique algorithm).

3.6 Supplemental Materials

3.6.1 Supplemental Tables

Supplemental Table 3.1: GWAS SNP-based subgroup cluster IDs, levels, and accessions. Please see attached spreadsheet "Furches_Chapter3_SuppTab3.1.xlsx".

3.6.2 Supplemental Text

3.6.2.1 SV Identification Algorithm Default Parameters

3.6.2.1.1 Delly Workflow

- delly call
    - -q, minimum paired-end mapping quality = 1
    - -r, minimum paired-end quality for translocation = 20

    - -s, deletion SV insert size cut-off, median+s*MAD = 9
    - -c, minimum clipping length = 25
    - -z, minimum paired-end/split-read clique size = 2
    - -m, minimum reference separation = 25
    - -n, maximum read separation = 40
- delly merge
    - -y, minimum SV site quality = 200
    - -u, maximum chunk size to merge groups of BCF files = 500
    - -a, minimum fractional ALT support = 0.150000006
    - -v, minimum coverage = 5
    - -m, minimum SV size = 0
    - -n, maximum SV size = 1000000
    - -b, maximum breakpoint offset = 1000
    - -r, minimum reciprocal overlap = 0.800000012
- delly filter
    - -y, min. SV site quality = 300
    - -a, min. fractional ALT support=0.0299999993
    - -m, min. SV size =0
    - -n, max. SV size =500000000
    - -r, min. fraction of genotyped samples =0.75
    - -q, min. median GQ for carriers and non-carriers =15
    - -e, max. read-depth ratio of carrier vs. non-carrier for a deletion=0.800000012
    - -u, min. read-depth ratio of carrier vs. non-carrier for a duplication=1.20000005
- delly classify
    - -m, min. CNV size=1000
    - -n, max. CNV size=500000000
    - -y, baseline ploidy=2
    - -q, min. site quality=50
    - -x, max. population SD=0. 150000006
- dicey chop
    - -l [--length] arg (=101) read length
    - -i [--insertsize] arg (=501) insert size
- bwa-mem2 mem
    - -k INT minimum seed length [19]
    - -w INT band width for banded alignment [100]
    - -d INT off-diagonal X-dropoff [100]
    - -r FLOAT look for internal seeds inside a seed longer than {-k} * FLOAT [1.5]
    - -y INT seed occurrence for the 3rd round seeding [20]
    - -c INT skip seeds with more than INT occurrences [500]
    - -D FLOAT drop chains shorter than FLOAT fraction of the longest overlapping chain [0.50]
    - -W INT discard a chain if seeded bases shorter than INT [0]

    - -m INT perform at most INT rounds of mate rescues for each read [50]
    - -A INT score for a sequence match, which scales options -TdBOELU unless overridden [1]
    - -B INT penalty for a mismatch [4]
    - -O INT[,INT] gap open penalties for deletions and insertions [6,6]
    - -E INT[,INT] gap extension penalty; a gap of size k cost '{-O} + {-E}*k' [1,1]
    - -L INT[,INT] penalty for 5'- and 3'-end clipping [5,5]
    - -U INT penalty for an unpaired read pair [17]
    - -T INT minimum score to output [30]
    - -h INT[,INT] if there are <INT hits with score >80% of the max score, output all in XA [5,200]
    - -I FLOAT[,FLOAT[,INT[,INT]]] specify the mean, standard deviation (10% of the mean if absent), max (4 sigma from the mean if absent) and min of the insert size distribution. FR orientation only. [inferred]
    - Important parameter settings:
        - BATCH_SIZE: 512
        - MAX_SEQ_LEN_REF: 256
        - MAX_SEQ_LEN_QER: 128
        - MAX_SEQ_LEN8: 128
        - SEEDS_PER_READ: 500
        - SIMD_WIDTH8 X: 32
        - SIMD_WIDTH16 X: 16
        - AVG_SEEDS_PER_READ: 64
- samtools sort
    - -K INT Kmer size to use for minimiser [20]
- dicey mappability
    - -r [--readlength] arg (=51) read length
    - -e [--maxEditDistance] arg (=1) max. edit distance
    - -x [--maxNeighborhood] arg (=10000) max. neighborhood size

3.6.2.1.2 Lumpy Workflow

- smoove call: identify SV in groups (cohorts)
    - -S SUPPORT minimum support required to report a variant. [default: 4]

References

Allaby, R. (2019). Clonal crops show structural variation role in domestication. In *Nature Plants* (Vol. 5, Issue 9, pp. 915–916). https://doi.org/10.1038/s41477-019-0515-8

Bastiaanse, H., Zinkgraf, M., Canning, C., Tsai, H., Lieberman, M., Comai, L., Henry, I., & Groover, A. (2019). A comprehensive genomic scan reveals gene dosage balance impacts on quantitative traits in Populus trees. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(27), 13690–13699. https://doi.org/10.1073/pnas.1903229116

Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J., & Edwards, D. (2020). Plant pan-genomes are the new reference. In *Nature Plants* (Vol. 6, Issue 8, pp. 914–920). Nature Research. https://doi.org/10.1038/s41477-020-0733-0

Bothwell, H. M., Cushman, S. A., Woolbright, S. A., Hersch-Green, E. I., Evans, L. M., Whitham, T. G., & Allan, G. J. (2017). Conserving threatened riparian ecosystems in the American West: Precipitation gradients and river networks drive genetic connectivity and diversity in a foundation riparian tree (Populus angustifolia). *Molecular Ecology*, *26*(19), 5114–5132. https://doi.org/10.1111/MEC.14281;REQUESTEDJOURNAL:JOURNAL:1365294X;PAGE:STRING:ARTICLE/CHAPTER

Cameron, D. L., Baber, J., Shale, C., Valle-Inclan, J. E., Besselink, N., van Hoeck, A., Janssen, R., Cuppen, E., Priestley, P., & Papenfuss, A. T. (2021). GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing. *Genome Biology*, *22*(1), 1–25. https://doi.org/10.1186/s13059-021-02423-x

Cameron, D. L., Di Stefano, L., & Papenfuss, A. T. (2019). Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nature Communications*, *10*(1), 3240. https://doi.org/10.1038/s41467-019-11146-4

Catacchio, C. R., Alagna, F., Perniola, R., Bergamini, C., Rotunno, S., Calabrese, F. M., Crupi, P., Antonacci, D., Ventura, M., & Cardone, M. F. (2019). Transcriptomic and genomic structural variation analyses on grape cultivars reveal new insights into the genotype-dependent responses to water stress. *Scientific Reports*, *9*(1).

https://doi.org/10.1038/s41598-019-39010-x

Chen, K., Wallis, J. W., McLellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., McGrath, S. D., Wendl, M. C., Zhang, Q., Locke, D. P., Shi, X., Fulton, R. S., Ley, T. J., Wilson, R. K., Ding, L., & Mardis, E. R. (2009). BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. *Nature Methods*, *6*(9), 677–681. https://doi.org/10.1038/nmeth.1363

Chen, S., Fang, J., Wang, Y., Wang, P., Zhang, S., Liao, Z., Lu, H., Zhang, X., Chen, S., Fang, J., Wang, Y., Wang, P., Zhang, S., Liao, Z., Lu, H., & Zhang, X. (2022). Evolutionary genomics of structural variation in the tea plant, *Camellia sinensis*. *Tropical Plants 2022 1:2*, *1*(1), 1–11. https://doi.org/10.48130/TP-2022-0002

Chhetri, H. B., Furches, A., Macaya-Sanz, D., Walker, A. R., Kainer, D., Jones, P., Harman-Ware, A. E., Tschaplinski, T. J., Jacobson, D., Tuskan, G. A., & DiFazio, S. P. (2020). Genome-Wide Association Study of Wood Anatomical and Morphological Traits in Populus trichocarpa. *Frontiers in Plant Science*, *11*, 545748. https://doi.org/10.3389/fpls.2020.545748

Chhetri, H. B., Macaya-Sanz, D., Kainer, D., Biswal, A. K., Evans, L. M., Chen, J., Collins, C., Hunt, K., Mohanty, S. S., Rosenstiel, T., Ryno, D., Winkeler, K., Yang, X., Jacobson, D., Mohnen, D., Muchero, W., Strauss, S. H., Tschaplinski, T. J., Tuskan, G. A., & DiFazio, S. P. (2019). Multitrait genome-wide association analysis of Populus trichocarpa identifies key polymorphisms controlling morphological and physiological traits. *New Phytologist*, *223*(1), 293–309. https://doi.org/10.1111/nph.15777

Chiang, C., Layer, R. M., Faust, G. G., Lindberg, M. R., Rose, D. B., Garrison, E. P., Marth, G. T., Quinlan, A. R., & Hall, I. M. (2015). SpeedSeq: Ultra-fast personal genome analysis and interpretation. *Nature Methods*, *12*(10), 966–968. https://doi.org/10.1038/nmeth.3505

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., & Davies, R. M. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), 1–4. https://doi.org/10.1093/GIGASCIENCE/GIAB008

Diaz-Papkovich, A., Anderson-Trocmé, L., Ben-Eghan, C., & Gravel, S. (2019). UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS Genetics*, *15*(11), e1008432. https://doi.org/10.1371/journal.pgen.1008432

Dobzhansky, T., & Sturtevant, A. H. (1938). Inversions in the chromosomes of Drosophila pseudoobscura. *Genetics*, *23*(1), 28–64. https://doi.org/10.1093/genetics/23.1.28

Dolatabadian, A., Bayer, P. E., Tirnaz, S., Hurgobin, B., Edwards, D., & Batley, J. (2020). Characterization of disease resistance genes in the Brassica napus pangenome reveals significant structural variation. *Plant Biotechnology Journal*, *18*(4), 969–982. https://doi.org/10.1111/pbi.13262

Du, Z. Z., He, J. B., & Jiao, W. B. (2024). A comprehensive benchmark of graph-based genetic variant genotyping algorithms on plant genomes for creating an accurate ensemble pipeline. *Genome Biology*, *25*(1), 1–24. https://doi.org/10.1186/s13059-024-03239-1

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., & Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. In *Nature Reviews Genetics* (Vol. 11, Issue 6, pp. 446–450). Nature Publishing Group. https://doi.org/10.1038/nrg2809

Escaramís, G., Docampo, E., & Rabionet, R. (2015). A decade of structural variants: Description, history and methods to detect structural variation. *Briefings in Functional Genomics*, *14*(5), 305–314. https://doi.org/10.1093/bfgp/elv014

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. www.aaai.org

Evans, L. M., Slavov, G. T., Rodgers-Melnick, E., Martin, J., Ranjan, P., Muchero, W., Brunner, A. M., Schackwitz, W., Gunter, L., Chen, J. G., Tuskan, G. A., & Difazio, S. P. (2014). Population genomics of Populus trichocarpa identifies signatures of selection and adaptive trait associations. *Nature Genetics*, *46*(10), 1089–1096. https://doi.org/10.1038/ng.3075

Excoffier, L., Smouse, P. E., & Quattro, J. M. (1992). Analysis of molecular variance

inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, *131*(2), 479–491. https://doi.org/10.1093/GENETICS/131.2.479

Fuentes, R. R., Chebotarov, D., Duitama, J., Smith, S., De la Hoz, J. F., Mohiyuddin, M., Wing, R. A., McNally, K. L., Tatarinova, T., Grigoriev, A., Mauleon, R., & Alexandrov, N. (2019). Structural variants in 3000 rice genomes. *Genome Research*, *29*(5), 870–880. https://doi.org/10.1101/gr.241240.118

Furches, A., Kainer, D., Weighill, D., Large, A., Jones, P., Walker, A. M., Romero, J., Gazolla, J. G. F. M., Joubert, W., Shah, M., Streich, J., Ranjan, P., Schmutz, J., Sreedasyam, A., Macaya-Sanz, D., Zhao, N., Martin, M. Z., Rao, X., Dixon, R. A., … Jacobson, D. (2019). Finding New Cell Wall Regulatory Genes in Populus trichocarpa Using Multiple Lines of Evidence. *Frontiers in Plant Science*, *10*. https://doi.org/10.3389/fpls.2019.01249

Gabur, I., Chawla, H. S., Snowdon, R. J., & Parkin, I. A. P. (2019). Connecting genome structural variation with complex traits in crop plants. *Theoretical and Applied Genetics*, *132*(3), 733–750. https://doi.org/10.1007/s00122-018-3233-0

Geraldes, A., Difazio, S. P., Slavov, G. T., Ranjan, P., Muchero, W., Hannemann, J., Gunter, L. E., Wymore, A. M., Grassa, C. J., Farzaneh, N., Porth, I., Mckown, A. D., Skyba, O., Li, E., Fujita, M., Klápště, J., Martin, J., Schackwitz, W., Pennacchio, C., … Tuskan, G. A. (2013). A 34K SNP genotyping array for Populus trichocarpa: Design, application to the study of natural populations and transferability to other Populus species. *Molecular Ecology Resources*, *13*(2), 306–323. https://doi.org/10.1111/1755-0998.12056

Geraldes, A., Farzaneh, N., Grassa, C. J., Mckown, A. D., Guy, R. D., Mansfield, S. D., Douglas, C. J., & Cronk, Q. C. B. (2014). Landscape genomics of populus trichocarpa: The role of hybridization, limited gene flow, and natural selection in shaping patterns of population structure. *Evolution*, *68*(11), 3260–3280. https://doi.org/10.1111/evo.12497

Göktay, M., Fulgione, A., & Hancock, A. M. (2021). A New Catalog of Structural Variants in 1,301 A. thaliana Lines from Africa, Eurasia, and North America

Reveals a Signature of Balancing Selection at Defense Response Genes. *Molecular Biology and Evolution*, *38*(4), 1498–1511. https://doi.org/10.1093/molbev/msaa309

Guo, J., Cao, K., Deng, C., Li, Y., Zhu, G., Fang, W., Chen, C., Wang, X., Wu, J., Guan, L., Wu, S., Guo, W., Yao, J. L., Fei, Z., & Wang, L. (2020). An integrated peach genome structural variation map uncovers genes associated with fruit traits. *Genome Biology*, *21*(1), 258. https://doi.org/10.1186/s13059-020-02169-y

Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M., & Kamatani, Y. (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biology*, *20*(1), 8–11. https://doi.org/10.1186/s13059-019-1720-5

Lagergren, J., Pavicic, M., Chhetri, H. B., York, L. M., Hyatt, D., Kainer, D., Rutter, E. M., Flores, K., Bailey-Bale, J., Klein, M., Taylor, G., Jacobson, D., & Streich, J. (2023). Few-Shot Learning Enables Population-Scale Analysis of Leaf Traits in Populus trichocarpa. *Plant Phenomics*, *5*. https://doi.org/10.34133/PLANTPHENOMICS.0072

Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). LUMPY: A probabilistic framework for structural variant discovery. *Genome Biology*, *15*(6), 1–19. https://doi.org/10.1186/gb-2014-15-6-r84

Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. https://arxiv.org/pdf/1303.3997

Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G. A., Zhang, H., Liu, Z., Shi, M., Huang, X., Li, Y., Zhang, M., Wang, Z., Zhu, B., Han, B., Liang, C., & Tian, Z. (2020). Pan-Genome of Wild and Cultivated Soybeans. *Cell*, *182*(1), 162-176.e13. https://doi.org/10.1016/j.cell.2020.05.023

McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences of the United States of America*, *36*(6), 344–355. https://doi.org/10.1073/pnas.36.6.344

Mérot, C., Oomen, R. A., Tigano, A., & Wellenreuther, M. (2020). A Roadmap for Understanding the Evolutionary Significance of Structural Genomic Variation. *Trends in Ecology and Evolution*, *35*(7), 561–572.

https://doi.org/10.1016/j.tree.2020.03.002

Mizrachi, E., Verbeke, L., Christie, N., Fierro, A. C., Mansfield, S. D., Davis, M. F., Gjersing, E., Tuskan, G. A., Van Montagu, M., Van De Peer, Y., Marchal, K., Myburg, A. A., de Peer, Y., & others. (2017). Network-based integration of systems genetics data reveals pathways associated with lignocellulosic biomass accumulation and processing. *Proceedings of the National Academy of Sciences*, *114*(5), 1195–1200. https://doi.org/10.1073/pnas.1620119114

Pinosio, S., Giacomello, S., Faivre-Rampant, P., Taylor, G., Jorge, V., Le Paslier, M. C., Zaina, G., Bastien, C., Cattonaro, F., Marroni, F., Morgante, M., Faivre-Rampant, P., Cattonaro, F., Taylor, G., Giacomello, S., Zaina, G., Jorge, V., Marroni, F., Pinosio, S., … Morgante, M. (2016). Characterization of the Poplar Pan-Genome by Genome-Wide Identification of Structural Variation. *Molecular Biology and Evolution*, *33*(10), 2706–2719. https://doi.org/10.1093/molbev/msw161

Pokrovac, I., & Pezer, Ž. (2022). Recent advances and current challenges in population genomics of structural variation in animals and plants. In *Frontiers in Genetics* (Vol. 13, p. 1060898). Frontiers Media S.A. https://doi.org/10.3389/fgene.2022.1060898

Prunier, J., Giguère, I., Ryan, N., Guy, R., Soolanayakanahally, R., Isabel, N., MacKay, J., & Porth, I. (2019). Gene copy number variations involved in balsam poplar (Populus balsamifera L.) adaptive variations. *Molecular Ecology*, *28*(6), 1476–1490. https://doi.org/10.1111/mec.14836

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R. R., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. W. W., Daly, M. J., Sham, P. C., & others. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., & Korbel, J. O. (2012). DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, *28*(18), i333–i339. https://doi.org/10.1093/bioinformatics/bts378

Sang, Y., Long, Z., Dan, X., Feng, J., Shi, T., Jia, C., Zhang, X., Lai, Q., Yang, G.,

Zhang, H., Xu, X., Liu, H., Jiang, Y., Ingvarsson, P. K., Liu, J., Mao, K., & Wang, J. (2022). Genomic insights into present local adaptation and future climate change vulnerability of a keystone forest tree species in East Asian. *BioRxiv*, 1–40. https://doi.org/10.1101/2022.04.04.486908

Sannigrahi, P., Ragauskas, A. J., & Tuskan, G. A. (2010). Poplar as a feedstock for biofuels: A review of compositional characteristics. *Biofuels, Bioproducts and Biorefining*, *4*(2), 209–226. https://doi.org/10.1002/bbb.206

Saxena, R. K., Edwards, D., & Varshney, R. K. (2014). Structural variations in plant genomes. *Briefings in Functional Genomics and Proteomics*, *13*(4), 296–307. https://doi.org/10.1093/bfgp/elu016

Schiessl, S. V., Katche, E., Ihien, E., Chawla, H. S., & Mason, A. S. (2019). The role of genomic structural variation in the genetic improvement of polyploid crops. In *Crop Journal* (Vol. 7, Issue 2, pp. 127–140). Crop Science Society of China/ Institute of Crop Sciences. https://doi.org/10.1016/j.cj.2018.07.006

Sibbesen, J. A., Maretty, L., & Krogh, A. (2018). Accurate genotyping across variant classes and lengths using variant graphs. *Nature Genetics*, *50*(7), 1054–1059. https://doi.org/10.1038/s41588-018-0145-5

Slavov, G. T., Difazio, S. P., Martin, J., Schackwitz, W., Muchero, W., Rodgers-Melnick, E., Lipphardt, M. F., Pennacchio, C. P., Hellsten, U., Pennacchio, L. A., Gunter, L. E., Ranjan, P., Vining, K., Pomraning, K. R., Wilhelm, L. J., Pellegrini, M., Mockler, T. C., Freitag, M., Geraldes, A., … Tuskan, G. A. (2012). Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree Populus trichocarpa. *New Phytologist*, *196*(3), 713–725. https://doi.org/10.1111/j.1469-8137.2012.04258.x

Slavov, G. T., Leonardi, S., Burczyk, J., Adams, W. T., Strauss, S. H., & Difazio, S. P. (2009). Extensive pollen flow in two ecologically contrasting populations of Populus trichocarpa. *Molecular Ecology*, *18*(2), 357–373. https://doi.org/10.1111/j.1365-294X.2008.04016.x

Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., others, Putnam, M., Ralph, S., Rombauts,

S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R. R. P. R., … Rokhsar, D. (2006). The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science*, *313*(5793), 1596–1604. https://doi.org/10.1126/science.1128691

Wang, J., & Zhang, Z. (2021). GAPIT Version 3: Boosting Power and Accuracy for Genomic Association and Prediction. *Genomics, Proteomics and Bioinformatics*, *19*(4), 629–640. https://doi.org/10.1016/j.gpb.2021.08.005

Wang, X., Zheng, Z., Cai, Y., Chen, T., Li, C., Fu, W., & Jiang, Y. (2017). CNVcaller: Highly efficient and widely applicable software for detecting copy number variations in large populations. In *GigaScience* (Vol. 6, Issue 12, pp. 1–12). Oxford Academic. https://doi.org/10.1093/gigascience/gix115

Weighill, D. A., Jones, P., Shah, M., Ranjan, P., Muchero, W., Schmutz, J., Sreedasyam, A., Macaya-Sanz, D., Sykes, R., Zhao, N., others, Martin, M. Z., DiFazio, S., Tschaplinski, T. J., Tuskan, G., Jacobson, D., & others. (2018). Pleiotropic and Epistatic Network-Based Discovery: Integrated Networks for Target Gene Discovery. *Frontiers in Energy Research*, *6*, 267997. https://doi.org/10.3389/fenrg.2018.00030

Weissensteiner, M. H., Bunikis, I., Catalán, A., Francoijs, K. J., Knief, U., Heim, W., Peona, V., Pophaly, S. D., Sedlazeck, F. J., Suh, A., Warmuth, V. M., & Wolf, J. B. W. (2020). Discovery and population genomics of structural variation in a songbird genus. *Nature Communications 2020 11:1*, *11*(1), 1–11. https://doi.org/10.1038/s41467-020-17195-4

Wellenreuther, M., & Bernatchez, L. (2018). Eco-Evolutionary Genomics of Chromosomal Inversions. *Trends in Ecology and Evolution*, *33*(6), 427–440. https://doi.org/10.1016/J.TREE.2018.04.002/ASSET/D4BC03CB-1B21-4BB5-823B-5B823234FC87/MAIN.ASSETS/GR2.JPG

Wellenreuther, M., Mérot, C., Berdan, E., & Bernatchez, L. (2019). Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification. *Molecular Ecology*, *28*(6), 1203–1209. https://doi.org/10.1111/mec.15066

Zhang, B., Zhu, W., Diao, S., Wu, X., Lu, J., Ding, C. J., & Su, X. (2019). The poplar

pangenome provides insights into the evolutionary history of the genus. *Communications Biology*, *2*(1). https://doi.org/10.1038/s42003-019-0474-7

Zhang, J., Yang, Y., Zheng, K., Xie, M., Feng, K., Jawdy, S. S., Gunter, L. E., Ranjan, P., Singan, V. R., Engle, N., Lindquist, E., Barry, K., Schmutz, J., Zhao, N., Tschaplinski, T. J., LeBoldus, J., Tuskan, G. A., Chen, J. G., & Muchero, W. (2018). Genome-wide association studies and expression-based quantitative trait loci analyses reveal roles of HCT2 in caffeoylquinic acid biosynthesis and its regulation by defense-responsive transcription factors in Populus. *New Phytologist*, *220*(2), 502–516. https://doi.org/10.1111/nph.15297

Zheng, Z., Zhu, M., Zhang, J., Liu, X., Hou, L., Liu, W., Yuan, S., Luo, C., Yao, X., Liu, J., & Yang, Y. (2024). A sequence-aware merger of genomic structural variations at population scale. *Nature Communications*, *15*(1), 1–9. https://doi.org/10.1038/s41467-024-45244-9

Zhou, Y., Minio, A., Massonnet, M., Solares, E., Lv, Y., Beridze, T., Cantu, D., & Gaut, B. S. (2019). The population genetics of structural variants in grapevine domestication. *Nature Plants*, *5*(9), 965–979. https://doi.org/10.1038/s41477-019-0507-8

4 Chapter 4: Identifying Climate-Associated SV Candidates in *Populus trichocarpa* using GWAS and Network Biology Methods

4.1 Abstract

Climate change threatens global agriculture and bioenergy production, necessitating the development of stress-resilient biofuel feedstock crops. Structural variants (SVs) can drive rapid environmental adaptation through gene dosage, loss of function, introduction of novel sequence, and suppression of recombination. Despite their adaptive potential, SVs remain underexplored in climate adaptation studies, particularly in forest trees. Genome-wide association studies (GWAS) are one method for finding climate-relevant genetic variants, but the stringent significance thresholding of traditional strategies can produce high false negative rates. Combining relaxed thresholds to increase sensitivity with network biology methods to reduce putative false positives can boost signal in a biologically meaningful manner. We combined GWAS using relaxed significance thresholds (Bonferroni 0.2) with network-based filtering that leveraged multiple lines of evidence in a multiplex network to identify SVs associated with 15 climate variables in 1,413 wild *Populus trichocarpa* accessions. GWAS analyses of Delly (325,372 SVs) and Lumpy (386,170 SVs) datasets identified 13,124 and 11,297 significant SV-climate associations, respectively. Different SV types captured distinct signals, with some climatype associations exhibiting patterns suggestive of seasonal variation in selection pressure. After mapping the union of significant SV to nearest neighbor genes, GRIN network filtering assessed gene connectivity in the multiplex network and retained 5,471 unique genes as biologically relevant candidates. Functional clustering using MENTOR produced 270 gene clades and identified three high-priority *PtKCS* gene candidates associated with 13-15 climate variables each. *PtKCS* (3-ketoacyl-CoA synthase) genes are located within an 80 kb region of complex SV produced by ancestral duplications, deletions, and translocations and encode enzymes essential for cuticular wax biosynthesis, which is critical for drought, temperature, and pathogen response. The integrated approach used here identified high priority, climate-relevant SV candidates in *P. trichocarpa*, providing actionable targets for gene-by-environment studies and validation efforts aimed at increasing resilience in bioenergy crops.

4.2 Introduction

Climate change poses unprecedented challenges to global agriculture and bioenergy production, with drought stress and extreme temperature fluctuations threatening crop yields and biomass productivity (Yang et al., 2018). The development of climate-resilient biofuel feedstocks has become critical for sustainable energy production, particularly in marginal lands where environmental stresses limit traditional crop cultivation. *Populus trichocarpa* (black cottonwood) serves as a model woody bioenergy crop due to its rapid growth, wide geographic distribution across diverse climatic gradients, and established genomic resources (Evans et al., 2014; Slavov et al., 2012). Understanding the genetic basis of climate adaptation in *P. trichocarpa* is essential for developing improved cultivars capable of maintaining productivity under increasingly variable environmental conditions (Porth et al., 2015).

Genome-wide association studies (GWAS) have emerged as a powerful approach for dissecting the genetic architecture of complex traits in plants (Storfer et al., 2018). By scanning thousands or millions of genetic variants across the genome, GWAS test whether variation in genetic traits is statistically associated with variation in phenotypic traits across a population. GWAS typically employ mixed linear models to control for population structure and cryptic relatedness while leveraging linkage disequilibrium between genetic markers to test for association with phenotypic data. Significant associations indicate genomic regions potentially harbor causal variants affecting the trait of interest. This approach is particularly effective for dissecting the genetic architecture of complex, polygenic traits where multiple loci contribute small to moderate effects.

In *P. trichocarpa*, GWAS has successfully identified loci underlying wood quality traits, disease resistance, and phenology (Evans et al., 2014; Porth et al., 2015). However, traditional GWAS approaches face several limitations when applied to climate adaptation studies. First, the use of stringent multiple testing corrections often results in high false negative rates, particularly for polygenic traits where many loci of small effect contribute to phenotypic variation (Rellstab et al., 2015). Second, arbitrary significance thresholds

fail to account for the biological relevance of genetic variants that may fall below statistical cutoffs but participate in important regulatory networks. These limitations are particularly pronounced when studying structural variants (SVs). Despite their potential for large phenotypic effects, SVs remain underrepresented in plant GWAS due to technical challenges in detection and genotyping from short-read sequencing data (Cameron et al., 2019).

Recent advances in network biology offer promising solutions to overcome traditional GWAS limitations. Network biology models biological systems as interconnected networks where nodes represent biological entities (genes, proteins, metabolites) and edges represent functional relationships (co-expression, protein interactions, regulatory connections). Network approaches reveal emergent properties and biological pathways that are not apparent from individual component analysis alone (for background information on network construction methods, see Chapter 2 and Chapter 4 Methods: 4.3.4 Network Biology Methods). By integrating association signals with gene co-expression networks, protein-protein interactions, and functional annotations, network-based approaches can identify biologically meaningful candidates that might be missed by arbitrary statistical thresholds. Methods such as Gene set Refinement through Interacting Networks (GRIN) and Multiplex Embedding of Networks for Team-Based Omics Research (MENTOR) leverage network topology whereby genes with similar network properties are likely to share functional roles (Kainer et al., 2024; Sullivan, Lane, et al., 2024; Sullivan, Miller, et al., 2024). This systems-level perspective is particularly valuable for understanding climate adaptation, where coordinated responses across multiple pathways determine plant survival under stress conditions (Batsale et al., 2021; Wang et al., 2020).

SVs include deletions (DEL), insertions (INS), duplications (DUP), inversions (INV), and translocations (TRA or BND), and represent a significant source of genetic diversity that can drive rapid adaptation in plants to environmental stress (Allaby, 2019; Fuentes et al., 2019; Gabur et al., 2019; Schiessl et al., 2019; Zmienko et al., 2020). SVs can

simultaneously affect multiple genes, alter gene expression through position effects, and suppress recombination to maintain favorable allelic combinations. In plants, SVs have been implicated in diverse adaptive traits including flowering time variation, disease resistance, and abiotic stress tolerance. Furthermore, different SV types capture distinct evolutionary signals: inversions often maintain locally adapted gene complexes by suppressing recombination, copy number variants (CNVs) facilitate rapid adaptation through gene dosage effects (whereby multiple copies of a gene lead to increased expression), and deletions and translocations primarily reflect neutral demographic processes.

In Chapter 3, I identified SV and population structure in the *P. trichocarpa* GWAS population. This chapter extends that work by investigating associations between SV and climate variables relevant to drought stress and seasonal acclimation. First, GWAS analyses identified SV associated with 15 climatic variables (such as soil moisture, temperature, drought severity) across seasonal timepoints and a relaxed significance threshold was used to increase sensitivity. Next, I utilized GRIN to filter out potential false positives based on multiplex network connectivity. MENTOR was then used to perform functional hierarchical clustering through which I identified three high priority candidates annotated as KCS (3-ketoacyl-CoA synthase) genes whose proteins perform very long chain fatty acid elongation integral to plant cuticular wax biosynthesis and stress response (Gonzales-Vigil et al., 2017; Jones and Furches et al., 2025; Lee et al., 2009). *PtKCS*-containing clades were then functionally characterized using the MENTOR-IA interpretation pipeline, and population scale RNA-seq and metabolomics data were used to further characterize the *PtKCS* gene region. By integrating population genomics, association mapping, and network biology methods in combination with increased sensitivity afforded by a relaxed GWAS threshold, I was able to identify biologically impactful gene candidates supported by multiple lines of evidence. This work provides a framework for identifying high-confidence variant candidates for climate-adaptive improvement of bioenergy crops.

4.3 Methods

A subset of analyses for this chapter were performed in collaboration with the following colleagues: Hari Chhetri (GWAS pipeline), Manesh Shah (GRIN/MENTOR pipeline), and Gabi Madrid (MENTOR-AI pipeline).

4.3.1 Genomic Data

The generation of genomic data and SV call sets was previously described in Chapter 3. In brief, Delly and Lumpy SV identification algorithms were used to call SV in the GWAS population, which consisted of analysis of short-read paired-end data derived from 1496 accessions collected from the native range of *P. trichocarpa*. SV calls were genotyped using each algorithm's native pipeline.

4.3.2 Climate Data

Climate data was downloaded from the terraclimate website (https://www.climatologylab.org/terraclimate.html) in netCDF format for 15 climate variables for each month (12 months/year) between 1958 to 2019, for a total of 10,980 data points (**Table 4.1**). An additional climate variable, temperature range (trange), was then created by subtracting minimum temperature (tmin) from maximum temperature (tmax). A custom script was used to extract climate data for each *P. trichocarpa* accession's given geolocation. Monthly data was averaged across years (e.g., for a given variable, all January data points across years were averaged to produce a single "January" value), resulting in each climate variable having 12 averaged monthly data points for a total of 180 data points. Prior to analysis, outlier data points that exceeded a Median Absolute Deviation (MAD) threshold of six and those with missing or invariable values were removed.

4.3.3 SV-Climatype GWAS Analyses

Prior to performing GWAS analyses, data preparation and filtering were performed on Delly and Lumpy SV call sets as previously described in Chapter 3. In brief, the following were removed from the data sets: clonal relatives (pairwise kinship > 0.177),

Table 4.1: Climate variables used in GWAS analyses.

Abbreviation	Climate variable description
aet	Actual evapotranspiration, derived using a one-dimensional soil water balance model
def	Climate water deficit, derived using a one-dimensional soil water balance model
PDSI	Palmer Drought Severity Index
pet	Reference evapotranspiration (ASCE Penman-Montieth)
ppt	Precipitation accumulation
q	Runoff, derived using a one-dimensional soil water balance model
soil	Soil moisture, derived using a one-dimensional soil water balance model
srad	Downward surface shortwave radiation
swe	Snow water equivalent, derived using a one-dimensional soil water balance model
tmax	Minimum temperature
tmin	Maximum temperature
trange	Temperature range (Max - Min temperature)
vap	Vapor pressure
vpd	Vapor pressure deficit
ws	Wind-speed at 10m

SV calls occurring in scaffolds, sites with missing data > 15% and MAF < 0.02. As previously described in Chapter 3, a dosage matrix was generated and PCA were performed using GAPIT on each call set. GAPIT was then used to run four GWAS analyses: one single locus method (MLM, or single locus mixed linear model, which tests one SV at a time) and three multi-locus methods (MLMM or multi-locus mixed model; FarmCPU, and BLINK) using the GWAS model

$$Y = SVs + GRM + 2SNP_PCs$$

where Y is the climate vector, SVs is the structural variant matrix, GRM is the genetic relationship matrix, and $2SNP_PCs$ are the first two principal components (PCs) of a PCA performed on the SNP data set described in Chapters 1 and 2, which were used as covariates. A relaxed Bonferroni significance threshold of 0.2 was used to avoid arbitrarily filtering out true positives.

4.3.4 Network Biology Analyses

SV positions from significant SV-climatype GWAS associations were mapped to the two physically nearest neighboring genes on the chromosome (i.e., the nearest upstream to the SV position and the nearest downstream) using a custom script to facilitate gene-based network biology and functional analyses. The GRIN method was then used to remove potential false positives from the GWAS gene sets by assessing interconnectivity of the genes within a biologically relevant multiplex network (Sullivan, Lane, et al., 2024).

4.3.4.1 Filtering false positives using GRIN

The generation of multiplex networks was previously discussed in Chapter 2. In brief, each layer consists of either gene to gene or gene to phenotype connections, in which genes or phenotypes are represented as nodes, and connecting values are represented as edges (Chapter 2: **Figure 2.1**). Edge values consist of the various relationships and values produced during generation of a given data layer. For example, a layer of pairwise gene expression correlations would consist of genes connected to another by the thresholded

pairwise correlation value. Here, multiplex network layers consisted of the following previously published *Populus* networks: a coexpression network, a knockout phenotype network, predictive expression networks (PEN) from leaf and xylem tissues, a metabolic pathway network derived from PoplarCyc, and a protein-protein interaction network derived from STRING-DB (Cope et al., 2023).

Step one of the GRIN method (https://github.com/sullivanka/GRIN) runs a RWR (Random Walk with Restart) analysis iteratively on *n -1* genes from the GWAS set (Kainer et al., 2024). In each round, the included genes are used to rank the left-out gene with regards to amount of connectivity to the rest of the GWAS gene set based on RWR analysis (i.e., leave-one-out cross validation, or LOOCV), resulting in a list of connectivity ranks for every gene in the GWAS set. Next, a null distribution is generated by running RWR-LOOCV on 100 randomly created gene sets of equivalent size to the input GWAS set, sorting each by rank, and retaining the median rank from each set. In step two, the GWAS gene set ranks are compared to null distribution ranks using a Wilcoxon rank sum test performed on a series of sliding windows that are 35% of the total size of the GWAS gene set. The resulting p-values are then plotted and the elbow point of the curve is identified, above which indicates significant difference from random. Genes with ranks above the elbow point are retained as likely true positives and those below the elbow point are removed. In this study, GRIN was applied to the gene-mapped GWAS results from each variant caller using default parameters, including a restart probability of 0.7 and a tau value of 1.

4.3.4.2 Identifying function using MENTOR

After removing potential false positives from each variant caller's GWAS results, MENTOR (Sullivan, Miller, et al., 2024) was used to functionally and hierarchically cluster the union of GRIN-retained Delly and Lumpy GWAS gene sets based on the multiplex network described above to facilitate overall biological interpretation, identification of highly impactful variants, and the comparison of the GWAS results produced by each variant caller. In brief, the first step of the MENTOR method creates

RWR-score vectors for each gene in the input (GWAS genes) set by scoring how frequently all other genes in the multiplex network are visited by the random walk when starting from a given input gene (in this case, input genes were those identified by the GWAS analyses). Score vectors are sorted in rank order, the mean RWR score (how often each gene was visited) at each rank is calculated, and the rank order vectors are truncated to retain the most frequently explored genes based on the elbow point in a mean scores-vs-ranks curve. In step two, a dendrogram of functionally related input genes is generated by computing Spearman's rank correlation coefficients (ρ) between all pairs of rank vectors, converting them to distances (1 – ρ), and performing agglomerative hierarchical clustering. The resulting dendrogram is segmented into a user-specified number of clades (k), with additional sub-clustering applied as needed to ensure that no clade exceeds the maximum allowed number of genes. Here, MENTOR was run on the union of GRIN-retained Delly and Lumpy GWAS gene sets using the following parameters: --restart 0.7, --tau 1, --distance-metric "spearman", --cluster-method "hierarchical-clustering", --linkage-method "complete", --k 3, --max-cluster-size 40. A rectangular dendrogram was created for visualization. In addition, a heatmap was produced indicating the source SV call set and SV-climatype associations mapped to each gene.

4.3.4.3 Functional Profiling Using MENTOR-IA

After producing the functionally clustered dendrogram of GRIN-retained Delly and Lumpy GWAS genes, a subset of genes were selected for further analysis using the MENTOR Interpretation Agent (MENTOR-IA) pipeline (Vlot et al., 2025). MENTOR-IA leverages large language model (LLM) and Retrieval-Augmented Generation (RAG) agents to retrieve and interpret diverse data sources and analytical approaches, generating a mechanistic framework for input gene sets.

In the case of plant analyses, MENTOR-IA derives its power from utilizing the extensive resources available for *Arabidopsis thaliana*, making it necessary to first map input genes to Arabidopsis orthologs shared orthology based on protein domains and sequence

identity (which is provided as part of the reference genome annotation data by Phytozome). Then, the MENTOR-IA multi-agent architecture performs a series of tasks for input gene sets. The first agent integrates gene function summaries from TAIR, Gene Ontology (GO) annotations pathway assignments via MyGene.info (Wu et al., 2013), and genetic association data. The second conducts targeted searches of EuropePMC (McWilliam et al., 2013) for input genes with associated PMIDs, extracts relevant scientific literature, and provides contextual insights into gene functions and associations. The third performs functional overrepresentation analyses using g:Profiler (Kolberg et al., 2023) to identify statistically significant enrichment of GO terms, pathways, and other functional categories within the gene set. Lastly, an aggregator agent integrates findings from all agents, highlighting consensus interpretations and resolving discrepancies across data sources.

In the Delly/Lumpy GWAS results dendrogram, genes having the highest number of significant GWAS associations with multiple climate variables were used to prioritize their respective clades of membership for MENTOR-IA analysis. All genes within selected clades were compiled to form input gene sets which were then mapped to their Arabidopsis orthologs using *P. trichocarpa* reference genome v4.1 annotation data (Goodstein et al., 2011; https://phytozome-next.jgi.doe.gov/) and custom scripts.

4.3.5 KCS expression profiles

To visualize divergence in expression profiles, a clustered heatmap of the *P. trichocarpa* RNA-seq Gene Atlas expression data in reference accession Nisqually-1 and other clones was created for *PtKCS* genes using the methods described in Chapter 2 (DOE Joint Genome Institute Plant Gene Atlas, Sreedasyam et al., unpublished data; available from phytozome.jgi.doe.gov; see Chapter 2 **Supplementle Table 1.1** for sample information).

Using methods described in (Jones et al., 2025), a clustered heatmap of RNA-seq data from the GWAS population was created and utilized to compare population variation in expression of *PtKCS* genes versus concentrations of the alkene metabolites pentacosene

and heptadecosene, both of which were measured in mature leaf tissue in which *PtKCS1* gene expression peaks (Gonzales-Vigil et al., 2017). *PtKCS*-relevant morphological classifications from Gonzales-Vigil et al. (2017) were mapped to matching GWAS population accessions to facilitate comparison to RNA-seq and metabolite data. Alkene "plus" and alkene "minus" designations capture the differences in leaf and whole tree morphology and physiology that are correlated with *PtKCS* gene expression as observed in the 2017 study.

4.4 Results

4.4.1 Genomic and Climate Data Sets

As discussed in Chapter 3, the Delly SV call set contained 863,241 variants and the Lumpy call set contained 896,882 variants. After PASS filtering, removal of 71 clonal relatives (leaving 1413 samples), SV calls located in scaffolds, calls with >15% missing and MAF < 0.2, a total of 325,372 Delly SV calls and 386,170 Lumpy SV calls remained for input to the GWAS pipeline. Filtering the climate data for outlier, missing, and invariable data resulted in a total of 169 variables for GWAS analysis after the following variables were removed: swe months 3 – 10, def months 1, 11 – 12.

4.4.2 SV-Climatype GWAS Analyses

PC1 and PC2 of the SNP-based PCA explained 3.77% and 1.49% of genetic variance. The PCA 2D projections shows roughly two subgroups and has a similar structure to the SV-based PCA projections observed in Chapter 3 (**Supplemental Figure 3.1**, Chapter 3: **Figure 3.4a**, **b**).

The GWAS results indicated 13,124 of 325,372 tested Delly SV-climate associations were significant across four GWAS models according to a Bonferroni 0.2 threshold, and 11,297 of 386,170 tested Lumpy SV-climate associations were significant (**Figure 4.1**; **Supplementary Tables 4.1**, **4.2**). See **Figure 4.2** for a selected example of a Manhattan plot and a QQ plot from GWAS MLM analysis of the climate variable "soil". Manhattan and QQ plots are standard diagnostic tools to assess potential bias and parameter settings

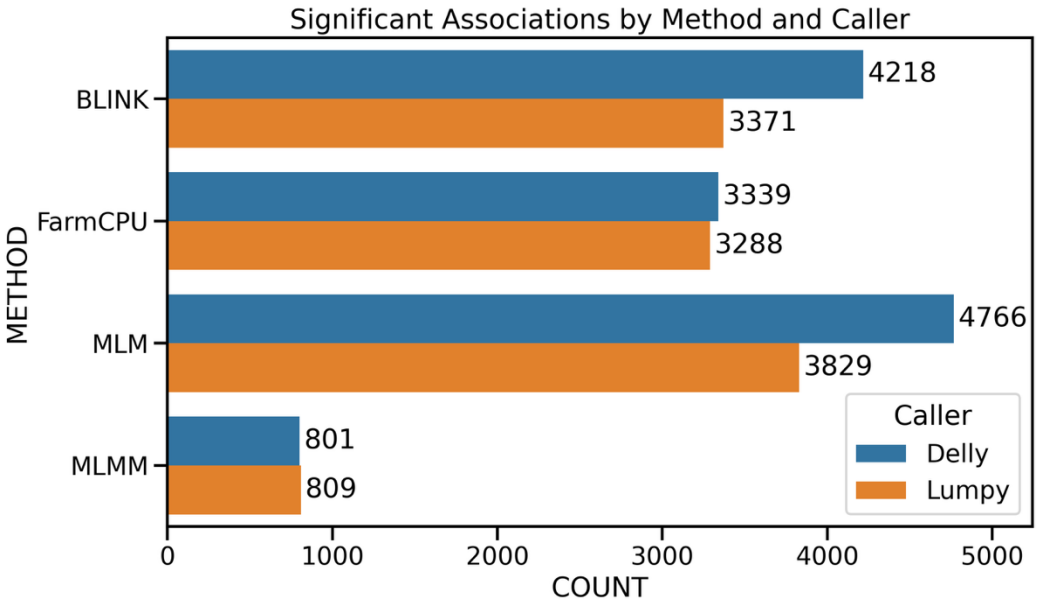


Figure 4.1: Total number of significant SV-climatype GWAS associations by GWAS method and SV caller.

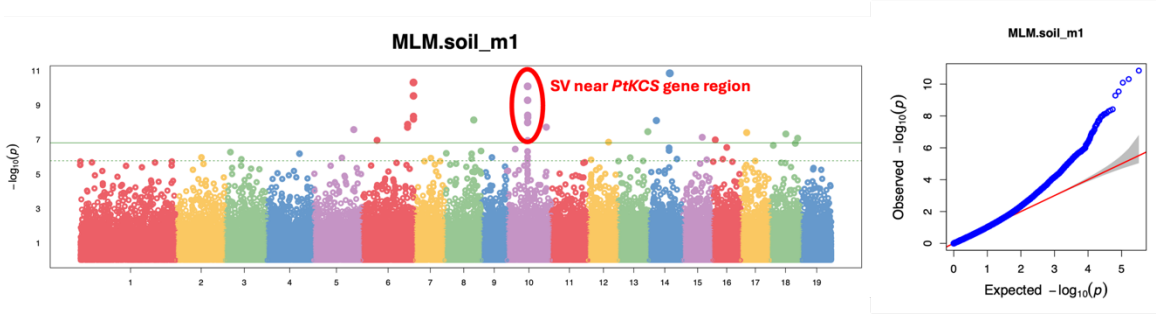


Figure 4.2: Selected example of Manhattan and QQ plots for MLM GWAS analysis of the "soil" climate variable. Manhattan and QQ plots are standard diagnostic tools to assess potential bias and parameter settings in GWAS analyses. Here, I selected one of many representative examples to demonstrate that SV-climatype analyses produced numerous highly significant results (data points above horizontal line in left most plot) and exhibit some evidence of *p* value inflation despite the incorporation of control for population structure and cryptic relatedness in the GWAS model (blue data points rising above the diagonal red line in the right most plot).

in GWAS analyses. Here, I selected one of many representative examples to demonstrate that SV-climatype analyses produced numerous highly significant results (data points above the horizontal line in left most plot) and exhibit some evidence of *p* value inflation despite the incorporation of control for population structure (SNP-based PCA) and cryptic relatedness (SNP-based GRM) in the GWAS model (blue data points rising above the diagonal red line in the right most plot), which is not uncommon to observe in climate GWAS (Voorman et al., 2011). For both variant callers, the greatest number of associations were significant in the MLM method results, and the fewest were significant in the MLMM results. There was moderate overlap among GWAS methods (**Figure 4.3**) within call sets, but each model identified unique associations as well.

The number of significant associations per climate variable by SV type (**Figure 4.4**) was similar to the total number of SV identified by each variant caller in those respective categories (Chapter 3: **Figure 3.1**). For example, both variant callers identified many more deletions (DEL) and translocations (BND) than other SV types and a greater number of significant GWAS associations was observed in those categories as well. However, this ratio does not hold across all climate variables, particularly in the Delly call set.

The distribution of significant associations by variant caller per climate variable, SV type, and month (**Figure 4.5**) shows notable differences among all partitions. Some SV-climate association distributions exhibit near uniform shape across months (for example, Delly INS-PDSI or Lumpy BND-PDSI) while others exhibit significant peaks (e.g., Delly INV-pet and Delly INV-ppt). Although there is overlap in how SV-climate associations are distributed between the variant call sets, seasonal patterns are more sharply delineated in the Delly data set.

SV-climate association effect sizes ranged from 45.64 (Lumpy DEL-soil month 11) to -42.20 (Delly DUP-soil month 1) and varied widely across climatypes and SV types, with positive values signifying an associated increase in the trait value with additional copies

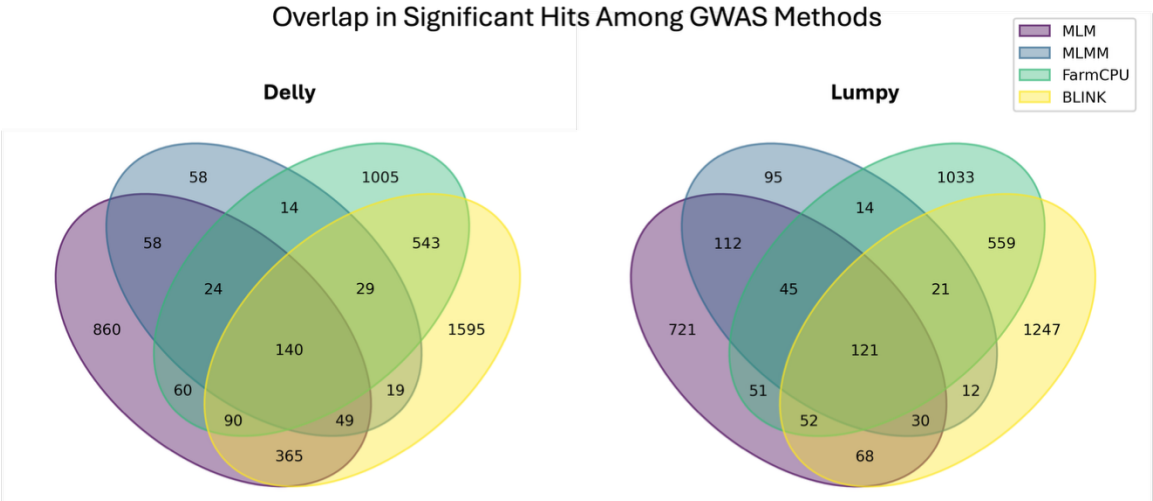


Figure 4.3: Overlap of significant associations among GWAS methods by SV caller.

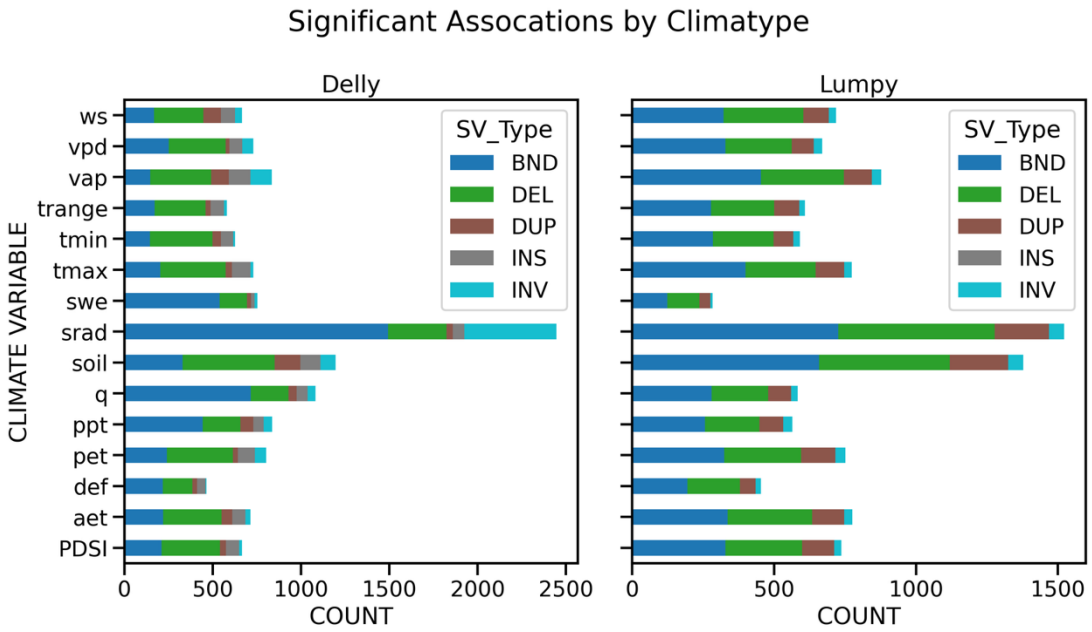


Figure 4.4: Number of significant GWAS associations per climate variable and SV type for each variant caller.

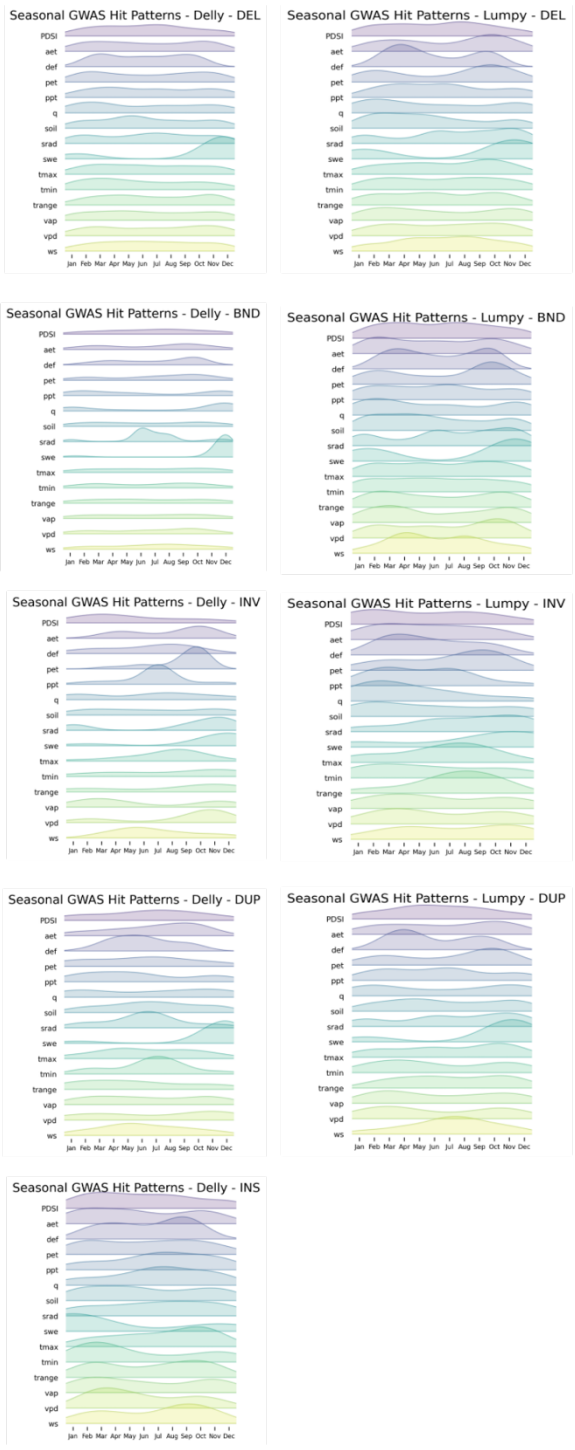


Figure 4.5: Distribution of significant GWAS associations across averaged months per variant caller (left column: Delly, right column: Lumpy), climate variable, and SV type. Note that the following climatypes were removed during preprocessing: swe months 3 – 10, def months 1, 11 – 12.

of alternate alleles and negative values indicating decreased trait values in association with additional copies of alternate alleles (**Figure 4.6**; **Supplementary Tables 4.1 - 4.2**). The 20 most extreme effect sizes (**Table 4.2**) are almost exclusively associated with soil and ppt climatypes, and dominated by DEL, DUP, and BND from both variant sets. Notably, variation in SV-climate association effect sizes by month generally does not correspond to the distribution of significant associations by month (**Figure 4.6**).

4.4.3 GRIN and MENTOR Analyses

Mapping SV from significant GWAS associations to nearest neighbor genes resulted in 7,292 unique genes in the Delly data set and 5,743 unique genes in the Lumpy data set (**Supplemental Files 4.1 - 4.2**). The union of the Delly and Lumpy data sets contained 10,833 unique genes, of which 5,471 (50%) unique genes were retained by GRIN as putative true positives for input to MENTOR (**Supplementary File 4.3**). MENTOR clustering produced a total of 270 functionally related gene clades. Although the resulting cladogram is too large to be visualized in its entirety here, a searchable PDF is provided in **Supplemental File 4.4**.

Genes having the highest number of significant associations with multiple climate variables across call sets were identified from the cladogram heatmap (**Figure 4.7**) and their clades of membership were analyzed using MENTOR-IA. Potri.010G079700 (Chr10:10641944..10643991) occurred in cladogram cluster 4 (15 total genes) and was mapped to significant associations with 15 of 15 climate variables, 11 of which were present in both variant GWAS results sets. Potri.010G080000 (Chr10:10683618..10685322) occurred in cluster 115 (37 genes) and was mapped to significant associations with 15 of 15 climate variables, 14 of which were present in both variant GWAS results sets. Potri.010G079500 (*PtKCS1*; Chr10:10600591..10602796) occurred in cluster 130 and was mapped to significant associations with 13 of 15 climate variables, nine of which were present in both variant GWAS results sets.

All three candidate genes are annotated as very-long-chain 3-ketoacyl-CoA synthases

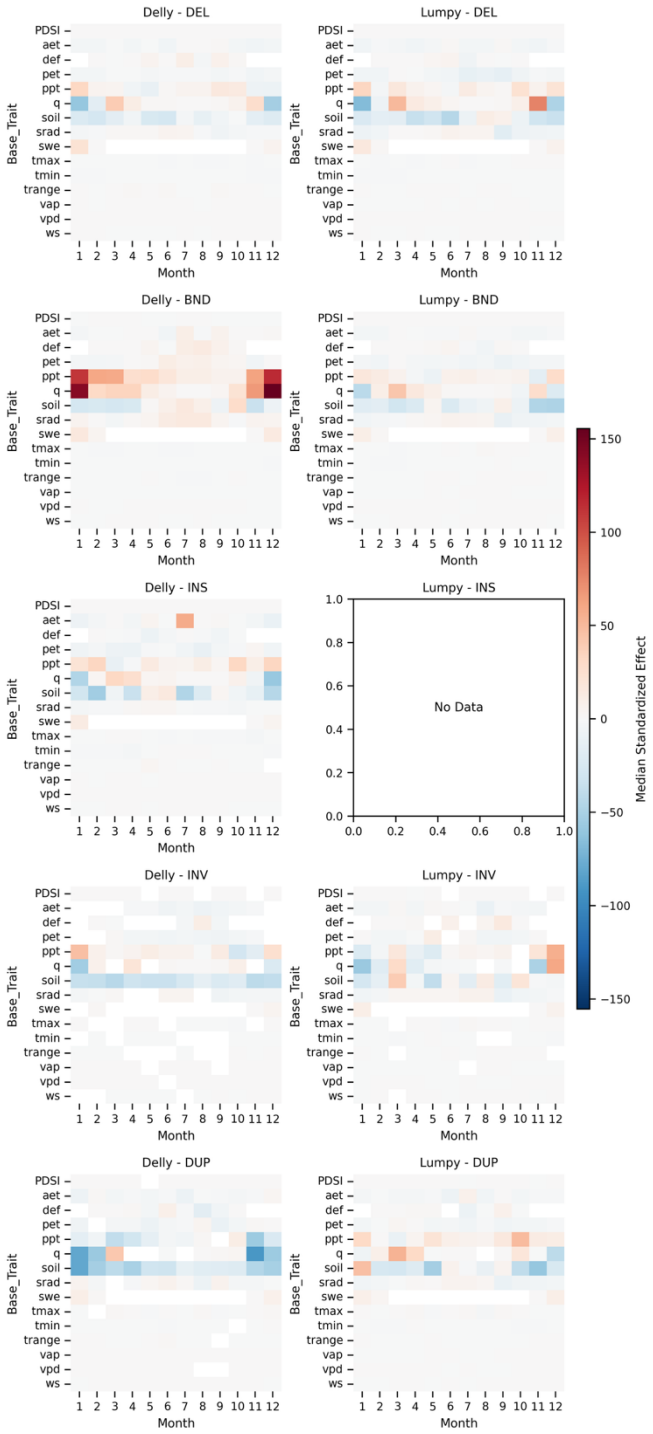


Figure 4.6: Heatmaps of median standardized effect sizes estimated during GWAS analysis by variant caller (left column: Delly, right column: Lumpy), climate variable, and SV type. Note that the following climatypes were removed during preprocessing: swe months 3 – 10, def months 1, 11 – 12.

Table 4.2: Twenty most extreme SV-climatype effect sizes. The "_m" suffix indicates averaged month (e.g., "_m1" indicates January).

Trait	SV_Type	Effect	SV Caller
soil_m11	DEL	45.639453	Lumpy
soil_m1	DUP	-42.202047	Delly
q_m12	DEL	-41.455232	Delly
q_m12	BND	-41.455232	Lumpy
q_m1	DEL	-41.166498	Delly
q_m1	DEL	-41.166498	Delly
soil_m2	DUP	-40.661073	Delly
q_m12	DUP	-39.881312	Lumpy
ppt_m1	BND	-39.633694	Delly
q_m1	BND	-39.541511	Lumpy
q_m1	DUP	-38.617638	Lumpy
q_m1	DUP	-38.617638	Lumpy
q_m1	DUP	-38.617638	Lumpy
soil_m12	DUP	-37.776573	Delly
q_m1	BND	-37.626701	Lumpy
soil_m5	BND	-36.729172	Lumpy
soil_m10	INV	36.381365	Lumpy
q_m11	BND	-35.904601	Delly
soil_m5	DEL	-35.900025	Lumpy
soil_m12	BND	35.802484	Delly

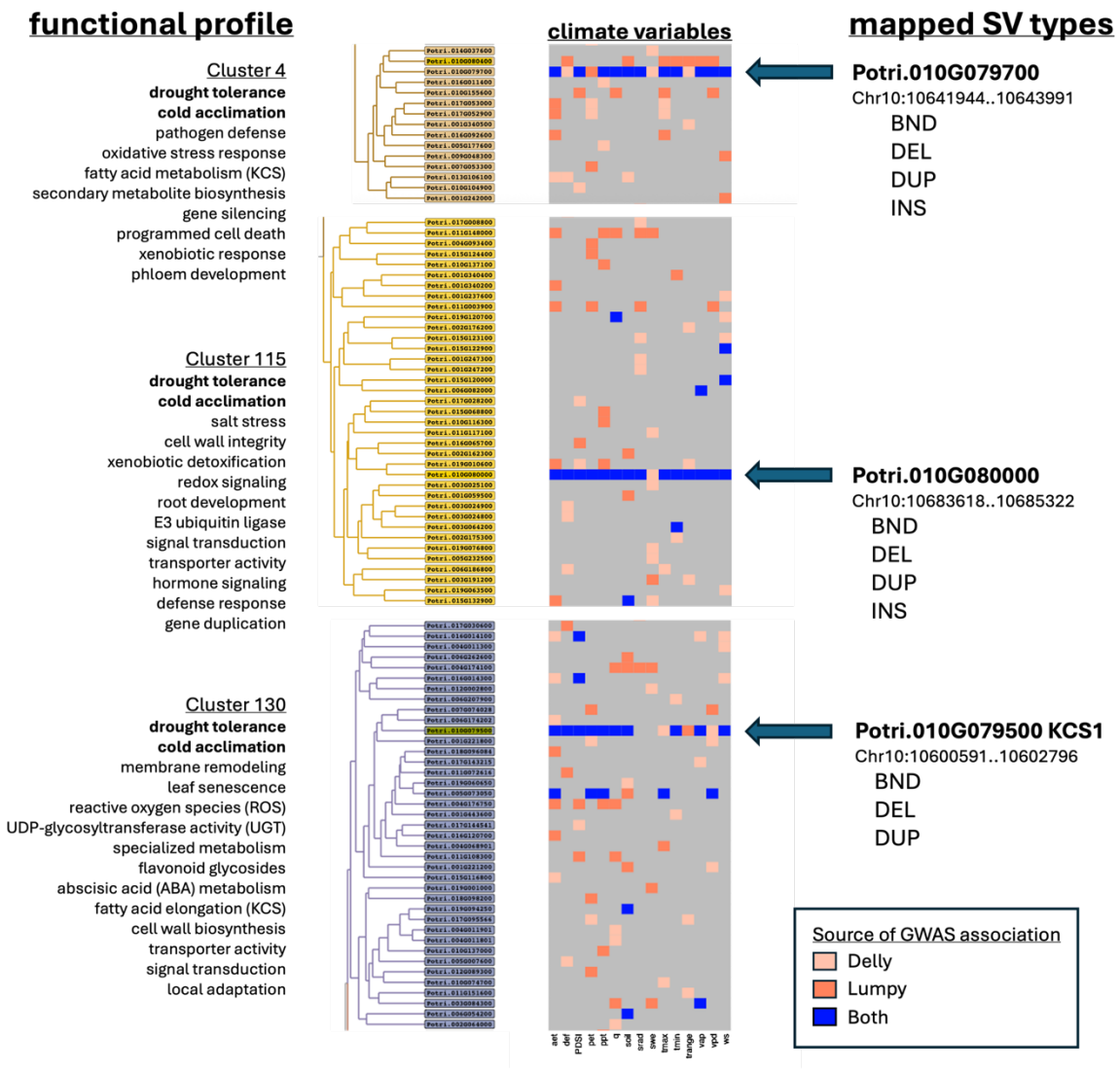


Figure 4.7: Selected clusters from the MENTOR cladogram that was produced from the union of Delly and Lumpy significant mapped-to-genes SV-climatype GWAS associations. Summaries of MENTOR-IA functional profiles are provided to the left of each cluster. The genes indicated with arrows were identified as high priority candidates based on having the highest numbers of significant climate associations and from both SV call sets, and were used to select cladogram clusters for follow up analyses. The SV types that each gene was mapped from are listed below the bold gene ID.

(KCS). Functional characterization of their respective clades shared a high degree of overlap, dominated by response to seasonal abiotic stress such as drought and cold acclimation and pathogen detection and response (**Figure 4.7**, **Supplemental Files 4.5** – **4.7**) which is consistent with expectations given the high number of GWAS associations with drought and temperature related climatypes by clade members.

4.4.4 PtKCS gene region

The *PtKCS* genes identified in the dendrogram are located within 80kb of one another on chromosome 10. Together with three other *PtKCS1*-like genes and *two PtKCS2*-like genes, they comprise the *P. trichocarpa KCS* gene region (Gonzales-Vigil et al., 2017). One additional translocated *PtKCS* gene occurs on Chromosome 8: Potri.008G160000. GWAS results for both variant callers contained multiple significant SV-climatype associations for BND, DUP, and DEL, and INS that mapped to all but one *PtKCS* gene: Potri.010G079900 (**Supplemental Files 4.1** - **4.2**). These observations of multiple climate-associated SV types and occurrences in and near the *PtKCS* region are consistent with previous studies that found complex SV in this region created by ancestral whole genome duplication and multiple occurrences of duplication, translocation, and deletion which were likely carried out by transposable elements such as Helitrons (Chen et al., 2022).

All *PtKCS* genes except for Potri.010G079900 were present in the retained GRIN gene set (**Supplemental File 3.4**), indicating strong evidence that *PtKCS* genes play important roles in plant response to the environment. Potri.010G080400 occurred in cluster 4 along with top candidate Potri.010G079700 (**Figure 4.7**). Potri.010G079300 and Potri.008G160000 were both located in cluster 217; Potri.010G079400 (cluster 250) Potri.010G080200 (cluster 260) occurred in unique clusters.

The Gene Atlas heatmap (**Figure 4.8**) reveals differential expression across tissues and developmental stages among *PtKCS* genes. Taken together, the occurrence of *PtKCS* genes in multiple MENTOR clades and divergent expression patterns are consistent with

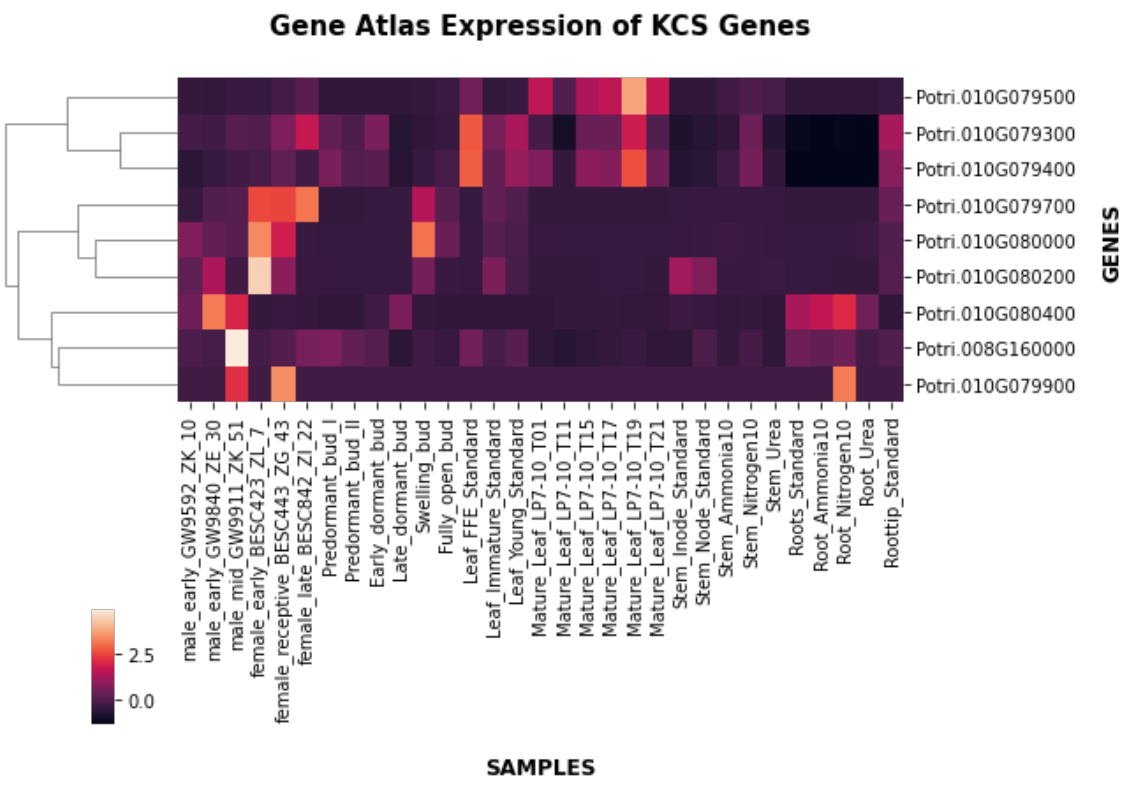


Figure 4.8: *Populus trichocarpa* reference genome Nisqually-1 gene atlas expression heatmap for *PtKCS* genes.

previous work that found evidence of differential *PtKCS* subfunctionalization following duplication (Chen et al., 2022; Gonzales-Vigil et al., 2017) The lack of GWAS associations mapped to Potri.010G079900 is potentially explained by specialization for a narrow range of reproductive tissue development which might not experience strong or direct climatic selection pressure.

The population-scale expression heatmap (**Figure 4.9**) shows that approximately 20% of individuals exhibit reduced or absent expression of *PtKCS1* (Potri.010G079500), which is strongly associated with minimal or undetectable levels of the alkene metabolites pentacosene and heptadecosene. Pentacosene and heptadecosene are enzymatically produced by the protein encoded by *PtKCS1* and are a significant component of mature leaf cuticle wax load (Gonzales-Vigil et al., 2017). These patterns align with the alkene plus and minus morphological mappings from the Gonzales-Vigil et al. (2017) publication. Notably, *PtKCS2* (Potri.010G079400) expression in the population is approximately opposite that of *PtKCS1*. Although mis-mapping of paired-end reads possibly contributes to the observed expression patterns because PtKCS genes are recently diverged, *PtKCS1*-like and *PtKCS2*-like genes exhibit greater divergence in sequence identity making mis-mapping less likely. Furthermore, the patterns observed here align with those observed in Gonzales-Vigil et al. (2017).

4.5 Discussion

4.5.1 Impact of SV calling methods

SV-climatype GWAS analyses identified similar numbers of significant associations per SV call set and exhibited similar patterns in the number of associations per GWAS method, climate variable, and SV type. However, clear differences were observed in the distribution of significant associations (**Figure 4.5**), effect sizes (**Figure 4.6**), and specific SV-climatype associations (heatmap in **Figure 4.7**) which might be related to differences in SV calling precision. For example, the Delly call set exhibits more sharply defined seasonal peaks in the ridgeline plots (**Figure 4.5**) and local maxima and minima in the effect size heatmaps (**Figure 4.6**). These differences could be explained by Delly's

Figure 4.9: Clustered heatmap and dendrogram of *PtKCS* gene expression in mature leaf tissue collected from the *P. trichocarpa* GWAS population. Rows represent accessions. Variation in gene expression levels are visualized in columns labeled with gene identifiers as a gradient from black (lowest or absent expression) to white (highest expression). Metabolite concentrations for the alkenes heptadecosene and pentacosene are visualized in the left-most columns as a gray-scale gradient from white (low or absent concentration) to black (highest concentration). The morphological classification of leaf phenotypes based on overlapping accessions with Gonzales-Vigil et al. (2017) is visualized in the blue column, with alkene "plus" accessions represented by a green bar, alkene "minus" by a yellow bar, and accessions with unknown morphology that were not present in the 2017 study represented by blue bars.

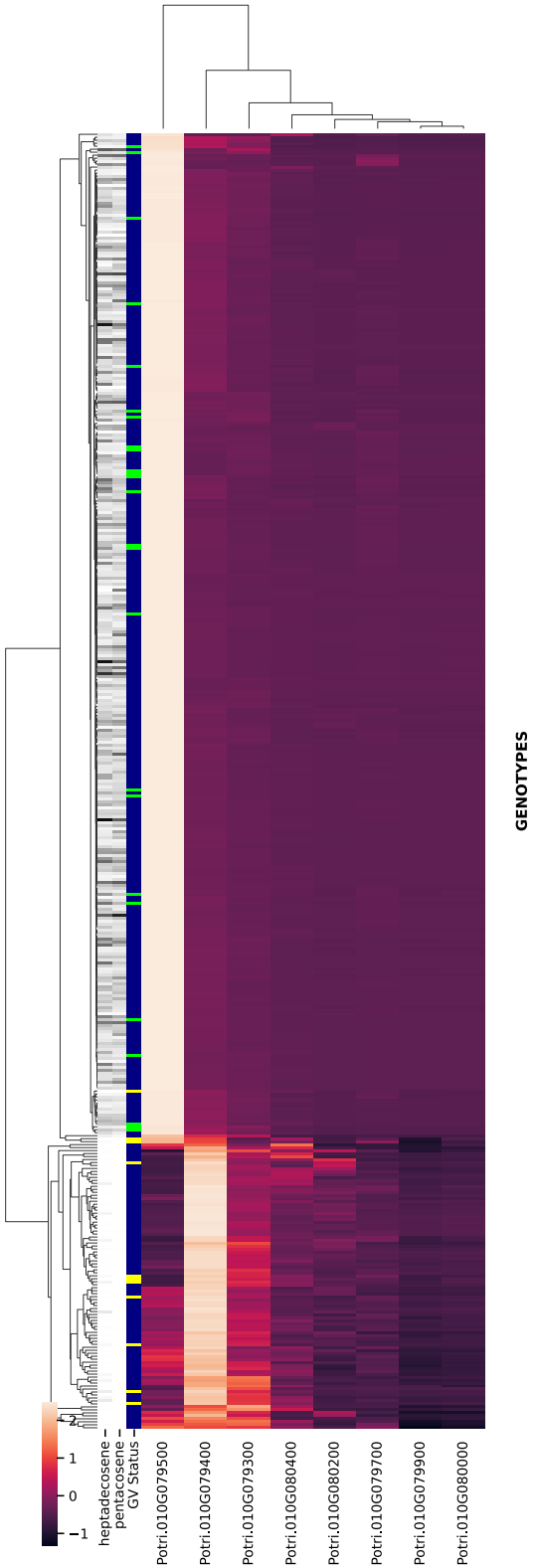


Figure 4.9 (Continued)

specialized soft-clip alignment algorithm which facilitates greater breakpoint precision when identifying SV locations (Cameron et al., 2019; Layer et al., 2014; Rausch et al., 2012).

However, differences in SV call sets, and therefore variant caller specific SV-climatype associations, are expected because independent SV calling algorithms typically produce a significant number of non-identical calls even when precision is similar due to general differences in identification approaches (Cameron et al., 2019; Kosugi et al., 2019). Because each variant caller has intrinsic strengths and biases, researchers typically employ multiple SV callers in order to increase sensitivity, after which they merge and filter the results such that only those SV calls supported by multiple variant callers are retained. However, merging practices have been shown to introduce substantial bias and inaccuracy to SV call sets through over-merging, artificial inflation of allele frequencies, and loss of unique allelic diversity approaches (Cameron et al., 2019; Kosugi et al., 2019). Furthermore, benchmarking studies demonstrate that widely used tools like SURVIVOR and bcftools suffer from systematic technical limitations that compromise downstream population genetic analyses (Zheng et al., 2024)

In Chapter 3, I avoided merging the Delly and Lumpy call sets to prevent introduction of such biases and instead relied upon group-based calling methods to help mitigate short-read biases. There is likely a significant degree of overlap among these SV call sets, therefore future use of this data should employ strategies to incorporate signals across variant callers into a unified SV call set while avoiding pitfalls associated with traditional methods. This could be accomplished by merging genotyped calls at population scale using the alignment-based re-genotyper PanPOP, followed by high throughput visualization to identify false positives using a tool like SV-plaudit (Belyeu et al., 2018; Zheng et al., 2024). PanPOP is also capable of resolving complex SV such as that in the *PtKCS* gene region. This protocol should yield improved recall, accuracy, and power, which would in turn improve support for biological insights.

4.5.2 Using SV candidates to improve biofuel feedstocks

In Chapter 3, I discussed the varying population genetic signals that are captured by different SV types. Inversions are under stronger purifying selection than other SV types because they suppress recombination between favorable combinations of alleles, CNVs reflect local adaptation through gene dosage effects, and deletions, duplications, and translocations generally capture neutral demographic patterns associated with migration and isolation-by-distance. Consistent with these expectations, the distribution of significant GWAS associations across climate variables and months for Delly-DEL, Lumpy-DEL, Delly-INS are roughly uniform (when missing data in swe and def are taken into account; **Figure 4.5**). In contrast, Delly-INV and Lumpy-INV exhibit seasonal patterns which indicate variation in climate conditions such as pet (evapotranspiration), ppt (precipitation accumulation), and tmax (maximum temperature) exert selection pressure on associated SV. Notably, seasonal patterns are also present in both call sets in BND distributions for srad (downward surface shortwave radiation) and DUP distributions for srad, tmin (minimum temperature), and ws (wind-speed). Although this may appear contrary to expectations, these patterns are logical when considering that DUP calls often overlap with the signals captured by CNV (because CNV are duplications) and translocations (BND) can result in suppressed recombination. Performing SV- and climate variable-specific population structure analyses would help tease apart signals of local adaptation versus demography. However, due to the north-south distribution of *P. trichocarpa*, careful selection of methodology is needed to distinguish local adaptation to latitudinal climate gradients from isolation-by-distance (Evans et al., 2014; Porth et al., 2015; Slavov et al., 2012).

The next steps towards using SV candidates for biofuel feedstock improvement are to quantify selection signals using genotype-environment association (GEA) analyses and validate adaptive properties using greenhouse and common garden experiments. The effect sizes observed in this study for soil, q, and ppt across all SV types and both call sets (**Figure 4.6**, **Table 4.2**) imply that soil moisture is exerting strong selection pressure on *P. trichocarpa*. Although the GWAS analyses provide a good starting point, direct

tests for selection on SV using Latent Factor Mixed Models (LFMM) and redundancy analyses (RDA) would provide stronger evidence (Rellstab et al., 2015; Storfer et al., 2018). These results could be used to inform common garden experiments with contrasting irrigation schemes and greenhouse experiments with controlled soil moisture content (for example, see Yang et al., 2018 and the references therein on drought experiments in *Populus*). Genotypes found to perform better under these conditions could then be selected for use in breeding programs, or the adaptive alleles could potentially be engineered into breeding stock.

4.5.3 SV candidates mapped to PtKCS gene region

PtKCS genes were identified as potentially high impact SV candidates by GRIN and MENTOR methods. *KCS* genes are very long chain fatty acid synthases (VLCFA) (Batsale et al., 2021). In *P. trichocarpa*, the final step in the elongation of VLCFAs produces alkanes and alkenes that are integrated into the protective wax layers of leaves (cuticle) and roots (suberin) (Gonzales-Vigil et al., 2017). The high number of climate associations from both call sets indicates the SV mapped to these genes may have potentially outsized roles in plant adaptation to climate variables. The MENTOR-IA functional characterization of their membership clades indicated the genes within play interacting roles in plant response to drought, cold, pathogens, and other stressors. These findings are consistent with previous work which showed that in addition to forming a hydrophobic barrier between the plant and the environment, alkenes and alkanes in the cuticle and suberin play direct roles in plant abiotic and biotic stress response, including pathogen defense (Batsale et al., 2021; Gonzales-Vigil et al., 2017; Jones and Furches et al., 2025; Lee et al., 2009; Rains et al., 2018; Wang et al., 2020).

The *PtKCS* gene region is characterized by complex SV that is thought to be the result of ancestral whole genome duplication and multiple occurrences of duplication, translocation, and deletion (Chen et al., 2022). Some *PtKCS* copies underwent subfunctionalization following duplication events, becoming specialized in function compared to the ancestral state (likely represented by *PtKCS2*-like genes,

Potri.010G079400 and Potri.010G079300). Consistent with this theory, the *P. trichocarpa* Gene Atlas heatmap (**Figure 4.8**) reflects divergent timing and location of gene expression among *PtKCS* genes. The placement by MENTOR of *PtKCS* genes in six unique clades indicates that these genes interact with different subsets of genes and phenotypes in the multiplex network and illustrate that network biology methods were able to properly capture the varying functional roles and spheres of *PtKCS* paralogs.

Due to varying levels of functional conservation among paralogs, special consideration must be taken when pursuing *PtKCS* genes and proximal SVs as candidates for crop improvement. For paralogs that retain a high degree of overlap in expression and function, redundancy prevents total loss of function if one copy is lost, therefore researchers must ensure expression of all related paralogs is eliminated. In the case of divergent paralogs, however, the silencing of a specialized gene can lead to higher expression in generalist paralogs which can produce unexpected and complex compensatory phenotypes. This has been observed to be the case for *PtKCS1*, which specializes in producing leaf cuticle alkenes and is expressed exclusively in mature leaves (**Figure 4.8**) (Gonzales-Vigil et al., 2017). Although *PtKCS2* and other *PtKCS* paralogs experience increased expression in genotypes with non- or low functioning *PtKCS1*, alkene production cannot be restored by overexpression of other paralogs (**Figure 4.9**). In plants lacking *PtKC1* expression, the total cuticle wax load is maintained but lacks alkenes, and the compositional difference has profound impacts on plant response to factors such as drought and pathogens, which in turn affects total biomass accumulation and quality (Gonzales-Vigil et al., 2017; Jones and Furches et al., 2025). Therefore, when validating the adaptative properties of Pt*KCS* paralogs to abiotic stress researchers should consider the legacy of complex SV and the resulting functional relationships in this region.

In addition to biological complexity, fully characterizing and validating candidates in the *PtKCS* region is complicated by technical complexity due to high sequence similarity inherent to duplications. Because such regions are challenging to accurately map with

short-read data, it is helpful to incorporate long-read data to improve the success rate of techniques like CRISPR for knock-out or overexpression.

4.5.4 Future Directions

In addition to the SV types analyzed here, investigating CNV will be essential for fully capturing signals of local adaption in *P. trichocarpa.* Plants can more rapidly adapt to local conditions through gene dosage effects produced by CNV, thereby avoiding fitness valleys that are inherent to new or disruptive mutations (Zmienko et al., 2020). A previous study in *Populus* demonstrated that gene dosage via CNV directly impacted traits of interest to biofuel improvement, including total biomass accumulation and phenology (Bastiaanse et al., 2019). In another study comparing two small populations of *P. balsamifera* from opposite ends of the species range, CNV exhibited differential adaptation to endemic selection as measured by pathogen resistance, water use efficiency, and photosynthetic assimilation rates (Prunier et al., 2019). Similar patterns of adaptation to water stress and pathogens have been observed in grapes and canola (Catacchio et al., 2019; Dolatabadian et al., 2020). Incorporating CNV data sets into LFMM and RDA analyses to quantify selection on SV by climate variables will help identify impactful variants for feedstock improvement.

4.5.5 Recent Work

This dissertation made use of work I previously did in other projects. First, I originally developed the SNP clustering method I utilized in Chapter 3 for prioritization of 60 *P. trichocarpa* accessions for generation of long-read sequencing data which was made possible by an award from the Joint Genome Institution (JGI). In collaboration with colleagues at Oregon State University, my analyses were used to direct field-collection of leaf tissue for extraction and sequencing. Forty-five accessions are now included in the long-read data set, with the remaining 15 assemblies still underway. The long-read data set has recently been used to create a *P. trichocarpa* pangenome matrix and perform long-read SV identification (Hyatt et al, in preparation). Future work will use the long-read data to validate and phase the short-read calls included here (as well as the

additional SV data sets that were not included here).

Another project of which I am a co-first author used GWAS and network biology methods to investigate the impact of *PtKCS* genes and related phenotypes on fungal endophyte communities (Jones and Furches et al., 2025). This project in turn utilized work I contributed to a literature review in which I described the role of microbial keystone taxa in plant hosts (Jones et al., 2019). Having previously studied the role of keystone microbes and community response was integral to my interpretation of the community dynamics on *P. trichocarpa* alkene phenotypes.

Finally, I have had the opportunity to participate in work in collaboration with Oregon State University to validate the effects of *PtKCS* gene expression on plant host-microbial community dynamics. In this project, CRISPR methods were used to knock out *PtKCS1* in some accessions and the entire *PtKCS* gene region in others. We utilized the long-read sequencing data described above to improve CRISPR designs. Ongoing greenhouse and field-based trials are collecting data on the impact of alkene phenotypes on the microbial community as well as changes in morphology and plant stress tolerance.

4.6 Supplemental Materials

4.6.1 Supplemental Tables

See attachment "Furches_Chapter4_SuppTables.xlsx" for the following:

Supplementary Table 4.1: Significant SV-climatype GWAS associations in the Delly call set.

Supplementary Table 4.2: Significant SV-climatype GWAS associations in the Lumpy call set.

4.6.2 Supplemental Files

Supplemental File 4.1: Delly significant SV-climatype GWAS association SV-to-genes mappings. Please see attached file "Furches_Chapter4_SuppFile4.1_snp2genes_DELLY.txt".

Supplemental File 4.2: Lumpy significant SV-climatype GWAS association SV-to-genes mappings. Please see attached file "Furches_Chapter4_SuppFile4.1_snp2genes_LUMPY.txt".

Supplemental File 4.3: Significant SV-climatype GWAS association SV mapped-to-genes unique gene set retained by GRIN for input to MENTOR. Please see attached tab-delimited text file "Furches_Chapter4_SuppFile4.3_GRINretained".

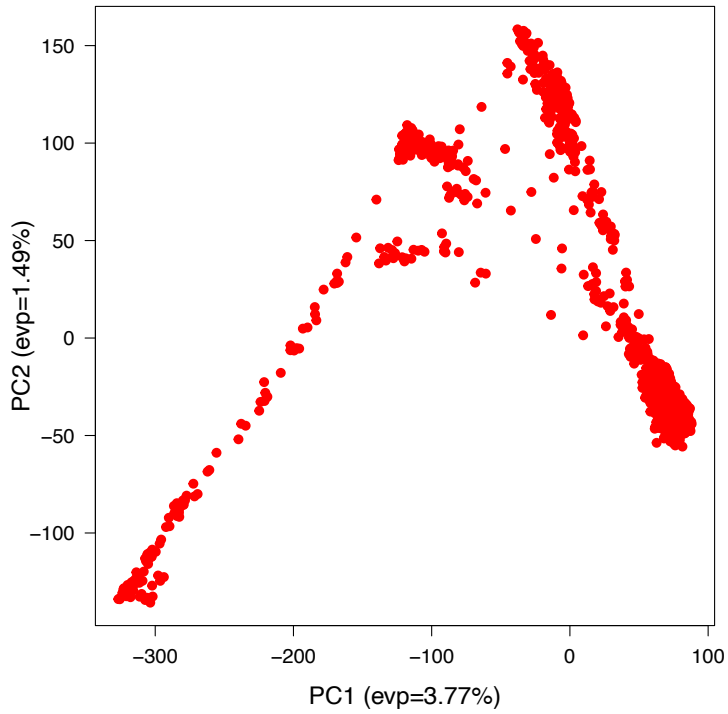
Supplemental File 4.4: MENTOR dendrogram of 5,471 GRIN-retained genes containing 270 functionally related clades. Source SV call set is indicated in the adjacent heatmap: Delly is red, Lumpy is orange, and occurrence in both the Delly and Lumpy call sets is indicated by blue. The associated climate variable is indicated by presence/absence of colorization in the corresponding columns, with climate variable identity labeled at the bottom of the heatmap. Please see attached file "Furches_Chapter4_SuppFile4.4_dendrogram_heatmap.pdf".

Supplemental File 4.5: MENTOR-IA synthesized results for cladogram cluster 4. Please see attachment "Furches_Chapter4_SuppFile4.5.pdf".

Supplemental File 4.6: MENTOR-IA synthesized results for cladogram cluster 115. Please see attachment "Furches_Chapter4_SuppFile4.6.pdf".

Supplemental File 4.7: MENTOR-IA synthesized results for cladogram cluster 130. Please see attachment "Furches_Chapter4_SuppFile4.7.pdf".

4.6.3 Supplemental Figures



Supplemental Figure 4.1: 2D projection of PCA of the GWAS population SNP data set.

References

Allaby, R. (2019). Clonal crops show structural variation role in domestication. In *Nature Plants* (Vol. 5, Issue 9, pp. 915–916). https://doi.org/10.1038/s41477-019-0515-8

Bastiaanse, H., Zinkgraf, M., Canning, C., Tsai, H., Lieberman, M., Comai, L., Henry, I., & Groover, A. (2019). A comprehensive genomic scan reveals gene dosage balance impacts on quantitative traits in Populus trees. *Proceedings of the National Academy of Sciences*, *116*(27), 13690–13699. https://doi.org/10.1073/pnas.1903229116

Batsale, M., Bahammou, D., Fouillen, L., Mongrand, S., Joubès, J., & Domergue, F. (2021). Biosynthesis and functions of very-long-chain fatty acids in the responses of plants to abiotic and biotic stresses. In *Cells* (Vol. 10, Issue 6, p. 1284). Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/cells10061284

Belyeu, J. R., Nicholas, T. J., Pedersen, B. S., Sasani, T. A., Havrilla, J. M., Kravitz, S. N., Conway, M. E., Lohman, B. K., Quinlan, A. R., & Layer, R. M. (2018). SV-plaudit: A cloud-based framework for manually curating thousands of structural variants. *GigaScience*, *7*(7), 1–7. https://doi.org/10.1093/GIGASCIENCE/GIY064

Cameron, D. L., Di Stefano, L., & Papenfuss, A. T. (2019). Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nature Communications*, *10*(1), 3240. https://doi.org/10.1038/s41467-019-11146-4

Catacchio, C. R., Alagna, F., Perniola, R., Bergamini, C., Rotunno, S., Calabrese, F. M., Crupi, P., Antonacci, D., Ventura, M., & Cardone, M. F. (2019). Transcriptomic and genomic structural variation analyses on grape cultivars reveal new insights into the genotype-dependent responses to water stress. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-39010-x

Chen, J. Y., Mumtaz, A., & Gonzales-Vigil, E. (2022). Evolution and molecular basis of substrate specificity in a 3-ketoacyl-CoA synthase gene cluster from Populus trichocarpa. *Journal of Biological Chemistry*, *298*(10).

https://doi.org/10.1016/J.JBC.2022.102496/ATTACHMENT/992C3D17-5642-406A-8993-368F34509534/MMC1.PDF

Cope, K. R., Prates, E. T., Miller, J. I., Demerdash, O. N. A., Shah, M., Kainer, D., Cliff, A., Sullivan, K. A., Cashman, M., Lane, M., Matthiadis, A., Labbé, J., Tschaplinski, T. J., Jacobson, D. A., & Kalluri, U. C. (2023). Exploring the role of plant lysin motif receptor-like kinases in regulating plant-microbe interactions in the bioenergy crop Populus. *Computational and Structural Biotechnology Journal*, *21*, 1122–1139. https://doi.org/10.1016/j.csbj.2022.12.052

Evans, L. M., Slavov, G. T., Rodgers-Melnick, E., Martin, J., Ranjan, P., Muchero, W., Brunner, A. M., Schackwitz, W., Gunter, L., Chen, J. G., Tuskan, G. A., & Difazio, S. P. (2014). Population genomics of Populus trichocarpa identifies signatures of selection and adaptive trait associations. *Nature Genetics*, *46*(10), 1089–1096. https://doi.org/10.1038/ng.3075

Fuentes, R. R., Chebotarov, D., Duitama, J., Smith, S., De la Hoz, J. F., Mohiyuddin, M., Wing, R. A., McNally, K. L., Tatarinova, T., Grigoriev, A., Mauleon, R., & Alexandrov, N. (2019). Structural variants in 3000 rice genomes. *Genome Research*, *29*(5), 870–880. https://doi.org/10.1101/gr.241240.118

Gabur, I., Chawla, H. S., Snowdon, R. J., & Parkin, I. A. P. (2019). Connecting genome structural variation with complex traits in crop plants. *Theoretical and Applied Genetics*, *132*(3), 733–750. https://doi.org/10.1007/s00122-018-3233-0

Gonzales-Vigil, E., Hefer, C. A., von Loessl, M. E., La Mantia, J., & Mansfield, S. D. (2017). Exploiting Natural Variation to Uncover an Alkene Biosynthetic Enzyme in Poplar. *The Plant Cell*, *29*(8), 2000–2015. https://doi.org/10.1105/tpc.17.00338

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., Rokhsar, D. S., & others. (2011). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, *40*(D1), D1178--D1186. https://doi.org/10.1093/NAR/GKR944

Jones, P., Furches, A., Leopold, D. R., Miller, J. I., Brown, S. P., Muchero, W., Chen, J., Tschaplinski, T. J., Busby, P., & Jacobson, D. (2025). Exploring the influence of host leaf cuticle on fungal phyllobiome community structure in Populus trichocarpa

using GWAS. *In Revision*.

Jones, P., Garcia, B. J., Furches, A., Tuskan, G. A., & Jacobson, D. (2019). Plant host-associated mechanisms for microbial selection. In *Frontiers in Plant Science* (Vol. 10). https://doi.org/10.3389/fpls.2019.00862

Kainer, D., Lane, M., Sullivan, K. A., Miller, J. I., Cashman, M., Morgan, M., Cliff, A., Romero, J., Walker, A., Blair, D. D., Chhetri, H., Wang, Y., Furches, A., Noshay, J., Drake, M., Landry, N., Ireland, A. J., & Missaoui, A. (2024). RWRtoolkit : multi-omic network analysis using random walks on multiplex networks in any species. *BioRxiv*, 2024.07.17.603975. https://doi.org/10.1101/2024.07.17.603975

Kolberg, L., Raudvere, U., Kuzmin, I., Adler, P., Vilo, J., & Peterson, H. (2023). g:Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic Acids Research*, *51*(W1), W207–W212. https://doi.org/10.1093/NAR/GKAD347

Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M., & Kamatani, Y. (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biology*, *20*(1), 8–11. https://doi.org/10.1186/s13059-019-1720-5

Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). LUMPY: A probabilistic framework for structural variant discovery. *Genome Biology*, *15*(6), 1–19. https://doi.org/10.1186/gb-2014-15-6-r84

Lee, S. B., Jung, S. J., Go, Y. S., Kim, H. U., Kim, J. K., Cho, H. J., Park, O. K., & Suh, M. C. (2009). Two Arabidopsis 3-ketoacyl CoA synthase genes, KCS20 and KCS2/DAISY, are functionally redundant in cuticular wax and root suberin biosynthesis, but differentially controlled by osmotic stress. *Plant Journal*, *60*(3), 462–475. https://doi.org/10.1111/j.1365-313X.2009.03973.x

McWilliam, H., Li, W., Uludag, M., Squizzato, S., Park, Y. M., Buso, N., Cowley, A. P., & Lopez, R. (2013). Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Research*, *41*(W1), W597–W600. https://doi.org/10.1093/NAR/GKT376

Porth, I., Klápště, J., McKown, A. D., La Mantia, J., Guy, R. D., Ingvarsson, P. K., Hamelin, R., Mansfield, S. D., Ehlting, J., Douglas, C. J., & El-Kassaby, Y. A.

(2015). Evolutionary quantitative genomics of Populus trichocarpa. *PLoS ONE*, *10*(11), e0142864. https://doi.org/10.1371/journal.pone.0142864

Prunier, J., Giguère, I., Ryan, N., Guy, R., Soolanayakanahally, R., Isabel, N., MacKay, J., & Porth, I. (2019). Gene copy number variations involved in balsam poplar (Populus balsamifera L.) adaptive variations. *Molecular Ecology*, *28*(6), 1476–1490. https://doi.org/10.1111/mec.14836

Rains, M. K., De Silva, N. D. G., & Molina, I. (2018). Reconstructing the suberin pathway in poplar by chemical and transcriptomic analysis of bark tissues. *Tree Physiology*, *38*(3), 340–361. https://doi.org/10.1093/treephys/tpx060

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., & Korbel, J. O. (2012). DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, *28*(18), i333–i339. https://doi.org/10.1093/bioinformatics/bts378

Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., & Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. In *Molecular Ecology* (Vol. 24, Issue 17, pp. 4348–4370). https://doi.org/10.1111/mec.13322

Schiessl, S. V., Katche, E., Ihien, E., Chawla, H. S., & Mason, A. S. (2019). The role of genomic structural variation in the genetic improvement of polyploid crops. In *Crop Journal* (Vol. 7, Issue 2, pp. 127–140). Crop Science Society of China/ Institute of Crop Sciences. https://doi.org/10.1016/j.cj.2018.07.006

Slavov, G. T., Difazio, S. P., Martin, J., Schackwitz, W., Muchero, W., Rodgers-Melnick, E., Lipphardt, M. F., Pennacchio, C. P., Hellsten, U., Pennacchio, L. A., Gunter, L. E., Ranjan, P., Vining, K., Pomraning, K. R., Wilhelm, L. J., Pellegrini, M., Mockler, T. C., Freitag, M., Geraldes, A., … Tuskan, G. A. (2012). Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree Populus trichocarpa. *New Phytologist*, *196*(3), 713–725. https://doi.org/10.1111/j.1469-8137.2012.04258.x

Storfer, A., Patton, A., & Fraik, A. K. (2018). Navigating the interface between landscape genetics and landscape genomics. In *Frontiers in Genetics* (Vol. 9, Issue

MAR, p. 68). Frontiers Media S.A. https://doi.org/10.3389/fgene.2018.00068

Sullivan, K. A., Lane, M., Cashman, M., Miller, J. I., Pavicic, M., Walker, A. M., Cliff, A., Romero, J., Qin, X., Mullins, N., Docherty, A., Coon, H., Ruderfer, D. M., Garvin, M. R., Pestian, J. P., Ashley-Koch, A. E., Beckham, J. C., McMahon, B., Oslin, D. W., … Kainer, D. (2024). Analyses of GWAS signal using GRIN identify additional genes contributing to suicidal behavior. *Communications Biology*, *7*(1), 1–16. https://doi.org/10.1038/s42003-024-06943-7

Sullivan, K. A., Miller, J. I., Townsend, A., Morgan, M., Lane, M., Pavicic, M., Shah, M., Cashman, M., & Jacobson, D. A. (2024). MENTOR: Multiplex Embedding of Networks for Team-Based Omics Research. *BioRxiv*, 2024.07.17.603821. https://doi.org/10.1101/2024.07.17.603821

Vlot, A., Lane, M. J., Madrid, G., Kruse, P., Dandy, J., Kaplanoglu, S., Sullivan, K., Prates, E., Shah, M., Walker, T., Nsoh, B., & Jacobson, D. (2025). Looking Forward: The MENTOR Interpretation Agent Accelerating Mechanistic Insights with LLM-Powered Automation. *The Center for Bioenergy Innovation Annual Science Meeting*, 1.

Voorman, A., Lumley, T., McKnight, B., & Rice, K. (2011). Behavior of QQ-plots and Genomic Control in studies of gene-environment interaction. *PLoS ONE*, *6*(5), e19416. https://doi.org/10.1371/journal.pone.0019416

Wang, X., Kong, L., Zhi, P., & Chang, C. (2020). Update on cuticular wax biosynthesis and its roles in plant disease resistance. In *International Journal of Molecular Sciences* (Vol. 21, Issue 15, pp. 1–15). MDPI AG. https://doi.org/10.3390/ijms21155514

Wu, C., MacLeod, I., & Su, A. I. (2013). BioGPS and MyGene.info: organizing online, gene-centric information. *Nucleic Acids Research*, *41*(D1), D561–D565. https://doi.org/10.1093/NAR/GKS1114

Yang, X., Tuskan, G. A., Garcia, B. J., Abraham, P. E., Engle, N., Jawdy, S. S., Jacobson, D. A., Hettich, R. L., Gunter, L. E., Tschaplinski, T. J., Garcia, B. J., Gunter, L. E., Jawdy, S. S., Engle, N., Yang, X., Jacobson, D. A., Hettich, R. L., Tuskan, G. A., & Tschaplinski, T. J. (2018). Quantitative proteome profile of water

deficit stress responses in eastern cottonwood (Populus deltoides) leaves. *PLOS ONE*, *13*(2), e0190019. https://doi.org/10.1371/journal.pone.0190019

Zheng, Z., Zhu, M., Zhang, J., Liu, X., Hou, L., Liu, W., Yuan, S., Luo, C., Yao, X., Liu, J., & Yang, Y. (2024). A sequence-aware merger of genomic structural variations at population scale. *Nature Communications*, *15*(1), 1–9. https://doi.org/10.1038/s41467-024-45244-9

Zmienko, A., Marszalek-Zenczak, M., Wojciechowski, P., Samelak-Czajka, A., Luczak, M., Kozlowski, P., Karlowski, W. M., & Figlerowicz, M. (2020). AthCNV: A map of DNA copy number variations in the arabidopsis genome. *Plant Cell*, *32*(6), 1797–1819. https://doi.org/10.1105/tpc.19.00640

5 Chapter 5: Conclusion

The overarching goal of this dissertation was to identify genetic variants that impact biofuel feedstock targets and plant stress resilience. These properties are inextricably linked because trees cannot maintain desirable growth and morphological characteristics without adequate stress tolerance. Here, SNPs and SVs (Chapter 3) across natural populations of *P. trichocarpa* were leveraged to gain insights into the genetic architecture of traits such as cell wall composition (Chapter 2) and abiotic stress response (Chapter 4), which are both essential for achieving growth on marginal lands using reduced resources.

The core methodology I used to find impactful genetic candidates throughout this dissertation calls on biological knowledge to guide variant discovery rather than sole reliance on statistical thresholds. Traditional GWAS approaches can miss true biological signals due to the conservative multiple testing corrections needed to reduce false positive rates. The network biology methods applied here offer an alternative through strategy the use of validated biological relationships to distinguish true positives from statistical noise. Chapter 2 demonstrated this approach using lines-of-evidence (LOE) integration across a multiplex network to identify previously unknown cell wall candidates. Chapter 4 applied RWR-based filtering strategies to identify impactful climate-associated SVs. These biological thresholding approaches allow researchers to increase sensitivity and target variants most likely to have functional importance. However, because relaxed GWAS thresholds can produce extensive candidate lists even after network-based biological thresholding, additional prioritization approaches are useful. The application of candidate LOE ranking metrics (Chapter 2) is a valuable strategy to help researchers identify the candidates with the strongest evidential support, ensuring wise investment of validation resources.

In addition to sharing improved methods for candidate discovery, a primary goal of this work is to provide fellow researchers with curated variant sets for use in their own investigations. Chapter 2 moves beyond well-characterized cell wall biosynthesis pathway components by contributing a refined list of previously unknown novel regulatory genes whose candidate status is well-supported by previous work in *P.*

trichocarpa despite having little to no prior direct annotation or research. Chapters 2 and 3 provide an extensive catalog of candidate SVs and their climate variable associations, functionally clustered using network approaches to facilitate interpretation. This population-scale SV dataset fills a critical gap, as previous *P. trichocarpa* SV studies were limited to small sample sizes. Together, these resources strengthen the role of *P. trichocarpa* as a model woody plant species by expanding the genomic toolkit available to researchers.

Beyond providing practical resources, this dissertation contributes to fundamental understanding in two key areas. Despite decades of research, the regulation of cell wall biosynthesis, particularly the regulatory networks controlling pathway flux, remains poorly understood. The novel candidates identified through network integration here represent new pieces of this complex puzzle. Second, the role of SVs in forest tree adaptation has largely been under characterized. The extensive climate associations discovered here suggest SVs may be particularly powerful for detecting and understanding abiotic stress adaptation, potentially capturing adaptive signals missed by SNP-based approaches alone.

However, the candidates identified in these chapters now require systematic validation. For cell wall candidates, this includes expression profiling, protein interaction studies, and functional validation through CRISPR or natural variant analysis. For climate-associated SVs, validation requires common garden experiments, greenhouse trials with controlled stressors, and field evaluation across environmental gradients. Some validation work is already underway. I collaborated in a project to functionally characterize of a selection of cell wall regulatory candidates (UGTs) through exogenous protein expression and enzyme affinity profiling. I have also participated in efforts to validate the impact of SV-generated *PtKCS* genes on microbial communities and tree growth through greenhouse and field experiments on CRISPR knock-out genotypes.

In addition to validation, additional work should be done to expand and improve the SV

dataset. Some of this work has already begun, while other pieces must be carried out in the future. First, population scale re-genotyping of expanded group-based call sets will increase sensitivity and accuracy of the data. Beyond the data included here, I performed SV identification in the GWAS population using three additional callers (Breakdancer, GRIDSS, CNVcaller) and performed additional group-based genotyping (BayesTyper, SVTyper) of the full 5-caller data set using two different methods. In this expanded data set, I also used SV call reassembly (via GRIDSS) to improve precision for imprecise callers (Breakdancer, Lumpy). Using this 5-caller data set after merging SV calls across calling algorithms using population-scale sequence realignment re-genotyping (PanPOP) would ensure the full SV landscape is sampled and the false positive rate is reduced. To further validate these methods, I also used a single-sample SV calling, genotyping, and merging ensemble pipeline (Parliament 2) in the GWAS population against which the improvements afforded by population-informed methods can be quantified. Second, additional biological and technical validation should be performed by comparing the genotyped SV call set against long-read sequencing data and SV call sets identified in family crosses. I collaborated in efforts to generate long-read sequencing data which have since been used by my colleagues to generate a pangenome matrix and perform long-read SV identification. Plans are underway to validate the short-read calls produced here against the long-read sets. I also performed extensive SV identification in a group of 56 reciprocally crossed *P. trichocarpa* families (using 5 callers and the single-sample ensemble pipeline) but due to a technical error that data could not be included here.

Third, genotype by environment (GxE) analyses are needed to directly and specifically quantify the selective forces acting on SVs and understand how potentially adaptive traits change across environments. Lastly, because different SV types have unique functional impacts and reflect distinct evolutionary dynamics, more work is needed to characterize the distribution of SV classes across subpopulations, which SV types are adaptive versus neutral, and how SV formation mechanisms operate in long-lived trees. The current work provides a foundation for these efforts but represents only an initial characterization. In particular, the data sets produced here should be expanded to include CNV because they

capture unique signals of local adaptation through gene dosage.

Beyond expansion, validation, and characterization of variant candidates, several challenges remain in this field. SV calling and genotyping remains computationally intensive, potentially excluding researchers with limited computational resources. Additionally, the dearth of new short-read based SV methods is concerning given that short-read sequencing will remain relevant for studies in non-model species due to cost constraints for some time to come. More emphasis is also needed on utilizing population-scale calling and genotyping methods to mitigate biases associated with short-read data, reference-based identification, and population structure effects.

This dissertation demonstrates how integrating established computational methods with comprehensive genomic resources can address important questions in forest tree biology. By applying network biology approaches to both SNP and SV datasets, this work expands our understanding of the genetic basis of bioenergy-relevant traits and plant response to abiotic conditions while providing practical resources for the research community. The emphasis on biological rather than purely statistical interpretation offers a framework for managing the complexity of modern genomic datasets. As we work toward developing climate-resilient bioenergy feedstocks, the combination of comprehensive variant detection, network-based prioritization, and systematic validation will be essential. The datasets and approaches presented here contribute to this collective effort, strengthening the foundation for genomics-enabled tree improvement and the available resources for *P. trichocarpa*.

Vita

Anna Furches completed a Bachelor of Science in Plant Biology, a Master of Science in Ecology & Evolutionary Biology with a focus in Plant Population Genetics and Phylogenetics, and a Doctor of Philosophy in Energy Science & Engineering with a focus in Computational and Systems Biology.